Supplementary Information: Riemannian Geometric Algebra Transformers

Zachary Nathan Joseph

February 6, 2026

Supplementary Information: Proofs

Result Index (S1–S14).

Label	Result
S1	Lemma: Sign invariance of the distance
S2	Lemma: Small-angle distance expansion
S3	Lemma: Softmax stability
S4	Theorem: Bridge Theorem (Euclidean limit)
S5	Theorem: GSM attention is a Markov diffusion operator
S6	Corollary: Non-expansive bounds
S7	Lemma: Exact truncation identity
S8	Corollary: Truncation bound
S9	Theorem: Gauge equivariance of GSM attention
S10	Theorem: Geodesic alignment gradient on S^3
S11	Corollary: Structural learning as geodesic alignment
S12	Lemma: Iterated BCH accumulation
S13	Theorem: Depth accumulates curvature
S14	Corollary: Standard attention approximates rotor flow

Note

Formal verification: Statements S1–S14 are mechanically verified in Lean (including the head-level and stack-level clauses of Theorem S4); see the repository documentation for a statement-to-file map and the archived software release (v0.1.2) at https://doi.org/10.5281/zenodo.18511210.

Definitions

Manifold and group structure. Let $q,k \in \mathrm{Spin}(3) \subset S^3$ be unit quaternions with sign-invariant similarity $s(q,k) = |\langle q,k\rangle|$, where $\langle q,k\rangle = \sum_{i=0}^{3} q_i k_i$ is the standard Euclidean inner product in $\mathbb{R}^4$. Define the geodesic distance used in RGAT as $d_{\mathrm{geo}}(q,k) = 2\arccos(s(q,k))$ with the principal branch; the factor of 2 ensures d_{geo} measures arc length on the unit 3-sphere. The principal log map Log_q is defined on $\mathrm{Spin}(3) \setminus \{-q\}$ (the cut locus) so that $k = \exp_q(\mathrm{Log}_q(k))$ and $\|\mathrm{Log}_q(k)\| = d_{\mathrm{geo}}(q,k)/2$. The *injectivity radius* of Spin(3) is π (half the diameter of S^3); all small-angle expansions require arguments to lie strictly within this radius.

Metric and exponential map. We use the standard bi-invariant (round) metric induced by the embedding $S^3 \subset \mathbb{R}^4$. The exponential map at the identity is defined by $\exp(u) = \cos\|u\| + (\sin\|u\|/\|u\|)\,u$ for $u \in \mathbb{R}^3$ (with the convention $\exp(0) = 1$), so $\|u\|$ is the half-angle of the corresponding rotation. All vector norms $\|\cdot\|$ are Euclidean norms induced by the chosen bi-invariant metric; operator norms use the induced ℓ_2 or row-wise ℓ_∞ conventions as stated. The notation $R(u) = \exp(u)$ denotes the rotor corresponding to generator $u \in \mathbb{R}^3$.

Temperature and scale parameters. The diffusion temperature $\tau > 0$ (or per-head τ_h) controls the bandwidth of the heat kernel: smaller τ yields sharper attention concentrated near geodesically close keys. We write $\tau_{\min} > 0$ for a lower bound on temperature when uniformity is required. The small-angle parameter $\varepsilon > 0$ bounds rotor generator norms, $\|u\|, \|v\| \le \varepsilon$, and the threshold $\varepsilon_0 > 0$ is chosen so that $\varepsilon \le \varepsilon_0$ lies strictly below the injectivity radius (typically $\varepsilon_0 \ll \pi$).

Lie algebra and bracket. The Lie algebra $\mathfrak{spin}(3) \cong \mathbb{R}^3$ consists of bivector generators under the commutator bracket. For $u, v \in \mathbb{R}^3$, the Lie bracket is the cross product $[u,v] = u \times v$, satisfying $\|[u,v]\| \le \|u\|\|v\|$. The *Baker–Campbell–Hausdorff (BCH) formula* expresses the product of exponentials as a single exponential: $\exp(u)\exp(v) = \exp(u + v + \frac{1}{2}[u,v] + \cdots)$, with higher-order terms involving nested brackets. Under the bi-invariant metric, ad_u is skew-adjoint, so $\langle u, [u,w] \rangle = 0$.

Sparse attention notation. For query i, let $S_i \subseteq \{1, \ldots, T\}$ be the candidate set of attended keys. Define the retained probability mass $p_i = \sum_{j \in S_i} P_{ij}$ and the dropped mass $\delta_i = 1 - p_i$. The truncated attention weights are $\tilde{P}_{ij} = P_{ij}/p_i$ for $j \in S_i$ (assuming $p_i > 0$), and the sparse output is $\tilde{y}_i = \sum_{j \in S_i} \tilde{P}_{ij} v_j$. We write $V_{\max} = \max_j \|v_j\|$ for the maximum value norm.

Sequence and layer indices. The sequence length (number of tokens/keys) is denoted T. In a depth-L stack, layers are indexed $\ell = 1, \ldots, L$. The Lipschitz constant of layer ℓ with respect to the $\|\cdot\|_{\infty,2}$ norm (max-row ℓ_2) is denoted L_ℓ; these constants are assumed uniformly bounded. The softmax function $\sigma : \mathbb{R}^T \to \mathbb{R}^T$ is defined component-wise by $\sigma(\ell)_i = \exp(\ell_i) / \sum_k \exp(\ell_k)$.

Group actions and equivariance. For $g \in \mathrm{Spin}(3)$, left multiplication acts on rotors by $q \mapsto gq$. An orthogonal representation $L(g) : \mathbb{R}^d \to \mathbb{R}^d$ satisfies $L(g)^\top L(g) = I$ and $L(g_1 g_2) = L(g_1)L(g_2)$. The tangent-space projector at $q \in S^3$ is $P_q = I - qq^\top$, projecting $\mathbb{R}^4$ onto $T_q S^3 = \{v \in \mathbb{R}^4 : \langle q, v \rangle = 0\}$. The rotor projection P_{rot} extracts the scalar+bivector components of an 8-dimensional multivector and normalizes to unit norm.

Absorbed constants. Several proofs involve constants that depend only on the bi-invariant metric and the Lie algebra structure constants. Specifically: C_{geo} bounds the geodesic expansion error (Lemma S2); C_{head} bounds head-level attention discrepancy (Theorem S4); $C_{\mathrm{stack}} = C_{\mathrm{head}} \prod_\ell L_\ell$ bounds stack-level discrepancy; and C_1, C_2 bound iterated BCH remainders (Lemma S12). These are finite, computable constants under the stated assumptions.

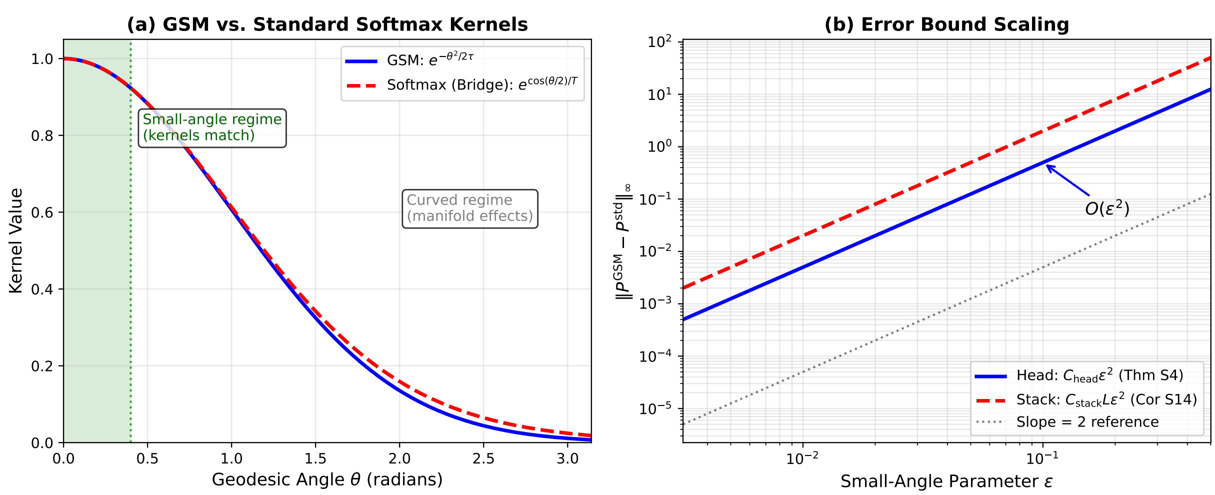

Figure 1: **Figure S1: The Bridge Theorem schematic.** (a) GSM heat kernel vs. standard softmax kernel as functions of geodesic angle; kernels match in the small-angle regime (shaded). (b) Error bound schematic: head-level error scales as $O(\varepsilon^2)$ (Theorem S4), stack-level error as $O(L\varepsilon^2)$ (Corollary S14).

Algorithm sketch (GSM attention). Inputs: rotor queries μ_i, rotor keys r_j, values v_j, temperature τ_h. 1) Project to rotor subspace and normalize: $\mu_i \leftarrow P_{\mathrm{rot}}(\mu_i)$, $r_j \leftarrow P_{\mathrm{rot}}(r_j)$. 2) Compute sign-invariant geodesic distances: $d_{ij} = d_{\mathrm{geo}}(\mu_i, r_j)$. 3) Form kernel logits: $\ell_{ij} = -(2\tau_h)^{-1} d_{ij}^2$. 4) Row-normalize: $P_{ij} = \exp(\ell_{ij}) / \sum_k \exp(\ell_{ik})$. 5) Mix values: $\tilde{v}_i = \sum_j P_{ij} v_j$. 6) If values live in rotor fibers, renormalize to the manifold after mixing.

Lemma S1 (Sign invariance of the distance)**.** *For any unit quaternions $q, k \in S^3$, the geodesic distance satisfies*

$$d_{\mathrm{geo}}(q,k) = d_{\mathrm{geo}}(-q,k) = d_{\mathrm{geo}}(q,-k).$$

Proof. By definition, $s(q,k) = |\langle q,k\rangle|$. Because $\langle -q,k\rangle = -\langle q,k\rangle$ and $\langle q,-k\rangle = -\langle q,k\rangle$, we have $s(-q,k) = s(q,-k) = s(q,k)$. Therefore

$$d_{\mathrm{geo}}(-q,k) = 2\arccos(s(-q,k)) = 2\arccos(s(q,k)) = d_{\mathrm{geo}}(q,k),$$

and similarly for $d_{\mathrm{geo}}(q,-k)$. □

Lemma S2 (Small-angle distance expansion)**.** *Let $u, v \in \mathbb{R}^3$ with $\|u\|, \|v\| \le \varepsilon$ and $0 < \varepsilon \le \varepsilon_0$, where ε_0 lies below the injectivity radius. Define $q = \exp(u)$ and $k = \exp(v)$ in* Spin(3)*. Then there exists $C_{\mathrm{geo}} > 0$ (depending only on the chosen bi-invariant metric and uniform bounds on the Lie bracket structure constants for $\mathfrak{spin}(3)$) such that*

$$\left| d_{\mathrm{geo}}(q,k)^2 - 4\|u-v\|^2 \right| \le C_{\mathrm{geo}}\,\varepsilon^4.$$

Proof. Let $q = \exp(u)$ and $k = \exp(v)$. The relative rotor is

$$q^* k = \exp(-u)\exp(v).$$

By the Baker–Campbell–Hausdorff expansion,

$$\exp(-u)\exp(v) = \exp\left(v - u + \tfrac{1}{2}[v,-u] + O(\varepsilon^3)\right).$$

Let w be the BCH log and write $a = v - u$, $c = \frac{1}{2}[v,-u]$, and r for the remainder with $\|r\| \le C_1 \varepsilon^3$. Then $w = a + c + r$ with $\|a\| = O(\varepsilon)$ and $\|c\| = O(\varepsilon^2)$. For the bi-invariant metric on $\mathfrak{spin}(3)$, ad_a is skew-adjoint, hence $\langle a, [a,z]\rangle = 0$ for all z; in particular $\langle a, c\rangle = 0$. Therefore

$$\|w\|^2 = \|a\|^2 + 2\langle a,r\rangle + \|c\|^2 + 2\langle c,r\rangle + \|r\|^2,$$

so $\|w\| = \|a\| + O(\varepsilon^3)$. Because $d_{\mathrm{geo}}(q,k) = 2\|w\|$, we obtain

$$d_{\mathrm{geo}}(q,k)^2 = 4\|w\|^2 = 4\|u-v\|^2 + O(\varepsilon^4),$$

which yields the stated bound after absorbing constants. □

Lemma S3 (Softmax stability)**.** *For any $\ell, \ell' \in \mathbb{R}^T$, the softmax σ satisfies*

$$\|\sigma(\ell) - \sigma(\ell')\|_\infty \le \tfrac{1}{2}\|\ell - \ell'\|_\infty.$$

Proof. The Jacobian of σ is $J(\ell) = \mathrm{diag}(\sigma(\ell)) - \sigma(\ell)\sigma(\ell)^\top$. Each row sums to zero and has entries bounded by $\sigma_i(1-\sigma_i) \le \frac{1}{4}$. The ℓ_∞ operator norm is the maximum absolute row sum, and row i has absolute sum $\sigma_i(1-\sigma_i) + \sum_{j\ne i}\sigma_i\sigma_j = 2\sigma_i(1-\sigma_i)$, hence

$$\|J(\ell)\|_{\infty\to\infty} \le \sup_i 2\sigma_i(1-\sigma_i) \le \tfrac{1}{2}.$$

By the mean value theorem applied along the line segment from ℓ to ℓ', we obtain

$$\|\sigma(\ell) - \sigma(\ell')\|_\infty \le \sup_{t\in[0,1]} \|J(\ell + t(\ell' - \ell))\|_{\infty\to\infty}\|\ell - \ell'\|_\infty \le \tfrac{1}{2}\|\ell - \ell'\|_\infty.$$

□

Theorem S4 (Bridge Theorem: Euclidean limit)**.** *Let $Q, K \in \mathbb{R}^{T\times 3}$ with $\|Q_i\|, \|K_j\| \le \varepsilon$ and $0 < \varepsilon \le \varepsilon_0$, $\tau \ge \tau_{\min} > 0$. Define rotors $R(Q_i), R(K_j) \in \mathrm{Spin}(3)$ and let $\|\cdot\|_{\infty,2}$ denote the max-row ℓ_2 norm on sequences. Let GSM logits be $\ell_{ij}^{\mathrm{GSM}} = -(2\tau)^{-1} d_{\mathrm{geo}}(R(Q_i), R(K_j))^2$ and standard logits be $\ell_{ij}^{\mathrm{std}} = \tau^{-1} Q_i^\top K_j$. Then there exists a constant C_{head} depending only on $\tau_{\min}$ and C_{geo} such that*

$$\|P^{\mathrm{GSM}} - P^{\mathrm{softmax}}\|_\infty \le C_{\mathrm{head}}\,\varepsilon^2,$$

and for a depth-L stack of Lipschitz layers, a constant C_{stack} such that

$$\|\mathscr{F}_{\mathrm{RGAT}} - \mathscr{F}_{\mathrm{Transformer}}\| \le C_{\mathrm{stack}}\,\varepsilon^2,$$

where each layer is L_ℓ-Lipschitz with respect to $\|\cdot\|_{\infty,2}$ and $C_{\mathrm{stack}} = C_{\mathrm{head}} \prod_\ell L_\ell$.

Proof. By Lemma S2, for each i, j,

$$d_{\mathrm{geo}}(R(Q_i), R(K_j))^2 = 4\|Q_i - K_j\|^2 + O(\varepsilon^4).$$

Expanding the square gives

$$\|Q_i - K_j\|^2 = \|Q_i\|^2 + \|K_j\|^2 - 2Q_i^\top K_j.$$

Substituting into the GSM logit yields

$$\ell_{ij}^{\mathrm{GSM}} = \frac{1}{\tau} Q_i^\top K_j - \frac{1}{2\tau}\|K_j\|^2 - \frac{1}{2\tau}\|Q_i\|^2 + O(\varepsilon^4).$$

For fixed i, the term $-\|Q_i\|^2/(2\tau)$ is constant across j and cancels inside the row-wise softmax, so

$$\|\ell_{i\cdot}^{\mathrm{GSM}} - \ell_{i\cdot}^{\mathrm{std}}\|_\infty \le C\varepsilon^2$$

for a constant C depending on $\tau_{\min}$, the metric choice, and the Lie-bracket bound encapsulated in C_{geo}, as well as the bound $\max_j \|K_j\|^2 \le \varepsilon^2$; the $O(\varepsilon^4)$ remainder is absorbed into $C\varepsilon^2$ for $\varepsilon \le 1$. Applying Lemma S3 row-wise gives

$$\|P^{\mathrm{GSM}} - P^{\mathrm{softmax}}\|_\infty \le \tfrac{1}{2} C\varepsilon^2 = C_{\mathrm{head}}\varepsilon^2.$$

For a depth-L stack of Lipschitz layers F_ℓ with constants L_ℓ, the discrepancy propagates as

$$\Delta_L \le \Big(\prod_{\ell=1}^{L} L_\ell\Big)\Delta_1,$$

yielding the stated stack-level bound with $C_{\mathrm{stack}} = C_{\mathrm{head}} \prod_{\ell=1}^{L} L_\ell$. □

Theorem S5 (GSM attention is a Markov diffusion operator)**.** *Let $K_{ij} = \exp(-d_{\mathrm{geo}}(\mu_i, r_j)^2/(2\tau))$ and $P_{ij} = K_{ij}/\sum_k K_{ik}$. Then each row of P is a probability distribution. For any value vectors $\{v_j\}$, the output $y_i = \sum_j P_{ij} v_j$ lies in the convex hull of $\{v_j\}$.*

Proof. For all i, j, $K_{ij} > 0$ because it is an exponential of a real number. Therefore

$$P_{ij} = \frac{K_{ij}}{\sum_k K_{ik}} \ge 0.$$

Summing over j gives

$$\sum_j P_{ij} = \frac{\sum_j K_{ij}}{\sum_k K_{ik}} = 1,$$

so each row is a probability distribution. Thus

$$y_i = \sum_j P_{ij} v_j$$

is a convex combination of the values and therefore lies in their convex hull. □

Corollary S6 (Non-expansive bounds)**.** *If $\|v_j\| \le V_{\max}$ for all j, then $\|y_i\| \le V_{\max}$ and $\|y_i - y'_i\| \le \max_j \|v_j - v'_j\|$ for two value sets $\{v_j\}, \{v'_j\}$.*

Proof. Because y_i is a convex combination,

$$\|y_i\| \le \sum_j P_{ij}\|v_j\| \le \sum_j P_{ij} V_{\max} = V_{\max}.$$

For two value sets, write

$$y_i - y'_i = \sum_j P_{ij}(v_j - v'_j).$$

Taking norms and using convexity,

$$\|y_i - y'_i\| \le \sum_j P_{ij}\|v_j - v'_j\| \le \max_j \|v_j - v'_j\|.$$

□

Lemma S7 (Exact truncation identity)**.** *Let S_i be a candidate set for query i, define $p_i = \sum_{j \in S_i} P_{ij}$ and $\delta_i = 1 - p_i$. For $p_i > 0$, define $\tilde{P}_{ij} = P_{ij}/p_i$ for $j \in S_i$ and $\mu_{S_i} = \sum_{j \in S_i} \tilde{P}_{ij} v_j$. For $\delta_i > 0$, define the complement mean $\mu_{S_i^c} = \delta_i^{-1} \sum_{j \notin S_i} P_{ij} v_j$. Then*

$$y_i - \tilde{y}_i = \delta_i(\mu_{S_i^c} - \mu_{S_i}),$$

with the convention $y_i - \tilde{y}_i = 0$ when $\delta_i = 0$.

Proof. Write the full output as

$$y_i = \sum_{j \in S_i} P_{ij} v_j + \sum_{j \notin S_i} P_{ij} v_j = p_i \mu_{S_i} + \delta_i \mu_{S_i^c}.$$

By definition, $\tilde{y}_i = \mu_{S_i}$. Therefore

$$y_i - \tilde{y}_i = (p_i \mu_{S_i} + \delta_i \mu_{S_i^c}) - \mu_{S_i} = \delta_i(\mu_{S_i^c} - \mu_{S_i}),$$

which proves the identity. □

Corollary S8 (Truncation bound)**.** *If $\|v_j\| \le V_{\max}$ for all j, then*

$$\|y_i - \tilde{y}_i\| \le 2V_{\max}\,\delta_i.$$

Proof. By Lemma S7,

$$\|y_i - \tilde{y}_i\| = \delta_i \|\mu_{S_i^c} - \mu_{S_i}\|.$$

Both $\mu_{S_i^c}$ and μ_{S_i} are convex combinations of values, so $\|\mu_{S_i^c}\| \le V_{\max}$ and $\|\mu_{S_i}\| \le V_{\max}$. Thus

$$\|\mu_{S_i^c} - \mu_{S_i}\| \le \|\mu_{S_i^c}\| + \|\mu_{S_i}\| \le 2V_{\max}.$$

Multiplying by δ_i yields the bound. □

Theorem S9 (Gauge equivariance of GSM attention)**.** *Assume d_{geo} is the geodesic distance induced by the bi-invariant metric on* Spin(3)*. Let $g \in \text{Spin}(3)$ act on rotors by left multiplication. Define transformed queries $\mu'_i = g\mu_i$ and keys $r'_j = gr_j$. Then $d_{\text{geo}}(\mu'_i, r'_j) = d_{\text{geo}}(\mu_i, r_j)$, and hence $P'_{ij} = P_{ij}$. If values transform linearly by an orthogonal representation $L(g)$ (so $\|L(g)v\| = \|v\|$), then $y'_i = L(g)y_i$.*

Proof. Because Spin(3) acts by isometries on itself,

$$d_{\text{geo}}(g\mu_i, gr_j) = d_{\text{geo}}(\mu_i, r_j).$$

Therefore

$$K'_{ij} = \exp\left(-\frac{d_{\text{geo}}(g\mu_i, gr_j)^2}{2\tau}\right) = \exp\left(-\frac{d_{\text{geo}}(\mu_i, r_j)^2}{2\tau}\right) = K_{ij},$$

and the row normalization is identical, yielding $P'_{ij} = P_{ij}$. For values,

$$y'_i = \sum_j P'_{ij} v'_j = \sum_j P_{ij} L(g) v_j = L(g) \sum_j P_{ij} v_j = L(g) y_i.$$

□

Theorem S10 (Geodesic alignment gradient on S^3)**.** *Let $q, r \in S^3$ with $q \neq -r$, and choose the sign of r so that $\langle q, r\rangle > 0$. Define $f(q) = \frac{1}{2} d_{\text{geo}}(q,r)^2$ with $d_{\text{geo}}(q,r) = 2\arccos(\langle q,r\rangle)$. Then the Riemannian gradient satisfies*

$$\nabla_R f(q) = -4\,\mathrm{Log}_q(r),$$

and the unique minimizers are $q = \pm r$.

Proof. We use the round metric induced by the embedding $S^3 \subset \mathbb{R}^4$, so the Riemannian gradient is obtained by projecting the Euclidean gradient onto the tangent space. Let $s = \langle q,r\rangle \in (0,1]$ and $d = 2\arccos(s)$, so $f(q) = \frac{1}{2}d^2$. By the chain rule,

$$\frac{\partial f}{\partial q} = d\,\frac{\partial d}{\partial q}.$$

Since $d = 2\arccos(s)$, we have

$$\frac{\partial d}{\partial s} = -\frac{2}{\sqrt{1-s^2}}, \qquad \frac{\partial s}{\partial q} = r,$$

so

$$\frac{\partial d}{\partial q} = -\frac{2}{\sqrt{1-s^2}}\,r \quad \Rightarrow \quad \frac{\partial f}{\partial q} = -\frac{2d}{\sqrt{1-s^2}}\,r.$$

The tangent projector on S^3 is $P_q = I - qq^\top$, hence

$$\nabla_R f(q) = P_q \frac{\partial f}{\partial q} = -\frac{2d}{\sqrt{1-s^2}}(r - sq).$$

Because $\sin(d/2) = \sqrt{1-s^2}$, the log map on S^3 satisfies

$$\mathrm{Log}_q(r) = \frac{d}{2\sin(d/2)}(r - sq).$$

Substituting yields

$$\nabla_R f(q) = -\frac{2d}{\sin(d/2)}(r - sq) = -4\,\mathrm{Log}_q(r).$$

The norm of $\mathrm{Log}_q(r)$ vanishes iff $q = \pm r$, hence these are the unique minimizers. The case $q = r$ follows by continuity. ☐

Corollary S11 (Structural learning as geodesic alignment)**.** *For a single-target energy $f(q) = \frac{1}{2} d_{\text{geo}}(q,r)^2$, the negative gradient flow $\dot{q} = -\nabla_R f(q)$ moves q along the geodesic toward r (up to sign), so learning structurally aligns rotors by minimizing geodesic distance.*

Proof. By Theorem S10, $\nabla_R f(q)$ is proportional to $\mathrm{Log}_q(r)$, which is the tangent vector of the unique minimal geodesic from q to r within the injectivity radius. Therefore the negative gradient flow follows that geodesic and converges to $q = \pm r$. ☐

Lemma S12 (Iterated BCH accumulation)**.** *Let $u_1, \dots, u_L \in \mathbb{R}^3$ satisfy $\|u_\ell\| \le \varepsilon$ and assume $0 < \varepsilon \le \varepsilon_0 / L$ so that $L\varepsilon$ lies below the injectivity radius and $L\varepsilon \le 1$. Then there exists $w_L \in \mathbb{R}^3$ such that*

$$\exp(u_1)\exp(u_2)\cdots\exp(u_L) = \exp(w_L),$$

and constants $C_1, C_2 > 0$ (depending only on the bi-invariant metric and Lie bracket constants) such that

$$\Big\| w_L - \sum_{\ell=1}^{L} u_\ell - \frac{1}{2}\sum_{1\le i<j\le L}[u_i,u_j]\Big\| \le C_1 L^3 \varepsilon^3 \quad \textit{and} \quad \|w_L - \sum_{\ell=1}^{L} u_\ell\| \le C_2 L^2 \varepsilon^2.$$

Proof. For $L = 2$, the BCH formula gives $\exp(u_1)\exp(u_2) = \exp(u_1 + u_2 + \frac{1}{2}[u_1,u_2] + R_2)$ with $\|R_2\| \le C\varepsilon^3$. Assume the statement holds for L with $\exp(u_1)\cdots\exp(u_L) = \exp(w_L)$ and $\|w_L\| \le L\varepsilon + C' L^2 \varepsilon^2$. Applying BCH to $\exp(w_L)\exp(u_{L+1})$ yields

$$\exp(w_L)\exp(u_{L+1}) = \exp\!\left(w_L + u_{L+1} + \tfrac{1}{2}[w_L, u_{L+1}] + R_{L+1}\right),$$

with $\|R_{L+1}\| \le C\|w_L\|^3 \le CL^3\varepsilon^3$. Expanding $[w_L, u_{L+1}]$ and collecting commutator terms adds $\frac{1}{2}\sum_{i=1}^{L}[u_i, u_{L+1}]$ plus a remainder controlled by $O(L^3\varepsilon^3)$. Induction gives the stated bounds after absorbing constants; the second bound follows because $L\varepsilon \le 1$ so $L^3\varepsilon^3 \le L^2\varepsilon^2$. ☐

Theorem S13 (Depth accumulates curvature)**.** *Let* $u_1, \dots, u_L \in \mathbb{R}^3$ *be small-angle generators with* $\|u_\ell\| \le \varepsilon$ *and let* $Q_L = \prod_{\ell=1}^{L} \exp(u_\ell)$*. Then there exists* w_L *such that* $Q_L = \exp(w_L)$ *and*

$$w_L = \sum_{\ell=1}^{L} u_\ell + \frac{1}{2} \sum_{1 \le i < j \le L} [u_i, u_j] + R_L, \quad \|R_L\| \le C L^3 \varepsilon^3.$$

Consequently, even when each step is small-angle (Euclidean regime), the composed motion includes commutator curvature of size $O(L^2 \varepsilon^2)$*.*

Proof. The expansion is immediate from Lemma S12. The commutator sum is $O(L^2\varepsilon^2)$ because each $[u_i, u_j]$ is $O(\varepsilon^2)$ and there are $O(L^2)$ pairs. □

Corollary S14 (Standard attention approximates rotor flow)**.** *Assume the Bridge Theorem hypotheses and that each layer operates in the small-angle regime with generators* u_ℓ*. Then a depth-L standard Transformer stack approximates the corresponding rotor flow with error* $O(L\varepsilon^2)$ *in attention weights (for uniformly bounded Lipschitz constants), while the effective generator includes commutator curvature as in Theorem S13.*

Proof. By Theorem S4, each layer's GSM attention differs from standard attention by $O(\varepsilon^2)$. Accumulating over L layers gives an $O(L\varepsilon^2)$ discrepancy when the layer Lipschitz constants are uniformly bounded. The effective rotor flow is the product of per-layer exponentials, whose generator expansion is given by Theorem S13. □

Acknowledgements

We thank the Python community for the libraries and tooling that make this work practical, and we extend special thanks to the Lean community for its culture of formal verification and collaborative proof development, which has supported the correctness and clarity of our formalization.