Equivariant Structured Positional Rotations

Krzysztof Choromanski — Google DeepMind, Columbia University

Daniel Hardesty Lewis — Columbia University

Abstract

We develop equivariant structured positional rotations that preserve relative displacements inside attention layers and provide variance-reduced positive random feature estimators of the softmax kernel. Section 1 outlines the motivation and the challenges encountered when moving beyond rotary encodings. Section 2 presents the algebraic framework of commuting skew generators that generate separable translations on the active subspace, while Section 3 introduces positive orthogonal random features that retain softmax-equivalence without sacrificing stability. Proofs and expanded technical results are deferred to Appendix 3.1.

1 Introduction

Positional encodings play a central role in modern attention architectures. Absolute embeddings introduce global references but fail to generalize reliably across varying context windows; relative embeddings, typically implemented through bias tables, improve extrapolation but complicate linear attention variants by requiring explicit attention matrices. Rotary position encodings (RoPE) address this tension by applying token-wise orthogonal rotations whose relative angles depend solely on the difference of token indices, endowing models with translation invariance and compatibility with key value caching. Yet RoPE implicitly assumes a two-dimensional block decomposition with fixed frequencies.

We revisit this design space from two complementary angles. First, we provide a comprehensive algebraic characterization of equivariant structured positional rotations: given commuting skew-symmetric generators $\{L_k\}$ we derive their joint invariant decomposition, identify the active subspace supporting non-trivial rotations, and show that exponentiating $A(r) = \sum_k r_k L_k$ recovers an $SO(d_h)$ action whose relative alignment depends only on displacement vectors $r_j - r_i$. This strictly generalizes the RoPE paradigm to arbitrary commuting families. We also give the block-identity condition on a Cayley post-rotation under which the relative property is exactly preserved on the active subspace.

Contributions.

1. We formalize the structured rotation model. Any commuting family of real skew-symmetric generators admits a joint real block decomposition into 2×2 rotational planes and a null subspace, which induces explicit orthogonal projectors onto an active subspace and its null complement.

2. We prove that the induced exponentials $R_{\mathrm{STR}}(r) = \exp(A(r))$ implement relative rotations: $R_{\mathrm{STR}}(r_i)^\top R_{\mathrm{STR}}(r_j) = R_{\mathrm{STR}}(r_j - r_i)$. When the post-rotation lies in the structured family $\mathcal{A}$ (i.e., it is identity on the active subspace), the projected attention logits depend only on the displacement $r_j - r_i$.

3. We derive stability bounds. If the generators only approximately commute and the post-rotation only approximately preserves the active subspace, then the deviation of the attention logits from perfect relative form is bounded by $\Delta_{ij} = C(\varepsilon|r_i|_1|r_j|_1 + \eta_{\text{mix}})\,|q_i^{(\text{act})}|\,|k_j^{(\text{act})}|$, i.e. it scales with commutator size ε, subspace leakage η_{mix}, and query/key magnitudes.

4. (Planned / forthcoming.) We integrate positive orthogonal random feature maps with these structured rotations to obtain linear-time softmax attention. We analyze unbiasedness, variance, and finite-dimensional tail bounds for orthogonal ensembles.

2 Structured Positional Rotations

We summarize the algebraic groundwork that underpins the equivariant rotations. Full proofs, including joint spectral decompositions and orthogonality lemmas, appear in Appendix 3.1.

2.1 Generators and Active Subspaces

Let $\{L_k\}_{k=1}^{d_c} \subset \mathbb{R}^{d_h \times d_h}$ be commuting skew-symmetric matrices satisfying $[L_a, L_b] = 0$ for all a, b. Define the linear combination $A(r) = \sum_{k=1}^{d_c} r_k L_k$ for $r \in \mathbb{R}^{d_c}$. Lemma 1 establishes that there exists an orthogonal basis U in which each L_k becomes block-diagonal with 2×2 rotation blocks $\lambda_{k,u} J$ and a residual null block. This yields complementary projectors Π_{act} and Π_{null} onto the active ($2m$-dimensional) and null subspaces. In this coordinate system $A(r)$ is block-diagonal with entries $\theta_u(r)J$, where $\theta_u(r) = \sum_k \lambda_{k,u} r_k$ encodes the displacement.

2.2 Exponentials and Relative Invariance

Exponentiating $A(r)$ produces

$$R_{\text{STR}}(r) = U \left[\left(\bigoplus_{u=1}^{m} R_2(\theta_u(r)) \right) \oplus I_{d_{\text{null}}} \right] U^\top,$$

an orthogonal matrix whose determinant equals one. Because $A(r)$ and $A(s)$ commute, their exponentials satisfy $R_{\text{STR}}(r)^\top R_{\text{STR}}(s) = R_{\text{STR}}(s - r)$, a property that directly ensures relative positional invariance within the active subspace. Moreover, $R_{\text{STR}}(r)$ commutes with Π_{act}, so the active components of queries and keys rotate coherently.

2.3 Cayley Post-Rotations

To allow learned post-rotations while preserving orthogonality, we employ Cayley transforms $P_{\text{sp}} = (I - S)(I + S)^{-1}$ with $S^\top = -S$. Such transforms remain in $SO(d_h)$ and, when restricted to the structured family

$$\mathcal{A} = \left\{ U \,\text{diag}(I_{2m}, R_{\text{null}}) U^\top : R_{\text{null}} \in SO(d_{\text{null}}) \right\},$$

they leave the active subspace invariant. Setting $R_{\text{sp}}(r) = R_{\text{STR}}(r) P_{\text{sp}}$ consequently preserves the relative rotation property in the active block.

3 Positive Orthogonal Random Features

¡¡¡¡¡¡¡ HEAD ======= We adapt the FAVOR$^+$ methodology to the structured rotation setting. The softmax kernel for queries x and keys y satisfies $\mathrm{SM}(x, y) = \exp(x^\top y)$ and admits representations based on positive random features. Appendix 3.1 proves that choosing

$$\phi(x) = \exp\left(\frac{\|x\|^2}{2}\right)\exp(w^\top x), \qquad w \sim \mathcal{N}(0, I_d),$$

produces an unbiased positive estimator with mean-squared error that scales linearly with $\mathrm{SM}(x, y)$. Replacing independent samples with orthogonalized ones further decreases variance by a factor that depends on d but not on asymptotic limits.

¿¿¿¿¿¿¿ origin/claude/relative-rotation-attention-011CUWhuEKcD3BCNeCNXEMzj

3.1 Relative Logits and Stability

Combining the structured rotations with the positive features yields the attention logit

$$\alpha_{ij} = \frac{1}{\sqrt{2m}}(q_i^{(\mathrm{act})})^\top \Big(\Pi_{\mathrm{act}} R_{\mathrm{STR}}(r_j - r_i)\Pi_{\mathrm{act}}\Big) k_j^{(\mathrm{act})},$$

provided $P_{\mathrm{sp}} \in \mathcal{A}$. Theorem 1 shows that α_{ij} depends solely on the relative displacement of tokens. When the commutators $[L_a, L_b]$ are small and P_{sp} nearly preserves the active subspace, the deviation from perfect relative invariance is bounded by $\Delta_{ij} = O(\varepsilon|r_i|_1|r_j|_1 + \eta_{\mathrm{mix}})$, where η_{mix} measures active-null leakage in P_{sp}.

¡¡¡¡¡¡¡ HEAD

======= ¿¿¿¿¿¿¿ origin/claude/relative-rotation-attention-011CUWhuEKcD3BCNeCNXEMzj

F Theoretical Results

This appendix supplies the algebraic details that justify the structural results stated in the main text. In keeping with the style of the referenced supplements, each proof is developed line by line, and intermediate equalities are recorded explicitly so that the chain of reasoning can be reconstructed without appealing to omitted computations. Throughout, $d_h, d_c, D \in \mathbb{N}$ denote the head, displacement, and token dimensions. For $k \in \{1, \dots, d_c\}$ we write $L_k \in \mathbb{R}^{d_h \times d_h}$ for skew-symmetric generators satisfying

$$L_k^\top = -L_k, \qquad [L_a, L_b] = L_a L_b - L_b L_a = 0 \quad \text{for all } a, b. \tag{1}$$

Given $r \in \mathbb{R}^{d_c}$ we define the linear combination and associated exponential

$$A(r) \coloneqq \sum_{k=1}^{d_c} r_k L_k, \qquad R_{\mathrm{STR}}(r) \coloneqq \exp\big(A(r)\big). \tag{2}$$

Unless otherwise stated, norms without qualification refer to the Euclidean vector norm and the corresponding operator norm on matrices. The 2×2 rotation matrix is written $R_2(\theta) = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$, and we use $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ for the canonical symplectic block.

F.1 Proof of Lemma 1 (Joint Block Structure)

Lemma 1. *There exist integers $m, d_{\mathrm{null}} \geq 0$ with $2m + d_{\mathrm{null}} = d_h$ and an orthogonal matrix $U \in O(d_h)$ such that, simultaneously for every k,*

$$U^\top L_k U = \left(\bigoplus_{u=1}^{m} \lambda_{k,u} J \right) \oplus 0_{d_{\mathrm{null}} \times d_{\mathrm{null}}}, \tag{3}$$

with real coefficients $\lambda_{k,u}$. Equivalently, the active subspace has dimension $2m$ and is spanned by the first $2m$ columns of U, while the null space has dimension d_{null}.

Proof. Each generator L_k is real skew-symmetric, hence normal, and the commuting family $\{L_k\}_{k=1}^{d_c}$ can be diagonalised simultaneously by a unitary matrix. Concretely, by the spectral theorem there exists a unitary $\mathcal{U}_c$ such that

$$\mathcal{U}_c^* L_k \mathcal{U}_c = \mathrm{diag}\left(i\lambda_{k,1}, \dots, i\lambda_{k,d_h}\right), \qquad k = 1, \dots, d_c, \tag{4}$$

with real eigenvalues $\lambda_{k,j}$. Writing the columns of $\mathcal{U}_c$ as $\mathcal{U}_c = [u_1, \dots, u_{d_h}]$ and separating real and imaginary parts, every non-real eigenvector occurs in conjugate pairs $(u_j, \overline{u_j})$. Extracting the real and imaginary components of each pair produces a real orthonormal basis:

$$\Re u_j = \frac{1}{2}(u_j + \overline{u_j}), \tag{5}$$

$$\Im u_j = \frac{1}{2i}(u_j - \overline{u_j}), \tag{6}$$

and for any k we compute

$$L_k(\Re u_j) = \Re(L_k u_j) = \Re(i\lambda_{k,j} u_j) = -\lambda_{k,j} \Im u_j, \tag{7}$$

$$L_k(\Im u_j) = \Im(L_k u_j) = \Im(i\lambda_{k,j} u_j) = \lambda_{k,j} \Re u_j. \tag{8}$$

Thus the two-dimensional plane spanned by $\{\Re u_j, \Im u_j\}$ is invariant under every L_k, and within that plane the action is represented by the block $\lambda_{k,j} J$. Purely real eigenvectors of $\mathcal{U}_c$ (if any) contribute to the null block. Collecting the real basis vectors in an orthogonal matrix U establishes the claimed decomposition (3). □

The block structure motivates the projections onto active and null subspaces:

$$\Pi_{\text{act}} \coloneqq U \begin{bmatrix} I_{2m} & 0 \\ 0 & 0 \end{bmatrix} U^\top, \qquad \Pi_{\text{null}} \coloneqq U \begin{bmatrix} 0 & 0 \\ 0 & I_{d_{\text{null}}} \end{bmatrix} U^\top, \tag{9}$$

which satisfy $\Pi_{\text{act}}^2 = \Pi_{\text{act}}$, $\Pi_{\text{null}}^2 = \Pi_{\text{null}}$, and $\Pi_{\text{act}}\Pi_{\text{null}} = 0$.

F.2 Proof of Lemma 2 (Structured Exponential and Relative Rotations)

Lemma 2. *Define $\theta_u(r) \coloneqq \sum_{k=1}^{d_c} \lambda_{k,u} r_k$. Then, for any $r, s \in \mathbb{R}^{d_c}$,*

$$U^\top A(r) U = \left(\bigoplus_{u=1}^{m} \theta_u(r) J \right) \oplus 0, \tag{10}$$

$$R_{\text{STR}}(r) = U \left[\left(\bigoplus_{u=1}^{m} R_2\big(\theta_u(r)\big) \right) \oplus I_{d_{\text{null}}} \right] U^\top, \tag{11}$$

$$R_{\text{STR}}(r)^\top R_{\text{STR}}(s) = R_{\text{STR}}(s - r). \tag{12}$$

Consequently $R_{\text{STR}}(r) \in SO(d_h)$ and $R_{\text{STR}}(r)\Pi_{\text{act}} = \Pi_{\text{act}} R_{\text{STR}}(r)$.

Proof. Equation (10) follows immediately by inserting (3) into the definition of $A(r)$ in (2). Because $J^2 = -I_2$ we have, for each u,

$$\exp\big(\theta_u(r) J\big) = \sum_{n=0}^{\infty} \frac{\theta_u(r)^n J^n}{n!} = \Big\{ \textstyle\sum_{k=0}^{\infty} \frac{(-1)^k \theta_u(r)^{2k}}{(2k)!} I_2 + \sum_{k=0}^{\infty} \frac{(-1)^k \theta_u(r)^{2k+1}}{(2k+1)!} J, \quad \text{if } n \text{ even/odd} = R_2\big(\theta_u(r)\big),$$

where the final equality collects the cosine and sine series. Conjugating back by U yields (11). In particular, each block has determinant 1 and orthogonal columns, so $R_{\text{STR}}(r) \in SO(d_h)$. Moreover, (10) implies $A(r)A(s) = A(s)A(r)$ for all r, s because the blocks commute, hence the Baker–Campbell–Hausdorff series truncates and we obtain

$$R_{\text{STR}}(r)^\top R_{\text{STR}}(s) = \exp\big(-A(r)\big) \exp\big(A(s)\big) = \exp\big(A(s) - A(r)\big) = R_{\text{STR}}(s - r), \tag{13}$$

which is (12). Finally, the block form (11) shows that $R_{\text{STR}}(r)$ acts as the identity on the null block, so it commutes with Π_{act}. □

F.3 Proof of Lemma 3 (Cayley Post-Rotations)

Lemma 3. *Let $S \in \mathbb{R}^{d_h \times d_h}$ be skew-symmetric and define the Cayley transform $P_{\text{sp}} = (I - S)(I + S)^{-1}$. Then $P_{\text{sp}} \in SO(d_h)$. Moreover, when P_{sp} belongs to the structured family*

$$\mathcal{A} \coloneqq \Big\{ U \operatorname{diag}(I_{2m}, R_{\text{null}})\, U^\top \mid R_{\text{null}} \in SO(d_{\text{null}}) \Big\}, \tag{14}$$

the active projection is preserved: $\Pi_{\text{act}} P_{\text{sp}} \Pi_{\text{act}} = \Pi_{\text{act}}$.

Proof. Because the eigenvalues of a real skew-symmetric matrix are purely imaginary, $1 + i\mu \neq 0$ for every eigenvalue $i\mu$ of S, so $(I + S)$ is invertible. We compute

$$P_{\text{sp}}^\top P_{\text{sp}} = (I + S)^{-\top} (I - S)^\top (I - S)(I + S)^{-1} \tag{15}$$

$$= (I - S^\top)^{-1} (I + S^\top)(I - S)(I + S)^{-1} \tag{16}$$

$$= (I + S)^{-1} (I + S)(I - S)(I + S)^{-1} = I, \tag{17}$$

where Skew-symmetry ($S^\top = -S$) and the commutativity of $I \pm S$ are used in the second equality. Likewise the determinant evaluates to

$$\det(P_{\mathrm{sp}}) = \frac{\det(I-S)}{\det(I+S)} = \prod_{u=1}^{d_h} \frac{1-i\mu_u}{1+i\mu_u} = 1, \tag{18}$$

showing $P_{\mathrm{sp}} \in SO(d_h)$. If $P_{\mathrm{sp}} \in \mathcal{A}$ its block form relative to U reads $\mathrm{diag}(I_{2m}, R_{\mathrm{null}})$, hence the active-portion is fixed pointwise and $\Pi_{\mathrm{act}} P_{\mathrm{sp}} \Pi_{\mathrm{act}} = \Pi_{\mathrm{act}}$. □

Given such a structured post-rotation we set

$$R_{\mathrm{sp}}(r) \coloneqq R_{\mathrm{STR}}(r) P_{\mathrm{sp}}, \tag{19}$$

which remains in $SO(d_h)$ and inherits the relative-rotation law on the active subspace by Lemma 2.

F.4 Proof of Theorem 1 (Relative Attention Logits)

Let $W_Q, W_K, W_V \in \mathbb{R}^{d_h \times D}$ be query, key, and value projections. For tokens $x_i, x_j \in \mathbb{R}^D$ and associated displacements $r_i, r_j \in \mathbb{R}^{d_c}$ define

$$q_i = W_Q x_i, \qquad k_j = W_K x_j, \qquad v_j = W_V x_j, \tag{20}$$

and project onto the active subspace to obtain

$$q_i^{(\mathrm{act})} = \Pi_{\mathrm{act}} q_i, \qquad k_j^{(\mathrm{act})} = \Pi_{\mathrm{act}} k_j. \tag{21}$$

The structured rotation and projection yield modified query/key pairs

$$\tilde{q}_i = \Pi_{\mathrm{act}} R_{\mathrm{sp}}(r_i) q_i^{(\mathrm{act})}, \qquad \tilde{k}_j = \Pi_{\mathrm{act}} R_{\mathrm{sp}}(r_j) k_j^{(\mathrm{act})}, \tag{22}$$

and the attention logit is $\alpha_{ij} = \frac{1}{\sqrt{d_{\mathrm{act}}}} \tilde{q}_i^\top \tilde{k}_j$ with $d_{\mathrm{act}} = 2m$.

Theorem 1. *If $P_{\mathrm{sp}} \in \mathcal{A}$, then*

$$\alpha_{ij} = \frac{1}{\sqrt{2m}} (q_i^{(\mathrm{act})})^\top \Big(\Pi_{\mathrm{act}}\, R_{\mathrm{STR}}(r_j - r_i)\, \Pi_{\mathrm{act}} \Big) k_j^{(\mathrm{act})}, \tag{23}$$

so the logit depends only on the displacement $r_j - r_i$.

Proof. Substituting (22) and using $\Pi_{\mathrm{act}}^\top = \Pi_{\mathrm{act}}$ yields

$$\alpha_{ij} = \frac{1}{\sqrt{2m}} (q_i^{(\mathrm{act})})^\top \big(\Pi_{\mathrm{act}} R_{\mathrm{sp}}(r_i) \big)^\top \Pi_{\mathrm{act}} R_{\mathrm{sp}}(r_j) k_j^{(\mathrm{act})}. \tag{24}$$

Lemma 3 ensures $\Pi_{\mathrm{act}} R_{\mathrm{sp}}(r) \Pi_{\mathrm{act}} = \Pi_{\mathrm{act}} R_{\mathrm{STR}}(r) \Pi_{\mathrm{act}}$ whenever $P_{\mathrm{sp}} \in \mathcal{A}$, so we may replace the post-rotations. Lemma 2 then gives the relative-rotation identity $R_{\mathrm{STR}}(r_i)^\top R_{\mathrm{STR}}(r_j) = R_{\mathrm{STR}}(r_j - r_i)$ and the commutation with Π_{act}, proving (23). □

F.5 Proof of Theorem 2 (Robustness to Approximate Symmetry)

To quantify deviations from the ideal assumptions we introduce the commutator bounds

$$\varepsilon_{ab} \coloneqq \big\|[L_a, L_b]\big\|, \qquad \varepsilon \coloneqq \max_{a,b} \varepsilon_{ab}, \tag{25}$$

and measure active/null mixing of the post-rotation by writing $P_{\mathrm{sp}} = U\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)U^\top$ and defining

$$\eta_{\mathrm{mix}} \coloneqq \|B\| + \|C\|. \tag{26}$$

The active components are again denoted $q_i^{(\mathrm{act})}$ and $k_j^{(\mathrm{act})}$.

Theorem 2. *There exists an absolute constant $C > 0$ such that, for all i, j,*

$$\left|\alpha_{ij} - \frac{1}{\sqrt{2m}}(q_i^{(\mathrm{act})})^\top \Pi_{\mathrm{act}} R_{\mathrm{STR}}(r_j - r_i)\Pi_{\mathrm{act}} k_j^{(\mathrm{act})}\right| \le C\Big(\varepsilon \|r_i\|_1 \|r_j\|_1 + \eta_{\mathrm{mix}}\Big)\|q_i^{(\mathrm{act})}\|\,\|k_j^{(\mathrm{act})}\|. \tag{27}$$

Proof. The Baker–Campbell–Hausdorff expansion gives an explicit series for the log of a product. Because $A(r)$ and $A(s)$ almost commute, we may bound the error by keeping the leading commutator term:

$$\begin{aligned}
\big\|R_{\mathrm{STR}}(r)^\top R_{\mathrm{STR}}(s) - R_{\mathrm{STR}}(s-r)\big\| &= \big\|\exp(-A(r))\exp(A(s)) - \exp(A(s) - A(r))\big\| && (28)\\
&\le \frac{1}{2}\|[A(r), A(s)]\| + C_1\|[A(r), A(s)]\|^2 && (29)\\
&\le \Big(\frac{1}{2} + C_1\|[A(r), A(s)]\|\Big)\sum_{a,b} |r_a|\,|s_b|\,\varepsilon_{ab} && (30)\\
&\le C_2\,\varepsilon\,\|r\|_1\|s\|_1, && (31)
\end{aligned}$$

for absolute constants C_1, C_2. Similarly, writing $S = S_- + E$ with $S_-^\top = -S_-$ and $\|E\| \le \eta < 1$, a Neumann-series argument shows that

$$\|(I+S)^{-1} - (I+S_-)^{-1}\| = \|(I+S)^{-1}(S - S_-)(I+S_-)^{-1}\| \le \frac{\eta}{1-\eta}, \tag{32}$$

and consequently the Cayley transforms satisfy

$$\|P_{\mathrm{sp}}(S) - P_{\mathrm{sp}}(S_-)\| \le C_3\eta \tag{33}$$

for some C_3. Because P_{sp} lies within η_{mix} of $\mathcal{A}$, the active block differs from the identity by at most $C_4\eta_{\mathrm{mix}}$, and the projection of $R_{\mathrm{sp}}(r)$ consequently differs from that of $R_{\mathrm{STR}}(r)$ by the same order. Inserting these estimates in the expression for α_{ij} and applying the triangle inequality produces (27), where C absorbs the universal constants. □

F.6 Discussion

Theorem 1 demonstrates that, under exact commuting generators and structured post-rotations, the attention logits depend solely on relative displacements. The robust estimate in Theorem 2 quantifies how violations of the commutation assumption and leakage between active and null subspaces perturb this behaviour: non-commuting generators introduce an error term proportional to $\varepsilon\,\|r_i\|_1\|r_j\|_1$, while imperfect post-rotations contribute an additive term controlled by η_{mix}.