# GAUGE SYMMETRIES IMPOSE LIMITS ON GLOBAL PARAMETER RECOVERY FOR LORA AND ATTENTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Low-Rank Adaptation (LoRA) updates and Multi-Head Attention (MHA) heads have inherent symmetries, so many weight arrays describe the same update or attention computation. In this work, we ask how many deterministic continuous candidate rules are needed to recover parameters up to these symmetries. Each rule takes a population distribution as input and returns a point in the (compact) quotient space and we require that, for every target, at least one rule returns a nearby point. We show that for even embedding dimension $d$ and even rank $r$, sufficiently accurate LoRA recovery requires at least $r(d-r)/4 + r/2$ candidates in the low rank-regime. So for example, at $d = 4096$ and $r = 64$, this gives that $64544$ continuous candidate rules are required for recovery. For even $d$ and $h$ attention heads with equal even key/value dimension $r$, we show a corresponding lower bound of $hr(d-r)/2 + (h+1)(r/2-1) + 1$. Thus stable, uniform, population-level proper recovery requires lists of size $\Omega(rd)$ for LoRA and $\Omega(hrd)$ for MHA. Our approach for the proof is to turn the success domains of the candidate rules into a topologically simple cover and then argue about its size through nonzero products in cohomology, which is measured by the cup-length. We develop a cup-length addition principle for fiber bundles admitting a global section under a freeness condition, and show that stronger hypotheses yield an exact formula. We do an explicit computation using tools from algebraic topology to identify the surviving classes for the quotient parameter space of LoRA, and symmetrized products handle head permutations for MHA. Together, these necessary list sizes establish an architecture-level lower bound to stable parameter reconstruction for both LoRA and MHA.

## 1 INTRODUCTION

Low-Rank Adaptation (LoRA) updates a pretrained weight matrix by a low-rank product $\Delta W = BA$, with $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times d}$ (Hu et al., 2022, §4.1). This description has an exact symmetry because every $g \in GL_r(\mathbb{R})$ sends $(B, A)$ to $(Bg, g^{-1}A)$ without changing $\Delta W$. Multi-head attention (MHA) has analogous symmetries within each head, along with the freedom to permute the heads (Vaswani et al., 2017). These transformations change the stored weights while preserving the modeled update or attention computation. We call these *architectural symmetries*.

A concrete problem here is *parameter recovery*. The above shows that the natural target for parameter recovery therefore should identify all weight descriptions related by these symmetries, which results in a *quotient parameter space*. For LoRA, a point in such a quotient space represents the update $\Delta W$ rather than just a chosen pair of factors. For MHA, it represents an unordered collection of the query-key and value-output products associated with the heads. So in this work, we will be concerned with such (compact) quotient spaces. Note that distinct points of these spaces are distinct modulo the architectural symmetries (the observation process may identify further points when two targets induce the same data distribution, we allow those without weakening our lower bounds).

Let us make the problem of recovery a little more concrete. Say we fix a base architecture and a distribution of prompts or inputs, and observe an unknown LoRA update or MHA component through the resulting population behavior. We then ask how many candidate rules are required to recover such a target (stably) over its entire compact core $X$. Let us write $\mathcal{P}(\mathcal{Z})$ for the space of probability distributions on the observation domain $\mathcal{Z}$. A target $w \in X$ determines a population distribution



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

through a continuous map $G : X \to \mathcal{P}(\mathcal{Z})$. A list of $L$ deterministic candidate rules consists of continuous maps $A_1, \ldots, A_L : \mathcal{P}(\mathcal{Z}) \to X$. We say that it achieves $\epsilon$-accurate proper recovery when every target has at least one successful candidate, that is,

$$\forall w \in X, \qquad \min_{1 \le k \le L} d_X\big(A_k(G(w)), w\big) < \epsilon$$

Here *proper* means that each candidate lies in the same quotient space as the target, and continuity expresses the intuition that a small change in the population distribution produces only a small change in the recovered (quotient) parameter. As an example, one can think of a finite-time gradient-flow procedure with fixed initialization under the usual smooth-dependence assumptions; we make this connection a bit more precise in Remark B.1. We should also remark here that we use population access to remove sampling limitations and study the obstruction imposed by the quotient topology under these global continuity and properness requirements.

One can now ask why this goal of recovery even matters. For this, note that it is useful whenever the generating component itself is the target - for example, teacher-student identification, reconstruction of an adapter from its behavior, comparison of learned components in symmetry-invariant coordinates, and inference over equivalence classes, all lead to a quotient-valued target. Our results give necessary list-size bounds for the parameter-reconstruction objective one can use downstream.

Before we delve into more details, the short version of our findings for the above problem of recovery is that it's hard: even with access to the full population distribution, sufficiently accurate recovery over the entire core quotient requires a large candidate list. That is, no small fixed family of stable rules can provide the universal guarantee studied here.

**How topology controls the list size.** Let us now see how topology enters this recovery problem. For each candidate rule, the information passes through the space of population distributions, that is, we have

$$X \xrightarrow{G} \mathcal{P}(\mathcal{Z}) \xrightarrow{A_k} X, \qquad w \longmapsto G(w) \longmapsto A_k(G(w))$$

The space $\mathcal{P}(\mathcal{Z})$ is convex and hence contractible, so each composition $A_k \circ G$ is null-homotopic (we can continuously deform it to a constant map). Now consider the open set of targets that the $k$th rule recovers within error $\epsilon$, which we call its success domain. On this set, the output stays close to the target. At accuracy below a uniform threshold, this lets us deform the restriction of $A_k \circ G$ to the inclusion of the success domain into $X$. The inclusion is therefore also null-homotopic, and we call such an open set *categorical*. We will show in the corresponding sections that the LoRA and MHA quotient parameter spaces admit such a threshold by virtue of their geometry.

If our list succeeds for every target, these $L$ success domains must cover $X$. We can now ask how many categorical open sets are needed to cover it. The smallest possible number is the Lusternik-Schnirelmann category $\mathrm{cat}(X)$ (we use the convention that counts the sets themselves). Thus $L \ge \mathrm{cat}(X)$. We make this reduction formal in Theorem 3.6.

To turn this into a much more computable numerical lower bound ($\mathrm{cat}$ is very hard to compute), we use cohomology. The cup-length $\mathrm{cup}(X;\mathbb{F})$ is the largest number of positive-degree cohomology classes over a field $\mathbb{F}$ whose product is nonzero. We can use the fact then that

$$L \ge \mathrm{cat}(X) \ge \mathrm{cup}(X;\mathbb{F}) + 1$$

(see Cornea et al. (2003, Proposition 1.5) for the second inequality, shifted by one to match our set-counting convention). As a simple example, a sphere has cup-length one and category two, so this second inequality is also exact.

### 1.1 Main results and contributions

Our first contribution is the reduction from stable population-level recovery to a covering invariant of the quotient parameter space, which we formalize in Theorem 3.6. It applies to deterministic continuous candidate rules and depends only on the information flow as we discussed above. This formulation counts candidate recovery branches. When an implementation produces one candidate per optimization run, the list size becomes a corresponding lower bound on the required runs. Procedures that generate several candidates jointly are also subject to the same output-list requirement.

Our second contribution is an addition principle for the cup-length of fiber bundles admitting a global section, which we prove in Theorem 4.1. In part (i) of Theorem 4.1, we show that $s$ positive-degree

classes extend a nonzero product from the base by $s$ factors when their square-free products remain independent over the base cohomology, giving us a lower bound on the cup-length. Under the stronger hypotheses in part (ii), we also obtain a matching upper bound, giving an exact formula. We believe both these results might be of independent interest.

Let us now mention some details about the two architectures we work on.

**LoRA.** We first bound the nonzero singular values away from zero and infinity, giving a compact model of the exact-rank update space. This deformation retracts onto

$$X_r = \big(V_r(\mathbb{R}^d) \times V_r(\mathbb{R}^d)\big)/O(r)$$

where $V_r(\mathbb{R}^d)$ is the space of orthonormal $r$-frames[1]. The space $X_r$ is a bundle over the real Grassmannian $\mathrm{Gr}(r, d)$. For even $d$ and $r$, the base contributes $r(d-r)/4$ factors to a nonzero cup product. By an explicit cohomology computation, we then show that $r/2 - 1$ additional fiber classes survive. Consequently, we obtain the cup-length estimate below and the list-size lower bound

$$\mathrm{cup}(X_r; \mathbb{Q}) \geq \frac{r(d-r)}{4} + \frac{r}{2} - 1, \qquad L \geq \frac{r(d-r)}{4} + \frac{r}{2}$$

in Theorem 5.7. So for example, for $d = 4096$ and $r = 64$, every sufficiently accurate stable recovery guarantee over the entire LoRA core $X_r$ requires at least $64544$ candidates. This thus gives a $L = \Omega(rd)$ in the low-rank regime.

**Multi-head attention.** A single attention head is represented by two bounded exact-rank products, one for query-key interaction and one for value-output interaction. For even $d$ and equal even key/value dimension $r$, its core is $X_r \times X_r$. An $h$-head layer is obtained from $(X_r \times X_r)^h$ by identifying permutations of its $h$ factors. The finite quotient generally has singular points where heads coincide. Working over $\mathbb{Q}$, we form permutation-invariant products from classes on the ordered space and use rational transfer to obtain nonzero classes on the quotient. This gives the cup-length and list-size bounds in Theorems 6.6 and 6.8, respectively

$$\mathrm{cup}(X_{\mathrm{MHA}}; \mathbb{Q}) \geq \frac{hr(d-r)}{2} + (h+1)\left(\frac{r}{2} - 1\right), \qquad L \geq \frac{hr(d-r)}{2} + (h+1)\left(\frac{r}{2} - 1\right) + 1$$

Again for an example, for $d = 768$, $r = 64$, and $h = 12$, the lower bound is $270740$. Hence MHA recovery requires lists of size $\Omega(hrd)$.

Let us remark here that our restriction to even $d$ and $r$ is because these cases admit the uniform rational-cohomology calculation used here. Odd cases require a separate calculation.

Our LS-category estimates also have the correct order in the low-rank regime and we give a dimension comparison in Remark 6.9.

To reiterate, these theorems give an overall structural negative result for global parameter reconstruction: sufficiently accurate recovery over the entire LoRA or MHA quotient space requires a large fixed list of stable candidate rules. If each candidate is generated by a separate expensive optimization run, the necessary list makes that implementation correspondingly burdensome. The measured quantity is list size; runtime, query complexity, and finite-sample complexity are complementary directions we do not study in this paper. In fact, the scope of the paper also points towards focused formulations that may avoid this lower bound - for example, restricting the target family, working locally, choosing a coarser accuracy, or posing the downstream objective directly in function space.

## 2 Brief Related Work and Comparisons

**Parameter recovery and low-rank geometry.** There are several algorithmic precedents for parameter recovery from behavior modulo symmetries. For example, the Expand-and-Cluster method of Martinelli et al. (2024) probes a target function, trains multiple overparameterized students, and clusters learned weights to recover incoming weight vectors up to activation-dependent sign or scaling. This provides a concrete example of parameter recovery using multiple training runs that we study.

[1] An orthonormal $r$-frame of $\mathbb{R}^d$ is an ordered list of $r$ orthonormal vectors in $\mathbb{R}^d$.

Our population-level theory isolates a uniform stability requirement for the LoRA and MHA quotient spaces. Geometrically, the LoRA target is a bounded part of the manifold of exact-rank matrices, connecting the result to fixed-rank matrix optimization (Vandereycken, 2013). The observation map in our problem is induced by model behavior, which distinguishes this setting from recovery through a prescribed collection of linear matrix measurements.

**Symmetry and mode connectivity.** Zhao et al. (2025) study how parameter symmetries connect selected zero-loss minima. Their exact component and permutation-connectivity results concern square full-rank linear networks with invertible weights and full-rank data, with width at least two required for the permutation-connectivity result; their explicit skip-connection count is a one-dimensional example and they leave the complete characterization of nonlinear minima as open. Our work and theirs therefore use complementary idealizations. Zhao et al. analyze connectivity among loss minima under network symmetries. Whereas, we analyze stable uniform recovery over the architecture-specific LoRA and MHA quotient spaces, idealizing learning through deterministic continuous branches that receive population observations.

**List replicability.** List replicability was introduced independently by Chase et al. (2023) and Dixon et al. (2023). Chase et al. established its equivalence with global stability and their lower-bound argument uses the Poincaré-Miranda theorem to obtain a topological obstruction in a learning-theoretic setting. We retain the central idea of a bounded output list and study a different source of complexity: nonzero cohomology products in continuous architecture-specific quotient spaces. This makes the resulting lower bounds depend explicitly on the embedding dimension, rank, and number of heads.

The broader literature also studies neural-network symmetries and identifiability (Simsek et al., 2021; Vlačić & Bölcskei, 2021; Bona-Pellissier et al., 2022), topology and mode connectivity in loss landscapes (Garipov et al., 2018; Draxler et al., 2018; Ainsworth et al., 2023), and the geometry of model and adapter merging (He et al., 2025; Shah et al., 2025; Zhang & Zhou, 2025). Our recovery theorem yields a formal implication for methods in these areas whenever their pipeline contains a stage that performs stable, uniform recovery of a quotient parameter from population observations; Corollary 7.1 states it precisely for LoRA. We give a fuller comparison and review in Section A.

## 3 Learning and Topological Preliminaries

Here we give the required definitions and the recovery-to-topology reduction. Let us start by formalizing mathematically the reconstruction problem introduced in Section 1. Let $(X, d_X)$ be a compact metric space of parameters (the points here are the targets for recovery) after quotienting by the architectural symmetries. Let $\mathcal{Z}$ be a metric space of observations and let $\mathcal{P}(\mathcal{Z})$ denote the Borel probability measures on $\mathcal{Z}$, equipped with the weak topology. This space is convex because for any $\mu_0, \mu_1 \in \mathcal{P}(\mathcal{Z})$ and $t \in [0, 1]$, the mixture $(1-t)\mu_0 + t\mu_1$ is again a probability measure. It is therefore contractible.

**Definition 3.1.** A *population map* is a continuous map $G : X \to \mathcal{P}(\mathcal{Z})$ that sends a target $w$ to its population observation law $G(w)$. A *stable recovery branch* is a continuous map $A : \mathcal{P}(\mathcal{Z}) \to X$.

Note that the map $G$ need not be injective because a chosen observation law may identify additional points after the modeled architectural symmetries have been removed. Our lower bound is conditional on the uniform recovery property below. Further observational non-identifiability can only make that property harder to satisfy so we are not losing anything (as mentioned previously, this does not weaken the lower bound).

**Definition 3.2.** Continuous branches $A_1, \dots, A_L : \mathcal{P}(\mathcal{Z}) \to X$ achieve *$L$-list $\epsilon$-accurate stable proper recovery* if, for every $w \in X$, at least one branch returns a point of $X$ within distance $\epsilon$ of $w$, that is, $d_X(A_k(G(w)), w) < \epsilon$ for some $k \in \{1, \dots, L\}$.

The word *proper* says that every candidate lies in the target space $X$. The formal quantity $L$ counts candidate recovery maps. We remark that a subsequent choice among these candidates may be discontinuous, since our recovery guarantee only requires the list to contain an accurate candidate.

**Definition 3.3.** The success domain of branch $A_k$ is

$$U_k := \{w \in X : d_X(A_k(G(w)), w) < \epsilon\}$$

Each $U_k$ is open by continuity. The uniform recovery condition says that $U_1, \dots, U_L$ cover $X$.

We next measure how many topologically simple open sets are required for such a cover. An open set $U \subseteq X$ is *categorical in* $X$ if its inclusion $U \hookrightarrow X$ is homotopic to a constant map.

**Definition 3.4.** The *Lusternik-Schnirelmann category* $\mathrm{cat}(X)$ is the least number of categorical open sets that cover $X$. We use the unnormalized convention, in which a nonempty contractible space has category 1.

The required accuracy is controlled by a local topological number. The following definition covers both smooth manifolds and their finite symmetry quotients.

**Definition 3.5.** A number $\rho > 0$ is a *uniform homotopy radius* for $(X, d_X)$ if, for every topological space $T$ and every pair of continuous maps $f_0, f_1 : T \to X$ satisfying $d_X(f_0(t), f_1(t)) < \rho$ for all $t \in T$, the maps $f_0$ and $f_1$ are homotopic.

Every compact Euclidean neighborhood retract (ENR)[2] has a positive uniform homotopy radius - this includes compact manifolds and quotients of compact smooth manifolds by smooth finite-group actions. For a compact Riemannian manifold, any positive number below its convexity radius is a uniform homotopy radius. We prove these facts in Section B.2. Now the main counting result.

**Theorem 3.6** (Recovery and LS category)**.** *Let $(X, d_X)$ be a compact metric ENR, let $\rho_X > 0$ be a uniform homotopy radius, and let $G : X \to \mathcal{P}(\mathcal{Z})$ be continuous. If $L$ continuous branches achieve $L$-list $\epsilon$-accurate stable proper recovery with $\epsilon < \rho_X$, then $L \geq \mathrm{cat}(X)$.*

**Remark 3.7.** Interestingly, if $G$ is injective and $\mathcal{P}(\mathcal{Z})$ is metrizable, this bound is actually attained by $\mathrm{cat}(X)$ continuous branches that collectively recover every target exactly. We give the proof in Proposition C.1.

We now use cohomology to turn this covering obstruction into a computable number. We use singular cohomology and its cup product as in Hatcher (2002, §3.2).

**Definition 3.8.** For a commutative coefficient ring $R$, the *cup-length* $\mathrm{cup}(X; R)$ is the greatest integer $q$ for which there are positive-degree classes $\alpha_1, \dots, \alpha_q \in H^{>0}(X; R)$ satisfying $\alpha_1 \smile \cdots \smile \alpha_q \neq 0$; equivalently, it is the number in Cornea et al. (2003, Definition 1.4).

We will also use this classical cup-length estimate, translated to our unnormalized convention, which states that $\mathrm{cat}(X) \geq \mathrm{cup}(X; R) + 1$, see for example Cornea et al. (2003, Proposition 1.5).

## 4 Cup-length in sectioned fiber bundles

We begin with an addition principle for cup-length in a sectioned fiber bundle. The idea is that independently surviving fiber classes lengthen a nonzero cup product pulled back from the base. There are some subtleties to make this idea work, so we characterize this in two forms. We first show that a module-freeness condition suffices for our lower bounds; then we give an exact formula under a stronger spectral-sequence hypotheses that produce a full exterior tower. For the proof, we use the classical Serre spectral sequence and Leray-Hirsch theorem, see Hatcher (n.d., Theorem 5.15) and Hatcher (2002, Theorem 4D.1). This might be of independent interest to a wide range of communities for computing/bounding the cup length of certain spaces.

**Theorem 4.1** (Cup-length addition)**.** *Let $p : X \to B$ be a fiber bundle of spaces having CW type over a path-connected base $B$, with fiber $F$ and a global section $s : B \to X$. Fix a field $\mathbb{F}$, and set $c_B = \mathrm{cup}(B; \mathbb{F}) < \infty$.*

(i) Split-tower lower bound. *Suppose there are homogeneous classes $u_1, \dots, u_t \in H^{>0}(X; \mathbb{F})$ for which the $H^*(B; \mathbb{F})$-linear map*

$$\Phi : H^*(B; \mathbb{F}) \otimes_{\mathbb{F}} \Lambda_{\mathbb{F}}(e_1, \dots, e_t) \longrightarrow H^*(X; \mathbb{F}), \qquad a \otimes e^K \longmapsto p^*(a) \smile u^K$$

*is injective, where $e^K$ and $u^K$ denote the corresponding square-free monomials. Then*

$$\mathrm{cup}(X; \mathbb{F}) \geq c_B + t$$

[2] Briefly, a *Euclidean neighborhood retract* (ENR) if it is homeomorphic to a subspace of a Euclidean space for which some open neighborhood admits a continuous retraction. See Definition B.2.

(ii) Equality under collapse. *Assume that the local coefficient system is trivial, the Serre spectral sequence collapses at $E_2$, and $H^*(F;\mathbb{F}) \cong \Lambda_{\mathbb{F}}(x_1,\dots,x_r)$ with every $x_i$ of positive degree. Then classes $u_1,\dots,u_r$ satisfying part* (i) *exist, and*

$$\operatorname{cup}(X;\mathbb{F}) = c_B + r$$

We prove this in Appendix C.

As we will see later, for LoRA, we can take $X = X_r$, $B = \mathrm{Gr}(r,d)$, and $F = V_r(\mathbb{R}^d)$. In Section 5, we construct a global section and the surviving classes needed for part (i), which add to the cup-length of the base.

Let us now compute the base contribution for LoRA. We give a self-contained derivation of the rational cup-length of the real Grassmannian, using its classical cohomology presentation and the comparison with the complex Grassmannian.

**Lemma 4.2.** *Let the embedding dimension $d$ and rank $r$ be positive even integers with $0 < r < d$, and write $r = 2k$ and $d = 2n$. The cup-length of the unoriented real Grassmannian* $\mathrm{Gr}(r,d)$ *is*

$$\operatorname{cup}(\mathrm{Gr}(r,d);\mathbb{Q}) = k(n-k) = \frac{r(d-r)}{4}$$

## 5 The LoRA Parameter Space and Lower Bounds

Let us now focus on LoRA for this section. LoRA, or Low-Rank Adaptation, reduces the computational burden of fine-tuning by constraining each weight update to a low-rank factorization. The update to a pretrained weight matrix $W_0$ is parameterized as $\Delta W = BA$, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times d}$, so the effective weight is $W_0 + BA$, see (Hu et al., 2022, §4.1). When both factors have rank $r$, the matrices $B$ and $A^\top$ belong to the non-compact Stiefel manifold of linearly independent $r$-frames, denoted $\widetilde{V}_r(\mathbb{R}^d)$.

This factorization is not unique because for any $g \in GL_r(\mathbb{R})$, the pair $(Bg, g^{-1}A)$ gives the same update. Equivalently, in frame coordinates, the action is $(B, A^\top)\cdot g = (Bg, A^\top g^{-\top})$. Quotienting the full-rank factor pairs by this action gives the smooth manifold $\mathcal{M}_r^\circ$ of $d\times d$ matrices of exact rank $r$ (Mishra et al., 2014, §3.1 and §4.1). The multiplication map $(B, A^\top) \mapsto BA$ is in fact a principal $GL_r(\mathbb{R})$-bundle projection (see also Remark 5.5 where we give a direct local trivialization).

This geometric perspective, viewing the LoRA parameter space as a fixed-rank matrix manifold, has been an object of study for a while now. It was developed for low-rank optimization by Vandereycken (2013), for matrix-manifold optimization by Absil et al. (2008), and for dynamical low-rank approximation by Billaud-Friess et al. (2022). A well-known algoithm, StelLA, applies Stiefel-manifold optimization directly to LoRA fine-tuning (Li et al., 2025). Let us now formalize this geometry which we will use in our recovery problem for LoRA (and by extension, MHA).

**Definition 5.1.** Let $d$ be the embedding dimension and $r < d$ be the LoRA rank. The *observable parameter space* of a rank-$r$ LoRA update, denoted $\mathcal{M}_r$, is the set of all matrices in $\mathbb{R}^{d\times d}$ of exact rank $r$ whose singular values $\sigma_1,\dots,\sigma_r$ satisfy $0 < c \le \sigma_i \le R$ for fixed constants $0 < c \le 1 \le R$.[3]

**Remark 5.2.** The lower bound $c$ keeps the target family inside one exact-rank stratum, while the upper bound $R$ controls its scale and makes the observable model compact. We put the condition $1 \in [c,R]$ so that the partial-isometry core is placed inside this model and ensures that the deformation in Lemma 5.3 stays within $\mathcal{M}_r$. The resulting homotopy type, and hence the list-size bound, is independent of the numerical values of $c$ and $R$ subject to these conditions.

**Lemma 5.3.** *The observable parameter space $\mathcal{M}_r$ strongly deformation retracts onto a core space that is canonically homeomorphic to the quotient manifold*

$$X_r := \frac{V_r(\mathbb{R}^d)\times V_r(\mathbb{R}^d)}{O(r)}$$

[3]The unrestricted fixed-rank space $\mathcal{M}_r^\circ$ is a smooth manifold of dimension $(2d-r)r$; see Vandereycken (2013, §2.1 and Proposition 2.1). The singular-value inequalities make $\mathcal{M}_r$ a compact semialgebraic subset of $\mathcal{M}_r^\circ$. Scalar clipping of the positive singular values to $[c,R]$ gives a strong deformation retraction $\mathcal{M}_r^\circ \to \mathcal{M}_r$, and Lemma 5.3 retracts $\mathcal{M}_r$ further onto the smooth compact core $X_r$.

*where $V_r(\mathbb{R}^d) = \{U \in \mathbb{R}^{d\times r} : U^\top U = I_r\}$ is the (compact) Stiefel manifold of orthonormal $r$-frames in $\mathbb{R}^d$, and the orthogonal group $O(r)$ acts freely on the product by the diagonal right action $(U, V) \cdot g = (Ug, Vg)$.*

**Remark 5.4.** In Figure 3 we illustrate the LoRA core for $r = 1$ and $d = 2$. In this case $V_1(\mathbb{R}^2) = S^1$, and we construct the quotient of $S^1 \times S^1$ by the diagonal $O(1) \cong \mathbb{Z}/2$ action from a fundamental domain.

**Remark 5.5.** The LoRA parameterization $(B, A) \mapsto BA$ realizes the fixed-rank manifold $\mathcal{M}_r^\circ$ as a quotient. The non-compact Stiefel manifold $\widetilde{V}_r(\mathbb{R}^d) = \{B \in \mathbb{R}^{d\times r} : \operatorname{rank} B = r\}$ parameterizes each frame factor. The fibers of the multiplication map $\mu(B, A) = BA$ are exactly the orbits of the free right action $(B, A) \cdot g = (Bg, g^{-1}A)$ (this quotient representation is also mentioned by Mishra et al. (2014, §3.1)).

For completeness, note that the principal-bundle charts can be written directly. Choose index sets $I, J \subseteq \{1, \dots, d\}$ of size $r$, and let $U_{I,J} \subseteq \mathcal{M}_r^\circ$ consist of matrices $Z$ for which the minor $Z_{I,J}$ is invertible. On this open set define $B_0(Z) = Z_{:,J}$ and $A_0(Z) = Z_{I,J}^{-1} Z_{I,:}$. The rank condition and the invertible pivot give $Z = B_0(Z)A_0(Z)$, so $Z \mapsto (B_0(Z), A_0(Z))$ is a smooth local section of $\mu$. Every factorization $Z = BA$ has a unique expression $B = B_0(Z)g$ and $A = g^{-1}A_0(Z)$, with $g = Z_{I,J}^{-1} B_{I,:} \in GL_r(\mathbb{R})$. Hence the map

$$\begin{aligned} U_{I,J} \times GL_r(\mathbb{R}) &\longrightarrow \mu^{-1}(U_{I,J}) \\ (Z, g) &\longmapsto (B_0(Z)g, g^{-1}A_0(Z)) \end{aligned}$$

is a smooth equivariant trivialization. The sets $U_{I,J}$ cover $\mathcal{M}_r^\circ$, showing that $\mu$ is a principal $GL_r(\mathbb{R})$-bundle. On the maximal compact subgroup $O(r) \subset GL_r(\mathbb{R})$, the frame-coordinate identity $g^{-\top} = g$ gives the diagonal action used to define $X_r$. We realize the corresponding compact core directly on the quotient matrix space using the same polar deformation in Lemma 5.3.

To compute the cup-length of this quotient space, we first show that the LoRA bundle admits a global continuous section, and then state our theorem.

**Lemma 5.6.** *The LoRA parameter manifold $X_r = (V_r(\mathbb{R}^d) \times V_r(\mathbb{R}^d))/O(r)$, viewed as a fiber bundle over the Grassmannian $\operatorname{Gr}(r, d)$ via the projection $\pi([U, V]) = \operatorname{span}(U)$, admits a continuous global section.*

We can now state our lower bound for LoRA.

First, note that the quotient $X_r$ is a compact smooth manifold and hence a compact ENR. Let $\rho_{X_r} > 0$ be a uniform homotopy radius supplied by Lemma B.3. For Frobenius distance between the matrices $UV^\top$ representing $[U, V] \in X_r$, we may take $\rho_{X_r} = 1$ and thus cover the $\epsilon < 1$ case completely in the theorem below. See Proposition C.2.

**Theorem 5.7** (LoRA recovery bound)**.** *Let $X_r$ be the rank-$r$ LoRA parameter manifold. Assume the embedding dimension $d$ and rank $r$ are positive even integers with $r < d$. Any family of continuous recovery branches achieving $L$-list $\epsilon$-accurate stable proper recovery on $X_r$ for $\epsilon < \rho_{X_r}$ must have a list size satisfying*

$$L \geq \frac{r(d-r)}{4} + \frac{r}{2}$$

We prove this in Section C.

The upshot of the above theorem can be seen via the following example. For $d = 4096$ and $r = 64$, the recovery guarantee over the entire LoRA core quotient $X_r$ requires at least $64544$ continuous candidate rules. Note that this is a necessary *list size*. Whether this bound is attained, and the runtime or sample complexity of a recovery family, need to be investigated separately.

## 6 The MHA Parameter Space and Lower Bounds

In this secton, we focus on Multi-Head Attention (MHA). MHA combines two architectural symmetries. The first is that within each head, changes of basis in the query-key and value-output factorizations leave the relevant matrix products unchanged. The second is that across heads, permuting the head labels leaves their summed output unchanged. We first identify the product pair associated with one head and then the permutation symmetry for an ordered collection of heads.

### 6.1 The Parameter Space of a Single Attention Head

Let us begin with one head. In the parameterization studied here, its attention scores and output depend on the weights through two low-rank matrix products, so the bounded exact-rank spaces from Definition 5.1 apply directly.

Let $d$ be the embedding dimension, and let $d_k, d_v < d$ be the query-key and value head dimensions. The raw weights of one head are $W^Q, W^K \in \mathbb{R}^{d\times d_k}$ and $W^V \in \mathbb{R}^{d\times d_v}$. Partitioning the rows of the standard multi-head output matrix into blocks gives a matrix $W_i^O \in \mathbb{R}^{d_v\times d}$ for the $i$th head. Let $\mathrm{Concat}(\mathrm{head}_1,\dots,\mathrm{head}_h)W^O$ be the sum of the terms $\mathrm{head}_i\, W_i^O$ (Vaswani et al., 2017, §3.2.2). Suppressing the head index for the moment, the two observable bilinear products are therefore $P = W^Q(W^K)^\top \in \mathbb{R}^{d\times d}$ and $M = W^V W^O \in \mathbb{R}^{d\times d}$.

**Definition 6.1.** For $s \in \{d_k, d_v\}$, let $\mathcal{M}_s$ be the bounded space of exact-rank-$s$ matrices from Definition 5.1. The observable parameter space of one attention head is $\mathcal{M}_{\mathrm{head}} = \mathcal{M}_{d_k} \times \mathcal{M}_{d_v}$.

**Lemma 6.2.** *For $s \in \{d_k, d_v\}$, the space $\mathcal{M}_s$ strongly deformation retracts onto $X_s = (V_s(\mathbb{R}^d) \times V_s(\mathbb{R}^d))/O(s)$. Consequently, $\mathcal{M}_{\mathrm{head}}$ strongly deformation retracts onto $X_{d_k} \times X_{d_v}$.*

### 6.2 The Full MHA Orbifold

We now assemble the full attention layer. For an input sequence $Z \in \mathbb{R}^{N\times d}$, the $i$th head contributes

$$\mathrm{Head}_i(Z) = \mathrm{softmax}\left(\frac{ZP_iZ^\top}{\sqrt{d_k}}\right) ZM_i$$

and the block output in this parameterization is $\mathrm{MHA}(Z) = \sum_{i=1}^h \mathrm{Head}_i(Z)$. The sum is unchanged when the pairs $(P_i, M_i)$ are reordered. Thus, after the within-head symmetries have been quotiented out, what remains is the unordered collection of these $h$ product pairs.

The permutation action has fixed points whenever two or more heads coincide. The resulting quotient is therefore generally singular along the corresponding diagonal strata. This is the reason why we will later use Theorem B.8 so that we can work directly with the finite quotient.

**Definition 6.3.** For an $h$-head Transformer, let $Y = (X_{d_k} \times X_{d_v})^h$ be the ordered core parameter space. The symmetric group $S_h$ acts on $Y$ by permuting its $h$ single-head factors. The modeled MHA quotient is $X_{\mathrm{MHA}} = Y/S_h$.

The product of invariant Riemannian metrics on the factors gives an $S_h$-invariant metric $d_Y$ on $Y$. We equip the quotient with $d_X([y_1],[y_2]) = \min_{\pi\in S_h} d_Y(y_1, \pi\cdot y_2)$. Because $Y$ is a compact smooth manifold and the action is finite, $X_{\mathrm{MHA}}$ is a compact global-quotient orbifold and a compact ENR by Proposition B.4.

**Remark 6.4.** The product of the single-head retractions in Lemma 6.2 is $S_h$-equivariant. It therefore descends to a deformation retraction from the bounded observable quotient $(\mathcal{M}_{\mathrm{head}})^h/S_h$ onto $X_{\mathrm{MHA}}$. Thus the core quotient has the same homotopy type as the bounded observable quotient.

**Remark 6.5.** We provide a couple of schematics for $X_{\mathrm{MHA}}$ in Figures 4 and 5. The first illustrates the quotient for $h = 2$ and its singular diagonal stratum. The second visualizes how head permutations identify the fundamental domains in the three-head case.

### 6.3 Lower bound for MHA

As we mentioned earlier briefly, for two or more heads, the quotient has singular strata where heads coincide, so we work directly with the finite quotient. Over $\mathbb{Q}$, Theorem B.8 identifies the cohomology of $X_{\mathrm{MHA}}$ with the $S_h$-invariant cohomology of the ordered space $Y$. This lets us construct a long nonzero product before descending it to the quotient.

**Theorem 6.6.** *Assume that $d$ and $r$ are positive even integers with $r < d$, and that $d_k = d_v = r$. Let $c_{\mathrm{base}} = \mathrm{cup}(\mathrm{Gr}(r,d);\mathbb{Q})$. Then the MHA quotient satisfies* $\mathrm{cup}(X_{\mathrm{MHA}};\mathbb{Q}) \geq 2h\, c_{\mathrm{base}} + (h+1)\left(\frac{r}{2}-1\right)$.

Let us quickly remark that $X_{\mathrm{MHA}}$ is a compact ENR.

**Remark 6.7.** The ordered core $Y$ is a compact smooth manifold and $S_h$ acts smoothly by permuting its factors. By Proposition B.4, the quotient $X_{\mathrm{MHA}} = Y/S_h$ is a compact ENR. It therefore has a positive uniform homotopy radius $\rho_{\mathrm{MHA}} > 0$ by Lemma B.3, so Theorem 3.6 applies.

We can now state our lower bound for MHA.

**Theorem 6.8** (MHA recovery bound)**.** *Assume that $d$ and $r$ are positive even integers with $r < d$, and let $X_{\mathrm{MHA}}$ be the quotient for $h$ heads with $d_k = d_v = r$. Let $\rho_{\mathrm{MHA}} > 0$ be a uniform homotopy radius as in Remark 6.7. Any family of continuous recovery branches achieving $L$-list $\epsilon$-accurate stable proper recovery on $X_{\mathrm{MHA}}$ for $\epsilon < \rho_{\mathrm{MHA}}$ must have*

$$L \geq \frac{hr(d-r)}{2} + (h+1)\left(\frac{r}{2} - 1\right) + 1$$

We prove this in Section C.

**Remark 6.9.** One can now ask how much room there is to improve these LS-category estimates. In the common low-rank regime, they already have the correct order. Under our convention, we have the estimate $\mathrm{cat}(X) \leq \dim(X) + 1$ (Cornea et al., 2003, Theorem 1.7). The LoRA core has dimension $2dr - (3r^2 + r)/2$, and a finite head-permutation quotient preserves the dimension of the ordered MHA core. For $r \leq \alpha d$ with fixed $\alpha < 1$, these upper bounds are $O(rd)$ and $O(hrd)$, respectively, matching our $\Omega(rd)$ and $\Omega(hrd)$ lower bounds up to constants.

## 7 Conclusion

In this paper, we studied stable recovery of LoRA updates and Multi-Head Attention parameters after identifying weight descriptions related by their architectural symmetries. We showed that for even embedding dimension $d$ and even rank $r$, the LoRA list size satisfies $L \geq r(d-r)/4 + r/2$; for MHA with $h$ attention heads and with equal even key/value dimension $r$, it satisfies $L \geq hr(d-r)/2 + (h+1)(r/2-1) + 1$. These list sizes grow as $\Omega(rd)$ and $\Omega(hrd)$, respectively.

Our results are mostly negative (for the stated problem), and is especially concrete for LoRA. At $d = 4096$ and $r = 64$, a universal stable recovery guarantee over the entire core quotient $X_r$ requires at least $64544$ candidates, even when the recovery rules receive the full population distribution. Note that the theorem lower bounds list-size, whereas runtime, query complexity, finite-sample complexity, and stochastic recovery are complementary quantitative questions that one can look at further.

We can also ask what these bounds imply for downstream alignment or merging. If a pipeline contains a stage that performs stable, uniform proper recovery, the same obstruction applies to that stage, regardless of whether its output is subsequently aligned, compared, or merged. The following corollary makes this precise and follows immediately from Theorem 3.6 and the cup-length estimate established in the proof of Theorem 5.7.

**Corollary 7.1.** *Let $X_r$ be the rank-$r$ LoRA quotient core for positive even integers $d$ and $r$ with $r < d$, and let $\rho_{X_r} > 0$ be as in Theorem 5.7. Suppose a downstream alignment or merging pipeline contains a continuous population map $G : X_r \to \mathcal{P}(\mathcal{Z})$ and continuous candidate rules $A_1, \ldots, A_L : \mathcal{P}(\mathcal{Z}) \to X_r$ such that, for every $w \in X_r$, some candidate satisfies $d_{X_r}(A_k(G(w)), w) < \epsilon$. If $\epsilon < \rho_{X_r}$, then*

$$L \geq \mathrm{cat}(X_r) \geq \frac{r(d-r)}{4} + \frac{r}{2}$$

Since alignment reconciles coordinate descriptions of the same modeled quotient point, while combination assembles components associated with potentially different quotient points, we can then ask how the picture changes for predictor-level combination, restricted target families, local recovery, coarser tolerances, or procedures that retain multiple candidates.

Finally, note that the quotient spaces and the recovery reduction exist in all parity cases. We briefly touched on this before. It would be interesting to see how the analysis of Euler and Pontryagin classes and orientation reversal goes in the odd case. Other natural directions for future work include incorporating torsion coefficients, extracting further permutation-invariant fiber contributions for MHA, and extending the population-level framework to stochastic and finite-sample recovery. Extending this analysis to even more architectures with richer symmetries should also be interesting. Last but not least, studying weaker stability requirements and their effect on the bounds could bring the theory closer to practice.

### AI USE STATEMENT

We used GPT-5.6 in Codex for checking the mathematical claims and proofs in the paper. We also used the same for verification of bibliograpic data and some LaTeX formatting. It also identified the identifiable upper bound (Proposition C.1) and the Frobenius radius statement (Proposition C.2) as improvements for the paper, which we found interesting enough to keep. It also suggested the proofs for these, which were correct but pretty clumsy, so we rewrote them for better exposition. We also used Claude 5.5 to independently verify the proofs and it did not find any issues (which we do not take as a proof of correctness, but as a heuristic). We reviewed the draft and citations ourselves independently. We did not use generative AI to develop the theoretical framework or formulate any of the main technical claims or proofs, which are the authors' own work. AI use for implementing methods is not applicable for us because this work is theoretical and has no computational implementation. Responsibility for the final manuscript, including all text and claims, rests with us, the authors.

## REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging Models modulo Permutation Symmetries, 2023. arXiv:2209.04836 [cs.LG].

Danil Akhtiamov and Matt Thomson. Connectedness of loss landscapes via the lens of Morse theory. In *Proceedings of the 1st NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 197, 2023.

Marie Billaud-Friess, Antonio Falcó, and Anthony Nouy. Principal Bundle Structure of Matrix Manifolds. *Mathematics*, 9(14):1669, 2021.

Marie Billaud-Friess, Antonio Falcó, and Anthony Nouy. A new splitting algorithm for dynamical low-rank approximation motivated by the fibre bundle structure of matrix manifolds. *BIT Numerical Mathematics*, 62(2), 2022.

Bruce Blackadar. The Homotopy Lifting Theorem for Semiprojective $C^*$-Algebras. *Mathematica Scandinavica*, 118(2), 2016.

Ari Blondal, Gao Shan, Hamed Hatami, and Pooya Hatami. Stability and List-Replicability for Agnostic Learners. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291, 2025.

Joachim Bona-Pellissier, François Malgouyres, and François Bachoc. Local identifiability of deep ReLU neural networks: the theory. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, art. 1998, 2022.

Armand Borel. Sur La Cohomologie des Espaces Fibres Principaux et des Espaces Homogenes de Groupes de Lie Compacts. *Annals of Mathematics*, 57, 1953.

Armand Borel. Kählerian Coset Spaces of Semisimple Lie Groups. *Proceedings of the National Academy of Sciences of the United States of America*, 40(12), 1954.

Glen E. Bredon. Chapter III: Homological Theory of Finite Group Actions. In *Introduction to Compact Transformation Groups*, volume 46. Elsevier, 1972.

Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability Is Stable: Connections between Replicability, Privacy, and Adaptive Generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 2023.

Zachary Chase, Shay Moran, and Amir Yehudayoff. Stability and Replicability in Learning. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, 2023.

Zachary Chase, Bogdan Chornomaz, Shay Moran, and Amir Yehudayoff. Local Borsuk-Ulam, Stability, and Replicability. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, 2024.

Anna Choromanska, MIkael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun. The Loss Surfaces of Multilayer Networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, 2015.

Octav Cornea, Gregory Lupton, John Oprea, and Daniel Tanré. *Lusternik-Schnirelmann category*. American Mathematical Society, Providence, RI, 2003.

Christopher De Sa, Satyen Kale, Jason D Lee, Ayush Sekhari, and Karthik Sridharan. From Gradient Flow on Population Loss to Learning with Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Peter Dixon, A. Pavan, Jason Vander Woude, and N. V. Vinodchandran. List and Certificate Complexities in Replicable Learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially No Barriers in Neural Network Energy Landscape. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 2018.

Eric Eaton, Marcel Hussing, Michael Kearns, and Jessica Sorrell. Replicable Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks. In *International Conference on Learning Representations*, 2022.

Hossein Esfandiari, Alkis Kalavasis, Amin Karbasi, Andreas Krause, Vahab Mirrokni, and Grigoris Velegkas. Replicable Bandits. In *The Eleventh International Conference on Learning Representations*, 2023.

Farber. Topological Complexity of Motion Planning. *Discrete & Computational Geometry*, 29(2), 2003.

Yves Félix, John Oprea, and Daniel Tanré. *Algebraic Models in Geometry*. Oxford University Press, 2008.

Rudolf Fritsch and Renzo Piccinini. *Cellular Structures in Topology*. Cambridge University Press, 1990.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew G Wilson. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Philip Hartman. *Ordinary differential equations*. Society for Industrial and Applied Mathematics (SIAM), second edition, 2002.

Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002. Online corrected version; theorem numbering follows the version available on the author's website.

Allen Hatcher. Spectral Sequences. Unpublished supplementary chapter to Algebraic Topology, n.d.

Chen He. Localization of equivariant cohomology rings of real Grassmannians, 2017. arXiv:1609.06243v3 [math.AT].

Yifei He, Yuzheng Hu, Yong Lin, Tong Zhang, and Han Zhao. Localize-and-Stitch: Efficient Model Merging via Sparse Task Arithmetic. *Transactions on Machine Learning Research*, 2025.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Dale Husemöller. *Fibre bundles*. Springer-Verlag, third edition, 1994.

Sören Illman. Smooth equivariant triangulations of G-manifolds for G a finite group. *Mathematische Annalen*, 233(3), 1978.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, 2022.

Amin Karbasi, Grigoris Velegkas, Lin Yang, and Felix Zhou. Replicability in Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Sungyoon Kim, Aaron Mishkin, and Mert Pilanci. Exploring The Loss Landscape Of Regularized Neural Networks Via Convex Duality. In *International Conference on Learning Representations*, volume 2025, 2025.

Zhizhong Li, Sina Sajadmanesh, Jingtao Li, and Lingjuan Lyu. StelLA: Subspace Learning in Low-rank Adaptation using Stiefel Manifold. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Flavio Martinelli, Berfin Şimşek, Wulfram Gerstner, and Johanni Brea. Expand-and-cluster: parameter recovery of neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

John Willard Milnor and James Stasheff. *Characteristic classes*. Princeton University Press,University of Tokyo Press, 1974.

Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and Riemannian low-rank optimization. *Computational Statistics*, 29(3), 2014.

Jeffrey Pennington and Yasaman Bahri. Geometry of Neural Network Loss Surfaces via Random Matrix Theory. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2017.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. ZipLoRA: Any Subject in Any Style by Effectively Merging LoRAs. In *Computer Vision – ECCV 2024*, 2025.

Eva Silverstein, Daniel Kunin, and Vasudev Shyam. Symmetry Breaking in Transformers for Efficient and Interpretable Training, 2026. arXiv:2601.22257 [cs.LG].

Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clement Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 2021.

Steve Smale. On the topology of algorithms, I. *Journal of Complexity*, 3(2), 1987.

Bart Vandereycken. Low-Rank Matrix Completion by Riemannian Optimization. *SIAM Journal on Optimization*, 23(2), 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Verner Vlačić and Helmut Bölcskei. Affine symmetries and neural network identifiability. *Advances in Mathematics*, 376:107485, 2021.

J. H. C. WHITEHEAD. CONVEX REGIONS IN THE GEOMETRY OF PATHS. *The Quarterly Journal of Mathematics*, os-3(1), 1932.

Tiankai Xie, Caleb Geniesse, Jiaqing Chen, Yaoqing Yang, Dmitriy Morozov, Michael W. Mahoney, Ross Maciejewski, and Gunther H. Weber. Evaluating Loss Landscapes from a Topology Perspective, 2024. arXiv:2411.09807 [cs.LG].

Andi Zhang, Xuan Ding, Haofan Wang, Steven McDonagh, and Samuel Kaski. Rethinking Inter-LoRA Orthogonality in Adapter Merging: Insights from Orthogonal Monte Carlo Dropout, 2025a. arXiv:2510.03262 [cs.LG].

Bohan Zhang, Michael Chen, A. Pavan, N. V. Vinodchandran, Lin F. Yang, and Ruosong Wang. List Replicable Reinforcement Learning, 2025b. arXiv:2512.00553 [cs.LG].

Haobo Zhang and Jiayu Zhou. Unraveling LoRA Interference: Orthogonal Subspaces for Robust Model Merging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Understanding Mode Connectivity via Parameter Space Symmetry. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, 2025.

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in Neural Network Parameter Spaces. *Transactions on Machine Learning Research*, 2026.

# A Extended related work

Let us now look more closely at the surrounding literature, building on the comparisons in Section 2.

## A.1 Parameter symmetries, identifiability, and recovery

Zhao et al. (2026) survey permutation, scaling, sign, and continuous change-of-basis symmetries through the parameter-to-function map, while Vlačić & Bölcskei (2021) study affine symmetries of neural-network realizations. For overparameterized feedforward networks, Simsek et al. (2021) show how permutation symmetries and additional invariances generate connected families of critical points and global minima. Bona-Pellissier et al. (2022) instead give geometric conditions for local identifiability of deep ReLU networks from a finite sample. This local viewpoint adds a useful distinction that identifying parameters near one realization does not by itself supply recovery rules over an entire parameter family.

For the geometry used in Section 5, a further connection is provided by Billaud-Friess et al. (2021). They construct explicit bundle charts for full-column-rank and fixed-rank matrix manifolds over Grassmannian bases. These descriptions provide a broader geometric setting for working with local coordinates and changes between matrix representations.

For attention specifically, Silverstein et al. (2026) study rotational redundancies and introduce query and value biases to change training dynamics. Their intervention modifies the architecture, so we should distinguish its effect from choosing a representative of a fixed symmetry orbit. The symmetries available for defining a quotient thus depend on the architecture being studied, which gives a complementary perspective on the geometries we study.

## A.2 Loss landscapes and mode connectivity

Let us next look at what is measured when topology is used to study a loss landscape. Early analyses connected neural-network losses with models from statistical physics (Choromanska et al., 2015; Pennington & Bahri, 2017). Mode-connectivity experiments exhibited low-loss curves between independently trained solutions (Garipov et al., 2018; Draxler et al., 2018), and Entezari et al. (2022) proposed that permutation alignment helps explain linear connectivity. Other approaches examine different features of the landscape, for example Xie et al. (2024) use topological data analysis, Akhtiamov & Thomson (2023) study connectedness through Morse theory, and Kim et al. (2025) analyze width-dependent changes in the topology and connectivity of optimal-solution sets.

We can now briefly sharpen the comparison with Zhao et al. (2025) made in Section 2. A connected space can still have large LS category. In particular, a path connecting two minima does not by itself supply a continuous recovery rule on a neighborhood or on the full quotient. Connectivity results and recovery obstructions can therefore coexist and the existence of paths leaves the complexity of a categorical cover as a nice open question.

**Continuous rules and topology.** The use of topology to bound the number of continuous local rules also has a classical history. Smale (1987) relates root-finding algorithms to covering invariants and cup products, while Farber (2003) develops related Schwarz-genus bounds for motion planning. We apply this categorical-cover strategy to stable population recovery on the quotient parameter spaces of LoRA and MHA.

## A.3 Adapter alignment and merging

We can make the connection with merging more concrete by looking at what the individual methods studied in the literature change. Git Re-Basin constructs permutation alignments before interpolating or merging models (Ainsworth et al., 2023). ZipLoRA studies joint subject-style composition of LoRA adapters (Shah et al., 2025), while Localize-and-Stitch uses sparse task arithmetic (He et al., 2025).

We should also distinguish the uses of orthogonality in LoRA merging. Zhang & Zhou (2025) constrain task-specific LoRA subspaces before fine-tuning to reduce interference. Zhang et al. (2025a) study orthogonality imposed at runtime through Orthogonal Monte Carlo Dropout, and

find experimentally that orthogonality alone need not yield semantic disentanglement or successful composition. These findings make the stage at which a constraint is imposed, and the behavior it is meant to preserve, relevant to the comparison.

For our purposes, the connection to a particular method depends on whether its pipeline contains a recovery stage satisfying Corollary 7.1. Since the ability to align or combine supplied weights does not by itself establish such a stage, we therefore need to examine what information the method receives and what its intermediate output is required to recover before applying the corollary.

### A.4 Replicability and list complexity

The word stability also takes a different form in the statistical literature. Impagliazzo et al. (2022) study reproducibility under shared internal randomness, where two independent samples should lead to the same output with high probability. In the list-replicability formulation of Chase et al. (2023), each underlying distribution has a small set containing the learner's single output with high probability. We should distinguish this distribution-dependent set of likely outputs from the fixed family of continuous candidate rules in Definition 3.2.

The dependence on accuracy matters as well. Blondal et al. (2025) characterize agnostic stability by finite Littlestone dimension when the stability parameter is allowed to depend on the excess error. Bun et al. (2023) establish statistical connections among replicability, approximate differential privacy, and perfect generalization, while also showing computational separations under cryptographic assumptions. These results illustrate why the stability requirement and the efficiency of a procedure need to be considered separately.

Beyond supervised learning, replicable bandits and reinforcement learning extend these questions to sequential decision problems (Esfandiari et al., 2023; Eaton et al., 2023; Karbasi et al., 2023). Zhang et al. (2025b) distinguish lists of final policies from lists of policy traces, which record the sequence of policies executed during learning. There are also further topological and complexity questions that have been studied with regards to replicability and list complexity. Chase et al. (2024) use local Borsuk-Ulam arguments to study limits on stability and list replicability, while Dixon et al. (2023) study list size and certificate length. These formulations give different meanings to what must remain consistent across runs, and hence to the quantity whose complexity is being bounded.

## B Extended preliminaries

Let us now look at the technical details deferred from Section 3. We check the continuity assumptions, prove the existence of a uniform homotopy radius, and provide the bundle and finite-quotient results used in the cohomology calculations.

### B.1 Population recovery and continuity

We first check why the contraction of $\mathcal{P}(\mathcal{Z})$ is continuous in the weak topology. Fix $\mu_0 \in \mathcal{P}(\mathcal{Z})$ and let $H(\mu, t) = (1-t)\mu + t\mu_0$. For every bounded continuous test function $\varphi : \mathcal{Z} \to \mathbb{R}$,

$$\int \varphi \, dH(\mu, t) = (1-t) \int \varphi \, d\mu + t \int \varphi \, d\mu_0$$

The right-hand side is continuous in $(\mu, t)$ by the definition of the weak topology, so $H$ is a continuous contraction. This gives the explicit homotopy used in the proof of Theorem 3.6.

The possible non-injectivity of $G$ in Section 3 also has a simple quantitative consequence. If $G(w) = G(w')$ and $d_X(w, w') \geq 2\epsilon$, the triangle inequality prevents one candidate from being within $\epsilon$ of both targets. These targets must therefore be covered by different candidates.

**Remark B.1.** Let us check two standard population-level constructions against the continuity requirement in Definition 3.1.

First, fix an initialization and a finite number of gradient-descent steps. If the population gradient $(\theta, \mu) \mapsto \nabla_\theta \mathcal{L}(\theta; \mu)$ is continuous, then every update $\theta_{j+1} = \theta_j - \eta_j \nabla_\theta \mathcal{L}(\theta_j; \mu)$ is continuous in $\mu$, and a finite composition of these updates defines a continuous branch when proper.

Second, fix an initialization and a time $T < \infty$, and solve the population gradient-flow equation $\dot{\theta}(t) = -\nabla_\theta \mathcal{L}(\theta(t); \mu)$. If the vector field is continuous in $(\theta, \mu)$, locally Lipschitz in $\theta$ with the uniform bounds needed on the relevant solution region, and solutions exist through time $T$, continuous dependence of ODE solutions on parameters implies that $\mu \mapsto \theta(T)$ is continuous (Hartman, 2002).

A fixed random seed defines a branch only when the resulting seed-conditioned procedure has this continuity property. A selector applied after the list is constructed may be discontinuous because the recovery premise is imposed on the candidates before selection. For a connection to discrete optimization, De Sa et al. (2022) derive GD and SGD guarantees from population gradient-flow convergence under their admissible-rate, Lyapunov-potential, and self-bounding assumptions with a gradient noise assumption. These additional analytic conditions need to be checked when using that connection.

The requirement that branches be defined on all of $\mathcal{P}(\mathcal{Z})$ matters here. If $X$ is an ENR, $G$ is injective, and $\mathcal{P}(\mathcal{Z})$ is metrizable, the ANR extension property gives a single exact proper decoder on a neighborhood of $G(X)$. Thus this neighborhood problem differs from our global one. Guarantees in probability, or allowing outputs in a larger rank-deficient family, also change the recovery problem.

### B.2 Uniformly close maps

We next prove the positive-radius statement used in Section 3.

**Definition B.2.** A space $X$ is a *Euclidean neighborhood retract* (ENR) if it is homeomorphic to a subspace $e(X) \subseteq \mathbb{R}^N$ for which some open neighborhood $W$ of $e(X)$ admits a continuous retraction $r : W \to e(X)$.

**Lemma B.3.** *Every compact metric ENR $(X, d_X)$ has a positive uniform homotopy radius in the sense of Definition 3.5.*

*Proof.* Choose an ENR embedding $e : X \hookrightarrow \mathbb{R}^N$, an open neighborhood $W$ of $e(X)$, and a retraction $r : W \to e(X)$. Since the space is compact, we have a $\delta > 0$ such that the open Euclidean $\delta$-neighborhood of $e(X)$ lies in $W$. Since $e$ is uniformly continuous, there is $\rho_X > 0$ such that $d_X(x, y) < \rho_X$ implies $\|e(x) - e(y)\| < \delta$.

Let $f_0, f_1 : T \to X$ be pointwise $\rho_X$-close. For every $(t, s) \in T \times [0, 1]$, the line segment $(1 - s)e(f_0(t)) + se(f_1(t))$ lies within Euclidean distance $\delta$ of $e(f_0(t))$ and hence lies in $W$. Therefore

$$H(t, s) := e^{-1}(r((1 - s)e(f_0(t)) + se(f_1(t))))$$

is a continuous homotopy from $f_0$ to $f_1$. □

**Proposition B.4.** *Every compact smooth manifold is an ENR. If a finite group $\Gamma$ acts smoothly on a compact smooth manifold $Y$, then the orbit space $Y/\Gamma$ is also a compact ENR.*

*Proof.* The manifold statement is given by Hatcher (2002, Corollary A.9). For the quotient, the equivariant triangulation theorem of Illman (1978, Theorem) gives a $\Gamma$-equivariant triangulation of $Y$. Compactness makes the triangulation finite, and after equivariant subdivision its orbit space is a finite CW complex. It is therefore an ENR by Hatcher (2002, Corollary A.10). □

**Remark B.5.** For a compact Riemannian manifold, we can use a more geometric threshold. The convex-neighborhood theorem of WHITEHEAD (1932) gives every point a basis of strongly convex normal neighborhoods. If we apply this along the compact diagonal in $X \times X$, these neighborhoods give a uniform positive distance below which a pair of points is joined by a unique minimizing geodesic that varies continuously with its endpoints. Equivalently, the global convexity radius $\rho_{\text{conv}}(X)$ is positive. If two maps are pointwise closer than this radius, these geodesics give a homotopy. Thus any $0 < \rho_X < \rho_{\text{conv}}(X)$ is a uniform homotopy radius. For a finite quotient, in Lemma B.3 and proposition B.4 we give the required positive threshold without requiring unique quotient geodesics near singular orbits.

### B.3 Bundles and the Serre spectral sequence

We now provide the bundle terminology used in Section 4 and the later cohomology calculations.

**Definition B.6.** A *fiber bundle* $F \hookrightarrow E \xrightarrow{p} B$ consists of a total space $E$, a base $B$, and a fiber $F$ such that every point of $B$ has a neighborhood $U$ with a homeomorphism $p^{-1}(U) \cong U \times F$ commuting with projection to $U$. A *global section* is a continuous map $s : B \to E$ satisfying $p \circ s = \mathrm{id}_B$.

**Definition B.7.** The Stiefel manifold $V_r(\mathbb{R}^d)$ consists of the orthonormal $r$-frames in $\mathbb{R}^d$, represented by matrices $U \in \mathbb{R}^{d\times r}$ satisfying $U^\top U = I_r$. The real Grassmannian $\mathrm{Gr}(r, d)$ consists of the $r$-dimensional linear subspaces of $\mathbb{R}^d$. Sending a frame to its span gives the principal bundle

$$O(r) \hookrightarrow V_r(\mathbb{R}^d) \longrightarrow \mathrm{Gr}(r, d)$$

For the interested reader, we refer to Milnor & Stasheff (1974, §5) for more details.

For a fibration $F \hookrightarrow E \to B$, the cohomological Serre spectral sequence organizes the cohomology of $E$ from that of $B$ and $F$. With coefficients in a field $\mathbb{F}$, a trivial local coefficient system, and with finite-dimensional fiber cohomology in each degree, its second page is $E_2^{p,q} \cong H^p(B; \mathbb{F}) \otimes H^q(F; \mathbb{F})$ (see for example Hatcher (n.d., Theorem 5.15)). We say that it *collapses at* $E_2$ when every differential from the $E_2$ page onward vanishes. Section 4 states precisely which surviving fiber classes are needed for the cup-length argument.

### B.4 Finite quotients and rational transfer

Finally, we provide the transfer statement used for the oriented covers in the LoRA calculation and the head-permutation quotient in the MHA calculation. The action is allowed to have fixed points.

**Theorem B.8** (Rational finite-group transfer)**.** *Let a finite group $\Gamma$ act simplicially on a finite simplicial complex $Y$. The quotient projection $\pi : Y \to Y/\Gamma$ induces a graded-ring isomorphism*

$$\pi^* : H^*(Y/\Gamma; \mathbb{Q}) \xrightarrow{\cong} H^*(Y; \mathbb{Q})^\Gamma$$

This is the transfer theorem of Bredon (1972, Chapter III, Theorem 2.4) and is also stated by Félix et al. (2008, Theorem 3.27).

## C Proofs

In this section, we provide the missing proofs. For each, we restate the claim and then give the proof.

**Theorem 3.6** (Recovery and LS category)**.** *Let $(X, d_X)$ be a compact metric ENR, let $\rho_X > 0$ be a uniform homotopy radius, and let $G : X \to \mathcal{P}(\mathcal{Z})$ be continuous. If $L$ continuous branches achieve $L$-list $\epsilon$-accurate stable proper recovery with $\epsilon < \rho_X$, then $L \geq \mathrm{cat}(X)$.*

*Proof.* Let $U_k$ be the success domain of branch $A_k$. By Definition 3.2, the open sets $U_1, \ldots, U_L$ cover $X$. Fix $k$ and write $i_k : U_k \hookrightarrow X$ for the inclusion and $f_k = (A_k \circ G)|_{U_k}$. For every $w \in U_k$, Definition 3.3 gives $d_X(i_k(w), f_k(w)) < \epsilon < \rho_X$. By the definition of a uniform homotopy radius (Definition 3.5), we have $i_k \simeq f_k$.

It remains to show that $f_k$ is null-homotopic. Choose any $\mu_0 \in \mathcal{P}(\mathcal{Z})$. Convexity gives a contraction $C(\mu, t) = (1 - t)\mu + t\mu_0$ of $\mathcal{P}(\mathcal{Z})$. Hence $A_k(C(G(w), t))$ is a homotopy from $f_k$ to the constant map with value $A_k(\mu_0)$. Thus $i_k$ is null-homotopic and $U_k$ is categorical in $X$. Every member of the $L$-set cover is categorical, so by the definition of LS category we have $L \geq \mathrm{cat}(X)$. □

**Proposition C.1.** *Let $X$ be a nonempty compact metric ENR, and suppose that $G : X \to \mathcal{P}(\mathcal{Z})$ is continuous and injective and that $\mathcal{P}(\mathcal{Z})$ is metrizable. Then there are $\mathrm{cat}(X)$ continuous proper branches that collectively recover every target exactly. Consequently, for every $0 < \epsilon < \rho_X$, the minimum list size in Definition 3.2 is exactly $\mathrm{cat}(X)$, where $\rho_X > 0$ is any uniform homotopy radius.*

*Proof.* Write $Y = \mathcal{P}(\mathcal{Z})$ and $m = \mathrm{cat}(X)$. Compact ENRs are locally contractible, so $m$ is finite. Choose a categorical open cover $U_1, \ldots, U_m$ of $X$. Since $X$ is compact metric, this cover has a closed shrinking, that is, there exists compact sets $V_i \subset U_i$ whose union is $X$.

Now since the continuous injection $G$ is a homeomorphism onto its image because $X$ is compact and $Y$ is Hausdorff, hence each $G(V_i)$ is closed in $Y$. On this closed subset, define $f_i(G(w)) = w$. Restricting a categorical contraction of $U_i$ to $V_i$ shows that $f_i : G(V_i) \to X$ is null-homotopic. Its terminal constant map extends over all of $Y$. Since an ENR is an ANR[4] for metric spaces (Fritsch & Piccinini, 1990, Propositions A.6.2-A.6.4), the Borsuk homotopy extension theorem (Blackadar, 2016, see Theorem 1.1 and the following paragraph) extends the reversed homotopy and hence $f_i$ to a continuous map $A_i : Y \to X$.

For every $w \in X$, some $V_i$ contains $w$, and then $A_i(G(w)) = w$. These $m$ branches give the upper bound for every positive accuracy. The matching lower bound comes from Theorem 3.6. □

**Theorem 4.1** (Cup-length addition)**.** *Let $p : X \to B$ be a fiber bundle of spaces having CW type over a path-connected base $B$, with fiber $F$ and a global section $s : B \to X$. Fix a field $\mathbb{F}$, and set $c_B = \mathrm{cup}(B;\mathbb{F}) < \infty$.*

(i) Split-tower lower bound. *Suppose there are homogeneous classes $u_1, \dots, u_t \in H^{>0}(X;\mathbb{F})$ for which the $H^*(B;\mathbb{F})$-linear map*

$$\Phi : H^*(B;\mathbb{F}) \otimes_{\mathbb{F}} \Lambda_{\mathbb{F}}(e_1, \dots, e_t) \longrightarrow H^*(X;\mathbb{F}), \qquad a \otimes e^K \longmapsto p^*(a) \smile u^K$$

*is injective, where $e^K$ and $u^K$ denote the corresponding square-free monomials. Then*

$$\mathrm{cup}(X;\mathbb{F}) \geq c_B + t$$

(ii) Equality under collapse. *Assume that the local coefficient system is trivial, the Serre spectral sequence collapses at $E_2$, and $H^*(F;\mathbb{F}) \cong \Lambda_{\mathbb{F}}(x_1, \dots, x_r)$ with every $x_i$ of positive degree. Then classes $u_1, \dots, u_r$ satisfying part* (i) *exist, and*

$$\mathrm{cup}(X;\mathbb{F}) = c_B + r$$

*Proof.* **Part (i): the split-tower lower bound.** Choose homogeneous positive-degree classes $\alpha_1, \dots, \alpha_{c_B} \in H^*(B;\mathbb{F})$ whose product $\sigma = \alpha_1 \smile \cdots \smile \alpha_{c_B}$ is nonzero, using the empty product $\sigma = 1$ when $c_B = 0$. The injectivity of $\Phi$ gives

$$0 \neq \Phi(\sigma \otimes e_1 \cdots e_t) = p^*(\sigma) \smile u_1 \smile \cdots \smile u_t$$

This is a product of $c_B + t$ positive-degree classes, proving part (i).

We now assume the stronger hypotheses of part (ii). Write $i : F \hookrightarrow X$ for the inclusion of a fiber.We proceed in steps to simplify the exposition. The steps first produce the full free tower and then control its multiplicative extensions.

**Step 1: Constructing the free tower.** Trivial local coefficients identify the second page as $E_2^{p,q} \cong H^p(B;\mathbb{F}) \otimes H^q(F;\mathbb{F})$. In the cohomological Serre spectral sequence of Hatcher (n.d., Theorem 5.15), the left-edge quotient $E_\infty^{0,q}$ identifies with the image of restriction $H^q(X;\mathbb{F}) \to H^q(F;\mathbb{F})$. Collapse at $E_2$ gives $E_\infty^{0,q} = E_2^{0,q} \cong H^q(F;\mathbb{F})$, so $i^* : H^*(X;\mathbb{F}) \to H^*(F;\mathbb{F})$ is surjective. Choose homogeneous classes $\widetilde{u}_i \in H^*(X;\mathbb{F})$ satisfying $i^*(\widetilde{u}_i) = x_i$.

The section gives $s^*p^* = \mathrm{id}$, so $p^*$ is injective. Let $J = \ker(s^*)$ and set $u_i = \widetilde{u}_i - p^*s^*(\widetilde{u}_i)$. Then $u_i \in J$ and $i^*(u_i) = x_i$, since the restriction of a positive-degree class pulled back from $B$ to a fiber is zero. Consequently, the square-free monomials $u^K = \prod_{k \in K} u_k$ restrict to the exterior basis $x^K$. Since $B$ is path connected, we can transport along base paths which identifies these restrictions with the restrictions on every other fiber, where they therefore also form a basis. The Leray-Hirsch theorem (Hatcher, 2002, Theorem 4D.1) now gives the direct-sum decomposition

$$H^*(X;\mathbb{F}) = \bigoplus_{K \subseteq \{1,\dots,r\}} p^*H^*(B;\mathbb{F}) \smile u^K$$

In particular, the classes $u_1, \dots, u_r$ satisfy the injectivity condition in part (i).

[4] A metrizable space $X$ is an *absolute neighborhood retract* (ANR) if, whenever $X$ is embedded as a closed subset of a metric space $Z$, some open neighborhood of its image in $Z$ admits a continuous retraction onto that image.

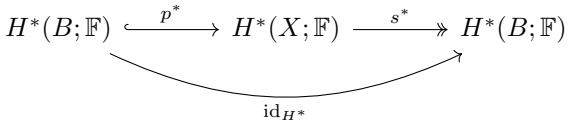



Figure 1: The section makes $s^*$ a left inverse of $p^*$, so base cohomology injects into the total space. Under the strong hypotheses of Theorem 4.1, Leray-Hirsch extends this splitting to the square-free module basis constructed in Step 1.

**Step 2: Controlling squares.** Because $J$ is an ideal, each $u_i^2$ lies in $J$. Its unique Leray-Hirsch expansion has the form

$$u_i^2 = \sum_{K \subseteq \{1,\dots,r\}} p^*(c_{i,K}) \smile u^K$$

where the coefficients $c_{i,K} \in H^*(B;\mathbb{F})$ are homogeneous. Applying $s^*$ gives $c_{i,\varnothing} = 0$: every term with $K \neq \varnothing$ vanishes because $s^*(u_i) = 0$, while $s^*p^* = \mathrm{id}$.

Restriction to the fiber gives $x_i^2 = 0$, directly from the assumed exterior-algebra presentation. A positive-degree base class restricts to zero on a fiber, whereas a scalar restricts to the same scalar. Since the classes $x^K$ are linearly independent, no nonzero $c_{i,K}$ can be a scalar. Thus, whenever $c_{i,K} \neq 0$, both $K \neq \varnothing$ and $c_{i,K} \in I_B := H^{>0}(B;\mathbb{F})$.

**Step 3: A multiplicative weight.** For nonzero $b \in H^*(B;\mathbb{F})$, let $w(b)$ be the largest integer $q$ for which $b \in I_B^q$; set $w(0) = \infty$. In particular, a nonzero scalar has weight zero. The definition of cup length gives $I_B^{c_B+1} = 0$, so every nonzero class has finite weight, and

$$w(ab) \geq w(a) + w(b)$$

Now use formal symbols $U_1, \dots, U_r$ only to record occurrences of the corresponding classes $u_i$. For a term $bU^\alpha$, let us define

$$Q(bU^\alpha) := w(b) + |\alpha|.$$

The square relation from Step 2 gives the reduction $U_i^2 \rightsquigarrow \sum_K c_{i,K} U^K$. Each resulting nonzero term satisfies

$$Q\big(bc_{i,K}U^\alpha U^K\big) \geq w(b) + 1 + |\alpha| + |K| \geq w(b) + |\alpha| + 2 = Q\big(bU^\alpha U_i^2\big)$$

Hence reduction does not decrease $Q$.

**Step 4: Termination and the square-free expansion.** Assign the symbol $U_i$ the cohomological degree of $u_i$, and let $D_{\mathrm{fib}}(U^\alpha) = \sum_i \alpha_i \deg(u_i)$. Homogeneity of the square relation gives

$$2\deg(u_i) = \deg(c_{i,K}) + \deg(u^K)$$

Every nonzero coefficient $c_{i,K}$ has positive degree by Step 2. Replacing $U_i^2$ therefore lowers $D_{\mathrm{fib}}$ by at least one in every resulting term and repeating this terminates because $D_{\mathrm{fib}}$ is a nonnegative integer. The resulting cohomology class has a unique square-free expansion because the classes $u^K$ form the Leray-Hirsch module basis from Step 1.

**Step 5: The upper bound of (ii).** Now let $N > c_B + r$ and let $\Gamma = \gamma_1 \smile \cdots \smile \gamma_N$, where each $\gamma_n$ is homogeneous and has positive degree. Expand each factor in the Leray–Hirsch basis. A term with no $u_i$ has a coefficient in $I_B$, while a term containing $|K| \geq 1$ fiber classes has a coefficient of weight at least zero. Thus every term in every factor has $Q \geq 1$.

After distributing the product, superadditivity of $w$ shows that every initial term has $Q \geq N$. Steps 3 and 4 reduce it to the square-free expansion

$$\Gamma = \sum_{J \subseteq \{1,\dots,r\}} p^*(b_J) \smile u^J$$

For a fixed $J$, every contribution to $b_J$ lies in $I_B^{N-|J|}$; therefore their sum $b_J$ lies in the same ideal power. Since $|J| \leq r$, we have $N - |J| > c_B$, so $b_J \in I_B^{c_B+1} = 0$. Hence $\Gamma = 0$, and

$$\mathrm{cup}(X;\mathbb{F}) \leq c_B + r$$

**Step 6: Exact equality.** Step 1 gives the injective free tower required in part (i), with $t = r$. Part (i) therefore gives $\mathrm{cup}(X;\mathbb{F}) \geq c_B + r$, while Step 5 gives the reverse inequality. This finishes the proof of (ii) □

**Lemma 4.2.** *Let the embedding dimension $d$ and rank $r$ be positive even integers with $0 < r < d$, and write $r = 2k$ and $d = 2n$. The cup-length of the unoriented real Grassmannian $\mathrm{Gr}(r,d)$ is*

$$\mathrm{cup}(\mathrm{Gr}(r,d);\mathbb{Q}) = k(n-k) = \frac{r(d-r)}{4}$$

*Proof.* To compute the cup-length over $\mathbb{Q}$, we determine the algebraic presentation of the cohomology ring $H^*(\mathrm{Gr}(2k,2n);\mathbb{Q})$. We organize the calculation into four steps for simplicity of exposition.

**Step 1: The finite homogeneous-space calculation.** We begin with the oriented double cover of the finite Grassmannian,

$$q : \widetilde{\mathrm{Gr}}(2k,2n) = SO(2n)/(SO(2k)\times SO(2(n-k))) \longrightarrow \mathrm{Gr}(2k,2n)$$

This passage allows us to work with connected compact Lie groups. Moreover,

$$\mathrm{rank}\, SO(2n) = n = k + (n-k) = \mathrm{rank}(SO(2k)\times SO(2(n-k))).$$

The subgroup therefore contains a maximal torus of $SO(2n)$, so the equal-rank homogeneous-space theorem of Borel (1953, Théorème 26.1) applies[5]. It therefore identifies the rational cohomology of $G/H$ with $H^*(BH;\mathbb{Q})$ modulo the ideal generated by the restrictions of the positive-degree classes in $H^*(BG;\mathbb{Q})$.

Let $S$ and $Q$ denote the oriented tautological bundle and its oriented orthogonal complement over $\widetilde{\mathrm{Gr}}(2k,2n)$. Write $p_i = p_i(S)$, $p'_j = p_j(Q)$, $e = e(S)$, and $e' = e(Q)$. For an oriented real bundle of rank $2a$, the rational cohomology of its classifying space is $H^*(BSO(2a);\mathbb{Q}) = \mathbb{Q}[p_1,\dots,p_{a-1},e]$, with $e^2 = p_a$ (Milnor & Stasheff, 1974, Theorem 15.9 and Corollary 15.8). Under the block-sum inclusion $SO(2k)\times SO(2(n-k)) \hookrightarrow SO(2n)$, the universal total Pontryagin class restricts to $p(S)p(Q)$ by the Whitney sum formula, while the universal Euler class restricts to $e(S)e(Q)$ (Milnor & Stasheff, 1974, Theorem 15.3 and Property 9.6). Borel's theorem consequently gives

$$H^*(\widetilde{\mathrm{Gr}}(2k,2n);\mathbb{Q}) \cong \frac{\mathbb{Q}[p_1,\dots,p_k,p'_1,\dots,p'_{n-k},e,e']}{\big(p(S)p(Q)=1,\ e^2=p_k,\ (e')^2=p'_{n-k},\ ee'=0\big)}$$

Here $p(S) = 1+p_1+\cdots+p_k$, $p(Q) = 1+p'_1+\cdots+p'_{n-k}$, and the relation $p(S)p(Q) = 1$ means that every positive-degree homogeneous component of this product vanishes.

We now pass back to the unoriented Grassmannian. Let $\tau$ be the nontrivial deck transformation of $q$. Reversing the orientation of the tautological plane also reverses the induced orientation of its orthogonal complement, because the ambient orientation remains fixed. Hence[6]

$$\tau^*(e) = -e, \qquad \tau^*(e') = -e', \qquad \tau^*(p_i) = p_i, \qquad \tau^*(p'_j) = p'_j$$

By Theorem B.8, the pullback along the double cover identifies the cohomology downstairs with the invariant subring

$$q^* : H^*(\mathrm{Gr}(2k,2n);\mathbb{Q}) \xrightarrow{\cong} H^*(\widetilde{\mathrm{Gr}}(2k,2n);\mathbb{Q})^{\langle\tau^*\rangle}$$

Let us now check that passing to invariants introduces no further relations among the Pontryagin classes. Let $A = \mathbb{Q}[p_i,p'_j]/(p(S)p(Q)=1)$, $a = p_k$, and $b = p'_{n-k}$. The top homogeneous Whitney relation already gives $ab = 0$. Before imposing $ee' = 0$, the algebra $A[e,e']/(e^2-a,(e')^2-b)$ is free over $A$ with basis $1, e, e', ee'$. The ideal generated by $ee'$ is the $A$-submodule generated by $ee', ae', be, ab$. Since $ab = 0$, quotienting by this ideal leaves

$$H^*(\widetilde{\mathrm{Gr}}(2k,2n);\mathbb{Q}) \cong A \oplus (A/(b))e \oplus (A/(a))e'$$

[5] Borel (1953, Théorème 26.1) is actually stated over $\mathbb{R}$. For us, the characteristic homomorphism and its defining relations are defined over $\mathbb{Q}$, so one needs a small argument to use it. For that, note that extending scalars to $\mathbb{R}$ gives an isomorphism, so its kernel and cokernel over $\mathbb{Q}$ vanish. This suffices for our purposes

[6] The signs of the Euler classes and the invariance of the Pontryagin classes under orientation reversal are the standard characteristic-class transformations, see for example (Milnor & Stasheff, 1974, Property 9.3 and §15).

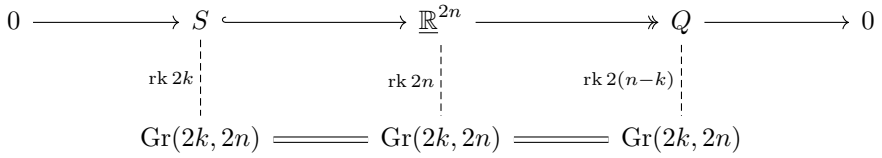
as an $A$-module. The deck involution fixes the first summand and negates the other two. Its invariant subring is therefore exactly the embedded copy of $A$, giving[7]

$$H^*(\mathrm{Gr}(2k,2n);\mathbb{Q}) \cong \frac{\mathbb{Q}[p_1,\dots,p_k,p'_1,\dots,p'_{n-k}]}{\big(p(S)p(Q)=1\big)}$$

**Step 2: Eliminating the complementary generators.** On the unoriented Grassmannian, the generators obtained in Step 1 are the Pontryagin classes of the tautological bundle $S$ and its orthogonal complement $Q$. The bundles have ranks $2k$ and $2(n-k)$ and satisfy $S \oplus Q \cong \underline{\mathbb{R}}^{2n}$. Because the total Pontryagin class of a trivial bundle is 1, and because the Whitney sum formula holds exactly over $\mathbb{Q}$ after its possible 2-torsion discrepancy is removed (Milnor & Stasheff, 1974, Theorem 15.3), we have

$$(1+p_1+\cdots+p_k)(1+p'_1+\cdots+p'_{n-k})=1,$$

where $p_i = p_i(S)$ and $p'_j = p_j(Q)$ are the rational Pontryagin classes of the tautological and quotient bundles, respectively. See Figure 2.




$$p(S)\cdot p(Q) \;=\; p\big(\underline{\mathbb{R}}^{2n}\big) \;=\; 1$$

Figure 2: The tautological short exact sequence of vector bundles over $\mathrm{Gr}(2k,2n)$. Here $\underline{\mathbb{R}}^{2n}$ denotes the *trivial* bundle $\mathrm{Gr}(2k,2n)\times\mathbb{R}^{2n}$, the fiber at each point $W \in \mathrm{Gr}(2k,2n)$ being the ambient space $\mathbb{R}^{2n}$ itself. The tautological subbundle $S$ has fiber $S_W = W$, and the quotient bundle $Q$ has fiber $Q_W = \mathbb{R}^{2n}/W$. The orthogonal complement gives a splitting $\underline{\mathbb{R}}^{2n} \cong S \oplus Q$, and the Whitney sum formula over $\mathbb{Q}$ yields the fundamental relation $p(S)\cdot p(Q) = 1$ in rational cohomology.

Step 1 proves that this Whitney relation gives the complete rational presentation. It allows us to express each $p'_j$ as a polynomial in $p_1,\dots,p_k$, and by expanding the formal power series inverse

$$1+p'_1+p'_2+\cdots = (1+p_1+\cdots+p_k)^{-1} = \sum_{j\ge 0} q_j(p_1,\dots,p_k)$$

we find that $p'_j = q_j$ for each $j$, where $q_j$ is the degree-$j$ homogeneous component (in the grading where $\deg p_i = i$) of the formal inverse series. Since $Q$ has rank $2(n-k)$, the Pontryagin classes $p'_j$ vanish for $j > n-k$, imposing the conditions $q_j(p_1,\dots,p_k)=0$ for all $j > n-k$.

The resulting presentation is thus

$$H^*(\mathrm{Gr}(2k,2n);\mathbb{Q}) \cong \mathbb{Q}[p_1,\dots,p_k] \,/\, (q_{n-k+1},\, q_{n-k+2},\, \dots)$$

where $|p_i| = 4i$ and $q_j$ denotes the degree-$j$ homogeneous component of the formal inverse $(1+p_1+\cdots+p_k)^{-1}$.

**Step 3: Comparison with the complex Grassmannian.** Now apply the same finite homogeneous-space calculation to $\mathrm{Gr}_{\mathbb{C}}(k,n) = U(n)/(U(k)\times U(n-k))$. The two groups have equal rank, so the theorem of Borel (1953, Théorème 26.1) identifies the rational cohomology ring with the quotient of $H^*(B(U(k)\times U(n-k));\mathbb{Q})$ by the positive-degree restrictions from $H^*(BU(n);\mathbb{Q})$. Milnor & Stasheff (1974, Theorem 14.5) identify each $H^*(BU(a);\mathbb{Z})$ as the polynomial ring in the universal Chern classes, and the Whitney product formula is equation (14.7) in Milnor & Stasheff (1974). Consequently, the Chern classes $c_1,\dots,c_k$ of the tautological bundle and $c'_1,\dots,c'_{n-k}$ of its complementary bundle satisfy the complete presentation

$$(1+c_1+\cdots+c_k)(1+c'_1+\cdots+c'_{n-k})=1$$

[7]This is also the explicit ordinary Leray-Borel presentation given in He (2017, Corollary 5.26).

The same presentation holds integrally by Borel (1953, Proposition 31.1). Over $\mathbb{Q}$, it gives

$$H^*(\mathrm{Gr}_{\mathbb{C}}(k,n);\mathbb{Q}) \cong \mathbb{Q}[c_1,\dots,c_k] \,/\, (\bar{q}_{n-k+1},\, \bar{q}_{n-k+2},\, \dots)$$

where $|c_i| = 2i$ and $\bar{q}_j$ is the degree-$j$ homogeneous component of $(1 + c_1 + \cdots + c_k)^{-1}$, defined by the identical formal power series as in the real case.

Let $\phi\colon H^*(\mathrm{Gr}_{\mathbb{C}}(k,n);\mathbb{Q}) \to H^*(\mathrm{Gr}(2k,2n);\mathbb{Q})$ be the $\mathbb{Q}$-algebra homomorphism defined on generators by $\phi(c_i) = p_i$. Because the generators and the ideal of relations are defined by identical polynomials, $\phi$ is an isomorphism of *ungraded* $\mathbb{Q}$-algebras. It doubles all topological degrees because $\phi(H^{2i}) \subseteq H^{4i}$.

**Step 4: Cup-length of the complex Grassmannian.** Recall that the cup-length of a space $X$ with coefficients in $\mathbb{Q}$ is the largest integer $\ell$ such that there exist classes $\alpha_1,\dots,\alpha_\ell \in \widetilde{H}^*(X;\mathbb{Q})$ (the augmentation ideal) with $\alpha_1 \cdots \alpha_\ell \neq 0$. Since $\phi$ is a ring isomorphism preserving the augmentation ideal, the rational cup-lengths of $\mathrm{Gr}(2k,2n)$ and $\mathrm{Gr}_{\mathbb{C}}(k,n)$ are equal. It remains to compute the latter. So let $m = \dim_{\mathbb{C}}(\mathrm{Gr}_{\mathbb{C}}(k,n)) = k(n-k)$.

*Lower bound.* Write the complex Grassmannian as $SU(n)/S(U(k)\times U(n-k))$. The denominator is the centralizer of a torus and quotienting the group and stabilizer by the finite center makes the action effective without changing the homogeneous space. The construction of Borel (1954, Theorem 2 and its proof) therefore supplies a homogeneous Kähler form $\omega$ whose cohomology class can be chosen integral. At any point, a unitary real coframe may be chosen so that $\omega = e^1\wedge f^1 + \cdots + e^m \wedge f^m$. Hence $\omega^m/m!$ is the positive Riemannian volume form. It follows that $\int_{\mathrm{Gr}_{\mathbb{C}}(k,n)} \omega^m > 0$, so $[\omega]^m \neq 0$. Since $[\omega]$ is integral, it also defines a rational class, and this product proves $\mathrm{cup}(\mathrm{Gr}_{\mathbb{C}}(k,n);\mathbb{Q}) \geq m$.

*Upper bound.* The complex-analytic cell-decomposition theorem of Borel (1954, Theorem 3) gives a finite decomposition of the complex Grassmannian into complex affine cells. Every cell therefore has even real dimension, so the integral cohomology vanishes in odd degrees. In particular, every nonzero class in the augmentation ideal has degree at least 2. Any cup product of $m+1$ such classes lies in degree at least $2(m+1)$, while the real dimension is $2m$. Thus the product vanishes and $\mathrm{cup}(\mathrm{Gr}_{\mathbb{C}}(k,n);\mathbb{Q}) \leq m$.

**Finishing the proof.** Combining the bounds, the rational cup-length of the complex Grassmannian is exactly its complex dimension:

$$\mathrm{cup}(\mathrm{Gr}(2k,2n);\mathbb{Q}) = \mathrm{cup}(\mathrm{Gr}_{\mathbb{C}}(k,n);\mathbb{Q}) = k(n-k)$$

Substituting $k = r/2$ and $n = d/2$ yields the stated formula $\frac{r(d-r)}{4}$. □

**Lemma 5.3.** *The observable parameter space $\mathcal{M}_r$ strongly deformation retracts onto a core space that is canonically homeomorphic to the quotient manifold*

$$X_r := \frac{V_r(\mathbb{R}^d) \times V_r(\mathbb{R}^d)}{O(r)}$$

*where $V_r(\mathbb{R}^d) = \{U \in \mathbb{R}^{d\times r} : U^\top U = I_r\}$ is the (compact) Stiefel manifold of orthonormal $r$-frames in $\mathbb{R}^d$, and the orthogonal group $O(r)$ acts freely on the product by the diagonal right action $(U,V)\cdot g = (Ug, Vg)$.*

*Proof.* For $W \in \mathcal{M}_r$, set $S(W) = W^\top W$ and note that its spectrum lies in $\{0\} \cup [c^2, R^2]$. Choose a continuous function $f : [0,R^2] \to \mathbb{R}$ satisfying $f(0) = 0$ and $f(\lambda) = \lambda^{-1/2}$ for $\lambda \in [c^2, R^2]$, and define the polar partial isometry

$$Q(W) = W f(S(W)) = W\big((W^\top W)^\dagger\big)^{1/2}$$

where the second expression uses the Moore-Penrose inverse on the rank-$r$ support. To see continuity directly, uniformly approximate $f$ on $[0,R^2]$ by polynomials. The corresponding polynomial matrix maps are continuous and converge uniformly in operator norm to $S \mapsto f(S)$. Hence $W \mapsto f(S(W))$, and therefore $W \mapsto Q(W)$, is continuous.

Now for any thin singular value decomposition $W = U\Sigma V^\top$, we have $Q(W) = UV^\top$. Thus $Q(W)$ has rank $r$ and all of its nonzero singular values equal 1. Let

$$H(W, t) = (1 - t)W + tQ(W)$$

The same decomposition gives $H(W, t) = U((1 - t)\Sigma + tI_r)V^\top$, whose nonzero singular values belong to $[c, R]$ because $1 \in [c, R]$. Therefore $H$ is a continuous homotopy in $\mathcal{M}_r$ from the identity to the map $Q$. If $W$ already has all nonzero singular values equal to 1, then $Q(W) = W$, so this homotopy fixes the core pointwise. Hence $H$ is a strong deformation retraction onto

$$\mathcal{C}_r = \{UV^\top : U, V \in V_r(\mathbb{R}^d)\}$$

It remains to identify this core with the stated quotient. Consider the continuous surjection $\pi : V_r(\mathbb{R}^d) \times V_r(\mathbb{R}^d) \to \mathcal{C}_r$ given by $\pi(U, V) = UV^\top$. If $U_1V_1^\top = U_2V_2^\top$, multiplication on the right by $V_1$ gives $U_1 = U_2g$ for $g = V_2^\top V_1$. Since both $U_1$ and $U_2$ have orthonormal columns, $g^\top g = I_r$, so $g \in O(r)$. Left multiplication of the original equality by $U_2^\top$ then gives $gV_1^\top = V_2^\top$, or $V_1 = V_2g$. Thus the fibers of $\pi$ are exactly the diagonal $O(r)$-orbits. The induced continuous bijection from the compact quotient to the Hausdorff space $\mathcal{C}_r$ is a homeomorphism, and therefore

$$\mathcal{C}_r \cong \frac{V_r(\mathbb{R}^d) \times V_r(\mathbb{R}^d)}{O(r)} = X_r \qquad \square$$

**Proposition C.2.** *For the Frobenius metric under $X_r \cong \mathcal{C}_r$, we may take $\rho_{X_r} = 1$.*

*Proof.* Let $f_0, f_1 : T \to \mathcal{C}_r$ be continuous with $\|f_0(x) - f_1(x)\|_F < 1$ for every $x \in T$. Let $P = f_0(x)^\top f_0(x)$. Since

$$\delta := \|f_0(x) - f_1(x)\|_{\mathrm{op}} \le \|f_0(x) - f_1(x)\|_F < 1,$$

we can join $f_0(x)$ to $f_1(x)P$, and then $f_1(x)P$ to $f_1(x)$, using

$$D_t = (1 - t)f_0(x) + tf_1(x)P, \qquad \text{and} \qquad E_t = f_1(x)\big((1 - t)P + tI\big), \quad \forall t \in [0, 1]$$

respectively. Now $D_t$ vanishes on $\ker P$, while for $v \in \operatorname{ran} P$, the triangle inequality gives

$$\|D_tv\| = \|f_0(x)v + t(f_1(x) - f_0(x))v\| \ge (1 - t\delta)\|v\|.$$

Thus $D_t$ is injective on the $r$-dimensional space $\operatorname{ran} P$ and hence has rank $r$. For $E_t$, note that for $t > 0$, the right factor in $E_t$ is invertible, so $E_t$ also has rank $r$, and at $t = 0$, this follows from $E_0 = f_1(x)P = D_1$.

Now apply the polar normalization $Q$ from the proof of Lemma 5.3, which is continuous on the fixed-rank locus by a similar argument as before. The resulting paths lie in $\mathcal{C}_r$, join $f_0(x)$ to $f_1(x)$, and depend continuously on $x$. Their concatenation therefore gives the required homotopy, proving that we may take $\rho_{X_r} = 1$. $\square$

**Lemma 5.6.** *The LoRA parameter manifold $X_r = (V_r(\mathbb{R}^d) \times V_r(\mathbb{R}^d))/O(r)$, viewed as a fiber bundle over the Grassmannian $\mathrm{Gr}(r, d)$ via the projection $\pi([U, V]) = \operatorname{span}(U)$, admits a continuous global section.*

*Proof.* First, observe that $\pi$ is well-defined on the quotient: if $(U, V) \sim (Ug, Vg)$, then $\operatorname{span}(Ug) = \operatorname{span}(U)$. The space $X_r$ is the associated fiber bundle $V_r(\mathbb{R}^d) \times_{O(r)} V_r(\mathbb{R}^d)$ to the principal $O(r)$-bundle $V_r(\mathbb{R}^d) \to \mathrm{Gr}(r, d)$, with typical fiber $V_r(\mathbb{R}^d)$; hence $\pi$ is a locally trivial fiber-bundle projection by the associated-bundle construction (see for example (Husemöller, 1994, Chapter 4, §§5 and 7)).

Now we construct the section explicitly. Define $s : \mathrm{Gr}(r, d) \to X_r$ by assigning to each $r$-dimensional subspace $P$ the equivalence class $s(P) = [U, U]$, where $U \in V_r(\mathbb{R}^d)$ is any orthonormal frame spanning $P$. To verify that this map is globally well-defined, suppose we choose a different orthonormal frame $U'$ for $P$. By definition, $U' = Ug$ for some orthogonal transition matrix $g \in O(r)$. Then $[U', U'] = [Ug, Ug] = [U, U]$ by the defining diagonal action, so $s(P)$ is independent of the chosen frame.

To establish continuity, consider the composite

$$V_r(\mathbb{R}^d) \xrightarrow{\Delta} V_r(\mathbb{R}^d) \times V_r(\mathbb{R}^d) \xrightarrow{q} X_r$$

where $\Delta(U) = (U, U)$ is the diagonal embedding and $q$ is the quotient map. Both maps are continuous, hence so is $q \circ \Delta$. The calculation above shows that $q \circ \Delta$ is constant on the fibers of the principal bundle projection $p : V_r(\mathbb{R}^d) \to \mathrm{Gr}(r, d)$. Since $p$ is a quotient map, the universal property of quotient spaces yields a unique continuous map $s : \mathrm{Gr}(r, d) \to X_r$ satisfying $s \circ p = q \circ \Delta$, and by construction $\pi \circ s = \mathrm{id}_{\mathrm{Gr}(r,d)}$. □

**Theorem 5.7** (LoRA recovery bound)**.** *Let $X_r$ be the rank-$r$ LoRA parameter manifold. Assume the embedding dimension $d$ and rank $r$ are positive even integers with $r < d$. Any family of continuous recovery branches achieving $L$-list $\epsilon$-accurate stable proper recovery on $X_r$ for $\epsilon < \rho_{X_r}$ must have a list size satisfying*

$$L \geq \frac{r(d-r)}{4} + \frac{r}{2}$$

*Proof.* By Theorem 3.6 and the cup-length estimate, $L \geq \mathrm{cat}(X_r) \geq \mathrm{cup}(X_r; \mathbb{Q}) + 1$. It remains to verify the split tower required by part (i) of Theorem 4.1.

Write $d = 2n$, $r = 2k$, and $d - r = 2m$, so $n = m + k$. We first pass to the oriented double covers

$$q_B : \widetilde{B} = SO(2n)/(SO(2m) \times SO(2k)) \longrightarrow B = \mathrm{Gr}(2k, 2n)$$

and

$$q_X : \widetilde{X} = \big(V_{2k}(\mathbb{R}^{2n}) \times V_{2k}(\mathbb{R}^{2n})\big)/SO(2k) \longrightarrow X_r$$

Writing $p : X_r \to B$ and $\widetilde{p} : \widetilde{X} \to \widetilde{B}$ for the bundle projections, these maps form the pullback square

$$\begin{array}{ccc} \widetilde{X} & \xrightarrow{q_X} & X_r \\ {\scriptstyle \widetilde{p}}\downarrow & \lrcorner & \downarrow{\scriptstyle p} \\ \widetilde{B} & \xrightarrow[q_B]{} & B \end{array}$$

Equivalently, $\widetilde{X} \cong X_r \times_B \widetilde{B}$. The diagonal section from Lemma 5.6 lifts to this pullback. The space $\widetilde{X}$ has the homogeneous-space description

$$\widetilde{X} \cong \frac{SO(2n) \times SO(2n)}{SO(2m) \times SO(2m) \times SO(2k)}$$

where $(h_1, h_2, g)$ maps to $(\mathrm{diag}(h_1, g), \mathrm{diag}(h_2, g))$.

At this point, we need to identify explicitly which classes survive on $\widetilde{X}$, so we compute them by hand. For that, we will use the Cartan-Weil model for a compact homogeneous space given by Félix et al. (2008, Theorem 2.71). So introduce the total Pontryagin polynomials

$$A(t) = 1 + A_1 t + \cdots + A_{m-1} t^{m-1} + a^2 t^m, \qquad B(t) = 1 + B_1 t + \cdots + B_{m-1} t^{m-1} + b^2 t^m$$

and

$$C(t) = 1 + C_1 t + \cdots + C_{k-1} t^{k-1} + c^2 t^k$$

Here $A_i, B_i, C_i$ have cohomological degree $4i$, while $a, b, c$ are the Euler classes of degrees $2m, 2m, 2k$, respectively. The identities $p_m = e^2$ and $p(E \oplus F) = p(E)p(F)$ over $\mathbb{Q}$, together with the polynomial presentation of $H^*(BSO(2q); \mathbb{Q})$ that we will use here, are given by Milnor & Stasheff (1974, Corollary 15.8; Theorem 15.3; and Theorem 15.9).

Now let $[t^i]P(t)$ denote the coefficient of $t^i$. The model has odd generators $x_i, y_i$ of degree $4i - 1$ for $1 \leq i \leq n - 1$ and Euler generators $\xi, \zeta$ of degree $2n - 1$, with

$$Dx_i = [t^i]A(t)C(t), \qquad Dy_i = [t^i]B(t)C(t), \qquad D\xi = ac, \qquad D\zeta = bc$$

These formulas are the restrictions of the universal characteristic classes under the two block-sum inclusions into $SO(2n)$ (for the Euler-class product formula, see Milnor & Stasheff (1974,

Property 9.6)). The subalgebra generated by $A_i, a, C_i, c, x_i, \xi$ is the Cartan model of the oriented Grassmannian $\widetilde{B}$.

Doing a change of variables $\omega_i = y_i - x_i$ and $z = \zeta - \xi$ gives us that

$$D\omega_i = [t^i](B(t) - A(t))C(t), \qquad Dz = (b-a)c$$

We spell out the reduction because we need to know what survives (note that just closedness does not suffice). Let $\Omega(t) := \sum_{i=1}^{n-1} \omega_i t^i$. Since $C(0) = 1$, its formal inverse exists. Define $\nu_i = [t^i](\Omega(t)C(t)^{-1})$ for $1 \le i \le n-1$. This is a triangular change of odd generators, with inverse $\Omega(t) \equiv (\sum_{i=1}^{n-1} \nu_i t^i)C(t) \pmod{t^n}$, and hence a differential graded algebra isomorphism. If $q_i = B_i - A_i$ for $i < m$ and $\delta = b^2 - a^2$, then

$$D\nu_i = q_i \ (1 \le i < m), \qquad D\nu_m = \delta, \qquad D\nu_i = 0 \qquad \text{for } m < i < n$$

No other differential contains $q_i$ or $\nu_i$ for $i < m$, so the model is the tensor product of a reduced model with the factors $(\mathbb{Q}[q_i] \otimes \Lambda(\nu_i), D\nu_i = q_i)$. Each such factor contracts onto $\mathbb{Q}$ by the following argument. On its positive-degree monomials, set $h(q_i^a) = q_i^{a-1}\nu_i$ for $a \ge 1$ and $h(q_i^a \nu_i) = 0$ for $a \ge 0$. Then $Dh + hD$ is the identity on the augmentation ideal. Deleting these factors therefore gives a quasi-isomorphism from the original Cartan model to the quotient by the differential ideal $(q_1, \nu_1, \dots, q_{m-1}, \nu_{m-1})$.

In this reduced model, $B_i = A_i$ for $i < m$, so $B(t) - A(t) = \delta t^m$. Denote the images of $\omega_m, \dots, \omega_{n-1}$ by $\eta_m, \dots, \eta_{n-1}$. The remaining differentials in the original $\omega$-coordinates are

$$D\eta_m = \delta, \qquad D\eta_{m+j} = \delta C_j \quad (1 \le j \le k-1), \qquad Dz = (b-a)c$$

Thus the elements

$$\theta_j = \eta_{m+j} - C_j \eta_m \quad \text{for } 1 \le j \le k-1$$

are closed. This follows by comparing coefficients of degree below $n$ in the truncated identity above, which gives $\theta_j = \nu_{m+j} + C_1\nu_{m+j-1} + \cdots + C_{j-1}\nu_{m+1}$ in the reduced model. Thus replacing the closed generators $\nu_{m+1}, \dots, \nu_{n-1}$ by $\theta_1, \dots, \theta_{k-1}$ is triangular and invertible. When $k = 1$, this list is empty, and the exterior factor below is understood to be $\mathbb{Q}$. The reduced model therefore decomposes as the actual tensor product of differential graded algebras

$$\mathcal{C}_{\text{red}} = \mathcal{C}_0 \otimes \Lambda(\theta_1, \dots, \theta_{k-1})$$

where $\mathcal{C}_0$ contains the oriented-Grassmannian submodel together with $b, \eta_m, z$, and $D\theta_j = 0$. Combining this decomposition with the preceding quasi-isomorphism and the Künneth theorem over $\mathbb{Q}$ gives

$$H^*(\widetilde{X}; \mathbb{Q}) \cong H^*(\mathcal{C}_0) \otimes \Lambda([\theta_1], \dots, [\theta_{k-1}])$$

Let $\widetilde{s} : \widetilde{B} \to \widetilde{X}$ be the lifted diagonal section. Since $\widetilde{p} \circ \widetilde{s} = \mathrm{id}_{\widetilde{B}}$, the identity $\widetilde{s}^*\widetilde{p}^* = \mathrm{id}$ makes $\widetilde{p}^*$ injective. The oriented-Grassmannian submodel lies inside $\mathcal{C}_0$, so the image of $\widetilde{p}^*$ lies in $H^*(\mathcal{C}_0) \otimes 1$. If $\sum_K \widetilde{p}^*(\alpha_K)[\theta]^K = 0$, the tensor-product decomposition forces every coefficient $\widetilde{p}^*(\alpha_K)$ to vanish, and injectivity forces every $\alpha_K$ to vanish. Thus the square-free monomials in the $[\theta_j]$ are nonzero and linearly independent over the pulled-back base cohomology. This is exactly the module freeness required in Theorem 4.1(i).

On the fiber, $z$ restricts to the generator of degree $2n - 1 = d - 1$, while $b$ has degree $2m = d - r$. In the total-space model, $Dz = (b-a)c$ can be nonzero. This Euler-paired generator is therefore not included in the surviving tower. The $k-1$ classes $[\theta_j]$, of degrees $4(m+j)-1$, are the surviving Pontryagin-type generators.

All that remains is to descend this tower to the unoriented spaces. So choose reflections $R_m \in O(2m)$ and $R_k \in O(2k)$, each of determinant $-1$, and let $t = \mathrm{diag}(R_m, R_k) \in SO(2n)$. In the homogeneous-space description of $\widetilde{X}$, right multiplication by $(t, t)$ induces the nontrivial deck transformation $\tau_X$ of $q_X$. It reverses the orientation of both $2k$-frames. On $\widetilde{B}$, the corresponding deck transformation $\tau_B$ reverses the orientation of the tautological $2k$-plane. The two deck transformations are compatible with the bundle projections

$$\begin{array}{ccc} \widetilde{X} & \xrightarrow{\tau_X} & \widetilde{X} \\ {\scriptstyle\widetilde{p}}\downarrow & & \downarrow{\scriptstyle\widetilde{p}} \\ \widetilde{B} & \xrightarrow[\tau_B]{} & \widetilde{B} \end{array}$$

To see how this deck transformation acts on the model, write $G = SO(2n) \times SO(2n)$, let $H$ be the stabilizer as above, and let $N = (t, t) \in N_G(H)$. On $G/H$, right translation factors as $R_N = L_N \circ \mathrm{Ad}_{N^{-1}}$. Since $G$ is connected, left translation $L_N$ is homotopic to the identity. Thus $\tau_X$ has the same cohomology action as conjugation by $N^{-1}$, an automorphism of the pair $(G, H)$. Functoriality of the Cartan-Weil construction represents this action by the following differential graded algebra involution $T$. Conjugation by $R_m$ reverses the orientations of the two complementary $2m$-planes, while conjugation by $R_k$ reverses the orientation of the common $2k$-plane. Accordingly[8],

$$T(a) = -a, \qquad T(b) = -b, \qquad T(c) = -c$$

$$T(A_i) = A_i, \qquad T(B_i) = B_i, \qquad T(C_i) = C_i$$

Conjugation by $t$ is an inner automorphism of $SO(2n)$ and fixes its universal characteristic classes. The induced Cartan-model map therefore fixes the corresponding odd generators $x_i, y_i, \xi, \zeta$. The formulae for $D$ also show directly that $TD = DT$.

We need things to be done $T$-equivariantly, so we know show that the reduction can be performed $T$-equivariantly. The variables $q_i = B_i - A_i$ are fixed, and the triangular change defining $\nu_i$ uses only the fixed coefficients $C_j$, so every eliminated pair is $T$-stable. It follows that the reduced generators $\eta_i$ are fixed. Since each $C_j$ is also fixed, $T(\theta_j) = \theta_j$. Consequently, every class $[\theta_j]$ and every square-free monomial $[\theta]^K$ belongs to the invariant subring $H^*(\widetilde{X}; \mathbb{Q})^{\langle \tau_X \rangle}$.

Now we apply Theorem B.8 to the free $\mathbb{Z}/2$-action generated by $\tau_X$. The pullback $q_X^*$ identifies $H^*(X_r; \mathbb{Q})$ with the invariant subring of $H^*(\widetilde{X}; \mathbb{Q})$. Therefore, for every $j$, there is a unique class $u_j \in H^*(X_r; \mathbb{Q})$ such that $q_X^*(u_j) = [\theta_j]$.

These descended classes form a free tower over $H^*(B; \mathbb{Q})$. To see this, suppose that a relation $\sum_K p^*(\alpha_K) u^K = 0$ holds, where $K$ ranges over subsets of $\{1, \dots, k-1\}$. Pulling it back along $q_X$ and using the first commutative square gives

$$\sum_K \widetilde{p}^* q_B^*(\alpha_K) [\theta]^K = 0$$

The freeness established above forces $q_B^*(\alpha_K) = 0$ for every $K$. A second application of Theorem B.8, now to $q_B$, shows that $q_B^*$ is injective. Hence every $\alpha_K$ is zero, which proves freeness downstairs.

We are now done. Part (i) of Theorem 4.1, together with Lemma 4.2, now yields

$$\mathrm{cup}(X_r; \mathbb{Q}) \geq \mathrm{cup}(\mathrm{Gr}(r, d); \mathbb{Q}) + (k-1) = \frac{r(d-r)}{4} + \frac{r}{2} - 1$$

This also covers $r = 2$: then $k = 1$, no classes $u_j$ occur, and the estimate is exactly the contribution pulled back from $B$. Finally,

$$L \geq \mathrm{cup}(X_r; \mathbb{Q}) + 1 \geq \frac{r(d-r)}{4} + \frac{r}{2}$$

□

**Lemma 6.2.** *For $s \in \{d_k, d_v\}$, the space $\mathcal{M}_s$ strongly deformation retracts onto $X_s = (V_s(\mathbb{R}^d) \times V_s(\mathbb{R}^d))/O(s)$. Consequently, $\mathcal{M}_{\mathrm{head}}$ strongly deformation retracts onto $X_{d_k} \times X_{d_v}$.*

*Proof.* For each $s \in \{d_k, d_v\}$, the space $\mathcal{M}_s$ is exactly the bounded exact-rank space of Definition 5.1. Hence Lemma 5.3 gives a strong deformation retraction $\mathcal{M}_s \to X_s$. Taking the product of the two retractions gives a strong deformation retraction of $\mathcal{M}_{\mathrm{head}}$ onto $X_{d_k} \times X_{d_v}$. □

**Theorem 6.6.** *Assume that $d$ and $r$ are positive even integers with $r < d$, and that $d_k = d_v = r$. Let $c_{\mathrm{base}} = \mathrm{cup}(\mathrm{Gr}(r, d); \mathbb{Q})$. Then the MHA quotient satisfies $\mathrm{cup}(X_{\mathrm{MHA}}; \mathbb{Q}) \geq 2h\, c_{\mathrm{base}} + (h+1)\left(\frac{r}{2} - 1\right)$.*

[8] This is the standard behavior of Euler and Pontryagin classes under orientation reversal, see for example (Milnor & Stasheff, 1974, Property 9.3 and §15).

*Proof.* Set $c = c_{\text{base}}$. In the proof of Lemma 4.2, the ungraded ring isomorphism with the complex Grassmannian sends its rational Kähler class to a class $\alpha_{\text{Gr}} \in H^4(\text{Gr}(r,d);\mathbb{Q})$ satisfying $\alpha_{\text{Gr}}^c \neq 0$. The class $\alpha_{\text{Gr}}^{c+1}$ has degree $4(c+1)$, while $\text{Gr}(r,d)$ has real dimension $r(d-r) = 4c$. Since its cohomology vanishes in degrees above $4c$, we have $\alpha_{\text{Gr}}^{c+1} = 0$. Let $p : X_r \to \text{Gr}(r,d)$ be the LoRA bundle projection and set $\alpha = p^*(\alpha_{\text{Gr}})$. The section in Lemma 5.6 makes $p^*$ injective, so $\alpha^c \neq 0$, while $\alpha^{c+1} = 0$.

The core of one head is $H = X_r \times X_r$. Let $x$ and $y$ be the pullbacks of $\alpha$ from its first and second factors. The Künneth isomorphism (Hatcher, 2002, Theorem 3.15) gives $x^c y^c \neq 0$. For the ordered $h$-head space $Y = H^h$, let $x_i$ and $y_i$ denote the corresponding classes on the $i$th head and define

$$S_x = \sum_{i=1}^{h} x_i, \qquad S_y = \sum_{i=1}^{h} y_i.$$

Both classes have degree four and are fixed by the permutation action of $S_h$.

We next show that their top balanced product is nonzero. In the multinomial expansion of $S_x^{hc}$, a term with exponents $e_1, \ldots, e_h$ vanishes if some $e_i > c$. Since $\sum_i e_i = hc$, the only surviving choice is $e_1 = \cdots = e_h = c$. Therefore

$$S_x^{hc} = \frac{(hc)!}{(c!)^h} \prod_{i=1}^{h} x_i^c$$

The same argument applies to $S_y$. The Künneth isomorphism and characteristic zero now give

$$S_x^{hc} S_y^{hc} = \left(\frac{(hc)!}{(c!)^h}\right)^2 \prod_{i=1}^{h} x_i^c y_i^c \neq 0$$

We can also use the surviving fiber classes from the LoRA proof. Set $s = r/2 - 1$. From the proof of Theorem 5.7, the odd-degree classes $u_1, \ldots, u_s$ have square-free products independent over the pulled-back base cohomology. In particular, $\alpha^c u_1 \cdots u_s \neq 0$. Let $a_{ij}$ and $b_{ij}$ be the copies of $u_j$ in the first and second $X_r$ factors of head $i$, and define the three sums

$$U_j = \sum_{i=1}^{h} a_{ij}, \qquad V_j = \sum_{i=1}^{h} b_{ij}, \qquad E_j = \sum_{i=1}^{h} a_{ij} b_{ij}$$

Note that these classes are invariant under head permutations. Now, recall that odd classes square to zero over $\mathbb{Q}$ while the products $a_{ij} b_{ij}$ are even. Thus looking at $E_j^{h-1}$, each surviving term of the product fills $h-1$ distinct heads and occurs $(h-1)!$ times. There's one unfilled head, and hence we have

$$U_j V_j E_j^{h-1} = h! \prod_{i=1}^{h} a_{ij} b_{ij}$$

where the factor of $h$ in $h! = h(h-1)!$ comes from the fact that there are $h$ possible choices of the unfilled head. Together with the base calculation above, this gives

$$\Gamma = S_x^{hc} S_y^{hc} \prod_{j=1}^{s} \left(U_j V_j E_j^{h-1}\right) \neq 0$$

Up to a nonzero rational scalar and an overall sign, this is the Künneth product of $\alpha^c u_1 \cdots u_s$ over all $2h$ copies of $X_r$. When $s = 0$, we use the empty product.

Let $q : Y \to X_{\text{MHA}} = Y/S_h$ be the quotient map. By Theorem B.8, $q^*$ identifies $H^*(X_{\text{MHA}};\mathbb{Q})$ with the invariant subring of $H^*(Y;\mathbb{Q})$. Every positive-degree factor of $\Gamma$ therefore descends to $X_{\text{MHA}}$. Their product is nonzero because its pullback is $\Gamma$ and hence counting the factors gives us

$$\text{cup}(X_{\text{MHA}};\mathbb{Q}) \geq 2hc + (h+1)s = 2h\, c_{\text{base}} + (h+1)\left(\frac{r}{2} - 1\right)$$

□

**Theorem 6.8** (MHA recovery bound)**.** *Assume that $d$ and $r$ are positive even integers with $r < d$, and let $X_{\mathrm{MHA}}$ be the quotient for $h$ heads with $d_k = d_v = r$. Let $\rho_{\mathrm{MHA}} > 0$ be a uniform homotopy radius as in Remark 6.7. Any family of continuous recovery branches achieving $L$-list $\epsilon$-accurate stable proper recovery on $X_{\mathrm{MHA}}$ for $\epsilon < \rho_{\mathrm{MHA}}$ must have*

$$L \geq \frac{hr(d-r)}{2} + (h+1)\left(\frac{r}{2} - 1\right) + 1$$

*Proof.* By Remark 6.7, $X_{\mathrm{MHA}}$ is a compact ENR with a positive uniform homotopy radius. Hence Theorem 3.6 and the cup-length estimate give

$$L \geq \mathrm{cat}(X_{\mathrm{MHA}}) \geq \mathrm{cup}(X_{\mathrm{MHA}}; \mathbb{Q}) + 1$$

By Theorem 6.6, the cup-length is at least $2h\, c_{\mathrm{base}} + (h+1)(r/2-1)$. The even-parity calculation in Lemma 4.2 gives $c_{\mathrm{base}} = r(d-r)/4$. Thus substituting the bounds gives us

$$L \geq \frac{hr(d-r)}{2} + (h+1)\left(\frac{r}{2} - 1\right) + 1$$ □

**Remark C.3.** It is interesting to note that the additional term in the MHA bound comes from the surviving odd-degree LoRA fiber classes and the Euler-paired generator is still not used. It would be interesting to strengthen the exact lower bound further by understanding how the Euler-paired generator interacts with the heads.

## D FIGURES

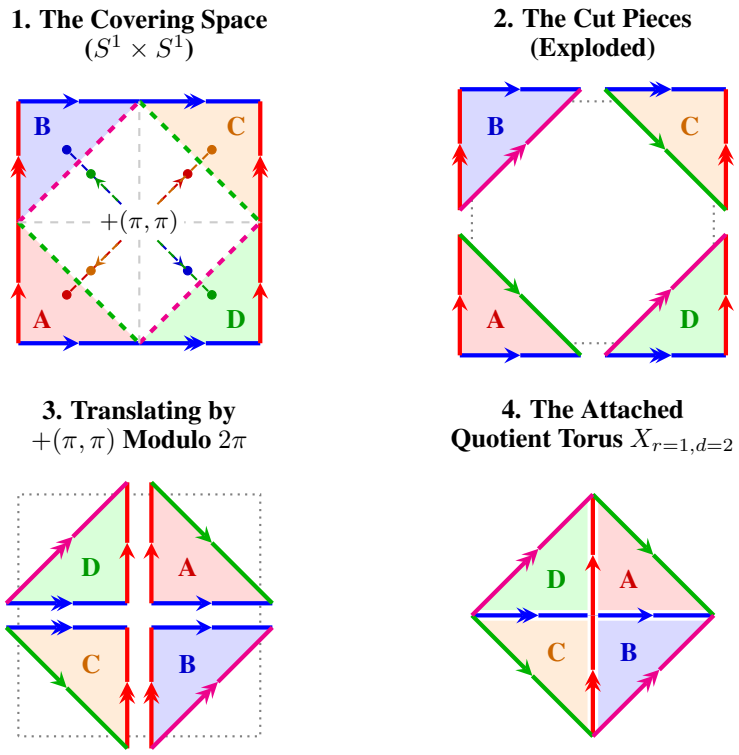



Figure 3: The cut-and-translate construction of $X_{r=1,d=2} = (S^1 \times S^1)/\mathbb{Z}_2$. **1.** The original $S^1 \times S^1$ covering space, with the central white diamond as a fundamental domain. The dashed arrows show the diagonal $+(\pi, \pi)$ action. **2.** The four corner pieces of the complementary fundamental domain are separated, retaining their boundary-identification labels. **3.** Translation by $+(\pi, \pi)$ modulo $2\pi$ moves the pieces into the fundamental domain. **4.** The reassembled pieces have the boundary identifications of a torus.

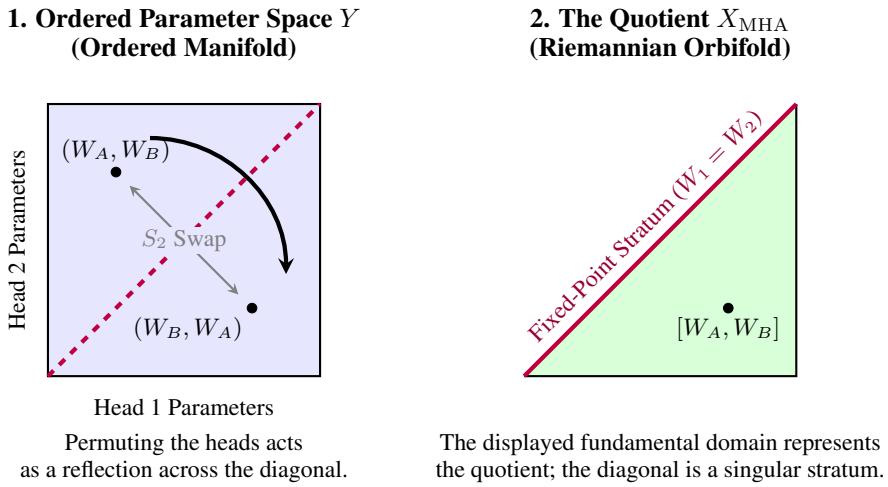



Figure 4: A schematic representation of the Multi-Head Attention quotient $X_{\mathrm{MHA}}$ for $h = 2$ heads. **1.** The ordered core $Y$ treats Head 1 and Head 2 as distinct. The permutation group $S_2$ acts by swapping the heads, shown schematically as reflection across the diagonal. **2.** Identifying this modeled head-permutation symmetry gives $X_{\mathrm{MHA}} = Y/S_2$, represented here by one fundamental domain. The fixed-point diagonal where the two heads coincide descends to a singular stratum of the quotient.

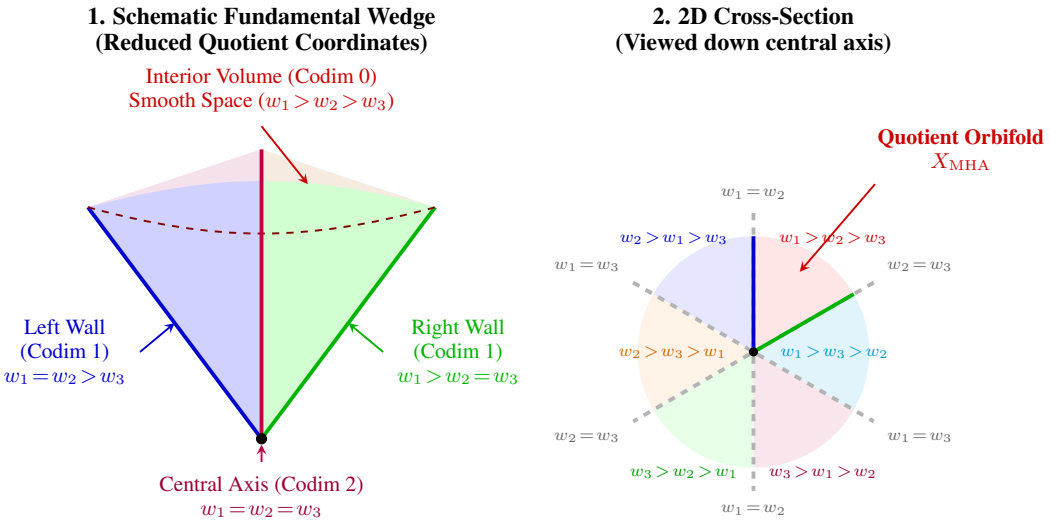



Figure 5: A schematic dimensional reduction of the Multi-Head Attention quotient $X_{\mathrm{MHA}}$ for $h = 3$ heads. **1.** One fundamental wedge represents the quotient in the reduced coordinates. Its mirror faces depict pairwise head-coincidence strata, and their common axis depicts the configuration where all three heads coincide. **2.** In the displayed ordered coordinate plane, the pairwise-equality loci divide the space into $3! = 6$ chambers. The permutation quotient identifies these chambers, so one chamber represents a fundamental domain. In the full parameter space, the codimensions of the coincidence strata depend on the dimension of a head.