Method 2: Parallel Transport on Hidden States

Gauranga

September 17, 2025

Preliminary

Consider a matrix X_ℓ which is a snapshot of all token representations in a batch at a specific layer ℓ. It is a numerical table where each row is a vector representing a single token (like a word or sub-word), and each column is a feature dimension. Let us trace the entire lifecycle of this matrix, from raw text to the final layer of a Transformer, to see how it's created and how it evolves.

Everything begins with the raw text you provide to the model. *Example Prompt:* `The cat sat on the mat.`. However, the model cannot work with raw text. Its first step is to break the sentence into a sequence of discrete units it recognizes, called tokens. For example, the Output Tokens would look like: $[\text{"The"}, \text{"cat"}, \text{"sat"}, \text{"on"}, \text{"the"}, \text{"mat"}]$. At this point, these are still just labels, not numbers. There are 6 tokens in our sequence.

This is where the matrix X is born. The model uses a giant lookup table, called an embedding table, which is essentially a dictionary that maps each token in its vocabulary to a high-dimensional vector. This process is called embedding. Assume our model has a hidden dimension of $d = 4$. Each token is converted into a 4-dimensional vector:

$$\begin{aligned}
&\text{"The"} \to \mathbf{h}_{0,1}^\top = [0.12, -0.45, 0.88, 0.31],\\
&\text{"cat"} \to \mathbf{h}_{0,2}^\top = [0.91, 0.23, -0.11, 0.54],\\
&\text{"sat"} \to \mathbf{h}_{0,3}^\top = [-0.20, 0.67, 0.34, -0.81],\\
&\text{... and so on for all 6 tokens.}
\end{aligned}$$

The matrix X_0 is formed by stacking these token vectors as rows:

$$X_0 = \begin{bmatrix} 0.12 & -0.45 & 0.88 & 0.31 \\ 0.91 & 0.23 & -0.11 & 0.54 \\ -0.20 & 0.67 & 0.34 & -0.81 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

This 6×4 matrix is our initial snapshot, X_0. At this stage, the meaning of each token is isolated; the vector for "cat" doesn't know that it "sat" on a "mat". The matrix X_0 is then passed through a series of Transformer layers (or blocks). Each layer acts as a processing unit that refines the token representations.

A Transformer layer's key component is the self-attention mechanism. In simple terms, self-attention allows every token to look at every other token in the sequence and decide which ones are most important for understanding its own meaning. For example, when processing the vector for "sat", the attention mechanism might learn that "cat" and "mat" are highly relevant. It then updates the vector for "sat" by mixing information from the vectors for "cat" and "mat".

This process happens for every token in parallel. The entire matrix X_0 is fed into Layer 1, resulting in a new matrix X_1:

$$X_0 \xrightarrow{\text{Layer 1}} X_1 \xrightarrow{\text{Layer 2}} X_2 \xrightarrow{\cdots} \cdots \xrightarrow{\text{Layer } L} X_L$$

Crucially, the shape of the matrix does not change; X_1 is also a 6×4 matrix. However, the values inside the vectors have changed. The first row of X_1 contains the new, context-aware representation for "The". The second row is the updated representation for "cat," which now contains information about the other words. As the matrix moves through layers ($X_1 \to X_2 \to X_3 \cdots$), the vectors become increasingly sophisticated and contextual. By the final layer, the representation for "cat" is deeply informed by the entire sentence.

What do We do With This Matrix ? The core idea of the analysis is to measure the transformation that each layer applies to the matrix X. By studying how X_ℓ becomes $X_{\ell+1}$, we seek to understand what work Layer $\ell + 1$ is doing. The goal is to Quantify the computation inside the "black box." Is a layer making large changes to the representations or only minor tweaks? Is it changing the direction of its processing?

Introduction

We look at a single forward pass through a depth-L network (e.g., a Transformer). For each depth $\ell \in \{0, \ldots, L-1\}$ let

$$X_\ell \in \mathbb{R}^{N \times d}$$

be the row-centered matrix of token representations we collect from some batch (stack all tokens from all sequences). Row i is a token vector $h_{\ell,i}^\top \in \mathbb{R}^d$. We assume the same N tokens are tracked across neighboring layers (same batch/order, masking applied consistently). Denote the (ridge-regularized) covariance at depth ℓ

$$\Sigma_\ell \;=\; \frac{1}{N} X_\ell^\top X_\ell \;+\; \lambda I_d, \qquad \lambda > 0.$$

This will be our *intrinsic metric tensor* at layer ℓ. Denote the cross-covariance between ℓ and $\ell + 1$:

$$C_\ell \;=\; \frac{1}{N} X_\ell^\top X_{\ell+1}.$$

We will define (i) a parallel transport map that removes spurious frame rotation between neighbors, (ii) a whitening map that puts distances in intrinsic units, and then compute the triad:

$$\text{thermodynamic length } L_\ell^{(\mathrm{PT})}, \qquad \text{curvature } \kappa_\ell^{(\mathrm{PT})}, \qquad \text{belief vector } \boldsymbol{v}_\ell^{(\mathrm{PT})}(c).$$

Orthogonal Procrustes

We wish to compare $X_{\ell+1}$ to X_ℓ in the *same* coordinate frame. Let $R_\ell \in O(d)$ solve the orthogonal Procrustes problem

$$R_\ell \;=\; \arg\min_{R \in O(d)} \left\| X_{\ell+1} R^\top - X_\ell \right\|_F^2.$$

Result (proved below) If $C_\ell = U_\ell S_\ell V_\ell^\top$ is the SVD of C_ℓ, then

$$R_\ell \;=\; U_\ell V_\ell^\top.$$

This is the unique minimizer when the top singular values are simple; otherwise any minimizer can be chosen with a canonical tie-break (e.g., closest to identity in the tied subspace). We now *transport* neighbors into the frame of ℓ:

$$X_{\ell+1\to\ell} := X_{\ell+1}R_\ell^\top, \qquad X_{\ell-1\to\ell} := X_{\ell-1}R_{\ell-1}.$$

This is laid down with the help of the following figures:

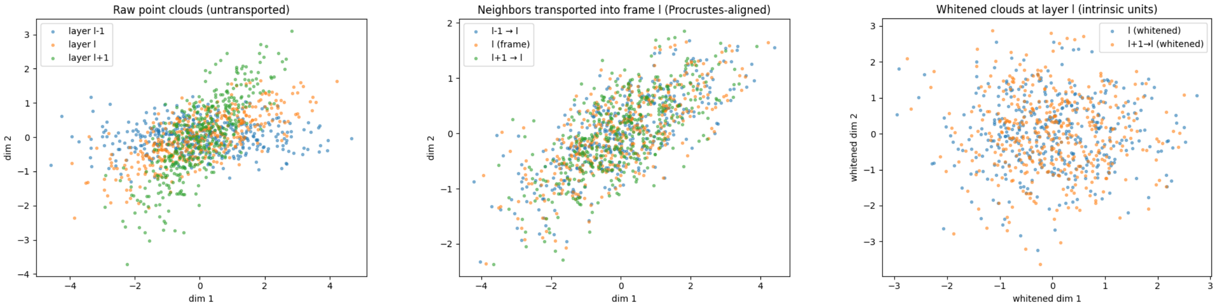

Figure 1: plot1: raw clouds, plot2: transport, plot3: whitened

Theorem 1 (Orthogonal Procrustes Problem)**.** *Let $A, B \in \mathbb{R}^{n\times n}$. The orthogonal Procrustes problem consists of finding an orthogonal matrix $\Omega \in O(n)$ that solves*

$$\min_\Omega \|\Omega A - B\|_F \quad \textit{subject to} \quad \Omega^T\Omega = I,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This is equivalent to finding the orthogonal matrix R closest to $M = BA^T$, i.e.

$$\min_R \|R - M\|_F \quad \textit{subject to} \quad R^TR = I.$$

The solution is given by

$$R = UV^T,$$

where $M = U\Sigma V^T$ is the singular value decomposition of M.

Proof. We begin with

$$R = \arg\min_\Omega \|\Omega A - B\|_F^2.$$

Expanding using the Frobenius inner product $\langle X, Y\rangle_F = \operatorname{tr}(X^TY)$, we have

$$\|\Omega A - B\|_F^2 = \|\Omega A\|_F^2 + \|B\|_F^2 - 2\langle \Omega A, B\rangle_F.$$

Since $\Omega \in O(n)$, it preserves the Frobenius norm, hence $\|\Omega A\|_F = \|A\|_F$. Thus,

$$R = \arg\max_\Omega \langle \Omega A, B\rangle_F.$$

By properties of the Frobenius inner product,

$$\langle \Omega A, B\rangle_F = \langle \Omega, BA^T\rangle_F.$$

Let $M = BA^T$. Suppose the singular value decomposition is

$$M = U\Sigma V^T.$$

Then

$$\langle \Omega, M\rangle_F = \langle \Omega, U\Sigma V^T\rangle_F = \langle U^T\Omega V, \Sigma\rangle_F.$$

Define $S = U^T\Omega V$. Since U, V are orthogonal and Ω is orthogonal, it follows that $S \in O(n)$. Therefore

$$\langle U^T\Omega V, \Sigma\rangle_F = \langle S, \Sigma\rangle_F.$$

The Frobenius inner product with Σ is maximized when $S = I$. Thus,

$$U^T RV = I \quad \Rightarrow \quad R = UV^T.$$

Hence, the optimal orthogonal approximation is

$$R = UV^T,$$

which minimizes $\|\Omega A - B\|_F$. □

Theorem 2 (Orthogonal Procrustes, rectangular)**.** *Let* $A, B \in \mathbb{R}^{N\times d}$ *and* $R \in O(d)$*. Then*

$$R^\star = \arg\min_{R\in O(d)} \|AR - B\|_F^2 \quad \Longleftrightarrow \quad R^\star = UV^\top,$$

where $B^\top A = U\Sigma V^\top$ *is the (thin) SVD.*

Finite differences in the frame of ℓ

With neighbors expressed in the same frame, we define the first and second covariant differences:

$$\Delta_\ell := X_{\ell+1\to\ell} - X_\ell, \qquad \Delta_\ell^2 := X_{\ell+1\to\ell} - 2X_\ell + X_{\ell-1\to\ell}.$$

These are the discrete analogues of the *velocity* and *acceleration* of the representation path, after removing frame spin. However, the representation cloud at ℓ can be highly anisotropic. To measure distances fairly we use the *Mahalanobis* metric induced by Σ_ℓ. Let $\Sigma_\ell^{-1/2}$ be the positive definite inverse square root (eigendecomposition or Cholesky). Define the whitened differences:

$$\widehat{\Delta}_\ell \;=\; \Delta_\ell\,\Sigma_\ell^{-1/2}, \qquad \widehat{\Delta^2}_\ell \;=\; \Delta_\ell^2\,\Sigma_\ell^{-1/2}.$$

This maps the local geometry to "unit variance" in all directions. All norms that follow are taken after whitening, so they are intrinsic and comparable across layers.

We can then define the layer-wise **Thermodynamic length** as well as the aggregate length as follows:

$$L_\ell^{(\mathrm{PT})} \;=\; \|\widehat{\Delta}_\ell\|_F \qquad \text{and} \qquad L_{0:L-1}^{(\mathrm{PT})} \;=\; \sum_{\ell=0}^{L-2} L_\ell^{(\mathrm{PT})}.$$

Statistical interpretation (Fisher/Mahalanobis). Write Δ_ℓ rowwise as token displacements $\delta_{\ell,i}^\top \in \mathbb{R}^d$. Then

$$\|\widehat{\Delta}_\ell\|_F^2 = \sum_{i=1}^{N} \delta_{\ell,i}^\top \Sigma_\ell^{-1} \delta_{\ell,i}.$$

If at depth ℓ token representations are modeled as $h \sim \mathcal{N}(\mu_\ell, \Sigma_\ell)$ and the next layer applies a small mean shift $\mu_\ell \mapsto \mu_\ell + \delta$, then the Fisher information metric for the mean parameter is $\langle u, v\rangle = u^\top \Sigma_\ell^{-1} v$. Consequently,

$$(L_\ell^{(\mathrm{PT})})^2 \;=\; \sum_{i=1}^{N} \|\delta_{\ell,i}\|^2_{\Sigma_\ell^{-1}}$$

is the *Fisher (Mahalanobis) work* of the step. This justifies the name "thermodynamic length" without invoking any extra thermodynamics formalism.

To define **Spectral Curvature**, let $s_\ell := \|\widehat{\Delta}_\ell\|_F$ be the local arc-length element (in intrinsic units). Define

$$\kappa_\ell^{(\mathrm{PT})} \;=\; \frac{\|\widehat{\Delta^2}{}_\ell\|_F}{(s_\ell^2 + \varepsilon)^{3/2}}, \qquad \varepsilon > 0.$$

This is the discrete, arc-length–normalized curvature of the representation path: it spikes when the direction of motion changes sharply between $\ell - 1 \to \ell$ and $\ell \to \ell + 1$. The normalization $(s_\ell^2)^{3/2}$ is the natural analogue of the continuous Frenet formula $\kappa = \|d\mathbf{T}/ds\|$, making κ dimensionless and scale-stable. If the path is locally straight (colinear steps), $\Delta_\ell^2 \approx 0$ and $\kappa_\ell^{(\mathrm{PT})}$ is small even when $L_\ell^{(\mathrm{PT})}$ is large.

Now, we define, **Belief vectors**. For a sample x, let $g_\ell(x) \in \mathbb{R}^d$ be a pooled gradient of $\log p(y \mid x)$ with respect to X_ℓ (CLS token or mean over tokens). Define the whitened, condition-averaged belief vector

$$\boldsymbol{v}_\ell^{(\mathrm{PT})}(c) \;=\; \mathbb{E}_{x \sim P(c)}\big[\,\Sigma_\ell^{-1/2}\, g_\ell(x)\,\big] \in \mathbb{R}^d.$$

The *direction* of $\boldsymbol{v}_\ell^{(\mathrm{PT})}$ is the intrinsic axis along which the objective most wants the representation cloud to move at depth ℓ. The *norm* $\|\boldsymbol{v}_\ell^{(\mathrm{PT})}(c)\|$ is the *steering pressure* for condition c.

Relation to natural gradient. In the metric $G_\ell = \Sigma_\ell^{-1}$, the natural gradient is $G_\ell^{-1}\nabla = \Sigma_\ell \nabla$. Here we use $\Sigma_\ell^{-1/2}$ to express the gradient in an *orthonormal basis of the intrinsic metric*, which is the appropriate way to compare magnitudes across layers. Thus $\boldsymbol{v}_\ell^{(\mathrm{PT})}$ can be read as a natural-gradient-consistent direction, expressed in intrinsic coordinates. Collecting the definitions, we have:

$$\boxed{\begin{aligned}
&\text{(i) Length:} && L_\ell^{(\mathrm{PT})} = \big\|\,(X_{\ell+1}R_\ell^\top - X_\ell)\,\Sigma_\ell^{-1/2}\big\|_F.\\
&\text{(ii) Curvature:} && \kappa_\ell^{(\mathrm{PT})} = \frac{\big\|\,(X_{\ell+1}R_\ell^\top - 2X_\ell + X_{\ell-1}R_{\ell-1})\,\Sigma_\ell^{-1/2}\big\|_F}{\Big(\big\|\,(X_{\ell+1}R_\ell^\top - X_\ell)\,\Sigma_\ell^{-1/2}\big\|_F^2 + \varepsilon\Big)^{3/2}}.\\
&\text{(iii) Belief:} && \boldsymbol{v}_\ell^{(\mathrm{PT})}(c) = \mathbb{E}_{x \sim P(c)}\big[\,\Sigma_\ell^{-1/2}\, g_\ell(x)\,\big].
\end{aligned}}$$

A composite *per-layer activity score* sometimes used is

$$F_\ell(c) \;=\; \kappa_\ell^{(\mathrm{PT})}\, L_\ell^{(\mathrm{PT})}\, \|\boldsymbol{v}_\ell^{(\mathrm{PT})}(c)\|,$$

and the nDNA aggregate $\sum_\ell \omega_\ell F_\ell(c)$ with nonnegative ω_ℓ (e.g., uniform or learned).

Gauge invariance

Imagine you have a trained neural network. One can go into the network and, between any two layers, insert a mathematical "spin"—a rotation matrix Q and its inverse $Q^\top$.

$$\cdots \xrightarrow{\text{Layer } \ell-1} X_{\ell-1} \xrightarrow{\text{Layer } \ell} X_\ell \xrightarrow{\text{Layer } \ell+1} X_{\ell+1} \to \cdots$$

Now if I insert a "spin" at layer ℓ:

$$\cdots \xrightarrow{\text{Layer } \ell-1} X_{\ell-1} \xrightarrow{\text{Layer } \ell} (X_\ell Q_\ell) \xrightarrow{Q_\ell^\top, \text{Layer } \ell+1} X_{\ell+1} \to \cdots$$

From the outside, the network's final output is exactly the same. The function it computes has not changed at all because the rotation Q_ℓ is immediately cancelled out by its inverse $Q_\ell^\top$. However, on the inside, the hidden representation matrix has changed: we now have $X'_\ell = X_\ell Q_\ell$. This is the problem. If our measurement tools are sensitive to this arbitrary spin, our results will be meaningless. We might measure a huge "change" between layers that is just a meaningless coordinate rotation, not a real change in the information being processed.

Gauge invariance is the property that our measurements are not fooled by this trick. A "gauge" is just a choice of coordinate system, and our metrics must be independent of or invariant to this choice. We provide a mathematical proof that they are. This models arbitrary changes of basis introduced by, e.g., LayerNorm-centered linear maps or any hidden orthogonal rotations. We prove that if we rotate the representations at every layer by an arbitrary orthogonal matrix Q_ℓ, the final length, curvature, and belief norm calculations remain exactly the same.

Theorem 3 (Gauge invariance)**.** *Let $X_\ell \in \mathbb{R}^{N\times d}$ be the row-centered matrix of token representations at layer ℓ. A gauge transform (spin) is a per-layer orthogonal change of coordinates*

$$X'_\ell = X_\ell Q_\ell, \qquad Q_\ell \in O(d).$$

Under this transformation,

$$L'_\ell = L_\ell, \quad \kappa'_\ell = \kappa_\ell, \quad \|\mathbf{v}'_\ell\|_2 = \|\mathbf{v}_\ell\|_2.$$

Proof. Note that Method 2 forms, at each ℓ, cross–covariance,

$$C_\ell := \frac{1}{N} X_\ell^\top X_{\ell+1},$$

the transport,

$$R_\ell \in O(d)$$

transported neighbors,

$$X_{\ell+1\to\ell} := X_{\ell+1} R_\ell^\top, \quad X_{\ell-1\to\ell} := X_{\ell-1} R_{\ell-1},$$

covariant differences,

$$\Delta_\ell := X_{\ell+1\to\ell} - X_\ell, \quad \Delta_\ell^2 := X_{\ell+1\to\ell} - 2X_\ell + X_{\ell-1\to\ell},$$

intrinsic metric

$$\Sigma_\ell := \frac{1}{N} X_\ell^\top X_\ell + \lambda I_d \quad (\lambda > 0),$$

whitening

$$\widehat{\Delta}_\ell := \Delta_\ell \, \Sigma_\ell^{-1/2}, \quad \widehat{\Delta^2}_\ell := \Delta_\ell^2 \, \Sigma_\ell^{-1/2},$$

Thermodynamic length

$$L_\ell := \|\widehat{\Delta}_\ell\|_F,$$

Spectral curvature

$$\kappa_\ell := \frac{\|\widehat{\Delta^2}_\ell\|_F}{(\|\widehat{\Delta}_\ell\|_F^2 + \varepsilon)^{3/2}} \quad (\varepsilon > 0),$$

belief vector

$$\mathbf{v}_\ell := \mathbb{E}_{x \sim P(c)} \big[\Sigma_\ell^{-1/2} g_\ell(x) \big],$$

where $g_\ell(x)$ is the pooled gradient $\nabla_{X_\ell} \log p(y \mid x)$ mapped to $\mathbb{R}^d$. Firstly, the proofs rely on three algebraic facts:

F1 (orthogonal polar equivariance). For any A and orthogonal Q, P,

$$\operatorname{polar}(Q^\top A P) = Q^\top \operatorname{polar}(A) \, P,$$

where $\operatorname{polar}(A) := A(A^\top A)^{-1/2} \in O(d)$ is the orthogonal polar factor (equal to the Procrustes solution $UV^\top$ where $A = USV^\top$ is an SVD).

F2 (SPD square-root similarity). For any SPD, Σ and orthogonal Q,

$$(Q^\top \Sigma Q)^{\pm 1/2} = Q^\top \, \Sigma^{\pm 1/2} \, Q,$$

where the principal (SPD) square roots are used.

F3 (Frobenius norm invariance). For any A and orthogonal Q,

$$\|AQ\|_F = \|A\|_F, \qquad \|QA\|_F = \|A\|_F.$$

All three are standard and will be used inline. Now the **(S1) Cross–covariance**, C'_ℓ is given by,

$$C'_\ell = \frac{1}{N} (X'_\ell)^\top X'_{\ell+1} = \frac{1}{N} (Q_\ell^\top X_\ell^\top)(X_{\ell+1} Q_{\ell+1}) = Q_\ell^\top C_\ell Q_{\ell+1}.$$

Since $C'_\ell = Q_\ell^\top C_\ell Q_{\ell+1}$,

$$R'_\ell = \operatorname{polar}(C'_\ell) = Q_\ell^\top R_\ell Q_{\ell+1}.$$

Therefore

$$X'_{\ell+1 \to \ell} = X'_{\ell+1} (R'_\ell)^\top = (X_{\ell+1} Q_{\ell+1})(Q_{\ell+1}^\top R_\ell^\top Q_\ell) = (X_{\ell+1} R_\ell^\top) Q_\ell = X_{\ell+1 \to \ell} \, Q_\ell,$$

and analogously $X'_{\ell-1 \to \ell} = X_{\ell-1 \to \ell} Q_\ell$.
The **(S4) Covariant differences and the (S5) Intrinsic covariance metric** would then be

$$\begin{aligned}
\Delta'_\ell &= X'_{\ell+1 \to \ell} - X'_\ell = (X_{\ell+1 \to \ell} Q_\ell) - (X_\ell Q_\ell) = (X_{\ell+1 \to \ell} - X_\ell) Q_\ell = \Delta_\ell Q_\ell, \\
\Delta_\ell^{2\prime} &= X'_{\ell+1 \to \ell} - 2X'_\ell + X'_{\ell-1 \to \ell} = (X_{\ell+1 \to \ell} - 2X_\ell + X_{\ell-1 \to \ell}) Q_\ell = \Delta_\ell^2 Q_\ell. \\
\Sigma'_\ell &= \frac{1}{N} (X'_\ell)^\top X'_\ell + \lambda I = Q_\ell^\top \left(\frac{1}{N} X_\ell^\top X_\ell + \lambda I \right) Q_\ell = Q_\ell^\top \Sigma_\ell Q_\ell.
\end{aligned}$$

By F2,

$$\Sigma_\ell'^{-1/2} = Q_\ell^\top \Sigma_\ell^{-1/2} Q_\ell.$$

The **(S6) Whitened differences** are given by,

$$\widehat{\Delta}_\ell' = \Delta_\ell' \Sigma_\ell'^{-1/2} = (\Delta_\ell Q_\ell)(Q_\ell^\top \Sigma_\ell^{-1/2} Q_\ell) = \Delta_\ell \Sigma_\ell^{-1/2} Q_\ell = \widehat{\Delta}_\ell Q_\ell,$$

and the same for the second difference:

$$\widehat{\Delta^2}'_\ell = \widehat{\Delta^2}_\ell Q_\ell.$$

To find out **Thermodynamic Length**, by F3 and (S6) we have,

$$L_\ell' = \|\widehat{\Delta}_\ell'\|_F = \|\widehat{\Delta}_\ell Q_\ell\|_F = \|\widehat{\Delta}_\ell\|_F = L_\ell.$$

For **Spetral Curvature**, we show that both its numerator and the arc-element in the denominator are invariant:

$$\|\widehat{\Delta^2}'_\ell\|_F = \|\widehat{\Delta^2}_\ell Q_\ell\|_F = \|\widehat{\Delta^2}_\ell\|_F, \quad \|\widehat{\Delta}_\ell'\|_F = \|\widehat{\Delta}_\ell\|_F,$$

hence

$$\kappa_\ell' = \frac{\|\widehat{\Delta^2}'_\ell\|_F}{\left(\|\widehat{\Delta}_\ell'\|_F^2 + \varepsilon\right)^{3/2}} = \frac{\|\widehat{\Delta^2}_\ell\|_F}{\left(\|\widehat{\Delta}_\ell\|_F^2 + \varepsilon\right)^{3/2}} = \kappa_\ell.$$

Thus L_ℓ and κ_ℓ are gauge-invariant.

Let $g_\ell(x) \in \mathbb{R}^d$ denote the pooled gradient w.r.t. the layer-ℓ representation in the original coordinates. Under the gauge $X_\ell' = X_\ell Q_\ell$, the chain rule gives the transformed pooled gradient

$$g_\ell'(x) = Q_\ell^\top g_\ell(x).$$

(Note that for any small perturbation δX_ℓ, we have

$$\langle g_\ell, \delta X_\ell \rangle = \langle g_\ell', \delta X_\ell' \rangle = \langle g_\ell', \delta X_\ell Q_\ell \rangle = \langle Q_\ell g_\ell', \delta X_\ell \rangle,$$

so $Q_\ell g_\ell' = g_\ell \Rightarrow g_\ell' = Q_\ell^\top g_\ell$.) Whiten and average:

$$\begin{aligned} \mathbf{v}_\ell' &= \mathbb{E}[\Sigma_\ell'^{-1/2} g_\ell'(x)] = \mathbb{E}[(Q_\ell^\top \Sigma_\ell^{-1/2} Q_\ell)(Q_\ell^\top g_\ell(x))] \\ &= Q_\ell^\top \mathbb{E}[\Sigma_\ell^{-1/2} g_\ell(x)] = Q_\ell^\top \mathbf{v}_\ell. \end{aligned}$$

Since Q_ℓ is orthogonal,

$$\|\mathbf{v}_\ell'\|_2 = \|\mathbf{v}_\ell\|_2.$$

Therefore, the belief-norm is gauge-invariant.

□

Sanity checks

Claim (A1) If $X_{\ell+1} = X_\ell$ and $X_{\ell-1} = X_\ell$, then

$$L_\ell = \kappa_\ell = 0.$$

Proof.

$$C_\ell = \frac{1}{N} X_\ell^\top X_\ell,$$

so

$$R_\ell = I$$

(polar of SPD is I). Then

$$X_{\ell+1\to\ell} = X_\ell, \quad X_{\ell-1\to\ell} = X_\ell,$$

hence

$$\Delta_\ell = 0, \quad \Delta_\ell^2 = 0,$$

implying

$$\widehat{\Delta}_\ell = \widehat{\Delta^2}_\ell = 0.$$

Therefore

$$L_\ell = 0, \quad \kappa_\ell = 0.$$

■

Claim $X_{\ell+1} = X_\ell Q, \quad X_{\ell-1} = X_\ell Q^\top$ with $Q \in O(d)$. Then,

$$L_\ell = \kappa_\ell = 0.$$

Proof.

$$C_\ell = \frac{1}{N} X_\ell^\top X_\ell Q \Rightarrow R_\ell = Q,$$

so

$$X_{\ell+1\to\ell} = X_\ell.$$

Similarly, $R_{\ell-1} = Q^\top \Rightarrow X_{\ell-1\to\ell} = X_\ell$. As in A1,

$$\Delta_\ell = \Delta_\ell^2 = 0.$$

■

Claim. If

$$X_{\ell+1\to\ell} - X_\ell = X_\ell - X_{\ell-1\to\ell}$$

(equal, colinear steps in the ℓ-frame), then

$$\kappa_\ell = 0.$$

Proof. Under the hypothesis,

$$\Delta_\ell^2 = (X_{\ell+1\to\ell} - X_\ell) - (X_\ell - X_{\ell-1\to\ell}) = 0,$$

hence

$$\widehat{\Delta^2}_\ell = 0,$$

and so

$$\kappa_\ell = 0.$$

■

Note: if steps are colinear but unequal (non-uniform "speed"), $\Delta_\ell^2 \neq 0$. In the small-step limit (uniform arclength sampling), $\kappa_\ell \to 0$, matching continuous Frenet curvature.

Claim (B1) For any $Q_\ell \in O(d)$ with $X'_\ell = X_\ell Q_\ell$, one has

$$L'_\ell = L_\ell, \quad \kappa'_\ell = \kappa_\ell, \quad \|\mathbf{v}'_\ell\|_2 = \|\mathbf{v}_\ell\|_2.$$

Proof. Cross-covariance transforms by

$$C'_\ell = \tfrac{1}{N}(X'_\ell)^\top X'_{\ell+1} = \tfrac{1}{N}(Q_\ell^\top X_\ell^\top)(X_{\ell+1}Q_{\ell+1}) = Q_\ell^\top C_\ell Q_{\ell+1}.$$

By (F1),

$$R'_\ell = \operatorname{polar}(C'_\ell) = Q_\ell^\top \operatorname{polar}(C_\ell)\, Q_{\ell+1} = Q_\ell^\top R_\ell\, Q_{\ell+1}.$$

Hence the transported neighbors satisfy

$$\begin{aligned} X'_{\ell+1\to\ell} &= X'_{\ell+1}(R'_\ell)^\top = (X_{\ell+1}Q_{\ell+1})(Q_{\ell+1}^\top R_\ell^\top Q_\ell) = (X_{\ell+1}R_\ell^\top)Q_\ell = X_{\ell+1\to\ell}\, Q_\ell,\\ X'_{\ell-1\to\ell} &= X_{\ell-1\to\ell}\, Q_\ell. \end{aligned}$$

Therefore

$$\Delta'_\ell = X'_{\ell+1\to\ell} - X'_\ell = (X_{\ell+1\to\ell} - X_\ell)Q_\ell = \Delta_\ell Q_\ell, \quad (\Delta^2)'_\ell = (\Delta^2)_\ell Q_\ell.$$

The intrinsic covariance transforms as

$$\Sigma'_\ell = \tfrac{1}{N}(X'_\ell)^\top X'_\ell + \lambda I = Q_\ell^\top \Sigma_\ell Q_\ell \quad \Rightarrow \quad \Sigma_\ell'^{-1/2} = Q_\ell^\top \Sigma_\ell^{-1/2} Q_\ell \quad \text{by (F2)}.$$

Thus the whitened differences obey

$$\widehat{\Delta}'_\ell = \Delta'_\ell \Sigma_\ell'^{-1/2} = (\Delta_\ell Q_\ell)(Q_\ell^\top \Sigma_\ell^{-1/2} Q_\ell) = \Delta_\ell \Sigma_\ell^{-1/2} Q_\ell = \widehat{\Delta}_\ell\, Q_\ell,$$

and similarly $\widehat{\Delta^2}'_\ell = \widehat{\Delta^2}_\ell\, Q_\ell$.
By (F3),

$$L'_\ell = \|\widehat{\Delta}'_\ell\|_F = \|\widehat{\Delta}_\ell Q_\ell\|_F = \|\widehat{\Delta}_\ell\|_F = L_\ell,$$

and likewise $\|\widehat{\Delta^2}'_\ell\|_F = \|\widehat{\Delta^2}_\ell\|_F$. Since also $\|\widehat{\Delta}'_\ell\|_F = \|\widehat{\Delta}_\ell\|_F$, the curvature denominator is unchanged, so

$$\kappa'_\ell = \frac{\|\widehat{\Delta^2}'_\ell\|_F}{(\|\widehat{\Delta}'_\ell\|_F^2 + \varepsilon)^{3/2}} = \frac{\|\widehat{\Delta^2}_\ell\|_F}{(\|\widehat{\Delta}_\ell\|_F^2 + \varepsilon)^{3/2}} = \kappa_\ell.$$

Under $X'_\ell = X_\ell Q_\ell$, the pooled gradient transforms as $g'_\ell(x) = Q_\ell^\top g_\ell(x)$ (chain rule). Hence

$$\mathbf{v}'_\ell = \mathbb{E}[\Sigma_\ell'^{-1/2} g'_\ell(x)] = \mathbb{E}[(Q_\ell^\top \Sigma_\ell^{-1/2} Q_\ell)(Q_\ell^\top g_\ell(x))] = Q_\ell^\top \mathbb{E}[\Sigma_\ell^{-1/2} g_\ell(x)] = Q_\ell^\top \mathbf{v}_\ell,$$

so $\|\mathbf{v}'_\ell\|_2 = \|\mathbf{v}_\ell\|_2$ by orthogonality of Q_ℓ. Combining the three parts yields the claim. □

Claim (B2) Let the transformation be a uniform scaling $X'_\ell = \alpha X_\ell$ for a scalar $\alpha \neq 0$, applied consistently across all layers. If and only if the ridge regularizer $\lambda = 0$, the thermodynamic length and curvature are invariant:

$$L'_\ell = L_\ell, \quad \text{and} \quad \kappa'_\ell = \kappa_\ell.$$

Proof. We analyze how each component of the calculation transforms under the scaling $X'_k = \alpha X_k$ for $k \in \{\ell-1, \ell, \ell+1\}$, under the condition that $\lambda = 0$.
Covariance and Cross-Covariance. The transformed cross-covariance C'_ℓ is:

$$C'_\ell = \frac{1}{N}(X'_\ell)^\top X'_{\ell+1} = \frac{1}{N}(\alpha X_\ell)^\top(\alpha X_{\ell+1}) = \alpha^2 \left(\frac{1}{N} X_\ell^\top X_{\ell+1}\right) = \alpha^2 C_\ell.$$

With $\lambda = 0$, the transformed intrinsic metric Σ'_ℓ is:

$$\Sigma'_\ell = \frac{1}{N}(X'_\ell)^\top X'_\ell = \frac{1}{N}(\alpha X_\ell)^\top(\alpha X_\ell) = \alpha^2 \left(\frac{1}{N} X_\ell^\top X_\ell\right) = \alpha^2 \Sigma_\ell.$$

Procrustes Rotation. The Procrustes rotation R_ℓ is the orthogonal polar factor of C_ℓ. Since α^2 is a positive scalar, it does not affect the rotational part of the polar decomposition.

$$R'_\ell = \mathrm{polar}(C'_\ell) = \mathrm{polar}(\alpha^2 C_\ell) = \mathrm{polar}(C_\ell) = R_\ell.$$

Similarly, $R'_{\ell-1} = R_{\ell-1}$. The rotation matrices are invariant to scaling. Transported Differences. The transported neighbors scale by α:

$$X'_{\ell+1\to\ell} = X'_{\ell+1}(R'_\ell)^\top = (\alpha X_{\ell+1})R_\ell^\top = \alpha(X_{\ell+1}R_\ell^\top) = \alpha X_{\ell+1\to\ell}.$$

Consequently, the covariant differences also scale by α:

$$\begin{aligned}\Delta'_\ell &= X'_{\ell+1\to\ell} - X'_\ell = \alpha X_{\ell+1\to\ell} - \alpha X_\ell = \alpha(X_{\ell+1\to\ell} - X_\ell) = \alpha\Delta_\ell,\\ (\Delta^2)'_\ell &= X'_{\ell+1\to\ell} - 2X'_\ell + X'_{\ell-1\to\ell} = \alpha(X_{\ell+1\to\ell} - 2X_\ell + X_{\ell-1\to\ell}) = \alpha(\Delta^2)_\ell.\end{aligned}$$

Whitening Transformation. The whitening matrix transforms by α^{-1}. Since $\Sigma'_\ell = \alpha^2\Sigma_\ell$:

$$(\Sigma'_\ell)^{-1/2} = (\alpha^2\Sigma_\ell)^{-1/2} = (\alpha^2)^{-1/2}\Sigma_\ell^{-1/2} = |\alpha|^{-1}\Sigma_\ell^{-1/2}.$$

Whitened Differences. The scaling factors cancel out perfectly, leaving the whitened differences invariant:

$$\begin{aligned}\widehat{\Delta}'_\ell &= \Delta'_\ell(\Sigma'_\ell)^{-1/2} = (\alpha\Delta_\ell)(|\alpha|^{-1}\Sigma_\ell^{-1/2}) = \frac{\alpha}{|\alpha|}(\Delta_\ell\Sigma_\ell^{-1/2}) = \mathrm{sgn}(\alpha)\widehat{\Delta}_\ell,\\ (\widehat{\Delta^2})'_\ell &= (\Delta^2)'_\ell(\Sigma'_\ell)^{-1/2} = (\alpha(\Delta^2)_\ell)(|\alpha|^{-1}\Sigma_\ell^{-1/2}) = \mathrm{sgn}(\alpha)(\widehat{\Delta^2})_\ell.\end{aligned}$$

Invariance of Metrics. The Frobenius norm is insensitive to the sign factor, so the length is invariant:

$$L'_\ell = \|\widehat{\Delta}'_\ell\|_F = \|\mathrm{sgn}(\alpha)\widehat{\Delta}_\ell\|_F = |\mathrm{sgn}(\alpha)|\|\widehat{\Delta}_\ell\|_F = L_\ell.$$

Both the numerator and the denominator of the curvature calculation are likewise invariant:

$$\|(\widehat{\Delta^2})'_\ell\|_F = \|(\widehat{\Delta^2})_\ell\|_F \quad\text{and}\quad (\|\widehat{\Delta}'_\ell\|_F^2 + \varepsilon)^{3/2} = (\|\widehat{\Delta}_\ell\|_F^2 + \varepsilon)^{3/2}.$$

Therefore, $\kappa'_\ell = \kappa_\ell$. This completes the proof for the case $\lambda = 0$.

□

Remark on the case $\lambda > 0$. The proof of invariance fails if the ridge regularizer $\lambda > 0$. In this case, the covariance matrix does not scale cleanly:

$$\Sigma'_\ell = \frac{1}{N}(\alpha X_\ell)^\top(\alpha X_\ell) + \lambda I = \alpha^2\left(\frac{1}{N}X_\ell^\top X_\ell\right) + \lambda I = \alpha^2(\Sigma_\ell - \lambda I) + \lambda I.$$

Since $\Sigma'_\ell \neq \alpha^2\Sigma_\ell$, the whitening matrix $(\Sigma'_\ell)^{-1/2}$ is not proportional to $\Sigma_\ell^{-1/2}$, and the cancellation in Step 5 does not occur. Therefore, the metrics are not invariant to scaling when regularization is used.

Claim. (B3) If a common permutation P (on rows) is applied to $X_{\ell-1}, X_\ell, X_{\ell+1}$, then $L_\ell, \kappa_\ell, \|\mathbf{v}_\ell\|$ are unchanged.

Proof. $C_\ell = \frac{1}{N}X_\ell^\top X_{\ell+1}$ is invariant to a common row permutation; likewise Σ_ℓ. Therefore R_ℓ, transported matrices, Δ-terms, whitening, and thus L_ℓ, κ_ℓ are the same. For belief, pooling over rows is permutation-invariant, hence $\|\mathbf{v}_\ell\|$ unchanged. ■

Claim. (B4) Adding a constant row vector $m^\top$ to all rows of any X_k does not change $L_\ell, \kappa_\ell, \|\mathbf{v}_\ell\|$ provided you row-center before forming C_k, Σ_k.
Proof. Row-centering subtracts the mean; any common offset cancels exactly before computing C_k, Σ_k. ■
Claim.(C1)

$$\|\widehat{\Delta}_\ell\|_F^2 = \sum_{i=1}^{N} \delta_{\ell,i}^\top \Sigma_\ell^{-1} \delta_{\ell,i},$$

where $\delta_{\ell,i}$ is the i-th row of Δ_ℓ.
Proof.

$$\|\Delta_\ell \Sigma_\ell^{-1/2}\|_F^2 = \operatorname{tr}(\Sigma_\ell^{-1/2} \Delta_\ell^\top \Delta_\ell \Sigma_\ell^{-1/2}) = \sum_i \delta_{\ell,i}^\top \Sigma_\ell^{-1} \delta_{\ell,i}.$$

■
Claim. (C2) $R_\ell = \operatorname{polar}(C_\ell)$ minimizes $\|X_{\ell+1} R^\top - X_\ell\|_F$ over $R \in O(d)$.
Proof.

$$\|X_{\ell+1} R^\top - X_\ell\|_F^2 = \|X_{\ell+1}\|_F^2 + \|X_\ell\|_F^2 - 2\operatorname{tr}(R X_{\ell+1}^\top X_\ell).$$

Maximizing $\operatorname{tr}(RA)$ over $R \in O(d)$ with $A := X_{\ell+1}^\top X_\ell = VSU^\top$ is achieved by $R^\star = UV^\top$ (von Neumann trace inequality); this equals $\operatorname{polar}(A^\top) = \operatorname{polar}(C_\ell)$. ■
Claim. (D1)

$$X_{\ell+1} = X_\ell + E_+, \quad X_{\ell-1} = X_\ell + E_-,$$

with $E_\pm$ i.i.d., rows $\sim \mathcal{N}(0, \sigma^2 I_d)$, independent of X_ℓ. For large N, $R_\ell \approx I$.

$$\mathbb{E}[L_\ell^2] = N\sigma^2 \operatorname{tr}(\Sigma_\ell^{-1}), \quad \mathbb{E}[\kappa_\ell] = \mathcal{O}(N^{-1}).$$

Proof.

$$\Delta_\ell \approx E_+ \implies \widehat{\Delta}_\ell \approx E_+ \Sigma_\ell^{-1/2}.$$

Then,

$$\mathbb{E}[L_\ell^2] = \mathbb{E}[\|E_+ \Sigma_\ell^{-1/2}\|_F^2] = \sum_{i=1}^{N} \mathbb{E}[e_i^\top \Sigma_\ell^{-1} e_i] = N\sigma^2 \operatorname{tr}(\Sigma_\ell^{-1}).$$

Similarly,

$$\Delta_\ell^2 \approx E_+ - E_-, \quad \implies \mathbb{E}[\|\widehat{\Delta^2}_\ell\|_F^2] = 2N\sigma^2 \operatorname{tr}(\Sigma_\ell^{-1}).$$

Using $\mathbb{E}\|\widehat{\Delta}_\ell\|_F^2 \asymp N$, the curvature scales like

$$\mathbb{E}[\kappa_\ell] \sim \frac{\sqrt{\mathbb{E}\|\widehat{\Delta^2}_\ell\|_F^2}}{\left(\mathbb{E}\|\widehat{\Delta}_\ell\|_F^2\right)^{3/2}} = \mathcal{O}\left(\frac{1}{N}\right).$$

■
Interpretation: with pure noise between layers, L_ℓ grows like $\sqrt{N}$ and $\kappa_\ell \to 0$. If you see large curvature under this null, something's wrong (or steps are not independent).
Claim. (E1) Define

$$T_\ell = \frac{\widehat{\Delta}_\ell}{\|\widehat{\Delta}_\ell\|_F}$$

and the alternative curvature

$$\widetilde{\kappa}_\ell := \frac{\|T_{\ell+1\to\ell} - T_\ell\|_F}{\|\widehat{\Delta}_\ell\|_F},$$

where $T_{\ell+1\to\ell}$ uses the same transport/whitening as Δ.If the path is sampled at (approximately) constant intrinsic arclength, then

$$\widetilde{\kappa}_\ell = \kappa_\ell + o(1) \quad \text{as step size} \to 0.$$

Proof. In the continuous-depth surrogate $X(s)$,

$$T = \frac{\dot{X}}{\|\dot{X}\|}, \quad \kappa = \left\| \frac{dT}{ds} \right\|.$$

Forward finite differences with uniform arclength step h give

$$\widehat{\Delta}_\ell \approx \dot{X}(s_\ell)h, \quad \widehat{\Delta^2}_\ell \approx \ddot{X}(s_\ell)h^2.$$

Then

$$\kappa_\ell = \frac{\|\ddot{X}\|}{\|\dot{X}\|^3} + o(1), \quad \widetilde{\kappa}_\ell = \left\| \frac{dT}{ds} \right\| + o(1),$$

which coincide with the Frenet formulas. ■

Claim. (F1) Let P be an orthogonal projector (e.g., top-k PCA after whitening).

$$\|\widehat{\Delta}_\ell P\|_F \le \|\widehat{\Delta}_\ell\|_F, \quad \|\widehat{\Delta^2}_\ell P\|_F \le \|\widehat{\Delta^2}_\ell\|_F.$$

Proof. For any matrix A and orthogonal projector P,

$$\|AP\|_F^2 = \operatorname{tr}(PA^\top AP) \le \operatorname{tr}(A^\top A) = \|A\|_F^2.$$

■

As one increases k, $L_\ell^{(k)}$ should be non-decreasing and upper-bounded by the full L_ℓ. (Curvature may not be monotone because its denominator also changes, but the numerator is non-increasing.)

Claim. (G1) Under $X'_\ell = X_\ell Q_\ell$,

$$\|\mathbf{v}'_\ell\| = \|\mathbf{v}_\ell\|.$$

Proof. From B1. ■

Claim. (G2) For any test direction $u \in \mathbb{R}^d$ in intrinsic units (i.e., compare u to $\mathbf{v}_\ell$ after whitening),

$$|\langle \mathbf{v}_\ell, u \rangle| \le \|\mathbf{v}_\ell\| \, \|u\|.$$

Proof. Use Cauchy–Schwarz. ■

Claim. (H1) If

$$\Sigma_\ell = \frac{1}{N} X_\ell^\top X_\ell + \lambda I, \quad \lambda > 0,$$

then

$$\|\Sigma_\ell^{-1/2}\|_2 \le \lambda^{-1/2}.$$

Proof. The smallest eigenvalue of Σ_ℓ is $\ge \lambda$. Thus the largest eigenvalue of $\Sigma_\ell^{-1/2}$ is $\le \lambda^{-1/2}$. ■

Claim. (H2) If

$$\tilde{X}_\ell = X_\ell + \mathbf{1} m^\top$$

(same m added to each row), then after row-centering $\tilde{X}_\ell$, you recover X_ℓ, hence all quantities are identical.

Proof. The row mean of $\tilde{X}_\ell$ is $m + \operatorname{mean}(X_\ell)$; subtracting it yields the centered X_ℓ. ■

Claim. (I1)

$$\|X_{\ell+1}R_\ell^\top - X_\ell\|_F = \|X_\ell R_\ell - X_{\ell+1}\|_F.$$

Proof.

$$\begin{aligned}
\|X_{\ell+1}R_\ell^\top - X_\ell\|_F &= \|(X_{\ell+1}R_\ell^\top - X_\ell)^\top\|_F \\
&= \|R_\ell X_{\ell+1}^\top - X_\ell^\top\|_F \\
&= \|R_\ell^\top (R_\ell X_{\ell+1}^\top - X_\ell^\top)\|_F \\
&= \|X_{\ell+1}^\top - R_\ell^\top X_\ell^\top\|_F \\
&= \|(X_\ell R_\ell - X_{\ell+1})^\top\|_F \\
&= \|X_\ell R_\ell - X_{\ell+1}\|_F.
\end{aligned}$$

■