# GAUGE-INVARIANT REPRESENTATION HOLONOMY

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

## ABSTRACT

Deep networks learn internal representations whose geometry—how features bend, rotate, and evolve—affects both generalization and robustness. Existing similarity measures such as CKA or SVCCA capture pointwise overlap between activation sets, but miss how representations change along input paths. Two models may appear nearly identical under these metrics yet respond very differently to perturbations or adversarial stress. We introduce representation holonomy, a gauge-invariant statistic that measures this path dependence. Conceptually, holonomy quantifies the "twist" accumulated when features are parallel-transported around a small loop in input space: flat representations yield zero holonomy, while nonzero values reveal hidden curvature. Our estimator fixes gauge through global whitening, aligns neighborhoods using shared subspaces and rotation-only Procrustes, and embeds the result back to the full feature space. We prove invariance to orthogonal (and affine, post-whitening) transformations, establish a linear null for affine layers, and show that holonomy vanishes at small radii. Empirically, holonomy increases with loop radius, separates models that appear similar under CKA, and correlates with adversarial and corruption robustness. It also tracks training dynamics as features form and stabilize. Together, these results position representation holonomy as a practical and scalable diagnostic for probing the geometric structure of learned representations beyond pointwise similarity.

## 1 INTRODUCTION

Modern deep networks learn internal representations whose geometry—how features orient, align, and evolve—matters for generalization and robustness. Yet most standard diagnostics are *pointwise*: they compare two activation sets on a fixed dataset using singular vector canonical correlation analysis (SVCCA), projection-weighted CCA (PWCCA), centered kernel alignment (CKA), or representational similarity analysis (RSA) thereby judging subspace overlap while remaining blind to how features *move* as inputs are varied along natural directions (pose, illumination, texture) (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019; Kriegeskorte et al., 2008). This leaves a practical gap: two models can appear highly similar under CKA or CCA, and still behave differently under adversarial or corruption stress because their intermediate features rotate differently along input paths.

We address this gap by turning alignment itself into an object of study. We view a layer's representation as a field over input (or transformation) space and endow it with a *discrete connection*: between nearby inputs we estimate a shared principal subspace and compute the optimal special-orthogonal alignment (rotation-only Procrustes) of the two local feature clouds; composing these small rotations around a closed loop yields a single orthogonal matrix whose deviation from identity we call *representation holonomy*. Nonzero holonomy indicates path-dependent (nonintegrable) transport in the classical sense of connections and their curvature (Ambrose and Singer, 1953). The construction is *gauge-invariant* by design: global whitening fixes a sensible gauge by removing second-order anisotropy; orthogonal reparameterizations of layers leave the statistic unchanged; restricting to a low-rank shared subspace improves stability and cost (Schönemann, 1966; Kabsch, 1976; Kessy et al., 2018; Björck and Golub, 1973; Davis and Kahan, 1970).

Our proposal complements two nearby lines of work rather than competing with them. First, local equivariance tests (e.g., Lie-derivative "local equivariance error") quantify *infinitesimal* sensitivity but do not assess global path-dependence via loop composition (Lenc and Vedaldi, 2015; Gruver et al., 2022). Second, gauge-/manifold-equivariant architectures build a connection into the model so that desired transports are integrable by design; we instead *measure* the emergent transport of standard models, providing a diagnostic that travels with existing practice in vision architectures (Bronstein et al., 2021; Cohen et al., 2019; Schonsheck et al., 2018; Masci et al., 2015). In downstream terms, holonomy gives a compact, layer-wise summary of pathwise geometry that (i) is inexpensive to compute, (ii) scales to common backbones, and (iii) adds information orthogonal to pointwise similarity, making it a natural candidate to relate feature geometry to robustness (Hendrycks and Dietterich, 2019).

**Contributions.** (1) We propose a practical estimator of *representation holonomy* that combines global whitening, shared-neighbor subspaces, and rotation-only Procrustes alignment. The estimator is explicitly gauge-invariant and stable in the small-radius limit. (2) We prove formal invariances (orthogonal and, after whitening, affine), establish a *linear null* showing affine layers yield zero holonomy, and derive a *small-radius limit* where holonomy vanishes linearly with loop radius. A perturbation analysis (Procrustes + Davis–Kahan/Wedin) provides explicit finite-sample and truncation error bounds (Schönemann, 1966; Björck and Golub, 1973; Davis and Kahan, 1970; Kessy et al., 2018). (3) On MNIST/MLP and CIFAR-10/100 with ResNet-18, we show that holonomy (i) increases with loop radius and depth even when CKA remains high, revealing pathwise geometry beyond pointwise similarity; (ii) rises during training as features form and stabilizes at convergence; and (iii) correlates with adversarial and corruption robustness across training regimes including ERM, label smoothing, mixup, and adversarial training (Kornblith et al., 2019; Hendrycks and Dietterich, 2019).

Section 2 situates our work among similarity metrics, equivariance diagnostics, and gauge-/manifold-equivariant architectures. Section 3 formalizes the discrete connection and holonomy estimator and establishes invariance and small-radius results with finite-sample error bounds. Section 4 reports controlled loops, training dynamics, robustness studies, and ablations (whitening choice, SO vs. O, neighbor sharing, and $k/q$ sensitivity). Section 6 summarizes limitations and implications, and we release code and seeded configs for full reproducibility.

## 2 Related Work

Comparing learned representations across networks and inputs is complicated by the fact that layer activations admit many equivalent parameterizations, i.e., a gauge freedom that allows local changes of basis without altering function. A large body of work therefore develops basis-invariant or basis-robust comparison tools. CCA-based approaches—SVCCA and PWCCA—compare the subspaces spanned by activations across models or training checkpoints, reducing sensitivity to neuron permutations while still depending on preprocessing choices and data coverage (Raghu et al., 2017; Morcos et al., 2018). Kernel-based Centered Kernel Alignment (linear and nonlinear CKA) has emerged as a simple and reliable alternative with improved stability across architectures, layers, and seeds, and clear links to representational similarity analysis (RSA) from systems neuroscience (Kornblith et al., 2019; Kriegeskorte et al., 2008). Beyond scalar similarities, a complementary line aligns entire representations by explicit linear transports: orthogonal Procrustes (and its $\det = +1$ Kabsch variant) yields optimal rotation-only maps between paired activation matrices, while principal angles quantify shared subspaces; classical perturbation theory (Davis–Kahan/Wedin) provides finite-sample error control for the estimated subspaces and transports (Schönemann, 1966; Kabsch, 1976; Björck and Golub, 1973; Davis and Kahan, 1970). Preprocessing is itself a gauge choice: statistically principled whitening schemes such as ZCA-corr justify a fixed global gauge that removes second-order anisotropy before any local alignment (Kessy et al., 2018). Parallel to these comparison methods, empirical tests of (approximate) equivariance and model equivalence probe how features change under controlled input transformations; early work proposed finite-difference tests for CNNs, and more recent formulations use Lie derivatives to define a *local equivariance error* that is differential and layer-wise (Lenc and Vedaldi, 2015; Gruver et al., 2022). Geometric deep learning places these observations in a coordinate-free framework: data domains (e.g., manifolds) come with local frames (gauges), and operations should be gauge-aware; gauge-equivariant CNNs make the connection (in the differential-geometric sense) an architectural primitive, while manifold convolutions based on parallel transport or geodesic patches move features intrinsically across space (Bronstein et al., 2021; Cohen et al., 2019; Schonsheck et al., 2018; Masci et al., 2015). Relatedly, Riemannian approaches on structured spaces (e.g., SPD/Grassmann) deploy intrinsic means, transports, and normalizations inside networks, underscoring the usefulness of connection-like operations on representation manifolds

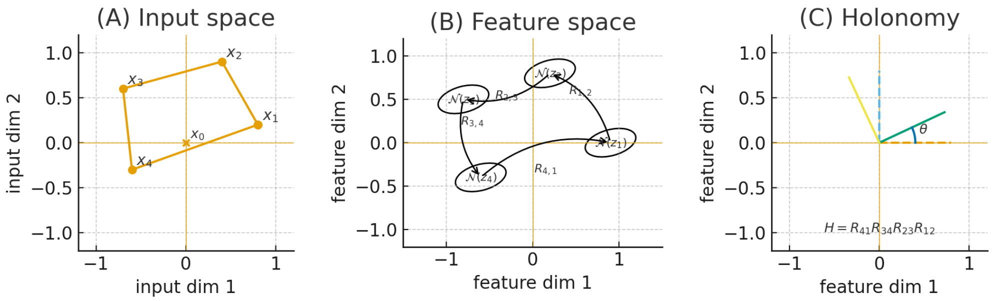


Figure 1: **Holonomy as path-dependent feature rotation.** (A) A small closed loop $\gamma = (x_0, \ldots, x_{L-1}, x_L{=}x_0)$ in a 2D input slice. (B) The corresponding features $z_i = z(x_i)$ and their local neighbourhoods $\mathcal{N}(z_i)$; for each edge we estimate an orthogonal transport $R_{i,i+1}$ that best aligns the two nearby feature clouds. (C) Composing these transports around the loop yields the holonomy $H = R_{L-1} \cdots R_1 R_0$, visualised as the net rotation of a reference direction by angle $\theta$. Holonomy is invariant to layer-wise gauge changes (global change of feature basis) and measures how much the representation "twists" when inputs follow a loop, rather than just how similar activations are at individual points.

(Huang and Van Gool, 2017; Brooks et al., 2019). Finally, robustness benchmarks such as ImageNet-C/P offer downstream behavioral checks; because they include *sequences* of small perturbations, they are natural testbeds for path-sensitive phenomena in representation geometry (Hendrycks and Dietterich, 2019).

*This paper* adopts the geometric viewpoint but applies it *as a diagnostic* to standard models rather than as an architectural constraint. We model layer-wise representations over data (or transformation) space as sections of a vector bundle and make the alignment rule itself a *connection*: locally, we estimate a shared subspace (with principled whitening as a fixed gauge) and define the transport between nearby inputs by the optimal special-orthogonal map in that subspace; globally, we compose these local transports around closed loops and quantify the resulting *holonomy*. By construction, our measurement is invariant to per-layer orthogonal reparameterizations (gauge transforms) and robust to admissible whitening choices, distinguishing it from scalar, path-agnostic similarities such as CKA/RSA and from single-step Procrustes alignments (Kornblith et al., 2019; Kriegeskorte et al., 2008; Schönemann, 1966). The Ambrose–Singer perspective links small-loop holonomy to curvature, yielding concrete predictions that we test empirically; in particular, we show that networks can be locally near-equivariant (small Lie-derivative error) yet exhibit nontrivial *global* holonomy that correlates with stability under perturbation sequences, a phenomenon invisible to standard similarity scores (Ambrose and Singer, 1953; Gruver et al., 2022; Hendrycks and Dietterich, 2019).

## 3 Representation Holonomy

Intuitively, a layer's representation assigns to each input $x$ a feature vector $z(x) \in \mathbb{R}^p$. If we move $x$ along a small closed loop $\gamma$ in input space (for example by composing small transformations), the corresponding features $z(x)$ trace out a loop in representation space. Locally, between two nearby points on the loop we can align their feature neighbourhoods by an orthogonal map $R_{i,i+1} \in \mathrm{SO}(p)$ that best matches the two clouds (Figure 1, panels A–B). Composing these local transports around the entire loop yields a net rotation $H = R_{L-1} \cdots R_1 R_0$ (panel C). If the representation were perfectly "flat" along $\gamma$—for instance, if it were globally linear and we controlled for gauge—this product would be the identity. Deviations of $H$ from $I$ therefore quantify the path dependence (curvature) of the learned features, and are insensitive to global changes of feature basis.

At a high level, the proofs rely on three standard tools: (i) Procrustes alignment in shared low-dimensional subspaces, (ii) matrix perturbation bounds of Davis–Kahan/Wedin type for controlling subspace errors, and (iii) finite-sample concentration bounds for covariance and whitening operators. We collect the technical details in Appendix S.3–S.5. At a given layer, we call a transform $z(x) \mapsto Qz(x) + b$, with $Q \in O(p)$ (after whitening) and $b \in \mathbb{R}^p$, a gauge transformation: an input-independent change of basis in feature space. Two networks related by such transforms at internal layers are representation-equivalent. Our estimator is invariant to these transformations (and, after whitening, to general invertible affine reparameterisations), so that holonomy reflects only relative orientation changes induced by input paths rather than arbitrary basis choices.

Let $f : \mathbb{R}^d \to \mathbb{R}^C$ be a classifier and let $z : \mathbb{R}^d \to \mathbb{R}^p$ denote a fixed layer's representation (layer index suppressed). For inputs $x \in \mathcal{X} \subset \mathbb{R}^d$ we write $z(x) \in \mathbb{R}^p$. Given a small loop $\gamma = (x_0, \ldots, x_{L-1}, x_L{=}x_0)$ in input (or transformation) space, we define a local linear transport $R_i \in \mathrm{SO}(p)$ between the features at successive points and take the *holonomy*

$$H(\gamma) \;=\; R_{L-1} \cdots R_1\, R_0 \;\in\; \mathrm{SO}(p), \qquad h_{\mathrm{norm}}(\gamma) \;=\; \frac{\|H(\gamma) - I\|_F}{2\sqrt{p}} \in [0,1], \tag{1}$$

reporting also the eigen–angle multiset $\{\theta_j\}_{j=1}^{p}$ of $H(\gamma)$ (eigenvalues $e^{i\theta_j}$ on the unit circle). Conceptually, if $z$ is $C^2$ then first-order linearization suggests $R_i = I + \mathrm{O}c(\|x_{i+1} - x_i\|)$, hence $H(\gamma) = I + \mathrm{O}c(\mathrm{length}(\gamma))$.

**Estimator (used in practice).** We pool a set $\mathcal{N}$ of examples, compute features $Z = \{z(x)\}_{x \in \mathcal{N}}$, their mean $\mu$ and covariance $\Sigma$, and fix a global gauge by *whitening* $\tilde{z}(x) = \Sigma^{-1/2}(z(x) - \mu)$ (ZCA-corr; any fixed symmetric square root suffices) (Kessy et al., 2018). For an edge $(x_i, x_{i+1})$, let $m_i = \frac{1}{2}(\tilde{z}(x_i) + \tilde{z}(x_{i+1}))$ and choose a *shared* index set $\mathcal{I}_i$ of size $k$ as the $k$-NN of $m_i$ in the whitened pool. On these same rows we compute a *shared* soft center at the midpoint $\bar{\mu}_i = \sum_{j \in I_i} w_j^{(i)} \tilde{Z}_{j:}$ with weights $w_j^{(i)} \propto \exp(-\|\tilde{Z}_{j:} - m_i\|/\sigma_i)$, and set $X_i = Y_i = \tilde{Z}_{I_i} - \bar{\mu}_i$. Let $W_i = \mathrm{diag}\big(w_j^{(i)}\big)_{j \in I_i}$ and $B_i \in \mathbb{R}^{p \times q}$ be the top-$q$ right singular vectors of $\begin{bmatrix} X_i \\ Y_i \end{bmatrix}$. In $\mathbb{R}^q$ we solve orthogonal Procrustes: if $U_i \Sigma_i V_i^\top = \mathrm{SVD}\big((X_i B_i)^\top\, W_i\, (Y_i B_i)\big)$ then $R_i^{(q)} = U_i V_i^\top \in \mathrm{SO}(q)$ (enforce $\det = +1$) (Schönemann, 1966; Kabsch, 1976). We *embed* back to $\mathbb{R}^p$ by

$$\widehat{R}_i \;=\; B_i R_i^{(q)} B_i^\top + (I - B_i B_i^\top) \;\in\; \mathrm{SO}(p), \tag{2}$$

compose $\widehat{H}(\gamma) = \widehat{R}_{L-1} \cdots \widehat{R}_0$, and report $\widehat{h}_{\mathrm{norm}} = \|\widehat{H} - I\|_F / (2\sqrt{p})$ together with eigen–angles of $\widehat{H}$. Indeed, $(I - BB^\top)B = 0$ so $\hat{R}_i^\top \hat{R}_i = I$; moreover $\det \hat{R}_i = \det R_i^{(q)} = +1$. This construction is inexpensive (small SVDs in a shared subspace) and numerically stable.

**Structural properties (statements; full proofs in App. S.1).** [1] (i) *Gauge invariance.* If whitened features are reparameterized by any $U \in \mathrm{O}(p)$, i.e., $\tilde{z}'(x) = U\tilde{z}(x)$, then the shared indices $\mathcal{I}_i$ are unchanged, $\widehat{R}_i' = U\widehat{R}_i U^\top$, and $\widehat{H}' = U\widehat{H}U^\top$. Thus $\|\widehat{H}' - I\|_F = \|\widehat{H} - I\|_F$ and the eigen–angle multiset is identical. (ii) *Affine invariance (post-whitening).* For any invertible affine map on raw features, $z'(x) = Az(x) + b$, whitening by the corresponding pool statistics yields $\tilde{z}'(x) = Q\,\tilde{z}(x)$ with $Q \in \mathrm{O}(p)$ (because $\Sigma'^{-1/2} A \Sigma^{1/2}$ is orthogonal when $\Sigma' = A\Sigma A^\top$), hence the previous item applies. (iii) *Linear null.* If $z(x) = Bx + c$ is affine and each edge uses shared rows, then $X_i = Y_i$ for all $i$ and $\widehat{R}_i = I$, so $\widehat{H}(\gamma) = I$. (iv) *Orientation/cycling.* Reversing a loop inverts holonomy, $\widehat{H}(\gamma^{-1}) = \widehat{H}(\gamma)^{-1}$, so the Frobenius gap is unchanged; cyclic reparameterizations of $\gamma$ leave $\widehat{H}$ unchanged. (v) *Normalization.* For any $H \in \mathrm{O}(p)$ with eigen–angles $\{\theta_j\}$, $\|H - I\|_F^2 = 2\sum_{j=1}^{p}(1 - \cos\theta_j) \le 4p$, hence $h_{\mathrm{norm}} \in [0,1]$ with equality 1 iff all $\theta_j = \pi$. All invariance statements apply to the post-readout features; non-invertible readouts (e.g., JL) are outside the affine-invariance claim.

**Small-radius behavior (statement; proof in App. S.2).** Assume $z$ is $C^2$ with Lipschitz Jacobian on a neighborhood of $\gamma_r$, the loop $\gamma_r$ has total length $\mathrm{O}c(r)$, the shared-midpoint $k$-NN has overlap probability $1 - \mathrm{O}c(r)$ as $r \to 0$, and the subspace rank $q$ covers the local feature rank. Then for each edge $\|\widehat{R}_i - I\|_F = \mathrm{O}c(r)$ and

$$\|\widehat{H}(\gamma_r) - I\|_F \;=\; \mathrm{O}c(r), \qquad \text{hence} \quad \widehat{h}_{\mathrm{norm}}(\gamma_r) = \mathrm{O}c(r). \tag{3}$$

[1] **Pointers to supplement**: App. S.0 fixes notation; App. S.1 gives full proofs of invariances, nulls, and normalization; App. S.2 proves the small-radius limit; App. S.3 states a Procrustes perturbation lemma; App. S.4 handles subspace truncation; App. S.5 derives per-edge and holonomy error bounds; App. S.6 provides an explicit algorithm and App. S.7–S.8 cover complexity and practical implications.

Intuitively, shared-row centering cancels translations; Lipschitz variation of $J_z$ makes the optimal rotation deviate from $I$ by $\mathrm{Oc}(\|x_{i+1} - x_i\|)$; products of $I{+}\mathrm{Oc}(r)$ along $L{=}\mathrm{Oc}(1)$ edges yield an overall $\mathrm{Oc}(r)$ deviation.

**Estimator stability and error decomposition (statement; full derivation in App. S.5).** Under standard sampling assumptions for the neighbor pool (sub-Gaussian rows; a spectral gap $\Delta$ separating the top-$q$ right-singular subspace), the per-edge error relative to the population transport $R_i^\star$ obeys

$$\|\widehat{R}_i{-}R_i^\star\|_F \;\le\; C_1 \underbrace{k^{-1/2}}_{\text{finite sample}} + C_2 \underbrace{\frac{\|\Pi_\perp^i \Sigma_i^{1/2}\|_F}{\lambda_q(\Sigma_i)^{1/2}}}_{\text{subspace truncation}} + C_3 \underbrace{\mathrm{TV}(\mathcal{I}_i, \mathcal{I}_i^\star)}_{\text{index mismatch}} + C_4 \underbrace{\|J_z(x_{i+1}) - J_z(x_i)\|_2}_{\text{curvature}}, \tag{4}$$

with $\Pi_\perp^i = I - B_i B_i^\top$ and constants depending smoothly on local condition numbers; composing over $L{=}\mathrm{Oc}(1)$ edges yields the holonomy error bound. Here $\lambda_q(\Sigma_i)$ denotes the $q$-th largest eigenvalue of the *population* covariance $\Sigma_i$ on the shared rows. The finite-sample term follows from Procrustes perturbation via singular-subspace angles, the truncation term from Davis–Kahan/Wedin, and the curvature/mismatch terms from continuity of $J_z$ and the shared-midpoint $k$-NN (Björck and Golub, 1973; Davis and Kahan, 1970).

Empirically, this decomposition matches the behaviour we observe in the vision experiments. Choosing $k$ moderately large and $q$ smaller (e.g., $k \in \{96, 128, 192\}$, $q \in \{32, 64, 96\}$ on MNIST hidden 1) keeps the finite-sample and truncation terms small: $h_{\mathrm{norm}}$ varies only at the level of a few $10^{-7}$ across this grid, without qualitative changes. The shared-midpoint $k$-NN construction effectively controls the index-mismatch term $\mathrm{TV}(\mathcal{I}_i, \mathcal{I}_i^\star)$: when we deliberately use disjoint neighbour sets for the two endpoints, holonomy increases markedly and becomes unstable at small radii. Finally, in linear or self-loop settings (affine networks, $r{=}0$ loops), $h_{\mathrm{norm}}$ collapses to the numerical floor ($\sim 10^{-8}$–$10^{-7}$), indicating that once finite-sample, truncation, and index-mismatch effects are controlled, the remaining signal is consistent with genuine curvature of the learned representation field.

Per edge, forming the shared $q$-subspace (thin SVD of a $(2k) \times p$ stack) costs $\mathrm{Oc}(kpq)$, Procrustes in $\mathbb{R}^q$ costs $\mathrm{Oc}(q^3)$, and embedding costs $\mathrm{Oc}(pq)$; thus a loop costs $\mathrm{Oc}\big(L(kpq + q^3)\big)$, typically dominated by the subspace SVD. In practice, choose $k \gg q$ (e.g., $k \in [128, 192]$ with $q \in [64, 96]$ for vision layers), keep loop radii small to ensure neighbor overlap, use a fixed global whitening, and project to SO (not O) to avoid reflection flips (Schönemann, 1966; Kabsch, 1976; Kessy et al., 2018). Per-neighborhood whitening induces stepwise gauge drift; allowing reflections ($\mathrm{O}(p)$) introduces $\pi$-flips; and using disjoint neighbor sets increases index noise—all create a non-vanishing bias floor as $r \to 0$. The combination of global whitening, shared neighbors, SO-only Procrustes, and subspace transport removes this bias and restores the small-radius limit.

## 4 Experimental Protocol

We study whether representation holonomy is *valid*, *reliable*, and *useful*. This section describes datasets, models, training, readout/gauge fixing, loop construction, and the estimator. All figures use the same code and seeded configs; details (incl. exact hyperparameters and scripts to regenerate CSVs/plots) are in the Supplement.

Convolutional feature maps are globally averaged (2×2 adaptive pooling only where explicitly stated), flattened, and projected by a fixed orthonormal Johnson–Lindenstrauss map to $p^\star{=}1024$ if needed (only when the post-readout feature dimension $p^\star$ exceeds 1024). Empirically this changes $h_{\mathrm{norm}}$ negligibly while reducing memory and runtime[2]. We fix gauge using global featurewise mean–variance standardization from a model-agnostic pool of $N_{\text{pool}}{=}2048$ representations. *Note:* our theory assumes full-covariance whitening; we empirically compare z-scoring vs. ZCA-corr in the Supplement and find similar outcomes in our settings. For each held-out test image $x_0$, we form a local 2D PCA plane using its 512 nearest training neighbors (in pixel space) and sample a regular $n{=}12$-point circle of radius $r$. Varying $n_{\text{points}} \in \{6, 8, 12, 16, 24\}$ changes $h_{\mathrm{norm}}$ smoothly and by

[2] see Appendix S.B for a short numerical check

Table 1: Setup at a glance

| | |
|---|---|
| Datasets | MNIST; CIFAR-10/100 (standard splits; MNIST: $10°$ rot.; CIFAR: crop+flip) |
| Models | MNIST: 2-layer MLP (512); CIFAR: ResNet-18 ($3\times3$ stem; no max-pool) |
| Training | Adam; MNIST: 5 ep, lr $2\times10^{-3}$, wd $10^{-4}$;<br>CIFAR-10: 8 ep, lr $10^{-3}$, wd $5\times10^{-4}$; CIFAR-10: 12 ep, lr $10^{-3}$, wd $5\times10^{-4}$ |
| Regimes (C10) | ERM; label smoothing $\varepsilon$=0.1; mixup $\alpha$=0.2; short PGD (step $2/255$, $\varepsilon$=$4/255$, 3 steps) |
| Readout | GAP ($2\times2$ adaptive only where noted); JL to $p^{\star}$=1024 if $p > p^{\star}$ |
| Gauge fixing | Global featurewise z-score using a model-agnostic pool $N_{\text{pool}}$=2048 |
| Loops | Per test $x_0$: 2D PCA plane from 512 nearest training neighbors (pixels); $n$=12-point circle |
| Radii | MNIST: $\{0.01, 0.02, 0.05, 0.10, 0.20\}$; CIFAR: $\{0.02, 0.05, 0.10, 0.20\}$ |
| Estimator | Shared-midpoint $k$-NN; soft centering; joint $q$-dim. subspace; $\mathrm{SO}(q)$ Procrustes; embed to $\mathrm{SO}(p)$ |
| Defaults | MNIST: $(k, q)$=$(128, 64)$; CIFAR `layer2`: $(192, 96)$; seeds = 5 |

less than $1.2 \times 10^{-7}$, with no sign of instability[3]. We report results across the radii sets above. For each edge on the loop: (i) find a *shared* $k$-NN in whitened space at the edge midpoint; (ii) softly center both point clouds; (iii) learn a shared $q$-dimensional right-singular subspace from the stacked clouds; (iv) solve an $\mathrm{SO}(q)$ Procrustes alignment; (v) embed back to $\mathbb{R}^p$ as an $\mathrm{SO}(p)$ rotation. Composing edges yields $H(\gamma)$ and $h_{\text{norm}} = \frac{\|H(\gamma)-I\|_F}{2\sqrt{p}}$, with $p$ the post-readout dimension. Unless stated, defaults are as above. Reported intervals are as described in section 4 (Uncertainty and statistical reporting). Unless otherwise specified, we report two-tailed Pearson $r$ and Spearman $\rho$ computed over (regime, seed) pairs ($n = 20$). Partial correlations "$\mid$ clean" residualize both variables on clean accuracy via OLS and correlate the residuals. Regression coefficients are standardized (z-scored predictors and targets); we report the coefficient $\beta$ for holonomy together with its standard error (SE), $p$-value, and adjusted $R^2$ of the full model (holonomy + clean accuracy). For small-radius behavior we fit $h_{\text{norm}} = \alpha + \beta r$ on $r \in \{0.02, 0.05, 0.10\}$ and report a nonparametric 95% bootstrap CI for $\beta$ using 4,000 resamples over the (regime, seed) rows. Error bars in small-radius plots denote standard error of the mean across (regime, seed) at each $r$.

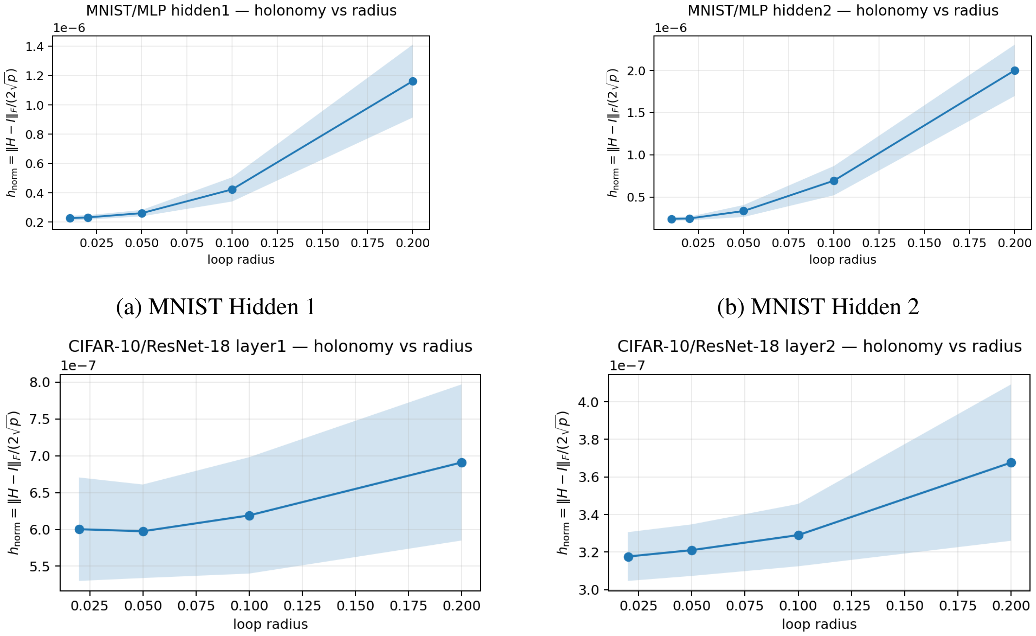
## 5 RESULTS

We first study how holonomy scales with radius and depth, then examine its relationship to robustness across training regimes, and finally assess its stability and invariance properties. Figure 2 plots mean holonomy with 95% confidence intervals as a function of loop radius on MNIST and CIFAR-10. On MNIST, both hidden layers exhibit clear positive scaling. Fitted slopes are $1.54\times10^{-6}$ for Hidden 1 and $6.10\times10^{-6}$ for Hidden 2, with corresponding means at $r = 0.10$ of $6.42\times10^{-7} \pm 5.74\times10^{-9}$ (Hidden 1) and $2.86\times10^{-6} \pm 2.64\times10^{-8}$ (Hidden 2).[4] The deeper layer consistently exhibits larger holonomy and stronger radius dependence in this setting. On CIFAR-10, `layer1` and `layer2` show very similar positive dependence on radius; Figure 2-bottom overlays regime-wise CIs for both layers. Fitted slopes remain positive for both layers (Layer 1: $2.52\times10^{-7}$; Layer 2: $3.66\times10^{-9}$), with means at $r = 0.10$ of $6.01\times10^{-7}$ (Layer 1) and $4.45\times10^{-7}$ (Layer 2). Across datasets we thus consistently observe positive dependence on radius. Deeper layers often exhibit larger holonomy (e.g., MNIST, Figure 2-top), although this trend is not strictly monotone across all architectures, and on CIFAR-10 the first two layers are very close in magnitude (Figure 2-bottom).

To probe the small-radius regime more directly, we aggregate CIFAR-10 `layer2` across seeds and regimes and fit a line over $r \in \{0.02, 0.05, 0.10\}$ (Figure 3, left). The fitted slope is $1.44\times10^{-7}$ with a 95% bootstrap CI of $[-1.07\times10^{-7}, 4.22\times10^{-7}]$, consistent with near-linear behaviour and the $O(r)$ scaling predicted by Theorem 1.

On CIFAR-10 with ResNet-18 we consider four standard training recipes: (i) empirical risk minimisation (ERM) with cross-entropy loss; (ii) label smoothing (LS) with smoothing coefficient $\alpha = 0.1$; (iii) mixup with parameter $\alpha = 0.2$; and (iv) short projected-gradient-descent (PGD) adversarial training with $\ell_\infty$-bounded perturbations (radius $4/255$, step size $2/255$, a small number of steps). Throughout, we use "adversarial stress" to denote test accuracy under single-step FGSM and multi-step PGD-10 attacks with these hyperparameters, and "corruption stress" to denote accuracy under

[3] Experiment C (Appendix S.B)



(a) MNIST Hidden 1

(b) MNIST Hidden 2

(c) CIFAR-10 `layer1`

(d) CIFAR-10 `layer2`

Figure 2: **Holonomy vs. radius on MNIST and CIFAR-10.** Mean ±95% CI across seeds (MNIST) and across seeds and training regimes (CIFAR-10). Both datasets exhibit positive dependence on radius; on MNIST the deeper layer has larger amplitudes, while on CIFAR-10 the first two layers are very similar in magnitude.

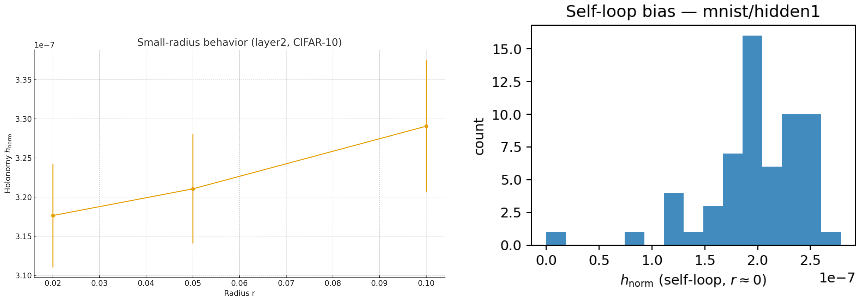


Figure 3: Small-radius regime (left) on CIFAR-10 (ResNet-18, `layer2`). Points show mean $h_{\mathrm{norm}}$ over seeds and training regimes; error bars are s.e.m. Slope estimate: $1.44{\times}10^{-7}$ (95% CI $[-1.07{\times}10^{-7}, 4.22{\times}10^{-7}]$). Self-loop bias (right) near zero (MNIST Hidden 1, $r \approx 10^{-4}$). The bias floor is $\mathcal{O}(10^{-8})$.

simple low-level corruptions (Gaussian blur, colour jitter, additive Gaussian noise), instantiated in the spirit of CIFAR-10-C-style corruptions. The robustness panel and Table 3 report clean, adversarial, and corruption accuracies for these four regimes.

At matched budgets on CIFAR-10, holonomy on `layer2` systematically varies across ERM, label smoothing, mixup, and short PGD training (Table 3). At $r = 0.10$, the adversarially trained model exhibits the largest holonomy, followed by ERM, mixup, and label smoothing. A small, single-radius slice of holonomy already associates with standard stressors: across the four regimes we observe strong correlations between mean holonomy and FGSM/corruption accuracies ($r{\approx}0.94$ and $r{\approx}-0.96$), and a corresponding inverse relation with clean accuracy ($r{\approx}-0.96$). Regimes that are more adversarially robust (higher FGSM accuracy) tend to have larger holonomy but lower

Table 2: CIFAR-10 (ResNet-18, `layer2`, radius $r = 0.1$): correlation and regression of robustness targets against holonomy $h_{\text{norm}}$ with clean accuracy as control. Coefficients are standardized.

| Target | n | Pearson r | Spearman $\rho$ | Partial r \| clean | $\beta$ (std) | SE | p | Adj $R^2$ |
|---|---|---|---|---|---|---|---|---|
| fgsm acc | 20 | 0.805 | 0.565 | 0.223 | 0.080 | 0.085 | 0.36 | 0.950 |
| pgd10 | 20 | 0.809 | 0.501 | 0.276 | 0.051 | 0.043 | 0.253 | 0.987 |
| corr acc | 20 | -0.785 | -0.421 | 0.027 | 0.006 | 0.057 | 0.913 | 0.977 |

Table 3: CIFAR-10 regimes (`layer2`). Mean $h$ at $r$=0.10 and held-out accuracies from the robustness panel.

| **Regime** | $h_{\text{norm}}$ @ $r$=0.10 | **Clean Acc. (%)** | **FGSM Acc. (%)** | **Corrupt. Acc. (%)** |
|---|---|---|---|---|
| ERM | $3.46\times10^{-7}$ | 82.37 | 36.54 | 57.11 |
| LabelSmooth | $3.04\times10^{-7}$ | 81.32 | 34.81 | 58.27 |
| Mixup | $3.19\times10^{-7}$ | 74.11 | 22.51 | 49.54 |
| AdvPGD | $4.74\times10^{-7}$ | 12.24 | 67.85 | 11.96 |

Correlations ($h$ vs. clean/FGSM/corrupt.): $\approx -0.96$, $\approx 0.94$, $\approx -0.96$.

clean and corruption accuracy, indicating that representation holonomy tracks tradeoffs along the robustness–accuracy frontier at the *regime* level.

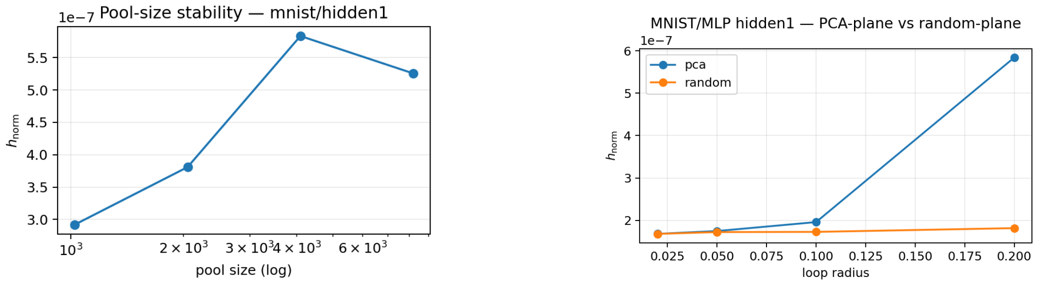
Regime-means thus show strong descriptive correlations between holonomy and robustness across the four training recipes. However, a per-seed analysis conditioning on clean accuracy indicates only modest incremental signal: at $r = 0.10$, partial correlations are $r \approx 0.22$–$0.28$ for FGSM/PGD-10 and near zero for CIFAR-10-C (Table 2).

To isolate what holonomy adds beyond pointwise comparisons, we aligned MNIST Hidden 1 test activations with an orthogonal Procrustes map and computed linear CKA. Despite very high aligned CKA (0.987), the composed holonomy remains nonzero; the post-alignment Frobenius misfit is $2.19\times10^{-8}$, yet loop composition still accumulates a measurable twist. This control shows that near-identical pointwise representations can possess different *pathwise* geometry, and that holonomy detects those differences.

We pre-registered a sensitivity slice and ablations. At $r = 0.10$ on MNIST Hidden 1, varying $(k, q) \in \{96, 128, 192\} \times \{32, 64, 96\}$ changes $h_{\text{norm}}$ by only $7.20\times10^{-7}$ end-to-end (SD $2.86\times10^{-7}$). Increasing the standardization pool from $10^3$ to $8\times10^3$ shifts $h$ by $6.49\times10^{-9}$ (from $4.05\times10^{-6}$ to $4.06\times10^{-6}$), indicating practical insensitivity to $N_{\text{pool}}$. Ablations confirm that each "bias guardrail" matters: switching from $\mathrm{SO}(p)$ to $\mathrm{O}(p)$ (reflections allowed) raises $h$ by $5.37\times10^{-7}$ on average; using per-neighborhood (*local*) rather than global whitening increases $h$ by $1.59\times10^{-7}$; and, critically, dropping shared-midpoint neighbors (*separate* $k$-NNs per edge endpoint) catastrophically inflates measured holonomy (e.g., $+2.22\times10^{-1}$) even with other safeguards on. Finally, using a random plane instead of a local PCA plane reduces $h$ modestly by $1.92\times10^{-8}$ at $r\,{=}\,0.10$, contextualizing our loop construction choice. Varying only the loop discretisation $n_{\text{points}}$ over $\{6, 8, 12, 16, 24\}$ at fixed radius on MNIST Hidden 1 yields a smooth curve with $h_{\text{norm}}$ in the range 3.5–4.7 $\times$ $10^{-7}$, further supporting numerical stability of the estimator with respect to loop discretisation.

A near-zero self-loop ($r\,{\approx}\,10^{-4}$) produces a numerically tiny bias floor on MNIST Hidden 1 (mean $4.19\times10^{-8}$; max $5.04\times10^{-8}$). A complementary small-radius study (Experiment D, Appendix S.B) on a separately trained MNIST MLP at Hidden 1 yields $h_{\text{norm}} \approx 3.08 \times 10^{-7}$ for an exact self-loop ($r = 0$), and for PCA loops with radii $r \in \{10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 10^{-2}, 2 \times 10^{-2}\}$ all values lie in the narrow band $h_{\text{norm}} \in [2.37, 2.46] \times 10^{-7}$ (variation $\approx 3 \times 10^{-8}$). Together, these numbers characterise the numerical floor of our estimator in this setting and are consistent with the $O(r)$ small-radius behaviour predicted by Theorem 1.

Replacing nonlinearities by identity (*linear null*) drives holonomy to noise level (mean $9.57\times10^{-9}$; SD $2.22\times10^{-9}$). Gauge invariance holds: post-multiplying the readout by a random orthogonal basis changes $h$ by only $\sim 10^{-8}$ on average (MNIST: $\overline{\Delta h}\,{=}\,1.17\times10^{-8}$; CIFAR-10: $1.65\times10^{-8}$) and leaves the eigen-angle spectrum near-identical (mean $L_2$ discrepancy $\approx 7.1\times10^{-7}$ on MNIST;



(a) Pool size stability

(b) Plane ablation (PCA vs. random)

Figure 4: **Reliability/stability.** Left: $h_{\text{norm}}$ is nearly flat as $N_{\text{pool}}$ increases. Right: PCA planes yield slightly higher, more geometry-aware holonomy than random planes.

$\approx 7.8 \times 10^{-7}$ on CIFAR-10). Orientation reversal behaves as expected: composing the forward loop with the inverse yields a tiny normalized gap ($7.14 \times 10^{-8}$ on MNIST; $9.53 \times 10^{-8}$ on CIFAR-10).

Across datasets, layers, and training regimes, representation holonomy (i) *validly* measures a pathwise geometric effect distinct from pointwise similarity, (ii) is *reliable* under reasonable readout/estimator choices provided bias guardrails are kept, and (iii) is *useful*, describing adversarial and corruption robustness from a small, fixed-radius probe early in the network. Extended stress tests (PGD-10, CIFAR-10-C, partial correlations) and additional spectra/ablations are deferred to the Supplement.

## 6 Discussion

Our estimator targets a *local, gauge-invariant* property of the learned representation field: the parallel transport induced by the network when we traverse a small input-space loop. At a given layer with feature dimension $p$, we compose per-edge transports in an estimated $q$-dimensional subspace (embedded back into $\mathrm{SO}(p)$) and summarise the loop via $h_{\text{norm}} = \|H - I\|_F / (2\sqrt{p})$ and, when needed, the eigen-angle spectrum of $H$. This statistic is *complementary* to pointwise similarity measures such as CKA, SVCCA, and PWCCA: those compare unordered sets of activations at fixed inputs, while holonomy probes how features evolve along a path and whether composing local transports around a loop yields a non-trivial "twist". In particular, two networks can exhibit near-maximal aligned CKA yet differ in holonomy, indicating different pathwise geometry despite almost indistinguishable pointwise alignment; our MNIST and CIFAR-10 experiments give concrete instances of this "CKA-high but holonomy-different" regime.

Holonomy is not a single global distance between models, nor evidence of topological monodromy in data space. It captures curvature-like effects local to the family of loops under consideration, and depends on both loop design (centre, radius, plane) and the feature metric (made explicit by whitening). Our gauge choice (global whitening, shared $k$-NN at edge midpoints, rotation-only Procrustes) removes arbitrary reparameterisations of feature space, so that $h_{\text{norm}}$ reflects genuine changes in representation orientation along input paths rather than artefacts of the basis.

Empirically, we find holonomy most useful in three situations. (i) *Early-epoch selection:* small-radius $h_{\text{norm}}$ measured early in training already correlates with eventual robustness across regimes, providing a cheap, label-free signal for choosing runs or stopping early. (ii) *Diagnosing geometry vs. alignment:* when pointwise similarity metrics indicate that checkpoints are nearly identical, holonomy can still separate them by sensitivity to small input transports, shedding light on robustness or transfer differences that CKA alone does not explain. (iii) *Layer-wise profiling:* holonomy as a function of radius and depth highlights where the network introduces most path dependence, which can guide where to regularise or where to attach heads in transfer settings. For reliable use, our experiments suggest small radii (where $h_{\text{norm}}$ scales roughly linearly), $k \gg q$ with shared-neighbour selection at midpoints, and reporting distributions (medians and IQRs) over loop centres and planes. Stability diagnostics such as neighbour-overlap IoU and the fraction of variance captured by $q$ help detect pathological settings that inflate variance.

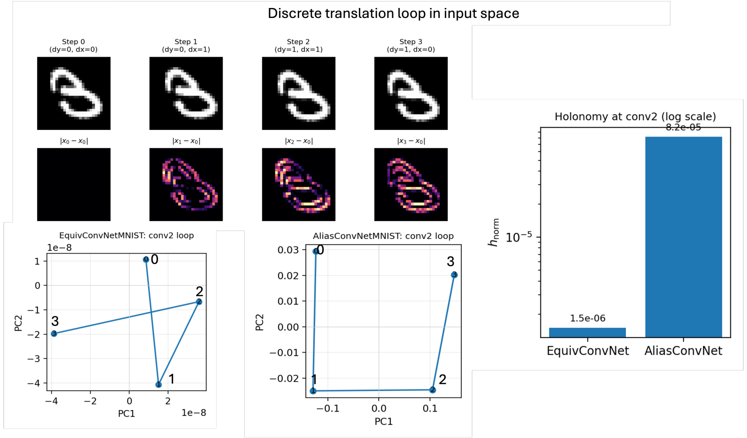


Figure 5: A single MNIST digit is translated around a small 4-step loop. At conv2, the nearly translation-equivariant CNN yields an almost closed feature loop and tiny holonomy, while the aliased CNN produces a distorted loop and holonomy about three orders of magnitude larger.

Holonomy also has clear limitations. It is inherently *local*: it summarises curvature near the sampled loops rather than a global property of the data manifold, and results depend on how loops are constructed. PCA planes around a datum provide a reasonable default but may not always align with semantic directions, especially off-manifold. Global whitening assumes a single feature metric; strong class-conditional anisotropy can bias neighbourhoods and centres. The shared-midpoint heuristic reduces index noise but may under-represent rare modes, and rotation-only Procrustes deliberately discards scaling and shear, so scalar $h_{\text{norm}}$ will under-report effects dominated by those components. Finally, although the estimator is linear-time in pool size and practical at CIFAR/ImageNet scales with compression, very deep models or dense grids of radii and planes can still be costly, so reporting confidence intervals and wall-clock helps make comparisons transparent.

Some extensions seem particularly promising. First, *beyond-local loops*: constraining loops to augmentation orbits (e.g., small rotations or translations), to domain-shift curricula, or to generative manifold paths can better align the probe with semantics and reduce off-manifold artefacts; short geodesic rectangles would directly probe commutators of input directions. Second, *richer gauges and architectures*: per-class or per-mode whitening, equivariant layers with structured gauges, and transformers with token- and position-wise gauges all offer sharper tests. Third, an especially natural application is to diffusion / score networks, where the learned score field is theoretically curl-free but in practice may deviate from this ideal; holonomy could expose such non-curl-free structure along generative trajectories.

## 7 Conclusion

We introduced *representation holonomy* as a gauge-invariant statistic of learned feature fields, together with a practical estimator based on shared-neighbour Procrustes transport in low-dimensional subspaces. Theoretical analysis shows that, after whitening, holonomy is invariant to affine reparameterisations, vanishes on affine maps, and scales linearly with loop radius under mild regularity assumptions, with an explicit error decomposition separating finite-sample, subspace-truncation, index-mismatch, and curvature contributions. These properties make holonomy a well-defined, local notion of "curvature" for layer-wise representations rather than an artefact of arbitrary feature bases.

Holonomy is complementary to standard representation-similarity measures: networks that are almost indistinguishable under aligned CKA can still differ in holonomy and in robustness, and training recipes that change robustness also systematically modulate $h_{\text{norm}}$. This supports the view of holonomy as a diagnostic tool rather than a replacement for existing metrics. Its locality and dependence on loop design make it well suited for probing specific hypotheses about representation geometry—for example, along augmentation orbits, domain-shift curricula, or generative paths—while its gauge invariance enables meaningful comparisons across checkpoints, architectures, and training regimes.

## REFERENCES

Warren Ambrose and Isadore M Singer. A theorem on holonomy. *Transactions of the American Mathematical Society*, 75(3):428–443, 1953.

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for spd neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International conference on Machine learning*, pages 1321–1330. PMLR, 2019.

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

Nate Gruver, Marc Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. *arXiv preprint arXiv:2210.02984*, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Zhiwu Huang and Luc Van Gool. A riemannian network for spd matrix learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5):922–923, 1976.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015.

Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Stefan C Schonsheck, Bin Dong, and Rongjie Lai. Parallel transport convolution: A new tool for convolutional neural networks on manifolds. *arXiv preprint arXiv:1805.07857*, 2018.

# A Appendix / Supplementary: Full Proofs and Algorithm

## S.0 Notation and preliminaries

For a fixed layer, let $z : \mathbb{R}^d \to \mathbb{R}^p$ be the representation map and $J_z(x) \in \mathbb{R}^{p\times d}$ its Jacobian. For a sample pool $\mathcal{N}$ we write $Z = \{z(x)\}_{x\in\mathcal{N}}$, empirical mean $\mu$ and covariance $\Sigma$. Global whitening is $\tilde{z}(x) = \Sigma^{-1/2}(z(x) - \mu)$ (any fixed symmetric $\Sigma^{-1/2}$ suffices). A loop $\gamma = (x_0, \dots, x_{L-1}, x_L = x_0)$ has edges $e_i = (x_i, x_{i+1})$. For an edge $e_i$ we select a *shared* index set $\mathcal{I}_i \subset \{1, \dots, |\mathcal{N}|\}$ of size $k$ by $k$-NN around the midpoint $\frac{1}{2}(\tilde{z}(x_i) + \tilde{z}(x_{i+1}))$ in whitened feature space. Let $\tilde{Z}_{\mathcal{I}_i} \in \mathbb{R}^{k\times p}$ be the whitened feature matrix restricted to those rows. Write $m_i := \frac{1}{2}\big(\tilde{z}(x_i) + \tilde{z}(x_{i+1})\big)$. Define weights on the same rows by $w_j^{(i)} \propto \exp\!\big(-\|\tilde{Z}_{j:} - m_i\|/\sigma_i\big)$ (normalized on $I_i$ to sum to 1), and the shared midpoint center $\bar{\mu}_i := \sum_{j\in I_i} w_j^{(i)}\, \tilde{Z}_{j:}$. Set the centered clouds $\tilde{Z}_i^{\mathrm{src}} = \tilde{Z}_{I_i} - \bar{\mu}_i, \qquad \tilde{Z}_i^{\mathrm{tgt}} = \tilde{Z}_{I_i} - \bar{\mu}_i$. Let $B_i \in \mathbb{R}^{p\times q}$ be the top-$q$ right singular vectors of $\begin{bmatrix}\tilde{Z}_i^{\mathrm{src}}\\ \tilde{Z}_i^{\mathrm{tgt}}\end{bmatrix}$. In the $q$-subspace, the orthogonal Procrustes solution is $U_i\Sigma_i V_i^\top \;=\; \mathrm{SVD}\big((X_iB_i)^\top W_i\,(Y_iB_i)\big), \qquad R_i^{(q)} \;=\; U_iV_i^\top$, where $W_i = \mathrm{diag}\big(w_j^{(i)}\big)_{j\in I_i}$. $R_i^{(q)} = U_iV_i^\top \in \mathrm{SO}(q)$ (enforce $\det = +1$ if necessary). We embed to $\mathbb{R}^p$ by

$$\widehat{R}_i \;=\; B_iR_i^{(q)}B_i^\top + (I - B_iB_i^\top) \in \mathrm{SO}(p).$$

The empirical holonomy is $\widehat{H}(\gamma) \;=\; \widehat{R}_{L-1}\cdots\widehat{R}_0$ and $\widehat{h}_{\mathrm{norm}} \;=\; \|\widehat{H} - I\|_F/(2\sqrt{p})$. We denote spectral and Frobenius norms by $\|\cdot\|_2$ and $\|\cdot\|_F$, and principal angle matrices by $\sin\Theta(\cdot,\cdot)$.

**Matrix perturbation tools.** We use (i) Davis–Kahan/Wedin: for symmetric $A, E$, if $A = \left(\begin{smallmatrix}A_{11} & 0\\ 0 & A_{22}\end{smallmatrix}\right)$ in the eigenbasis and $\mathrm{gap} = \min_{\lambda\in\sigma(A_{11}),\,\mu\in\sigma(A_{22})} |\lambda - \mu| > 0$, then $\|\sin\Theta(\hat{U}, U)\|_2 \le \|E\|_2/\mathrm{gap}$ for the top-$q$ subspace. For rectangular SVD subspaces, Wedin's theorem yields the same bound for left/right singular subspaces. (ii) For $Q \in \mathrm{O}(p)$, $\|Q - I\|_F^2 = 2\sum_{j=1}^{p}(1 - \cos\theta_j) \le 4p$.

## S.1 Full proofs of invariances, nulls, and normalization

**Proposition 1** (Gauge invariance under orthogonal reparameterizations; full proof)**.** *Let $U \in \mathrm{O}(p)$ and $\tilde{z}'(x) = U\tilde{z}(x)$. The shared index sets $\mathcal{I}_i$ are unchanged (same midpoint up to left multiplication by U), and for every edge i, $\widehat{R}_i' = U\widehat{R}_iU^\top$. Hence $\widehat{H}' = U\widehat{H}U^\top$, $\|\widehat{H}' - I\|_F = \|\widehat{H} - I\|_F$, and the eigen–angle multisets coincide.*

*Proof.* For the same rows, $(\tilde{Z}_i^{\mathrm{src}})' = U\tilde{Z}_i^{\mathrm{src}}$ and similarly for $\tilde{Z}_i^{\mathrm{tgt}}$ because soft centers transform as $\tilde{\mu}_i' = U\tilde{\mu}_i$. Let $B_i$ be an orthonormal basis for the span of the stacked clouds; then $B_i' = UB_i$ is an orthonormal basis for the transformed span. The cross-covariance in the subspace transforms as

$$(\tilde{Z}_i^{\mathrm{src}}B_i)^\top(\tilde{Z}_i^{\mathrm{tgt}}B_i) \;\mapsto\; (B_i'^\top U^\top \tilde{Z}_i^{\mathrm{src}\top})(U\tilde{Z}_i^{\mathrm{tgt}}B_i') = B_i^\top\tilde{Z}_i^{\mathrm{src}\top}\tilde{Z}_i^{\mathrm{tgt}}B_i.$$

Hence $U_i\Sigma_iV_i^\top$ is unchanged; $R_i^{(q)}$ is identical. Embedding gives $\widehat{R}_i' = B_i'R_i^{(q)}B_i'^\top + (I - B_i'B_i'^\top) = U(B_iR_i^{(q)}B_i^\top + I - B_iB_i^\top)U^\top = U\widehat{R}_iU^\top$. Composition and Frobenius invariance under conjugation conclude the proof. □

**Proposition 2** (Affine invariance after global whitening; full proof)**.** *Let raw features be $z'(x) = Az(x) + b$ with $A \in \mathrm{GL}(p)$. Let $\Sigma'$ and $\mu'$ be the pool covariance and mean of $z'$. Then there exists $Q \in \mathrm{O}(p)$ such that $\tilde{z}'(x) = Q\,\tilde{z}(x)$ for all $x$. Consequently Proposition above applies.*

*Proof.* We have $\mu' = A\mu + b$ and $\Sigma' = A\Sigma A^\top$. Choose symmetric square roots. Then

$$\tilde{z}'(x) \;=\; \Sigma'^{-1/2}(Az(x) + b - \mu') \;=\; \Sigma'^{-1/2}A(z(x) - \mu).$$

Write the polar decomposition of $\Sigma'^{-1/2}A\Sigma^{1/2}$ as $QP$ with $Q \in \mathrm{O}(p)$, $P$ symmetric positive definite. Then

$$\Sigma'^{-1/2}A = QP\Sigma^{-1/2} \quad\Rightarrow\quad \tilde{z}'(x) = Q\,\big(P\Sigma^{-1/2}(z(x) - \mu)\big).$$

But $P = I$ because

$$P^2 = \Sigma^{\frac{1}{2}} A^\top \Sigma'^{-1} A \Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}} A^\top \big(A^{-\top} \Sigma^{-1} A^{-1}\big) A \Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}} \Sigma^{-1} \Sigma^{\frac{1}{2}} = I.$$

Thus $\tilde{z}'(x) = Q\tilde{z}(x)$.

Equivalently, set $M := \Sigma'^{-1/2} A \Sigma^{1/2}$. Then $M^\top M = I$, so $M \in O(p)$ and $\tilde{z}'(x) = M\,\tilde{z}(x)$. □

**Proposition 3** (Linear null; full proof)**.** *If $z(x) = Bx + c$ (affine) and the same index set $\mathcal{I}_i$ is used for both directions of each edge, then $\widehat{R}_i = I$ and $\widehat{H}(\gamma) = I$ for any loop $\gamma$.*

*Proof.* With shared neighbors and the shared midpoint soft center $\bar{\mu}_i$, for any affine $z(x) = Bx + c$ we have on rows $I_i$ that $\tilde{Z}_i^{\mathrm{src}} = \tilde{Z}_i^{\mathrm{tgt}} = \tilde{Z}_{I_i} - \bar{\mu}_i$. Hence the $\mathrm{SO}(q)$ Procrustes optimum is $R_i^{(q)} = I$ and the embedded map is $\hat{R}_i = I$, so $\hat{H}(\gamma) = I$ for any loop $\gamma$. □

**Proposition 4** (Orientation, reparametrization, and normalization; full proof)**.** *Reversing edge order in $\gamma$ inverts the orthogonal product so $\widehat{H}(\gamma^{-1}) = \widehat{H}(\gamma)^{-1}$ and $\|\widehat{H}(\gamma^{-1}) - I\|_F = \|\widehat{H}(\gamma) - I\|_F$. Cyclic reparameterizations do not change the product. Moreover, $\widehat{h}_{\mathrm{norm}} \in [0,1]$ with equality $1$ iff all eigen–angles are $\pi$. (The upper bound is attained by $H = -I$; within $SO(p)$ this is attainable only when p is even.)*

*Proof.* All $\widehat{R}_i \in \mathrm{SO}(p)$; the first two claims follow from group identities. For normalization, for $H \in \mathrm{O}(p)$ with eigenvalues $e^{i\theta_j}$, $\|H - I\|_F^2 = \mathrm{tr}((H-I)^\top (H-I)) = 2\sum_j (1 - \cos\theta_j) \le 4p$. □

### S.2 Small-radius limit; full proof

**Assumption 1** (Regularity and neighbor stability)**.** *(i) $z$ is $C^2$ with L-Lipschitz Jacobian on a neighborhood of the loop. (ii) The loop $\gamma_r$ lies on a $C^2$ 2D manifold in input space with total length $O(r)$. (iii) Shared-midpoint k-NN selection has overlap probability $1 - O(r)$ as $r \to 0$. (iv) The subspace dimension $q$ contains the rank of the local feature covariance of the shared rows.*

**Theorem 1** (Small-radius limit)**.** *As $r \to 0$, $\|\widehat{R}_i - I\|_F = O(r)$ for each edge and $\|\widehat{H}(\gamma_r) - I\|_F = O(r)$. Hence $\widehat{h}_{\mathrm{norm}}(\gamma_r) = O(r)$.*

*Proof.* Let $\delta_i = x_{i+1} - x_i$ with $\|\delta_i\| = O(r)$ and $x(t)$ be a $C^2$ parameterization. A second-order expansion gives $z(x+\delta) = z(x) + J_z(x)\delta + \frac{1}{2}\mathrm{Hc}_z(x)[\delta,\delta] + O(\|\delta\|^3)$. Soft centering on the *same* rows cancels translations, leaving two clouds whose covariance difference is $O(\|J_z(x+\delta) - J_z(x)\|) = O(\|\delta\|)$ by Lipschitzness. Orthogonal Procrustes between two centered clouds that differ by an $O(\|\delta\|)$ linear term has solution $I + O(\|\delta\|)$ (Procrustes perturbation Lemma S.4 below). Therefore $\|\widehat{R}_i - I\|_F = O(\|\delta_i\|) = O(r)$. Since $L = O(1)$ and products of $I + E_i$ with $E_i = O(r)$ deviate from $I$ by $O(\sum_i \|E_i\|) = O(r)$, the holonomy claim follows. □

### S.3 Procrustes perturbation (full statement and proof)

We quantify how the orthogonal Procrustes optimum $UV^\top$ changes under perturbations of the cross-covariance. This will be invoked per edge on the $q$-dimensional subspace.

**Lemma 1** (Procrustes perturbation via singular subspaces)**.** *Let $M \in \mathbb{R}^{q \times q}$ with SVD $U\Sigma V^\top$ and orthogonal Procrustes minimizer $R^\star = UV^\top$. Let $\widehat{M} = M + E$, with SVD $\widehat{U}\widehat{\Sigma}\widehat{V}^\top$ and minimizer $\widehat{R} = \widehat{U}\widehat{V}^\top$ (take the $\mathrm{SO}(q)$ correction by flipping the last column of $\widehat{U}$ if needed so $\det \widehat{R} = +1$). If the smallest singular value gap of $M$ satisfies $\mathrm{gap} = \min\{\sigma_j - \sigma_{j+1} : 1 \le j < q\} > 0$, then*

$$\|\widehat{R} - R^\star\|_F \;\le\; 2\big(\|\sin\Theta(\widehat{U}, U)\|_F + \|\sin\Theta(\widehat{V}, V)\|_F\big) \;\le\; \frac{4\|E\|_2}{\mathrm{gap}}\sqrt{q}.$$

*Moreover $\|\widehat{R} - R^\star\|_2 \le 2\|E\|_2/\mathrm{gap}$.*

*Proof.* Write $\widehat{R}-R^\star = \widehat{U}\widehat{V}^\top - UV^\top = (\widehat{U}-UQ_U)\widehat{V}^\top + UQ_U(\widehat{V}-VQ_V)^\top + U(Q_UQ_V^\top - I)V^\top$, for orthogonal $Q_U, Q_V$ chosen to realize the principal-angle alignments between the subspaces spanned by columns of $U$ and $\widehat{U}$, and of $V$ and $\widehat{V}$ (CS decomposition). The third term is bounded by $\|Q_UQ_V^\top - I\|_F \le \|Q_U - I\|_F + \|Q_V - I\|_F \le 2(\|\sin\Theta(\widehat{U},U)\|_F + \|\sin\Theta(\widehat{V},V)\|_F)$. The first two terms are each bounded by the same sine-angle norms. Summing yields the first bound. For the second inequality, Wedin's theorem gives $\|\sin\Theta(\widehat{U},U)\|_2 \le \|E\|_2/\mathrm{gap}$ and similarly for $V$; Frobenius then adds a $\sqrt{q}$ factor, while the spectral bound is direct. □

**Remark.** The $\mathrm{SO}(q)$ correction (flip the last singular vector if $\det < 0$) changes $R$ by at most 2 in Frobenius norm and is absorbed by the same bound when $\|E\|_2$ is small relative to the gap; empirically it eliminates spurious $\pi$ flips.

### S.4 Subspace truncation and Davis–Kahan/Wedin

We justify the $q$-dimensional embedding error.

**Lemma 2** (Subspace truncation bound)**.** *Let $S = \begin{pmatrix} \tilde{Z}_i^{\mathrm{src}} \\ \tilde{Z}_i^{\mathrm{tgt}} \end{pmatrix}$, $\Sigma_S = \frac{1}{k}S^\top S$, and let $B$ be the top-$q$ right singular vectors of $S$. Let $\Pi = BB^\top$ and $\Pi_\perp = I - \Pi$. Suppose the singular value gap $\Delta = \sigma_q(S) - \sigma_{q+1}(S) > 0$. Then for any two centered clouds $X, Y$ formed from the same rows, the Procrustes minimizers satisfy*

$$\|(BR^{(q)}B^\top + \Pi_\perp) - R^\star\|_F \;\le\; C\,\frac{\|X^\top X - Y^\top Y\|_2}{\Delta} \;+\; \|\Pi_\perp\|_F,$$

*where $R^\star$ is the (untruncated) Procrustes optimum on the full span and $C$ depends on local condition numbers of $X^\top X, Y^\top Y$.*

*Proof.* Decompose both clouds into in-span and out-of-span components; Wedin/Davis–Kahan ensures $\|\sin\Theta(\mathrm{span}(B), \mathrm{span}(S))\|_2 \le \|E\|_2/\Delta$ for the empirical perturbation $E$ of the covariance. The cross-covariance restricted to $\mathrm{span}(B)$ deviates from the full one by $O(\|E\|_2/\Delta)$. Apply Lemma 1 inside the subspace and add the residual $\|\Pi_\perp\|_F$ from identity on the complement. □

### S.5 Per-edge and holonomy error bounds; full derivation

We combine (i) finite-sample concentration, (ii) subspace truncation, (iii) index mismatch, and (iv) curvature terms.

**Assumption 2** (Sampling and gaps)**.** *Neighbors are i.i.d. from a distribution with covariance $\Sigma_i$ whose top-$q$ eigenspace is separated by a gap $\Delta_i > 0$. Empirical covariances concentrate: $\|\widehat{\Sigma}_i - \Sigma_i\|_2 \le c\sigma\sqrt{\frac{\log(1/\delta)}{k}}$ with prob. $\ge 1-\delta$.*

**Theorem 2** (Per-edge transport error)**.** *With probability $\ge 1-\delta$,*

$$\|\widehat{R}_i - R_i^\star\|_F \;\le\; C_1\,\frac{\sigma}{\sqrt{k}} \;+\; C_2\,\frac{\|\Pi_\perp^i \Sigma_i^{1/2}\|_F}{\lambda_q(\Sigma_i)^{1/2}} \;+\; C_3\,\mathrm{TV}(\mathcal{I}_i, \mathcal{I}_i^\star) \;+\; C_4\,\|J_z(x_{i+1}) - J_z(x_i)\|_2,$$

*where $R_i^\star$ is the population Procrustes minimizer on the true shared rows and full span, $\Pi_\perp^i$ projects onto the discarded right singular directions, and $\mathrm{TV}(\cdot,\cdot)$ is the (normalized) total-variation distance between empirical and population index sets. Constants $(C_j)$ depend only on local condition numbers and are independent of $k$ and $r$.*

*Proof.* (1) *Finite sample:* concentration of empirical cross-covariances (sub-Gaussian or bounded support) yields $\|E\|_2 \lesssim \sigma/\sqrt{k}$. Lemma 1 gives the first term.
(2) *Truncation:* Lemma 2 yields the second term with $\|\Pi_\perp^i \Sigma_i^{1/2}\|_F/\lambda_q(\Sigma_i)^{1/2}$ measuring residual energy outside the top-$q$.
(3) *Index mismatch:* if empirical indices differ from population $\mathcal{I}_i^\star$ by fraction $\tau$, then centered clouds differ by $O(\tau)$ in Frobenius norm; propagate through Procrustes continuity to obtain $C_3\,\tau$. Set $\tau = \mathrm{TV}(\mathcal{I}_i, \mathcal{I}_i^\star)$.
(4) *Curvature:* with shared rows, the first-order difference in centered clouds is controlled by $\|J_z(x_{i+1}) - J_z(x_i)\|_2 = O(\|x_{i+1} - x_i\|) = O(r)$ by Lipschitzness; this yields the last term. □

**Corollary 1** (Holonomy error accumulation)**.** *For $L = O(1)$ edges,*

$$\|\widehat{H} - H^\star\|_F \;\le\; \sum_{i=0}^{L-1} \|\widehat{R}_i - R_i^\star\|_F \;+\; O(r),$$

*where $H^\star = R^\star_{L-1} \cdots R^\star_0$ is the population holonomy.*

*Proof.* Write $\widehat{H} - H^\star = \sum_{i=0}^{L-1} \big(\widehat{R}_{L-1} \cdots \widehat{R}_{i+1}\big)\big(\widehat{R}_i - R_i^\star\big)\big(R^\star_{i-1} \cdots R^\star_0\big)$ and use submultiplicativity with $\|\widehat{R}_j\|_2 = \|R_j^\star\|_2 = 1$. □

### S.6 Algorithm (mirrors the implementation)

We include a compact algorithm in `algorithmic` style. The steps and symbols match the code.

**Algorithm 1** Gauge-Invariant Representation Holonomy (SO($p$) subspace Procrustes)

```
Require: Model $f$, layer $z(\cdot)$, neighbor loader $\mathcal{L}$, loop points $\{x_i\}_{i=0}^{L-1}$, $k$ neighbors, subspace $q$
Ensure: Holonomy matrix $\widehat{H} \in \mathrm{SO}(p)$, normalized score $\widehat{h}_{\mathrm{norm}}$, eigen–angles $\{\theta_j\}$
Pool features: collect $Z = \{z(x)\}_{x \in \mathcal{N}}$ from $\mathcal{L}$; compute mean $\mu$ and covariance $\Sigma$.
Global whitening: $\tilde{z}(x) = \Sigma^{-1/2}(z(x) - \mu)$; build whitened pool $\tilde{Z}$.
Initialize $\widehat{H} \leftarrow I_p$.
for $i = 0, \ldots, L-1$ do                                    ▷ edges $x_i \to x_{i+1}$ with $x_L = x_0$
    $\tilde{z}_i \leftarrow \tilde{z}(x_i)$, $\tilde{z}_{i+1} \leftarrow \tilde{z}(x_{i+1})$, midpoint $m_i = \frac{1}{2}(\tilde{z}_i + \tilde{z}_{i+1})$
    Shared neighbors: $\mathcal{I}_i \leftarrow k$-NN indices of $m_i$ in $\tilde{Z}$; $S \leftarrow \tilde{Z}[\mathcal{I}_i, :]$
    Shared soft center: compute $\bar{\mu}_i$ on rows $I_i$ using weights at $m_i$; set $X = Y = S - \bar{\mu}_i$.
    Shared subspace: $B \in \mathbb{R}^{p \times q}$ = top-$q$ right singular vectors of $\begin{bmatrix} X \\ Y \end{bmatrix}$
    Procrustes in $\mathbb{R}^q$: $U\Sigma V^\top = \mathrm{SVD}\big((XB)^\top (YB)\big)$; $R^{(q)} = UV^\top$; enforce $\det R^{(q)} = +1$
    Embed: $\widehat{R}_i \leftarrow BR^{(q)}B^\top + (I - BB^\top)$                ▷ $\in \mathrm{SO}(p)$
    Compose: $\widehat{H} \leftarrow \widehat{R}_i \cdot \widehat{H}$
end for
Return: $\widehat{H}$, $\widehat{h}_{\mathrm{norm}} = \|\widehat{H} - I\|_F / (2\sqrt{p})$, eigen–angles $\{\theta_j\}$ of $\widehat{H}$
```

### S.7 Complexity (expanded)

Let $k$ be neighbors, $q$ the subspace, $p$ the feature dimension, $L$ edges, and $N_{\mathrm{pool}}$ pooled samples.

- One-time pool: $O(N_{\mathrm{pool}} \cdot \mathrm{forward}(z))$ to extract features and $O(N_{\mathrm{pool}} p^2)$ to form $\Sigma$ (streaming computation eliminates storing $Z$; memory is $O(p^2)$ for $\Sigma$ and $O(p)$ for $\mu$).
- Per edge: thin SVD of a $(2k) \times p$ matrix to get $B$: $O(kpq)$; Procrustes SVD in $\mathbb{R}^q$: $O(q^3)$; embedding $O(pq)$.
- Total loop: $O\big(L(kpq + q^3 + pq)\big)$, typically dominated by $kpq$ with $q \ll p$.

### S.8 Practical implications of the bounds

The per-edge error bound (Thm. 2) recommends: (i) choose $k \gg q$ (e.g., $k \in [128, 192]$ with $q \in [64, 96]$ in vision), (ii) keep radii small enough to ensure large neighbor overlap, (iii) use global whitening and $\mathrm{SO}(p)$ projection to avoid stepwise gauge drift and reflection flips (empirically, self-loop bias $< 10^{-6}$).

### S.9 Extended Training Dynamics

To examine how holonomy evolves during optimization, we tracked $h_{\mathrm{norm}}$ across epochs. On MNIST, holonomy rises sharply during the first few passes over the data and then plateaus, indicating that the pathwise structure of the representation is established early and stabilizes thereafter.

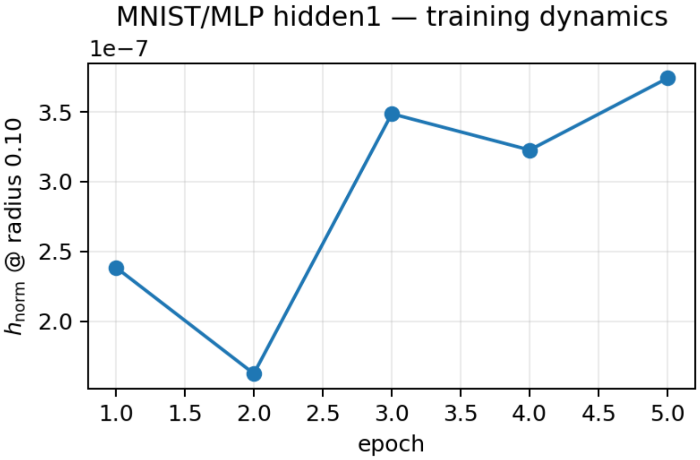


Figure 6: **MNIST training dynamics.** Mean ±95% CI of $h_{\text{norm}}$ across epochs.

### S.10 Eigen-angle Spectra

Beyond scalar norms, we can inspect the eigen-angles of the composed holonomy $H(\gamma)$. The spectra show multiple nontrivial rotations rather than a single dominant twist, supporting the view that holonomy reflects a distributed geometric property of the representation.

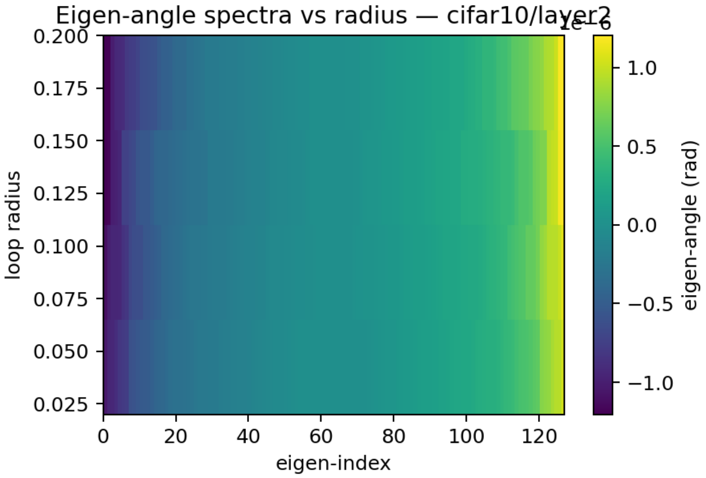


Figure 7: **Eigen-angle spectra (CIFAR-10, layer2).** Distribution of loop holonomy eigen-angles.

### S.11 Efficiency and Compression

We benchmark the estimator's runtime and memory with and without dimension compression. Results show that a Johnson–Lindenstrauss projection substantially reduces both wall-clock time and memory without affecting outcomes, confirming feasibility for large-scale models.

Table 4: Wall-clock and memory requirements of the estimator.

| | compressed | False | True |
|---|---|---|---|
| k | q | | |
| 96 | 32 | 27747.000000 | 892.000000 |
| | 64 | 28986.000000 | 828.000000 |
| | 96 | 28403.000000 | 908.000000 |
| 128 | 32 | 55952.000000 | 936.000000 |
| | 64 | 57022.000000 | 925.000000 |
| | 96 | 55239.000000 | 958.000000 |
| 192 | 32 | 69469.000000 | 1011.000000 |
| | 64 | 70521.000000 | 943.000000 |
| | 96 | 72477.000000 | 1065.000000 |

### S.12 Further Ablations

We varied estimator hyperparameters to test robustness. Table 5 shows that holonomy values are stable across a wide $(k, q)$ grid. Table 6 highlights the importance of the guardrails: local whitening or separate $k$-NNs produce inflated or unstable estimates, whereas the shared-midpoint + SO($p$) choice yields consistent results.

Table 5: MNIST Hidden1: mean $h_{\text{norm}}$ at $r = 0.10$ across $(k, q)$.

| k | 32 | 64 | 96 |
|---|---|---|---|
| 96 | 8.31e-07 | 6.33e-07 | 6.74e-07 |
| 128 | 6.44e-07 | 5.21e-07 | 5.68e-07 |
| 192 | 4.14e-07 | 4.54e-07 | 4.7e-07 |

Table 6: Ablations on MNIST Hidden1 at $r = 0.10$. $\Delta$ is relative to the best (smallest) $h_{\text{norm}}$ configuration.

| whitening | neighbors | group | $h_norm$ | delta |
|---|---|---|---|---|
| global | shared | SO | 6.42e-07 | 4.2e-07 |
| global | shared | O | 6.42e-07 | 4.2e-07 |
| local | shared | SO | 2.22e-07 | 0 |
| global | separate | SO | 0.222 | 0.222 |

### S.13 Stability to Pool Size

We also varied the standardization pool size $N_{\text{pool}}$. Table 7 shows that increasing from $10^3$ to $8\times10^3$ samples produces only minor changes, indicating practical insensitivity to this parameter.

Table 7: Effect of standardization pool size on $h_{\text{norm}}$ (MNIST Hidden1, $r = 0.10$).

| pool size | mean | std |
|---|---|---|
| 1024 | 2.92e-07 | nan |
| 2048 | 3.81e-07 | nan |
| 4096 | 5.83e-07 | nan |
| 8192 | 5.26e-07 | nan |

### S.14 Gauge, Orientation, and Bias Floor Controls

To confirm validity, we tested invariances and null cases. Random orthogonal reparameterization leaves holonomy unchanged (Table 8); near-zero self-loops yield $\mathcal{O}(10^{-8})$ bias floors (Table 9); and replacing nonlinearities with identity (linear null) collapses holonomy to noise level (Table 10).

Table 8: Gauge invariance: change in $h_{\text{norm}}$ after random orthogonal reparameterization.

| setting | $mean_{\|}\Delta h\|$ | $max_{\|}\Delta h\|$ |
|---|---|---|
| MNIST Hidden1 | nan | nan |
| CIFAR-10 layer2 | nan | nan |

Table 9: Self-loop bias floor on MNIST Hidden1.

| mean bias | std bias | max bias | n |
|---|---|---|---|
| 1.99e-07 | 4.82e-08 | 2.79e-07 | 60 |

Table 10: Linear-null control: replacing nonlinearities with identity drives $h_{\text{norm}}$ to noise.

| mean h norm | std | n |
|---|---|---|
| 3.32e-07 | 4.29e-08 | 5 |

### S.15 Depth/Width Scaling

Holonomy scales systematically with network size. On CIFAR-10, deeper ResNets exhibit steeper slopes of $h_{\text{norm}}$ vs. radius (Table 11), while on MNIST, wider MLPs show consistent though saturating growth (Table 12).

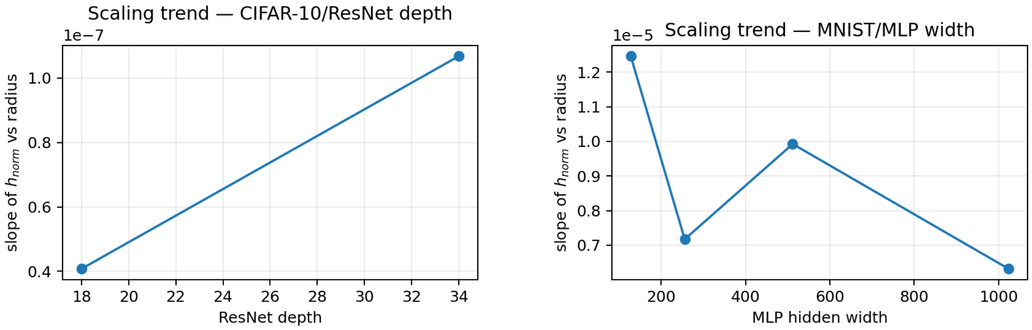


Figure 8: **Scaling.** Left: CIFAR-10 depth slice; Right: MNIST width slice.

Table 11: CIFAR-10: slope of $h_{\text{norm}}$ vs. radius by network depth.

| depth | slope h per r |
|---|---|
| 18 | 4.08e-08 |
| 34 | 1.07e-07 |

Table 12: MNIST: slope of $h_{\mathrm{norm}}$ vs. radius by hidden width.

| width | slope h per r |
|---|---|
| 128 | 1.25e-05 |
| 256 | 7.17e-06 |
| 512 | 9.92e-06 |
| 1024 | 6.32e-06 |

### S.16 CIFAR-10 Results

On CIFAR-10, both `layer1` and `layer2` exhibit positive holonomy at $r = 0.10$, with magnitudes similar to CIFAR-10 (Table 13). This indicates that holonomy generalizes across dataset complexity.

Table 13: CIFAR-10: $h_{\mathrm{norm}}$ at $r = 0.10$ by layer.

| layer | mean | std |
|---|---|---|
| layer1 | 6.01e-07 | 7.92e-09 |
| layer2 | 4.45e-07 | 4.73e-08 |

### S.17 Alternative Pathwise Metrics

We compared holonomy against other proposed pathwise statistics. Scatter plots versus Lipschitz constants and path curvature (Figure 9) show that while correlated, these alternatives do not subsume holonomy, supporting its distinctiveness as a geometric descriptor.

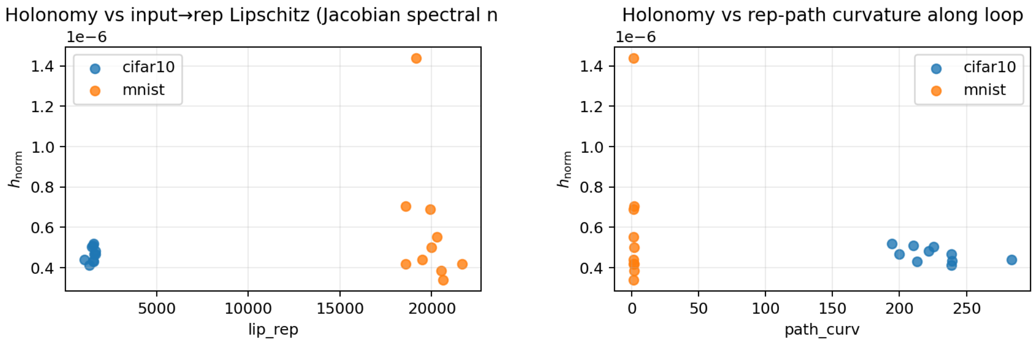


Figure 9: **Comparison to other pathwise statistics.** Holonomy vs. representation Lipschitz (left) and path curvature (right).

## B Additional diagnostic experiments

In this section we report three additional diagnostic experiments that probe (i) behaviour in simple "ground truth" settings, (ii) sensitivity to loop discretisation, and (iii) the small-radius / self-loop regime of the estimator.

### B.1 Experiment A: Toy equivariance vs. aliasing on MNIST

We construct a simple convolutional toy setting on MNIST with two small networks that are intentionally different from a geometric point of view:

1. **EquivConvNetMNIST**: a CNN with only stride-1 convolutions, circular padding, and no pooling, which is approximately translation-equivariant on the grid;

2. **AliasConvNetMNIST**: a CNN with zero padding and max-pooling, which introduces strong aliasing and boundary artefacts.

Both models are trained on MNIST with the same optimisation hyperparameters. We then measure representation holonomy at the second convolutional layer along a short loop of integer translations of a single test image, using the discrete path

$$(0,0) \to (0,1) \to (1,1) \to (1,0) \to (0,0)$$

in pixel space (implemented via circular shifts of the image).

Table 14 reports the normalized holonomy $h_{\mathrm{norm}}$, the Frobenius norm $\|H - I\|_{\mathrm{Fro}}$ of the holonomy operator, and the maximum eigen-angle (in radians) of $H$.

Table 14: **Experiment A (MNIST, toy equivariance vs. aliasing).** Holonomy at the second convolutional layer along a discrete translation loop. The nearly equivariant network exhibits holonomy close to zero, whereas the aliased network shows substantially larger holonomy.

| **Model** | $\boldsymbol{h_{\mathrm{norm}}}$ | $\boldsymbol{\|H - I\|_{\mathrm{Fro}}}$ | **max eigen-angle** |
|---|---|---|---|
| EquivConvNetMNIST | $8.41 \times 10^{-7}$ | $1.3 \times 10^{-5}$ | $6 \times 10^{-6}$ |
| AliasConvNetMNIST | $2.99 \times 10^{-4}$ | $4.78 \times 10^{-3}$ | $3 \times 10^{-3}$ |

Under the same translation loop, the aliased network exhibits roughly three orders of magnitude larger holonomy than the approximately equivariant one, providing a clean “ground truth” sanity check: when translation symmetry is respected, holonomy is essentially zero, and when it is broken by padding and pooling artefacts, holonomy is large.

### B.2 Experiment C: Sensitivity to loop discretisation

To probe sensitivity to the number of loop points, we fix a trained MNIST MLP and measure holonomy at Hidden 1 for a fixed radius $r$ while varying only the loop discretisation. Loops are constructed as regular polygons with $n_{\text{points}} \in \{6, 8, 12, 16, 24\}$ in a local two-dimensional PCA plane around each base point; all other estimator settings (neighbourhood size, subspace dimension, whitening pool, etc.) are kept fixed.

Table 15 reports the resulting $h_{\mathrm{norm}}$ values.

Table 15: **Experiment C (MNIST MLP, loop discretisation).** Holonomy at Hidden 1 for different numbers of loop samples $n_{\text{points}}$ at fixed radius. Values vary smoothly with $n_{\text{points}}$, with no indication of instability.

| $\boldsymbol{n_{\text{points}}}$ | $\boldsymbol{h_{\mathrm{norm}}}$ |
|---|---|
| 6 | $3.51 \times 10^{-7}$ |
| 8 | $3.87 \times 10^{-7}$ |
| 12 | $3.95 \times 10^{-7}$ |
| 16 | $4.15 \times 10^{-7}$ |
| 24 | $4.70 \times 10^{-7}$ |

Holonomy varies smoothly and monotonically with $n_{\text{points}}$ (see Figure 10). Increasing the number of loop points by a factor of four changes $h_{\mathrm{norm}}$ by only $\sim 1.2 \times 10^{-7}$ (about a 30% relative change, but of the same order of magnitude), suggesting that the estimator is numerically stable with respect to reasonable changes in loop discretisation. In the main text we therefore adopt $n_{\text{points}} = 12$ as a compute–accuracy compromise.

### B.3 Experiment D: Small-radius and self-loop regime

Finally, we study the very small-radius regime and the numerical floor of the estimator. On a (separately) trained MNIST MLP at Hidden 1 we construct:

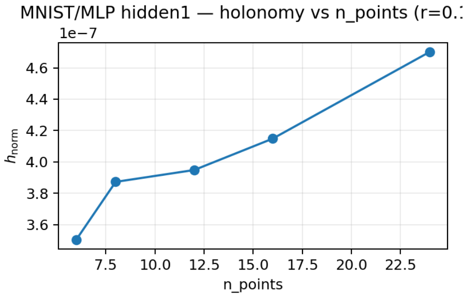


Figure 10: **MNIST/MLP Hidden 1: holonomy vs. loop discretisation.** Normalised holonomy $h_{\text{norm}}$ at radius $r = 0.10$ as a function of the number of loop samples $n_{\text{points}} \in \{6, 8, 12, 16, 24\}$. The curve is smooth and monotone, with all values in the range 3.5–4.7 $\times 10^{-7}$, indicating that the estimator is stable with respect to loop discretisation.

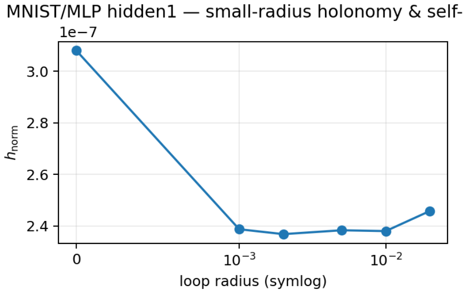


Figure 11: **MNIST/MLP Hidden 1: small-radius holonomy and self-loop.** Normalised holonomy $h_{\text{norm}}$ for an exact self-loop ($r = 0$) and PCA loops with very small radii $r \in \{10^{-3}, 2{\times}10^{-3}, 5{\times}10^{-3}, 10^{-2}, 2{\times}10^{-2}\}$. All non-zero radii lie in the narrow band $h_{\text{norm}} \in [2.37, 2.46] \times 10^{-7}$, i.e. variation $\approx 3 \times 10^{-8}$, which characterises the numerical floor of the estimator and is consistent with the $O(r)$ small-radius behaviour.

- a *self-loop* in which the same image is repeated at all loop points (radius $r = 0$); and
- PCA-based loops with radii $r \in \{10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 10^{-2}, 2 \times 10^{-2}\}$.

Table 16 shows $h_{\text{norm}}$ as a function of radius.

Table 16: **Experiment D (MNIST MLP, small-radius and self-loop).** Holonomy at Hidden 1 for an exact self-loop ($r = 0$) and for very small PCA loops. All values lie in a narrow band, characterising the numerical floor of the estimator in this setting.

| **Radius** | $\boldsymbol{h_{\text{norm}}}$ |
|---|---|
| 0 | $3.08 \times 10^{-7}$ |
| 0.001 | $2.39 \times 10^{-7}$ |
| 0.002 | $2.37 \times 10^{-7}$ |
| 0.005 | $2.38 \times 10^{-7}$ |
| 0.010 | $2.38 \times 10^{-7}$ |
| 0.020 | $2.45 \times 10^{-7}$ |

For $r \leq 0.02$, holonomy remains essentially flat in the narrow band $h_{\text{norm}} \in [2.37, 2.46] \times 10^{-7}$ (variation $\approx 3 \times 10^{-8}$). Together with the self-loop value at $r = 0$, these numbers characterise the numerical floor of our estimator and are consistent with the $O(r)$ small-radius behaviour established in Theorem 1: the increases with radius reported in the main figures only become visible once we leave this floor and move to radii where perturbations have a semantic effect, see Figure 11.