Gauge Fiber Bundle Geometry of Transformers

Hong Wang HONG.WANG@INTEL.COM
Intel Corporation

Kelly Wang KELLY.WANG@IEEE.ORG
The Harker School

Abstract

We give a geometry-first account of Transformers with GeLU. Building on a companion NeurReps paper that *completely characterizes* the head-wise gauge symmetries of multi-head attention [23], we treat the maximal head-wise symmetry group as given and study the induced geometry on the resulting quotient of functionally distinct models. On a generic regular set of parameters, this symmetry group acts freely and properly, so the parameter space fibers over a quotient manifold with gauge orbits as fibers. We establish an Ehresmann connection using the ambient Euclidean metric, which resolves the degeneracy of the Fisher–Rao (FR) metric along gauge directions. This framework clarifies that the natural gradient is the *horizontal Riesz representative* of the Euclidean gradient with respect to the FR geometry on the quotient. We show the connection has generically nonzero curvature, implying path-dependent holonomy in parameter updates. We also clarify the roles of the Attention (MHA) and FFN blocks: while MHA parameters possess gauge symmetry, FFN gradients are strictly horizontal as the FFN parameters are invariant under the MHA gauge group. We turn these ideas into practical diagnostics–a gauge-aware gradient split and a small-loop holonomy estimator–and report consistency checks aligning with the theory. Architectural choices such as RoPE appear as principled gauge reductions (e.g., per-head Q/K dimension from d_k^2 to d_k).

1. Introduction

Standard multi–head self–attention layers admit systematic head–wise changes of basis on the key/query and value channels, together with permutations of heads, that leave the realized function unchanged. In this paper we analyze the resulting gauge group, the induced quotient manifold of functionally distinct parameters, and learning dynamics on that quotient.

Preliminaries. Let $\pi : \Theta_0 \to \mathcal{Q}$ denote the parameter–function map on a regular (Zariski–open) subset Θ_0 of parameters. The fibers of π are gauge orbits of a Lie group G. The tangent space $T_\theta\Theta$ admits a decomposition $T_\theta\Theta = \mathcal{V}_\theta \oplus \mathcal{H}_\theta$ into vertical directions $\mathcal{V}_\theta = \ker d\pi_\theta$ and horizontal directions $\mathcal{H}_\theta$. An Ehresmann connection specifies $\mathcal{H}_\theta$ smoothly. Because the empirical Fisher (Gauss–Newton) metric g_θ is degenerate along $\mathcal{V}_\theta$, we use the ambient Euclidean inner product to define the canonical (mechanical) connection via $\mathcal{H}_\theta = \mathcal{V}_\theta^{\perp_{\mathrm{Euc}}}$. Under a mild Fisher-regularity assumption on the data (Assumption 1), the only degeneracies of g_θ come from gauge directions, so its restriction to $\mathcal{H}_\theta$ is positive definite and induces a Riemannian geometry on the quotient $\mathcal{Q}$ by identifying horizontal vectors along gauge orbits.

© 2025 H. Wang & K. Wang.

Companion paper and scope. A separate companion paper [23] provides a complete classification of gauge symmetries in Transformer architectures. It proves that, on a generic stratum satisfying mild rank and distinctness conditions, the maximal head-wise gauge group for a standard multi-head attention (MHA) block is

$$G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h,$$

and develops variants for RoPE, grouped/multi–query attention, LayerNorm placement, and layer-wise factorization. In this *geometry-focused* paper we take those symmetry results as given, recall only the minimal statements we need (Section 2), and concentrate on the induced principal-bundle structure, canonical connection, curvature, diagnostics, and optimization implications.

This paper makes the following contributions:

- **Principal-bundle geometry.** On the generic stratum Θ_0 (Definition 1), we recall that the maximal head-wise gauge group $G_{\max}$ from [23] acts freely and properly, so $\pi : \Theta_0 \to \mathcal{Q} := \Theta_0/G_{\max}$ is a principal $G_{\max}$–bundle (Theorem 3). This identifies vertical directions with gauge orbits and sets up the quotient geometry.

- **Canonical connection and quotient Fisher geometry.** We define a canonical (Euclidean) connection by $\mathcal{H}_\theta = \mathcal{V}_\theta^{\perp_{\mathrm{Euc}}}$ (Section 3) and show that the empirical Fisher metric induces a nondegenerate Riemannian metric on the quotient. In this geometry, the natural gradient is the *horizontal* Riesz representative of the Euclidean gradient (Theorem 9).

- **Curvature, holonomy, and path dependence.** We analyze the curvature of the canonical connection, prove that it is generically nonzero (Theorem 10), and relate it to measurable small-loop holonomy in parameter space (Theorem 12), yielding a geometric account of path-dependent training dynamics.

- **MHA/FFN separation.** Within this framework, we show that gauge symmetry is localized to the MHA block, while FFN parameters lie entirely in the horizontal, function-changing directions, reflecting their invariance under the MHA gauge group (Proposition 11). This gives a clean geometric separation of roles between attention and feed-forward components.

- **Diagnostics, complexity, and empirical checks.** We turn the geometry into practical diagnostics: a least–squares vertical/horizontal gradient split and a small–loop holonomy estimator (Section 5), with complexity analysis that explains why we use Euclidean proxies in practice. Section 7 reports consistency checks that read existing Euclidean measurements through this quotient lens.

In this paper, we work with standard Transformer blocks with GeLU activations and fixed widths. Our results are proved on a generic stratum Θ_0 where the relevant projection operators have full column rank; outside Θ_0 stabilizers can grow and the $G_{\max}$–action stratifies by orbit type. Architectural variants fit naturally into this framework: rotary position embeddings (RoPE) restrict the Q/K factor to the plane–wise commutant, reducing the per–head gauge dimension from d_k^2 to d_k, while grouped/multi–query attention ties symmetry factors across heads.

For the remainder of the paper, Section 2 recalls the maximal head-wise symmetry and principal–bundle structure implied by $G_{\max}$ and its RoPE/GQA/MQA variants. Section 3 develops the canonical Euclidean connection and the induced Fisher–Rao geometry on the quotient. Section 4 analyzes the curvature of this connection and the separation between MHA and FFN gradients. Section 5 turns the geometry into practical diagnostics and discusses their computational complexity. Section 6 gives a Morse–Bott view of optimization on the quotient, and Section 7 reports Euclidean–proxy measurements read through this geometric lens. Section 8 records architectural variants and limitations, Section 9 situates our work within the broader literature, and Section 10 concludes. Proofs of maximality and gauge classification appear in [23]; proofs of the geometric statements appear in the appendices.

2. Maximal Gauge Symmetry and the Principal Bundle

We write h for the number of heads, d_k and d_v for key/query and value widths, and d_{model} for the model dimension. A single multi-head attention (MHA) layer is parameterized by

$$\theta \;=\; \big\{(W_Q^{(i)}, W_K^{(i)}, W_V^{(i)})_{i=1}^{h},\; W_O\big\},$$

and a depth-L model by $(\theta_1, \dots, \theta_L)$. The parameter manifold is an open set $\Theta \subset \mathbb{R}^D$, and $\pi : \Theta \to \mathcal{Q}$ denotes the parameter–function map to the quotient of functionally distinct models.

Definition 1 (Generic stratum) *Let $\Theta_0 \subset \Theta$ be the set of parameters such that, for every head i in every layer:*

(G1) *(*Full column rank*)* $\operatorname{rank}(W_Q^{(i)}) = \operatorname{rank}(W_K^{(i)}) = d_k$ *and* $\operatorname{rank}(W_V^{(i)}) = d_v$.

(G2) *(*Full row rank outputs*)* *Writing W_O in block rows $W_{O,1}, \dots, W_{O,h} \in \mathbb{R}^{d_v \times d_{model}}$, the i-th block row has full row rank:*

$$\operatorname{rank}(W_{O,i}) = d_v.$$

(G3) *(*Canonical dimensions*)* $d_{model} = h\, d_v$.

(G4) *(*Head-wise controllability/identifiability*)* $\operatorname{rank}([W_Q^{(i)} \mid W_K^{(i)}]) = 2d_k$ *and* $d_{model} \geq 2d_k$.

(G5) *(*Head distinctness*)* *For any $i \neq j$, the pairs*

$$\big(W_Q^{(i)}(W_K^{(i)})^\top,\; W_V^{(i)} W_{O,i}\big) \quad \textit{and} \quad \big(W_Q^{(j)}(W_K^{(j)})^\top,\; W_V^{(j)} W_{O,j}\big)$$

are distinct. Equivalently, no two heads realize the same attention map and value/output composition.

These conditions define a Zariski–open subset of the full parameter space Θ and coincide (up to minor notational differences) with the generic stratum used in the companion maximality paper [23].

Gauge group and action. On this generic stratum, the head-wise gauge group for standard MHA is

$$G_{\max} \;=\; \Big((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h\Big) \rtimes S_h.$$

Two families of changes of basis preserve the realized function: invertible transforms in the Q/K channels for each head, and invertible transforms in the V channels paired with a compensating block in W_O; heads may also be permuted. The group $G_{\max}$ acts by

$$(W_Q^{(i)}, W_K^{(i)}) \mapsto \big(W_Q^{(i)} A_i,\; W_K^{(i)} A_i^{-\top}\big), \qquad (W_V^{(i)}, W_{O,i}) \mapsto \big(W_V^{(i)} C_i,\; C_i^{-1} W_{O,i}\big),$$

for $(A_i) \in (\mathrm{GL}(d_k))^h$, $(C_i) \in (\mathrm{GL}(d_v))^h$, and by $\sigma \in S_h$ permuting heads.

Theorem 2 (Maximal gauge group for a single attention layer) *On Θ_0, the group of all parameter transformations that preserve the multi-head attention block equals*

$$G_{\max} \;=\; \Big((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h\Big) \rtimes S_h,$$

acting head-wise as above. No additional continuous or discrete symmetries exist.

Proof reference. This is a specialization of the global maximality theorem proved in the companion paper [23], restricted to a single layer with conditions (G1)–(G5) (matching assumptions A1–A4 and A6–A8 in [23]). For full proofs, including Lie-algebra, identifiability, and factorization arguments, see Theorem 2 and Theorem 22 in [23].

Theorem 3 (Principal bundle on the generic stratum) *Let $G_{\max}$ act as above. On the Zariski–open regular stratum Θ_0 the action is* free and proper*; hence $\pi : \Theta_0 \to \mathcal{Q} := \Theta_0/G_{\max}$ is a principal $G_{\max}$–bundle.*

Proof sketch. Freeness of the continuous $((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h)$ part follows from the maximality and identifiability results in [23]: on Θ_0, preserving all $Q_i K_i^\top$ and $V_i W_{O,i}$ forces the query–key and value–output transformations to be trivial. Condition (G5) in Definition 1 (head distinctness) rules out nontrivial stabilizing permutations in S_h: if $\sigma \neq \mathrm{id}$ and $\sigma \cdot \theta = \theta$, then some pair of heads would share both $W_Q^{(i)}(W_K^{(i)})^\top$ and $W_V^{(i)} W_{O,i}$, contradicting (G5). Thus the $G_{\max}$–action on Θ_0 is free. Properness of the explicit linear action is standard for products of GL-actions on full-rank matrix manifolds (Stiefel-type actions) together with a finite permutation group, and the action is algebraic (hence smooth and continuous) in the matrix entries. A short argument is given in Appendix A. The principal-bundle structure then follows from general results on free, proper Lie-group actions [11; 12].

Quotient geometry and bundle picture. Theorem 3 identifies the parameter–function map $\pi : \Theta_0 \to Q := \Theta_0/G_{\max}$ as a principal $G_{\max}$–bundle on the regular stratum. Geometrically, Θ_0 is highly redundant: many parameter points $\theta \in \Theta_0$ realize the same function class $[\theta] \in Q$. The gauge group $G_{\max}$ acts freely and properly, and each orbit $G_{\max} \cdot \theta$ is a fiber of π:

$$G_{\max} \hookrightarrow \Theta_0 \xrightarrow{\pi} Q := \Theta_0/G_{\max}.$$

In this picture, moving along a fiber (a gauge orbit) changes the parameterization but leaves the realized function unchanged, while moving across fibers changes the function.

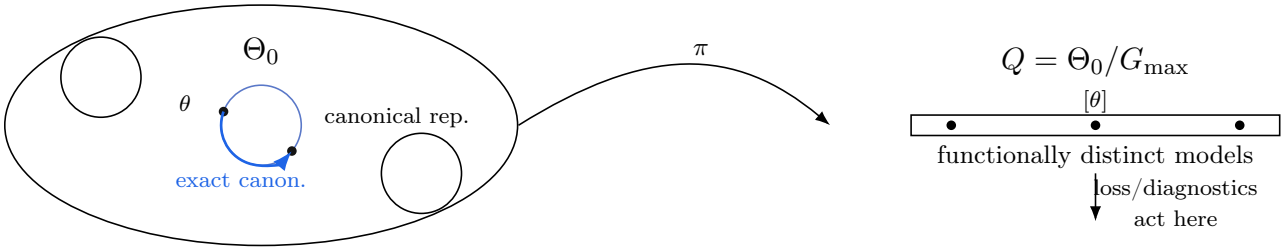

Figure 1: Quotient geometry schematic. Gauge orbits in Θ_0 collect function–equivalent parameterizations; quotienting by $G_{\max}$ yields the base manifold $Q = \Theta_0/G_{\max}$ of functionally distinct models. Exact canonicalization moves *along* an orbit (fiberwise, function–preserving), while gauge–invariant losses and diagnostics depend only on the quotient class $[\theta] \in Q$.

Equipping Θ_0 with the empirical Fisher metric g_θ and the vertical spaces $V_\theta = \ker d\pi_\theta$, the canonical Euclidean connection chooses the horizontal complement $H_\theta = V_\theta^{\perp_{\text{Euc}}}$ (Section 3). Under the Fisher-regularity assumption (Assumption 1), g_θ is nondegenerate on H_θ, so it induces a Riemannian metric on the quotient Q by identifying horizontal lifts along fibers. Infinitesimal parameter updates split into vertical, gauge-preserving directions in V_θ and horizontal, function-changing directions in H_θ; gauge-invariant quantities such as the quotient loss and the diagnostics of Section 5 depend only on $[\theta] \in Q$ and are therefore constant along fibers.

Figures 1 and 2 provide a schematic view of the principal-bundle geometry introduced above. In Figure 1, the regular parameter stratum Θ_0 is pictured as being foliated by gauge orbits: each orbit collects parameters related by the head-wise gauge group $G_{\max}$ that all implement the same function, and the quotient $Q = \Theta_0/G_{\max}$ collapses each orbit to a single class $[\theta]$. Moving along a circle in Θ_0 corresponds to an exact gauge transformation (e.g., canonicalization) that preserves the network function, while moving on the line $\mathcal{M}$ / Q changes the realized function. Figure 2 reorganizes the same objects in the "bundle over base" layout: the total space Θ_0 lives above the base Q, vertical arrows are gauge fibers, and horizontal arrows depict horizontal lifts of base moves under the Euclidean connection $H_\theta = V_\theta^{\perp_{\text{Euc}}}$. This cartoon is the geometric backdrop for our later results: natural gradients live in the horizontal directions of this bundle, and curvature of the connection shows up as path-dependent holonomy when we follow small loops in Q.

Corollary 4 (Head sharing (GQA/MQA)) *If the h heads are partitioned into g key/-value groups with shared (W_K, W_V) per group, then on the corresponding generic stratum the continuous symmetry reduces to $(\mathrm{GL}(d_k))^g \times (\mathrm{GL}(d_v))^g$ tied per group, with permutations $S_h \times S_g$ acting discretely.*

Corollary 5 (RoPE reduction) *Under rotary position embeddings with a nondegenerate frequency schedule and even d_k, the Q/K factor reduces on each 2×2 plane to the real commutant $\{aI + bJ\} \cong \mathrm{GL}(1, \mathbb{C})$, giving*

$$G_{\text{RoPE}} \;=\; \big((C_{\text{RoPE}})^h \times (\mathrm{GL}(d_v))^h\big) \rtimes S_h, \quad C_{\text{RoPE}} \cong \mathrm{GL}(1,\mathbb{C})^{d_k/2}.$$

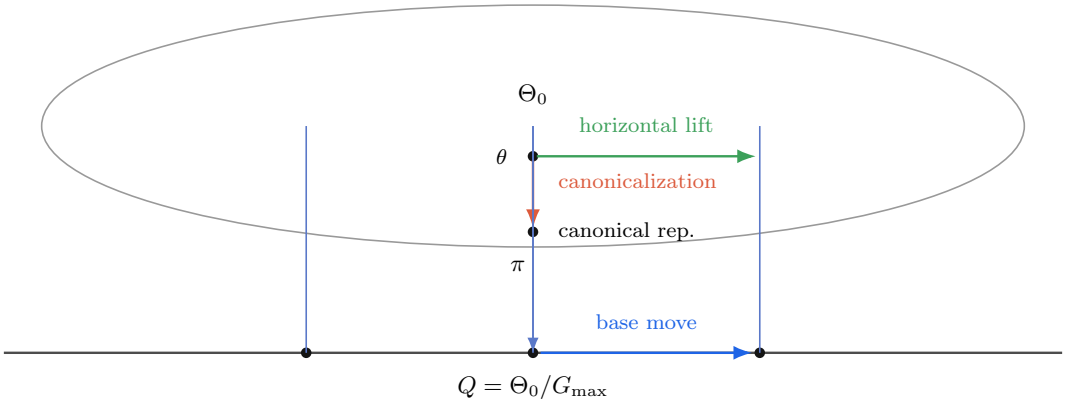

Figure 2: Fiber-bundle view with the **parameter space** Θ_0 (total space, top) and the **function space** $Q = \Theta_0/G_{\max}$ (base, bottom). Vertical fibers are gauge orbits; vertical arrows depict fiberwise canonicalization, and the projection π carries parameters down to their function class. Horizontal arrows in Θ_0 represent horizontal lifts of function-changing moves on Q, as determined by the canonical connection.

Consequently, the per-head Q/K *gauge dimension drops from* d_k^2 *to* d_k.

Corollary 6 (Layerwise product) *For a depth-L model without cross-layer parameter sharing, the model-level gauge group is the direct product*

$$G_{\text{model}} \cong \prod_{\ell=1}^{L} G_{\max}^{(\ell)} \quad (\text{or } \prod_{\ell=1}^{L} G_{\text{share}}^{(\ell)}, \prod_{\ell=1}^{L} G_{\text{RoPE}}^{(\ell)} \text{ in the corresponding variants}).$$

Corollary 7 (Continuous dimension) *For a single layer,* $\dim_{\mathbb{R}} G_{\max}^{\circ} = h(d_k^2 + d_v^2)$*; with RoPE,* $\dim_{\mathbb{R}} G_{\text{RoPE}}^{\circ} = h(d_k + d_v^2)$*; with head sharing into g groups, replace h by g in these counts.*

Proof references for Corollaries. These corollaries follow directly from the RoPE commutant analysis, grouped/multi–query head sharing, and layerwise product structure developed in [23] (see, e.g., Theorem 2, Theorem 8, Theorem 31, and Theorem 33 there). We do not re-derive those arguments here.

Geometric consequences. Theorem 3 identifies the vertical space $\mathcal{V}_\theta = \ker d\pi_\theta$ as the tangent to the gauge orbit through θ. In Section 3 we define the canonical connection using the ambient Euclidean inner product, taking $\mathcal{H}_\theta = \mathcal{V}_\theta^{\perp_{\text{Euc}}}$; the empirical Fisher metric g_θ is then restricted to $\mathcal{H}_\theta$ to induce a nondegenerate Riemannian metric on the quotient $\mathcal{Q}$. This horizontal/vertical calculus underpins the optimization and curvature results that follow.

3. The Connection and Empirical Fisher Geometry

By Theorem 3, the parameter–to–function map $\pi : \Theta_0 \to \mathcal{Q}$ is a principal bundle. At $\theta \in \Theta_0$, the *vertical* space

$$\mathcal{V}_\theta \;=\; \ker d\pi_\theta$$

is the tangent to the gauge orbit through θ. An Ehresmann connection requires choosing a complementary *horizontal* subspace $\mathcal{H}_\theta$ such that $T_\theta\Theta_0 = \mathcal{V}_\theta \oplus \mathcal{H}_\theta$.

Degeneracy of the empirical Fisher metric. We aim to study the geometry induced by the empirical Fisher (Gauss–Newton) metric g_θ on Θ_0. However, g_θ is degenerate on the total space. Since the network function is invariant along gauge orbits, every vertical direction lies in the null space (radical) of g_θ:

$$g_\theta(v, w) = 0 \quad \forall v \in \mathcal{V}_\theta,\ \forall w \in T_\theta\Theta_0.$$

In principle the data distribution $\mathcal{D}$ could introduce additional degeneracies beyond gauge directions. Throughout we therefore work under the Fisher-regularity Assumption 1, which states that $\ker g_\theta$ coincides with $\mathcal{V}_\theta$. Under this assumption, the restriction of g_θ to the horizontal space $\mathcal{H}_\theta$ is positive definite, so it induces a Riemannian metric on the quotient $\mathcal{Q}$.

Canonical (Euclidean) connection. To obtain a well-posed Ehresmann decomposition, we use the ambient Euclidean metric $\langle\cdot,\cdot\rangle$ on Θ_0 and define the canonical connection by

$$\mathcal{H}_\theta \ := \ \mathcal{V}_\theta^{\perp_{\mathrm{Euc}}}. \tag{3.1}$$

This yields a smooth, complementary distribution satisfying $T_\theta\Theta_0 = \mathcal{V}_\theta \oplus \mathcal{H}_\theta$; it is exactly the standard *mechanical connection* associated with this free isometric $G_{\max}$–action.

Lemma 8 (Gauge-null gradient directions) *If $L : \Theta_0 \to \mathbb{R}$ is gauge-invariant, then $dL_\theta[v] = 0$ for all $v \in \mathcal{V}_\theta$. Equivalently, the Euclidean gradient ∇L (the Riesz representative of dL_θ with respect to the Euclidean metric) is orthogonal to $\mathcal{V}_\theta$, i.e. $\nabla L \in \mathcal{H}_\theta$.*

Fisher geometry on the quotient. While the connection is defined via the Euclidean metric, the relevant geometry for optimization is the one induced by the empirical Fisher metric on the quotient $\mathcal{Q}$. Under Assumption 1, the degenerate form g_θ on Θ_0 induces a unique, nondegenerate Riemannian metric on $\mathcal{Q}$: informally, g_θ becomes nondegenerate once we quotient out gauge directions. Concretely, this corresponds to restricting g_θ to the horizontal subspace $\mathcal{H}_\theta$, where it is positive definite on Θ_0. When g_θ coincides with the classical Fisher–Rao metric (e.g., for log-likelihood losses), this recovers the usual Fisher–Rao geometry on the quotient.

The natural (Riemannian) gradient $\widetilde{\nabla} L$ is defined with respect to this quotient geometry.

Theorem 9 (Natural gradient as horizontal Riesz representative) *At $\theta \in \Theta_0$, the natural gradient is the unique vector $\widetilde{\nabla} L \in \mathcal{H}_\theta$ such that*

$$g_\theta(\widetilde{\nabla} L,\, w) = dL_\theta[w] \quad \forall\, w \in \mathcal{H}_\theta.$$

In coordinates, since $\nabla L \in \mathcal{H}_\theta$ by Lemma 8, the natural gradient is given by

$$\widetilde{\nabla} L \ = \ (G_{\theta|\mathcal{H}_\theta})^{-1}\, \nabla L, \tag{3.2}$$

where $G_{\theta|\mathcal{H}_\theta}$ is the Fisher information restricted to $\mathcal{H}_\theta$.

The decomposition of a vector $u \in T_\theta\Theta_0$ into vertical and horizontal components is now the standard Euclidean orthogonal projection. Given a vertical generator set $\{v_j\}_{j=1}^m$ stacked into a matrix A, the projection solves $(A^\top A)c = A^\top u$ and

$$u_{\text{vert}} = Ac, \qquad u_{\text{hor}} = u - u_{\text{vert}}.$$

Section 5 formalizes this procedure.

4. Geometry of the Connection

The canonical (Euclidean) connection from Section 3 provides the Ehresmann decomposition $T_\theta\Theta_0 = \mathcal{V}_\theta \oplus \mathcal{H}_\theta$. We now analyze the geometry induced by this connection, focusing on its curvature and the behavior of the Attention (MHA) and Feed-Forward (FFN) blocks within this framework.

4.1. Curvature and path dependence in parameter space

The curvature of an Ehresmann connection measures the extent to which the horizontal distribution $\mathcal{H}_\theta$ fails to be integrable. Geometrically, it quantifies how the space of function-changing directions (the horizontal space) twists as we move through Θ_0.

Interpretation of curvature: holonomy. Nonzero curvature implies path dependence (holonomy). Lifting a closed loop in the quotient $\mathcal{Q}$ horizontally to Θ_0 starting at θ results in an endpoint $\theta' = g \cdot \theta$. This gauge transformation g is the holonomy of the loop.

In the context of optimization, this means the order of infinitesimal horizontal parameter updates matters for the final gauge configuration. Applying a horizontal update u then v versus v then u results in the same functional change (to second order) but different final parameters, related by a gauge transformation determined by the curvature $\Omega(u, v)$.

Theorem 10 (Connection curvature is generically nonzero) *On the generic stratum Θ_0 (Definition 1), the curvature two-form Ω of the canonical connection on the parameter bundle $\pi : \Theta_0 \to \mathcal{Q}$ is generically nonzero.*

Proof reference. Appendix C. The conditions for flatness (integrability of $\mathcal{H}_\theta$) impose algebraic constraints on the vertical generators that are not structurally satisfied by the MHA architecture; hence, nonvanishing is generic.

This geometric property underpins the holonomy estimator (Algorithm 2), which directly measures this path dependence.

4.2. Attention and FFN gradients

We examine how gradients of the MHA and FFN blocks behave relative to the Euclidean horizontal/vertical split. We assume the standard Transformer block structure where MHA and FFN parameters are disjoint, leading to an orthogonal decomposition of the tangent space $T\Theta = T\Theta_{\text{MHA}} \oplus T\Theta_{\text{FFN}}$.

Proposition 11 (FFN horizontality and MHA/FFN separation) *Let $G_{\max}$ be the MHA gauge group acting only on Θ_{MHA}. Then:*

1. *The vertical space is contained entirely within the MHA subspace: $\mathcal{V}_\theta \subset T\Theta_{\text{MHA}}$.*

2. *The FFN tangent space is strictly horizontal:* $T\Theta_{\mathrm{FFN}} \subset \mathcal{H}_\theta$. *In particular, for any gauge-invariant loss* L *we have* $\nabla_{\mathrm{FFN}} L \in \mathcal{H}_\theta$.

Proof (1) By definition, $G_{\max}$ acts trivially on Θ_{FFN}. The tangent space to the orbits (the vertical space $\mathcal{V}_\theta$) therefore has zero component in the $T\Theta_{\mathrm{FFN}}$ directions, so $\mathcal{V}_\theta \subset T\Theta_{\mathrm{MHA}}$.

(2) The horizontal space $\mathcal{H}_\theta$ is the Euclidean orthogonal complement of $\mathcal{V}_\theta$. Since $T\Theta_{\mathrm{FFN}}$ is orthogonal to $T\Theta_{\mathrm{MHA}}$ in the product Euclidean metric and $\mathcal{V}_\theta \subset T\Theta_{\mathrm{MHA}}$, we have $T\Theta_{\mathrm{FFN}} \subset \mathcal{H}_\theta$. For any gauge-invariant L, Lemma 8 implies that ∇L is horizontal; its FFN block $\nabla_{\mathrm{FFN}} L$ therefore lies in $T\Theta_{\mathrm{FFN}} \subset \mathcal{H}_\theta$. ■

Implications. This proposition rigorously confirms that FFN gradients contribute purely to functional change and are geometrically separated from the MHA gauge redundancy within this Ehresmann framework: gauge symmetry is localized to the attention mechanism, while FFN parameters live entirely in the horizontal, function-changing subspace.

5. Diagnostics and Complexity

Two concrete procedures make the geometry from Sections 3 and 4 operational. The first resolves a vector into its gauge (vertical) and function-changing (horizontal) parts with respect to the Fisher–Rao metric. The second turns curvature into a measurable small-loop effect. We use the first with *Euclidean* inner products in Section 7 as a scalable proxy; the second is presented here with guarantees and left for future large-scale Fisher–Rao evaluation.

Gauge–aware decomposition. Let $\{v_j\}_{j=1}^m$ span the vertical space $\mathcal{V}_\theta$ and stack them into a matrix $A = [\,\mathrm{vec}(v_1)\ \cdots\ \mathrm{vec}(v_m)\,]$. For a vector u (e.g., $u = \nabla L$), the canonical (Euclidean) projection onto the vertical space is obtained by solving the least–squares problem $(A^\top A + \lambda I)c = A^\top u$, giving $u_{\mathrm{vert}} = Ac$ and $u_{\mathrm{hor}} = u - u_{\mathrm{vert}}$. In practice we stabilize the solution using thin–QR, Cholesky, or CG with column–pivoting; these choices do not change the canonical connection.

Algorithm 1 Gauge–aware gradient decomposition (Canonical connection)

Require: parameter $\theta \in \Theta_0$; vertical generators $\{v_j\}_{j=1}^m$ stacked in A; vector u

1: Form the Gram matrix $G \leftarrow A^\top A$ and right-hand side $b \leftarrow A^\top u$
2: Solve $(G + \lambda I)c = b$ (least squares; optional $\lambda \geq 0$)
3: $u_{\mathrm{vert}} \leftarrow Ac, \quad u_{\mathrm{hor}} \leftarrow u - u_{\mathrm{vert}}$
4: **return** $(u_{\mathrm{vert}}, u_{\mathrm{hor}})$ and vertical fraction $\|u_{\mathrm{vert}}\|/\|u\|$

From curvature to holonomy. For g_θ-orthonormal $u, v \in \mathcal{H}_\theta$, consider the horizontal loop that moves by $+\varepsilon u$, $+\varepsilon v$, then returns by $-\varepsilon u$, $-\varepsilon v$.

Theorem 12 (Small-loop holonomy scaling) *The induced gauge displacement* $\Delta_{\square_\varepsilon}(u, v)$ *obeys*

$$\|\Delta_{\square_\varepsilon}(u, v)\| \;=\; \varepsilon^2\, \|\Omega_\theta(u, v)\| \;+\; O(\varepsilon^3),$$

where Ω_θ *is the curvature two-form of the canonical connection at* θ.

Table 1: Cost at a glance for one layer (parameter dim D, generators $m = h(d_k^2+d_v^2)$; with RoPE: $m = h(d_k+d_v^2)$).

Procedure	Leading cost	Notes
Euclidean vertical split	form $A^\top A$: $O(m^2 D)$; solve: $O(m^3)$	CG: $O(mD)$ per matvec
FR vertical split	form Gram G: $O(m^2 \cdot \texttt{FisherEval})$; solve: $O(m^3)$	Heavy for $m \gtrsim 10^4$
Holonomy (loop + LS)	4 flows + LS on $\mathfrak{g}$	Richardson reduces $O(\varepsilon^3)$ bias

Algorithm 2 Holonomy estimator with Richardson extrapolation

Require: θ; g_θ-orthonormal $u, v \in \mathcal{H}_\theta$; steps $\varepsilon > \varepsilon' > 0$

1: For each $\delta \in \{\varepsilon, \varepsilon'\}$, traverse the horizontal loop using FR projection at each leg
2: Compute $\Delta_{\square_\delta}(u, v)$ (Lie-algebra coordinates via a log map)
3: $h(\delta) \leftarrow \|\Delta_{\square_\delta}(u, v)\|/\delta^2$
4: **return** $h^* \leftarrow \dfrac{4h(\varepsilon/2) - h(\varepsilon)}{3}$ (Richardson; bias $O(\varepsilon^3)$)

Cost at a glance. Let $m = h(d_k^2+d_v^2)$ (RoPE reduces $d_k^2 \to d_k$ per head). Forming $A^\top A$ costs $O(m^2 D)$ for parameter dimension D, and solving costs $O(m^3)$ (or CG with matvec $O(mD)$). At h=12, d_k=d_v=64, one has $m \approx$ 98,304, so FR projectors and holonomy become heavy; we therefore use Euclidean proxies in Section 7 and expose FR-exact procedures here for future evaluation.

These two diagnostics tie the abstract calculus to practice: the first reports how much of a step is "just gauge," the second quantifies the path dependence predicted by curvature. In Section 7 we read existing Euclidean measurements through this lens; FR-exact holonomy is left for future large-scale runs.

6. Optimization on the Quotient: A Morse–Bott View

Gauge symmetry means many parameter settings realize the same function. On the total space Θ_0 this appears as flat directions tangent to gauge orbits; on the quotient $\mathcal{Q} = \Theta_0/G_{\max}$ those directions disappear and the landscape reflects genuine functional change. The right language is Morse–Bott: critical sets in Θ_0 are manifolds (orbits), while the induced problem on $\mathcal{Q}$ is Morse once we restrict to horizontal directions.

Theorem 13 (Gauge orbits as critical manifolds; Morse behavior on the quotient) *Let $L : \Theta_0 \to \mathbb{R}$ be gauge-invariant and let $\theta \in \Theta_0$ be critical. Then the entire orbit $G_{\max} \cdot \theta$ lies in the critical set, the Hessian $\nabla^2 L(\theta)$ vanishes on the vertical space $\mathcal{V}_\theta = \ker d\pi_\theta$, and its horizontal restriction on $\mathcal{H}_\theta$ is well defined. Writing $\ell : \mathcal{Q} \to \mathbb{R}$ for the induced loss, $[\theta]$ is critical for ℓ and has nondegenerate Hessian equal to the horizontal restriction of $\nabla^2 L(\theta)$. In particular, ℓ is Morse at $[\theta]$ whenever the horizontal Hessian is nondegenerate.*

Proof. Appendix D. The argument uses the slice construction around a free, proper orbit and the horizontal/vertical split from Section 3.

Table 2: Gauge invariance (relative errors across 100 trials). Outputs remain invariant up to machine precision.

Metric	Value	Interpretation
Mean relative difference	2.68×10^{-15}	Machine precision
Maximum relative difference	2.86×10^{-15}	$\approx 13\,\varepsilon_{\text{mach}}$
Std. deviation	9.41×10^{-17}	Highly consistent
Minimum relative difference	2.52×10^{-15}	$\approx 11\,\varepsilon_{\text{mach}}$

Two consequences are worth keeping in mind. First, methods that suppress vertical components–see Algorithm 1–are aligned with the true second-order structure of the quotient problem: they follow directions that actually change the function. Second, small horizontal steps can move far in Euclidean parameter norm while staying close in function space, which explains why Euclidean distances often overstate functional change and why seemingly "distant" minima in Θ_0 can sit in the same basin on $\mathcal{Q}$.

The result is local to the generic stratum: near Θ_0 the bundle picture applies cleanly; away from it stabilizers may grow and the space stratifies. Our empirical reading in Section 7 stays within Θ_0 and interprets existing (Euclidean-proxy) measurements through this quotient lens.

7. Empirical Consistency Checks (Euclidean Proxies)

All measurements in this section use *Euclidean* inner products as computationally tractable *proxies* for the Fisher–Rao geometry. They provide conservative validation of the bundle predictions: Euclidean angles *lower bound* Fisher–Rao angles, and Euclidean vertical fractions *upper bound* horizontal leakage. These proxies are consistent with the theory but *do not constitute full Fisher–Rao validation.* FR–exact procedures appear in Section 5 and proofs in the appendix; we do not add new runs beyond the existing computations reported here.

7.1. Gauge invariance

A basic check is functional invariance under the head-wise symmetry of Section 2. We apply controlled transforms from $G_{\max}$ to MHA layers (h=12, d_k=d_v=64), including random head permutations and well-conditioned Q/K and V/O changes of basis generated by thin–QR with prescribed singular values; outputs are compared on fixed inputs.

As a control, transformations *outside* $G_{\max}$ (e.g., cross–head mixing) induce $\mathcal{O}(1)$ changes, consistent with maximality. The same precision holds after 1,000 gradient steps on a reconstruction objective.

7.2. Gauge–aware gradient split (Euclidean proxy)

The bundle theory (Lemma 8, Theorem 9) predicts horizontally aligned gradients in Fisher geometry. We test the *proxy* claim by replacing the Fisher inner product in Algorithm 1 with dot products: stack vertical generators into A and solve $(A^\top A)c = A^\top \nabla L$. On 50 independent samples, we report the Euclidean vertical fraction $\|P_{\mathcal{V}}^{(\text{Euc})} \nabla L\| / \|\nabla L\|$.

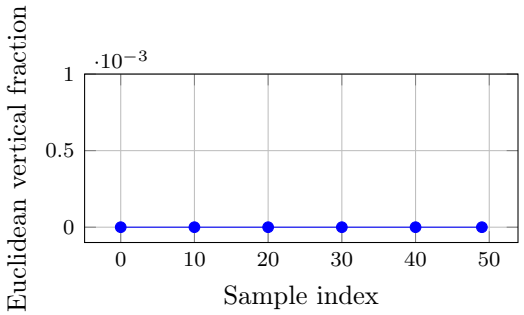

Figure 3: Gauge–aware split (Euclidean proxy). The Euclidean vertical fraction stays below a 10^{-4} threshold across 50 samples, consistent with horizontality predicted by Theorem 9.

The Figure 3 is uniform across layers and seeds: vertical fractions remain at numerical precision, and the horizontal component accounts for essentially all of the norm. This matches the Riesz characterization in Theorem 9–in practice, gradient steps align with the quotient geometry.

The near-zero (Euclidean) vertical fractions indicate gradients lie almost entirely in function-changing (horizontal) directions, consistent with Theorem 9 for gauge-invariant objectives.

Note on scope. Angles between attention and FFN gradients are likewise reported with Euclidean inner products; they should be read as lower bounds on Fisher angles. FR–exact angle and holonomy measurements are enabled by Section 5 and left for future large-scale evaluation.

8. Architectural Variants and Limitations

Real systems add positional structure and sharing patterns that slightly change the symmetry–and with it, the bundle–without altering the main thread of our results. We record the two common cases and then state the limits of our analysis in plain terms.

Rotary position embeddings (RoPE). RoPE rotates Q/K channels in 2×2 planes determined by frequencies. On each plane the admissible head-wise transform collapses to the real commutant $\{aI + bJ\}$ of that rotation, so the Q/K factor of $G_{\max}$ reduces plane-wise. A useful way to remember the effect is dimensional: per head, the Q/K gauge dimension drops from d_k^2 to d_k, while the V/O factor stays at d_v^2. All statements in Sections 2–6 remain true after substituting this reduced group and restricting the generic stratum accordingly; the canonical connection remains $\mathcal{H}_\theta = \mathcal{V}_\theta^{\perp_{\mathrm{Euc}}}$ with the (now smaller) vertical space, and the Fisher–Rao metric is restricted to $\mathcal{H}_\theta$ to induce the quotient geometry as before.

Grouped and multi–query attention (GQA/MQA). When heads share key/value projections, the layer symmetry is no longer a direct product across individual heads but across groups that share parameters. Concretely, one replaces $G_{\max}$ by a product over groups, with the shared $\mathrm{GL}(d_k)$ or $\mathrm{GL}(d_v)$ factor acting jointly on the grouped channels. The principal-bundle theorem and the free–proper argument are unchanged under this

substitution, and the diagnostics of Section 5 apply verbatim once the vertical generator set respects the sharing pattern.

Limitations. Our claims are local and precise by design. They are proved on a Zariski–open regular stratum Θ_0 where stabilizers are trivial; away from Θ_0 the orbit type changes and the ambient space stratifies, so we avoid global topological statements (e.g., global connectivity of minima). The Fisher–Rao connection depends on the evaluation batch (standard in information geometry); consequently, numerical angles or vertical fractions are batch–specific. For scalability, the empirical section reports *Euclidean* proxies; these are consistent with the theory but *do not constitute full Fisher–Rao validation*. The Fisher–Rao–exact procedures are presented in Section 5 for future large-scale evaluation. We restrict to smooth activations (GeLU) so that the curvature expansion and Fisher calculus are well-defined; non-smooth activations such as ReLU require a stratified or Clarke–generalized treatment and fall outside our scope here. Finally, Fisher–Rao projectors and holonomy are costly at scale (they solve Gram systems and perform horizontal projections along loops); we therefore expose the methods and guarantees but do not add new measurements beyond the existing Euclidean computations.

Takeaway. RoPE and GQA/MQA change the structure group in controlled ways–shrinking or tying Q/K symmetries–while the principal-bundle picture and the underlying geometric framework remain intact: the canonical connection is still defined by $\mathcal{H}_\theta = \mathcal{V}_\theta^{\perp_{\mathrm{Euc}}}$ (with a correspondingly adjusted vertical space), and the Fisher–Rao metric is still restricted to $\mathcal{H}_\theta$ to induce the quotient geometry. The limitations above mark exactly where our guarantees apply and explain why we pair Euclidean–proxy measurements with Fisher–Rao–exact algorithms.

9. Related Work

Symmetries in neural networks. Permutation invariances in multilayer perceptrons have been recognized for decades [9; 1; 8], and convolutional networks admit translation symmetries naturally modeled by group actions [6]. For Transformers, recent analyses have documented partial symmetries and superposition effects [10; 7], but a complete account of the *maximal* head–wise gauge and its consequences has been missing. Our work fills this gap by proving a principal-bundle structure on a generic stratum (see 3) and developing a connection–curvature calculus on the quotient.

Information geometry and optimization. Natural gradient methods [2; 4; 3; 14] and their scalable approximations [18; 16; 5] bring Riemannian structure to learning, typically as an algorithmic tool. Here the Fisher metric plays a different role: it *selects a connection* on the principal bundle, and the natural gradient is the Riesz representative restricted to the horizontal subspace (Theorem 9). Prior studies of Transformer optimization and loss geometry [13; 22] report phenomena–flat directions, easy-to-traverse basins–that our quotient viewpoint explains cleanly: gauge orbits give Morse–Bott critical manifolds on Θ_0, while the induced loss on the base is Morse (Theorem 13).

Interpretability and model merging. Mechanistic interpretability proceeds by reverse engineering circuits and features [17; 15]. The bundle picture complements that line of work: fibers formalize "functionally the same" models, while horizontal directions capture genuine functional change. Practical procedures like canonicalization and model merging [1; 20] then

acquire a geometric reading–deterministic gauge-fixes become *sections* of the bundle rather than ad hoc normal forms (cf. 5).

Curvature, holonomy, and parameter–space transport. Curvature has entered machine learning through Hessian structure and flatness surrogates, but here it arises from the principal-bundle structure and the Fisher–Rao geometry on the quotient. We formalize attention as an Ehresmann connection on the principal bundle $\pi : \Theta_0 \to \mathcal{Q}$ using the ambient Euclidean metric, and the Fisher–Rao metric then induces a Riemannian geometry on the quotient. We show that the curvature of this connection is generically nonzero (Theorem 10). A small-loop expansion then relates this curvature to a measurable holonomy in parameter space (Theorem 12), which in turn induces path-dependent transport of representations along a training trajectory. All of these statements are phrased purely in the parameter–space principal bundle, without requiring an explicit representation-bundle structure over the input space. For background, see [11; 12].

A complementary, more local viewpoint would treat attention as a connection on an associated representation bundle over $\mathcal{Q} \times T$, where T indexes token positions and the fibers carry value vectors. Systematically developing this associated-bundle geometry and relating its curvature and holonomy to the Fisher–Rao principal curvature studied here is an interesting direction for future work.

Remark 14 (Architectural variants and gauge structure) *Architectural choices adjust the structure group in predictable ways. Rotary position embeddings (RoPE) [21] impose position-dependent rotations after linear projections, constraining any $A \in \mathrm{GL}(d_k)$ to commute with the rotation blocks; for standard 2×2 planes, this reduces the query–key factor to the commutant $\mathcal{C}_{RoPE} \cong (\mathrm{GL}(1, \mathbb{C}))^{d_k/2}$ (Proposition 15). Multi-query attention [19] couples heads by sharing key–value projections, shrinking the gauge degrees of freedom from $h(d_k^2+d_v^2)$ to $h \cdot d_k^2 + d_v^2$ while maintaining expressivity. Both variants preserve the principal-bundle perspective on the appropriate generic stratum; see 8.*

Position in the literature. Most geometric treatments fix a parameterization and study its local properties, or impose a data-dependent canonical form before analysis. We take the opposite route: start from the maximal head–wise gauge, prove a principal-bundle structure, and use the empirical Fisher metric to define a connection. This yields a horizontal/vertical calculus that explains optimization behavior (Theorem 9), clarifies architectural roles (attention curvature vs. horizontal FFN gradients; Theorems 10 and 13, Proposition 11), and supplies operational diagnostics (Algorithm 1, Algorithm 2)–all within a single, coordinate-free framework.

10. Conclusion

We presented a coordinate–free account of Transformer geometry built on a maximal head–wise gauge and a principal–bundle structure on a generic stratum. Equipped with the empirical Fisher metric and the canonical Euclidean connection, this yields a clean horizontal/vertical calculus and a precise statement that the natural gradient is the horizontal Riesz representative with respect to the quotient Fisher geometry (Theorem 9). On the parameter side, the canonical connection has generically nonzero curvature (Theorem 10), while FFN gradients

are strictly horizontal (Proposition 11). These ingredients lead to practical diagnostics–gauge–aware gradient splitting and a small–loop holonomy estimator (Algorithm 1, Algorithm 2, Theorem 12)–and a Morse–Bott view that explains why apparent "mode" gaps in parameter space collapse on the quotient (Theorem 13).

This framework organizes several geometric ingredients in a form that can be directly applied to analysis and tooling for Transformers.

References

[1] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. arXiv:2209.04836.

[2] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10 (2):251–276, 1998. doi: 10.1162/089976698300017746.

[3] Shun-ichi Amari. *Information Geometry and its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, Tokyo, 2016.

[4] Shun-ichi Amari and Shigeru Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 2000. ISBN 978-0-8218-0531-2.

[5] Alberto Bernacchia, Máté Lengyel, and Guillaume Hennequin. Exact natural gradient in deep linear networks and its application to the nonlinear case. In *Advances in Neural Information Processing Systems*, volume 31, pages 5945–5954, 2018. URL https://papers.nips.cc/paper/2018/hash/7f018eb7b301a66658931cb8a93fd6e8-Abstract.html.

[6] Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999. PMLR, 2016.

[7] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. Transformer Circuits Thread, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

[8] 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. arXiv:2110.06296.

[9] Robert Hecht-Nielsen. On the algebraic structure of feedforward network weight spaces. In *Advanced Neural Computers*, pages 129–135. Elsevier, 1990.

[10] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, et al. Superposition, memorization, and double descent. Transformer Circuits / Anthropic technical report, 2023. arXiv:2301.05101.

[11] Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of Differential Geometry, Vol. I*. Interscience Publishers, New York, 1963. ISBN 978-0471472503.

[12] John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, 2 edition, 2013. ISBN 978-1441999826. doi: 10.1007/978-1-4419-9982-5.

[13] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5763. Association for Computational Linguistics, 2020.

[14] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL http://jmlr.org/papers/v21/20-441.html.

[15] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. arXiv preprint, 2023. arXiv:2301.05217.

[16] Yann Ollivier. Riemannian metrics for neural networks i: Feedforward networks. arXiv preprint, 2015. URL https://arxiv.org/abs/1303.0818. arXiv:1303.0818.

[17] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. Transformer Circuits Thread, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[18] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. arXiv preprint, 2013. URL https://arxiv.org/abs/1301.3584. arXiv:1301.3584.

[19] Noam Shazeer. Fast transformer decoding: One write-head is all you need. arXiv preprint, 2019. arXiv:1911.02150.

[20] Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, pages 22045–22055, 2020.

[21] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. doi: 10.1016/j.neucom.2023.127063.

[22] Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 331–335. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1032.

[23] Hong Wang and Kelly Wang. Complete characterization of gauge symmetries in transformer architectures. In *Symmetry and Geometry in Neural Representations (NeurReps 2025), Proceedings Track*, 2025. Accepted for publication.

Appendix A. Group Action and Maximal Gauge: Summary of Results

This appendix summarizes the gauge symmetry structure used in the main text and records the group action for completeness. All *proofs* of maximality, RoPE commutant structure, and layerwise factorization are deferred to the companion NeurReps paper [23]; here we only restate the statements we need.:contentReference[oaicite:4]index=4

Gauge group and action. We work with the head-wise gauge group

$$G_{\max} \;=\; \big((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h\big) \rtimes S_h.$$

An element $g \in G_{\max}$ is a triple $g = ((A_i)_{i=1}^h, (C_i)_{i=1}^h, \sigma)$ with $A_i \in \mathrm{GL}(d_k)$, $C_i \in \mathrm{GL}(d_v)$ and $\sigma \in S_h$. A single multi-head attention layer is parameterized by

$$\theta \;=\; \big\{(W_Q^{(i)}, W_K^{(i)}, W_V^{(i)})_{i=1}^h,\; W_O\big\},$$

where $W_Q^{(i)}, W_K^{(i)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_V^{(i)} \in \mathbb{R}^{d_{\text{model}} \times d_v}$, and $W_O \in \mathbb{R}^{h d_v \times d_{\text{model}}}$. We write W_O in block rows as $W_O^\top = [W_{O,1}^\top \;\cdots\; W_{O,h}^\top]$, with each $W_{O,i} \in \mathbb{R}^{d_v \times d_{\text{model}}}$.

On the generic stratum Θ_0 (Definition 1), the left action $g \cdot \theta = \theta'$ is defined componentwise by

$$\begin{aligned}
W_Q'^{(i)} &= W_Q^{(\sigma(i))}\, A_i, \\
W_K'^{(i)} &= W_K^{(\sigma(i))}\, A_i^{-\top}, \\
W_V'^{(i)} &= W_V^{(\sigma(i))}\, C_i, \\
W'_{O,i} &= C_i^{-1}\, W_{O,\sigma(i)}.
\end{aligned}$$

We index (A_i, C_i) by the *target* head i, so this action coincides with the standard wreath-product action used in Definition 1 of [23] and yields the usual semidirect-product composition law.

Summary of maximality results (from [23]). The companion paper [23] shows that:

- On Θ_0, the full symmetry group of the MHA block is exactly $G_{\max}$ (Theorem 2 in [23]); no additional continuous or discrete symmetries exist.
- The Lie algebra of infinitesimal symmetries is

$$\mathfrak{g}_{\max} \;=\; \bigoplus_{i=1}^h \mathfrak{gl}(d_k) \;\oplus\; \bigoplus_{i=1}^h \mathfrak{gl}(d_v),$$

  with generators $\delta W_Q^{(i)} = W_Q^{(i)} X_i$, $\delta W_K^{(i)} = -W_K^{(i)} X_i^\top$, $\delta W_V^{(i)} = W_V^{(i)} Y_i$, $\delta W_{O,i} = -Y_i W_{O,i}$.
- Under RoPE with even d_k, the Q/K factor reduces to a block-diagonal commutant subgroup $C_{\text{RoPE}} \cong (\mathrm{GL}(1,\mathbb{C}))^{d_k/2}$ on each 2×2 plane, so the per-head Q/K gauge dimension drops from d_k^2 to d_k while the V/O factor remains d_v^2 (Theorem 33 in [23]).
- For depth-L models without cross-layer sharing, the full gauge group factors as a direct product over layers, $G_{\text{model}} \cong \prod_{\ell=1}^L G_{\max}^{(\ell)}$, with analogous products for grouped/multi–query attention and RoPE (Theorem 8 and Theorem 31 in [23]).

- On the generic stratum, the stabilizer of a parameter under the $G_{\max}$–action is trivial: if $g \cdot \theta = \theta$ and $\theta \in \Theta_0$, then g is the identity (this is implicit in the identifiability and factorization arguments underlying Theorem 2 and Theorem 22 in [23]).

In this paper we do not re-derive these results; we use them as structural input for the connection, curvature, and quotient-geometry analysis developed in Sections 3–6.

Proposition 15 (RoPE commutant and gauge reduction) *On the generic stratum Θ_0, under rotary position embeddings with a nondegenerate frequency schedule and even d_k, the head-wise gauge group of the Q/K block reduces to the plane-wise commutant*

$$C_{\mathrm{RoPE}} \;\cong\; \mathrm{GL}(1,\mathbb{C})^{d_k/2},$$

so that the per-head Q/K gauge dimension drops from d_k^2 to d_k while the value/output factor remains $(\mathrm{GL}(d_v))^h$. In particular,

$$G_{\mathrm{RoPE}} \;=\; \big((C_{\mathrm{RoPE}})^h \times (\mathrm{GL}(d_v))^h\big) \rtimes S_h.$$

Proof reference. This proposition is a restatement of the RoPE commutant structure proved in [23] (see, for example, Propositions 32–33 there). We include it here for completeness but do not rederive the proof.

Free and proper action. On the generic stratum Θ_0 the stabilizer of the $G_{\max}$–action is trivial: if $g \cdot \theta = \theta$ for some $\theta \in \Theta_0$, then g must be the identity. For the continuous $((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h)$ part this follows from the full-rank and controllability assumptions (G1)–(G4) and the identifiability arguments in [23]. Condition (G5) in Definition 1 (head distinctness) rules out nontrivial permutations in S_h that fix θ: if $\sigma \neq \mathrm{id}$ and $\sigma \cdot \theta = \theta$, then at least two heads would share both $W_Q^{(i)}(W_K^{(i)})^\top$ and $W_V^{(i)}W_{O,i}$, contradicting (G5). Thus the $G_{\max}$–action is free on Θ_0.

Properness of the action is standard for products of GL-actions on full-rank matrix manifolds. Concretely, each factor such as

$$(W_Q^{(i)}, A_i) \;\mapsto\; W_Q^{(i)} A_i, \qquad W_Q^{(i)} \in \mathbb{R}^{d_{\mathrm{model}} \times d_k} \text{ full column rank}, \; A_i \in \mathrm{GL}(d_k),$$

is the right action of $\mathrm{GL}(d_k)$ on a Stiefel-type manifold of full-rank matrices. Finite head permutations S_h act properly as well. The combined $G_{\max}$–action is algebraic in the matrix entries and hence smooth and continuous; finite products and semidirect products of smooth, free, proper actions are again smooth, free, and proper. The principal-bundle statement in Theorem 3 then follows from standard results on free, proper Lie-group actions [11; 12].

Appendix B. Empirical Fisher Geometry and Natural Gradient

This appendix spells out the empirical Fisher (Gauss–Newton) geometry behind Section 3 and its associated mechanical connection. Recall that the vertical space $\mathcal{V}_\theta = \ker d\pi_\theta$ is tangent to the $G_{\max}$–orbit through θ, and the canonical connection uses the ambient Euclidean metric to define the horizontal space $\mathcal{H}_\theta = \mathcal{V}_\theta^{\perp_{\mathrm{Euc}}}$, so that $T_\theta\Theta_0 = \mathcal{V}_\theta \oplus \mathcal{H}_\theta$.

The empirical Fisher (or Gauss–Newton) metric at θ is the symmetric, positive–semidefinite form

$$g_\theta(u,v) \;=\; \mathbb{E}_{x\sim\mathcal{D}}\big[\langle J_\theta(x)u,\; J_\theta(x)v\rangle\big], \tag{B.1}$$

where $J_\theta(x)$ is the Jacobian of model outputs with respect to parameters at input x.

Gauge invariance always implies that every vertical direction lies in the null space of g_θ, i.e. $\mathcal{V}_\theta \subseteq \ker g_\theta$. To ensure that there are no *additional* degeneracies coming from the data distribution, we make the following standing assumption.

Assumption 1 (Fisher-regular data) *For every $\theta \in \Theta_0$, the empirical Fisher (Gauss–Newton) metric g_θ has null space exactly equal to the vertical space:*

$$\ker g_\theta \;=\; \mathcal{V}_\theta.$$

Equivalently, if $g_\theta(u,u) = 0$ then u is a gauge direction. This holds, for example, when the data distribution $\mathcal{D}$ is sufficiently rich that every non-gauge functional perturbation is excited with nonzero probability.

Under Assumption 1, g_θ is positive–definite when restricted to the horizontal subspace $\mathcal{H}_\theta$, and induces a genuine Riemannian metric on the quotient $\mathcal{Q} = \Theta_0/G_{\max}$.

Gauge-null gradient directions. We first record the gauge-null property of the loss differential.

Lemma 16 (Gauge-null gradient directions) *If $L : \Theta_0 \to \mathbb{R}$ is gauge-invariant, then $dL_\theta[v] = 0$ for all $v \in \mathcal{V}_\theta$. Equivalently, the Euclidean gradient ∇L satisfies $\nabla L \in \mathcal{H}_\theta$.*

Proof Gauge invariance means $L(g\cdot\theta) = L(\theta)$ for all $g \in G_{\max}$ and $\theta \in \Theta_0$. Differentiating the curve $t \mapsto \exp(tX)\cdot\theta$ at $t = 0$ for any vertical generator $X \in \mathfrak{g}_{\max}$ gives

$$dL_\theta[X_\theta] \;=\; 0,$$

where $X_\theta \in \mathcal{V}_\theta$ is the infinitesimal action. Since $\mathcal{V}_\theta$ is spanned by such generators, we have $dL_\theta[v] = 0$ for all $v \in \mathcal{V}_\theta$. With respect to the Euclidean inner product, $dL_\theta[w] = \langle \nabla L, w\rangle$ for all w, so orthogonality of ∇L to every $v \in \mathcal{V}_\theta$ is equivalent to $\nabla L \in \mathcal{H}_\theta$. ■

Fisher–Rao geometry on the quotient. Gauge invariance always implies that the vertical space $\mathcal{V}_\theta$ is contained in the null space of the empirical Fisher (Gauss–Newton) metric g_θ. Under Assumption 1, this inclusion is an equality, so the restriction of g_θ to the horizontal space $\mathcal{H}_\theta$ is positive–definite and induces a unique Riemannian metric on the quotient $\mathcal{Q} = \Theta_0/G_{\max}$. Concretely, we restrict both the domain and codomain of g_θ to $\mathcal{H}_\theta$ and view

$$G_{\theta|\mathcal{H}_\theta} : \mathcal{H}_\theta \times \mathcal{H}_\theta \to \mathbb{R}, \qquad G_{\theta|\mathcal{H}_\theta}(u,v) := g_\theta(u,v), \tag{B.2}$$

as a positive–definite bilinear form. This defines the Fisher–Rao metric on $\mathcal{Q}$ via the identification of $\mathcal{H}_\theta$ with $T_{[\theta]}\mathcal{Q}$.

The natural gradient $\widetilde{\nabla} L$ is by definition the Riemannian gradient of L with respect to this quotient Fisher–Rao metric.

Theorem 17 (Natural gradient as horizontal Riesz representative) *At $\theta \in \Theta_0$, the natural gradient is the unique vector $\widetilde{\nabla} L \in \mathcal{H}_\theta$ such that*

$$g_\theta(\widetilde{\nabla} L,\, w) = dL_\theta[w] \quad \textit{for all } w \in \mathcal{H}_\theta. \tag{B.3}$$

Equivalently, if we identify g_θ with the linear operator $G_{\theta|\mathcal{H}_\theta} : \mathcal{H}_\theta \to \mathcal{H}_\theta$ via the Euclidean inner product, then

$$\widetilde{\nabla} L \;=\; (G_{\theta|\mathcal{H}_\theta})^{-1}\, \nabla L. \tag{B.4}$$

Proof By Lemma 16, $\nabla L \in \mathcal{H}_\theta$. The Riemannian gradient $\widetilde{\nabla} L$ on the quotient satisfies

$$g_\theta(\widetilde{\nabla} L,\, w) = dL_\theta[w] = \langle \nabla L, w \rangle \quad \text{for all } w \in \mathcal{H}_\theta.$$

Identifying g_θ with the positive–definite operator $G_{\theta|\mathcal{H}_\theta}$ via the Euclidean inner product, this says

$$\langle G_{\theta|\mathcal{H}_\theta}\, \widetilde{\nabla} L,\; w \rangle = \langle \nabla L, w \rangle \quad \forall\, w \in \mathcal{H}_\theta,$$

so $G_{\theta|\mathcal{H}_\theta}\, \widetilde{\nabla} L = \nabla L$. Invertibility of $G_{\theta|\mathcal{H}_\theta}$ on $\mathcal{H}_\theta$ then gives the stated coordinate formula and uniqueness. ■

Appendix C. Curvature and Holonomy of the Connection

This appendix details the curvature of the canonical (Euclidean) connection on the parameter bundle $\pi : \Theta_0 \to \mathcal{Q}$, defined by the orthogonal split $T_\theta \Theta_0 = \mathcal{V}_\theta \oplus \mathcal{H}_\theta$.

Definition of curvature. Let X and Y be two horizontal vector fields on Θ_0 (i.e. $X(\theta), Y(\theta) \in \mathcal{H}_\theta$ for all θ). The curvature two-form Ω is a $\mathfrak{g}_{\max}$-valued 2-form defined by the vertical component of the Lie bracket of horizontal vector fields:

$$\Omega(X, Y) \;=\; -[X, Y]^{\mathrm{vert}}. \tag{C.1}$$

By the Frobenius theorem, the horizontal distribution $\mathcal{H}$ is integrable (locally forms a foliation tangent to the base) if and only if the curvature Ω vanishes identically.

Remark 18 (Clarification on curvature calculation) *It is crucial to note that the curvature is defined by the noncommutativity of the horizontal vector fields themselves ($[X, Y]$),* not *by the noncommutativity of mixed partial derivatives of the network output. For smooth functions (like MHA with GeLU/Softmax), mixed partial derivatives always commute by Schwarz's theorem, so analyzing the Hessian of the output cannot reveal the curvature of the connection.*

Theorem 19 (Connection curvature is generically nonzero, restated) *On a Zariski–open (hence full–measure) subset of Θ_0, the curvature two–form Ω of the canonical connection is nonzero.*

Proof [Proof (sketch)] The canonical connection is the Euclidean mechanical connection associated with the vertical distribution $\mathcal{V}_\theta$ and the ambient inner product. Its curvature

vanishes if and only if the horizontal distribution $\mathcal{H}$ is integrable, i.e. tangent to a foliation whose leaves project diffeomorphically to open sets in $\mathcal{Q}$.

The horizontal space $\mathcal{H}_\theta$ is defined by Euclidean orthogonality to $\mathcal{V}_\theta$. Both the vertical generators (derived from the action of $G_{\max}$) and the Euclidean structure depend smoothly and algebraically on θ. The condition $\Omega \equiv 0$ imposes a system of polynomial constraints on these generators and their derivatives, expressing closure of $\mathcal{H}$ under Lie brackets. In the MHA architecture there are no structural symmetries forcing these constraints to hold identically across Θ_0. Consequently, the flatness conditions define a proper algebraic subvariety of Θ_0. Thus Ω is nonzero on a Zariski–open (hence full–measure) subset of Θ_0, so nonvanishing is generic. ■

Holonomy for small horizontal loops. We relate the curvature to the measurable holonomy around small horizontal loops, justifying Algorithm 2.

Let $u, v \in \mathcal{H}_\theta$ be Euclidean-orthonormal and let X_u, X_v be the corresponding horizontal vector fields. Let $\Phi_{t,W}$ be the flow along W. Consider the small horizontal loop

$$\square_\varepsilon(u,v) \;=\; \Phi_{\varepsilon,X_v} \circ \Phi_{\varepsilon,X_u} \circ \Phi_{-\varepsilon,X_v} \circ \Phi_{-\varepsilon,X_u}.$$

Using the Baker–Campbell–Hausdorff (BCH) expansion and the definition of curvature $\Omega(X_u, X_v) = -[X_u, X_v]^{\text{vert}}$, the holonomy (the net displacement in the fiber) is given, to leading order, by:

Proposition 20 (BCH expansion and holonomy) *The holonomy $\Delta_{\square_\varepsilon}(u,v) \in G_{\max}$ around the small horizontal loop is given, in Lie–algebra coordinates (via the exponential map), by*

$$\log(\Delta_{\square_\varepsilon}(u,v)) \;=\; -\varepsilon^2\, \Omega_\theta(X_u, X_v) \;+\; O(\varepsilon^3).$$

Consequently, $\|\Delta_{\square_\varepsilon}(u,v)\| = \varepsilon^2 \|\Omega_\theta(u,v)\| + O(\varepsilon^3)$ for any suitable norm on $G_{\max}$.

Proof This is a standard application of the Ambrose–Singer theorem relating holonomy to curvature, localized via the BCH expansion. ■

Holonomy scaling and Richardson extrapolation. Proposition 20 shows that, for a small rectangular loop of side length ε traced by horizontal directions $u, v \in \mathcal{H}_\theta$, the resulting holonomy has magnitude $\|\Delta_{\square_\varepsilon}(u,v)\| = \varepsilon^2 \|\Omega_\theta(u,v)\| + O(\varepsilon^3)$. In practice, we estimate $\|\Omega_\theta(u,v)\|$ by running the loop at two scales ε and $\varepsilon/2$, computing the corresponding holonomies, and forming a Richardson-extrapolated estimate $h^* = \frac{4h(\varepsilon/2) - h(\varepsilon)}{3}$, where $h(\delta) := \|\Delta_{\square_\delta}(u,v)\|/\delta^2$. This yields an $O(\varepsilon^3)$-bias estimator for the curvature norm along the pair (u,v) and directly motivates the holonomy estimator in Algorithm 2.

Appendix D. Morse–Bott Structure and the Quotient Loss

This section gives the proof of Theorem 13 from the main text.

Proof [Proof of Theorem 13] Gauge–invariance gives $dL_\theta[v] = 0$ for every $v \in \mathcal{V}_\theta$ (Lemma 16), so the entire orbit $G_{\max} \cdot \theta$ lies in the critical set and the Hessian $\nabla^2 L(\theta)$ annihilates $\mathcal{V}_\theta$. Since the $G_{\max}$–action on Θ_0 is free and proper (Theorem 3), the slice theorem yields a submanifold S through θ with $T_\theta S = \mathcal{H}_\theta$ and a $G_{\max}$–equivariant diffeomorphism from a

neighborhood of θ onto a neighborhood of the orbit modeled on $G_{\max} \times S$. Restricting L to S freezes vertical directions, so the Hessian of $L|_S$ at θ equals the horizontal block of $\nabla^2 L(\theta)$. The quotient map identifies S with a chart of $\mathcal{Q}$ around $[\theta]$; because $\ell \circ \pi = L$, the Hessian $\nabla^2 \ell([\theta])$ matches the horizontal restriction of $\nabla^2 L(\theta)$. Nondegeneracy of the horizontal block is therefore equivalent to ℓ being Morse at $[\theta]$. ■

Remarks. (i) The argument is local to the regular stratum Θ_0; outside it, stabilizers may grow and the space stratifies. (ii) The identification of $\nabla^2 \ell([\theta])$ with the horizontal block is independent of the slice, since all slices share $T_\theta S = \mathcal{H}_\theta$.

Appendix E. Algorithms and Reproducibility

This appendix gives the concrete procedures behind Section 5 and the minimal choices needed to reproduce our figures. We keep the presentation compact; all routines are drop–in and numerically stable at the scales reported in Section 7.

Vertical generators (respecting RoPE and sharing). We obtain a numerically independent spanning set of the vertical space $\mathcal{V}_\theta$ by differentiating the $G_{\max}$–action. For $X \in \mathfrak{gl}(d_k)$,

$$\delta^{(i)}_{QK}(X) : \ (W^{(i)}_Q, W^{(i)}_K) \mapsto \big(W^{(i)}_Q X, \, -W^{(i)}_K X^\top\big),$$

and for $Y \in \mathfrak{gl}(d_v)$,

$$\delta^{(i)}_{VO}(Y) : \ (W^{(i)}_V, W_O) \mapsto \big(W^{(i)}_V Y, \ \text{insert} \ -Y \ \text{in the } i\text{th } d_v\text{–block of } W_O\big).$$

Head permutations generate a discrete symmetry and do not contribute to $\mathcal{V}_\theta$. With RoPE (even d_k), we restrict X plane–wise to the commutant $\{aI + bJ\}$ on each $2{\times}2$ rotation block; with GQA/MQA we tie the corresponding X or Y across shared heads. In practice we assemble candidates, vectorize, and perform thin–QR with column pivoting to remove near–collinear directions (tolerance 10^{-10}), yielding a well–conditioned basis $\{v_j\}_{j=1}^m$.

FR and Euclidean projectors (implementation). Given $\{v_j\}$ and a vector u (e.g., $u = \nabla L$), the FR–orthogonal decomposition solves the Gram system

$$G_{ij} = g_\theta(v_i, v_j), \qquad b_i = g_\theta(v_i, u), \qquad (G + \lambda I)c = b,$$

with optional Tikhonov $\lambda \in [10^{-10}, 10^{-6}]$ for stability. We then set

$$u_{\text{vert}} = \sum_j c_j v_j, \qquad u_{\text{hor}} = u - u_{\text{vert}}.$$

The Euclidean proxy replaces g_θ by the dot product: stack $A = [\,\text{vec}(v_1) \ \cdots \ \text{vec}(v_m)\,]$ and solve $(A^\top A + \lambda I)c = A^\top u$. We report the *vertical fraction* $\|u_{\text{vert}}\|/\|u\|$ and the residual $\|Ac - u\|/\|u\|$ (Euclidean) or $\|(G + \lambda I)c - b\|/\|b\|$ (FR) as fit diagnostics.

Algorithm 3 Vertical/Horizontal Decomposition (FR and Euclidean)

Require: basis $\{v_j\}_{j=1}^m$ for $\mathcal{V}_\theta$; vector u; metric handle $\texttt{inner}(\cdot,\cdot)$

1: $G_{ij} \leftarrow \texttt{inner}(v_i, v_j), \quad b_i \leftarrow \texttt{inner}(v_i, u)$
2: Solve $(G + \lambda I)c = b$ (Cholesky if well–conditioned, else CG with stopping on relative residual 10^{-10})
3: $u_{\text{vert}} \leftarrow \sum_j c_j v_j, \quad u_{\text{hor}} \leftarrow u - u_{\text{vert}}$
4: **return** $(u_{\text{vert}}, u_{\text{hor}})$, vertical fraction $\|u_{\text{vert}}\|/\|u\|$, residual

Holonomy estimator (procedure and bias control). For g_θ–orthonormal $u, v \in \mathcal{H}_\theta$, we consider the horizontal four–segment loop in Θ_0 with vertices $\{\pm\varepsilon u, \pm\varepsilon v\}$, traversed in the order $+\varepsilon u \to +\varepsilon v \to -\varepsilon u \to -\varepsilon v$, projecting back to $\mathcal{H}$ at the end of each segment and measuring the resulting holonomy.

Algorithm 4 Holonomy Estimator with Richardson Extrapolation

Require: θ; orthonormal $u, v \in \mathcal{H}_\theta$; steps $\varepsilon > \varepsilon' = \varepsilon/2$

1: **for** $\delta \in \{\varepsilon, \varepsilon'\}$ **do**
2: Flow horizontally by $+\delta u, +\delta v, -\delta u, -\delta v$, projecting to $\mathcal{H}$ at each leg
3: Compute $\Delta_{\square_\delta}(u, v)$ via least–squares alignment to Lie generators; set $H(\delta) \leftarrow \|\Delta_{\square_\delta}(u, v)\|/\delta^2$
4: **end for**
5: **return** $H^\star = \dfrac{4H(\varepsilon/2) - H(\varepsilon)}{3}$ (bias $O(\varepsilon^3)$); report $|H^\star - H(\varepsilon/2)|$ as an error proxy

By Theorem 12, the Richardson combination $H^\star$ has bias $O(\varepsilon^3)$ as $\varepsilon \to 0$.

Complexity at a glance. Let $m = h(d_k^2 + d_v^2)$; RoPE reduces $d_k^2 \to d_k$ per head. Forming $A^\top A$ costs $O(m^2 D)$ for parameter dimension D, and solving costs $O(m^3)$ (dense) or $O(mD)$ per CG iteration. At h=12, d_k=d_v=64 one has $m \approx$98,304, so FR projectors and holonomy become expensive; this is why Section 7 uses Euclidean proxies while Section 5 exposes FR–exact routines.

Deterministic gauge–fix (for reproducibility, not theory). To make measurements repeatable, we select a canonical representative on each orbit by: (i) thin–QR with positive diagonals for V/O and the induced block update on W_O; (ii) plane–wise Gram balancing for Q/K within the RoPE commutant; (iii) deterministic head ordering by a fixed tie–break rule (e.g., lexicographic on row–major W_O blocks). This is a *section choice*–it does not constrain the theory–and affects only how we display or cache parameters.

Minimal reproducibility notes. We fix random seeds (42) and reuse a single evaluation batch for all Fisher and holonomy measurements so that the reported quantities are directly comparable across configurations.

Appendix F. Experimental Configuration and Procedures

This note collects the minimal details needed to reproduce the figures in Section 7. We keep the setup intentionally simple and fixed: one hardware/software stack, one evaluation batch, and a deterministic gauge–fix used only for display consistency.

Environment and common settings. Unless stated otherwise, computations use `float64` on a single H100 (95 GB), CUDA 12.1, PyTorch 2.4.1, with TF32 disabled. The global seed is 42. A *single* evaluation batch is used throughout to define Fisher–Rao quantities and to keep all comparisons on identical data. The architecture matches Section 2 (e.g., h=12, d_k=d_v=64 in the invariance experiment). Thin–QR with column pivoting is used with a pivot tolerance of 10^{-10}; linear solves stop at relative residual 10^{-10}.

Deterministic gauge–fix (for repeatability, not theory). To make outputs bytewise identical across runs, we pick a canonical representative per gauge orbit: (i) thin–QR with positive diagonals for V/O and the induced block update in W_O; (ii) plane–wise Gram balancing for Q/K within the RoPE commutant; (iii) a fixed head order by a simple tie–break rule. This is a *section choice* only; it does not constrain the theory or diagnostics (see Section E).

Gauge invariance (Table 2). We sample $(A_i, C_i) \in \mathrm{GL}(d_k) \times \mathrm{GL}(d_v)$ per head with prescribed condition numbers (via thin–QR on random draws), optionally permute heads, apply

$$(W_Q^{(i)}, W_K^{(i)}) \mapsto (W_Q^{(i)} A_i,\, W_K^{(i)} A_i^{-\top}), \qquad (W_V^{(i)}, W_O) \mapsto (W_V^{(i)} C_i,\, C_i^{-1} W_O),$$

evaluate on the fixed batch, and report $\|Y' - Y\|/\|Y\|$. Controls outside $G_{\max}$ (e.g., cross–head mixing) yield $\Theta(1)$ changes.

Gauge–aware split (Figure 3). We compute per–sample gradients for a mean–squared reconstruction objective and apply the *Euclidean proxy* of Algorithm 1: stack numerically independent vertical generators into a matrix A (thin–QR with pivoting), solve $(A^\top A + \lambda I)c = A^\top \nabla L$ (with optional $\lambda \in [10^{-10}, 10^{-6}]$), and report the normalized Euclidean vertical fraction $\|Ac\|/\|\nabla L\|$ along with residuals $\|Ac - \nabla L\|/\|\nabla L\|$ as a fit diagnostic.

Fisher–Rao counterparts (for future scale). The FR–exact projector and the small–loop holonomy estimator are given in Algorithm 1 and Algorithm 4, with derivations in §B and §C. They require Gram systems on the vertical basis and horizontal projections along the loop, which are costly at large $m = h(d_k^2 + d_v^2)$ (RoPE reduces $d_k^2 \to d_k$ per head). We therefore use the Euclidean proxy for Section 7 and provide the FR procedures as ready-made routines for future large-scale evaluation.