Equivariant Structured Positional Rotations: Complete Algebraic Derivation

Abstract

This document provides a complete, boxed derivation of the relative-position property for equivariant structured positional rotations in attention mechanisms. Each step is recorded as a labelled statement $[T\cdot]$ with explicit dependencies so that the argument can be followed mechanically.

Contents

1 Tier 0: Linear-Algebra Primitives

$\boxed{[T0.1 : \mathbb{N}, \mathrm{Idx}]}$ $\mathbb{N} \stackrel{\varnothing}{=} \{1, 2, 3, \dots\}$ [deps : $\varnothing$]

$d_h, d_c, D \in \mathbb{N}$

$H \stackrel{\varnothing}{=} \{1, \dots, d_h\}, \quad K \stackrel{\varnothing}{=} \{1, \dots, d_c\}, \quad J_D \stackrel{\varnothing}{=} \{1, \dots, D\}$

$\boxed{[T0.2 : I_n, \delta_{ij}]}$ $\delta_{ij} \stackrel{\varnothing}{=} \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$ [deps : $T0.1$]

$I_n \stackrel{\varnothing}{=} [\delta_{ij}]_{i,j=1}^{n}$

$\boxed{[T0.3 : \mathbb{R}^{m\times n}, (\cdot)^\top, (\cdot)^*, \mathrm{tr}]}$ $\mathbb{R}^{m\times n} \stackrel{\varnothing}{=}$ set of $m \times n$ real matrices [deps : $T0.1$]

$\mathbb{C}^{m\times n} \stackrel{\varnothing}{=}$ set of $m \times n$ complex matrices

$A^\top \in \mathbb{R}^{n\times m} \quad (A \in \mathbb{R}^{m\times n})$

$B^* \stackrel{\varnothing}{=} \overline{B}^\top \quad (B \in \mathbb{C}^{m\times n})$

$\mathrm{tr}(M) \stackrel{\varnothing}{=} \sum_{i=1}^{n} M_{ii} \quad (M \in \mathbb{C}^{n\times n})$

$\boxed{[T0.4 : \det]}$ $\det : \mathbb{C}^{n\times n} \to \mathbb{C}$ [deps : $T0.3$]

$\det(AB) = \det(A)\det(B)$

$\det(A^\top) = \det(A)$

U invertible $\Rightarrow \det(UAU^{-1}) = \det(A)$

(multiplicativity, transpose invariance, similarity invariance)

$\boxed{[T0.5 : \oplus, \mathrm{blk}, \det\oplus]}$ $\bigoplus_{u=1}^{m} M_u \stackrel{\varnothing}{=} \mathrm{diag}(M_1, \dots, M_m)$ [d

$\det\left(\bigoplus_{u=1}^{m} M_u\right) = \prod_{u=1}^{m} \det(M_u)$

(block triangular matrices have determinant equal to product of block determinants)

$\boxed{[T0.6 : O, SO, U]}$ $O(n) \stackrel{\varnothing}{=} \{Q \in \mathbb{R}^{n\times n} \mid Q^\top Q = I_n\}$ [deps : $T0.2, T0.3, T0.4$]

$SO(n) \stackrel{\varnothing}{=} \{Q \in O(n) \mid \det Q = 1\}$

$U(n) \stackrel{\varnothing}{=} \{W \in \mathbb{C}^{n\times n} \mid W^*W = I_n\}$

$Q \in O(n) \Rightarrow Q^{-1} = Q^\top$

$\boxed{[T0.7 : \det O]}$ $Q \in O(n) \Rightarrow (\det Q)^2 = \det(Q^\top Q) = \det(I_n) = 1$ [deps : $T0.6, T0.4$]

$\Rightarrow \det Q \in \{\pm 1\}$

$\boxed{[T0.8 : \sigma, \sigma_{\min}, \text{normal}]}$ $\sigma(M) \stackrel{\varnothing}{=} \{\lambda \in \mathbb{C} \mid \exists x \neq 0 : Mx = \lambda x\}$ [deps : $T0.3$]

$M^*M = MM^* \Rightarrow M$ normal

Singular values $s_i(M) \stackrel{\varnothing}{=} \sqrt{\sigma_i(M^*M)}$

$\sigma_{\min}(M) \stackrel{\varnothing}{=} \min_i s_i(M)$

M normal $\Rightarrow s_i(M) = |\lambda_i|$ for $\lambda_i \in \sigma(M)$

(unitary diagonalisation of normal matrices)

$\boxed{[T0.9 : |\cdot|, |\cdot|_1]}$ $|x| \stackrel{\varnothing}{=} \left(\sum_u x_u^2\right)^{1/2}$ $(x \in \mathbb{R}^m)$ [deps : $T0.8$]

$|x|_1 \stackrel{\varnothing}{=} \sum_u |x_u|$

$|M| \stackrel{\varnothing}{=} \sup_{v \neq 0} \frac{|Mv|}{|v|}$ $(M \in \mathbb{R}^{m \times n})$

$|AB| \leq |A|\,|B|, \quad |A + B| \leq |A| + |B|$

$|a| \stackrel{\varnothing}{=} |a|$ $(a \in \mathbb{R})$

(overloaded notation chooses scalar, vector, or operator norm from context)

$\boxed{[T0.10 : [A, B]]}$ $[A, B] \stackrel{\varnothing}{=} AB - BA$ [deps : $T0.3$]

$\boxed{[T0.11 : \exp, \log, \exp \text{ props}]}$ $\exp(M) \stackrel{\varnothing}{=} \sum_{t \geq 0} \frac{1}{t!} M^t$ [deps : $T0.3, T0.4, T0.5, T$

$\log(X)$ locally defined as $\exp^{-1}(X)$ near I_n

$[A, B] = 0 \Rightarrow \exp(A + B) = \exp(A)\exp(B)$

$(\exp M)^\top = \exp(M^\top)$

U invertible $\Rightarrow \exp(UMU^{-1}) = U \exp(M) U^{-1}$

$\det(\exp M) = \exp(\operatorname{tr} M)$

$\exp\left(\bigoplus_{u=1}^{m} M_u\right) = \bigoplus_{u=1}^{m} \exp(M_u)$

(all properties follow from the power-series definition)

$\boxed{[T0.12 : \mathcal{C}, O(\cdot), \approx_\varepsilon]}$ $\mathcal{C} \stackrel{\varnothing}{=} \{C, C', C'', \cdots \mid C \geq 0 \text{ finite constants}\}$ [deps : $T0.9$]

$f = O(g) \Leftrightarrow \exists C \in \mathcal{C} : f \leq Cg$

$X \approx_\varepsilon Y \Leftrightarrow |X - Y| \leq \varepsilon$

$\boxed{[T0.13 : \text{BCH}]}$ $\exp(M)\exp(N) = \exp\big(M + N + \frac{1}{2}[M, N] + R(M, N)\big)$ [deps : $T0.11, T0.10, T0.12, T0.9$]

$|R(M, N)| = O(|[M, N]|^2)$

(Baker–Campbell–Hausdorff series (local version))

$\boxed{[T0.14 : \ker, \mathrm{im}, \mathrm{rank}]}$ $\ker(M) \stackrel{\varnothing}{=} \{x \mid Mx = 0\}$ [deps : $T0.3, T0.1$]

$\mathrm{im}(M) \stackrel{\varnothing}{=} \{Mx \mid x\}$

$\mathrm{rank}(M) \stackrel{\varnothing}{=} \dim(\mathrm{im}(M))$

$\boxed{[T0.15 : \mathrm{proj}]}$ $P^2 = P$ and $P^\top = P \Rightarrow P$ orthogonal projector [deps : $T0.14, T0.3$]

P, Q orthogonal projectors and $PQ = 0 \Rightarrow \mathrm{im}(P) \perp \mathrm{im}(Q)$

(characterisation via images)

$\boxed{[T0.16 : \mathrm{skew}_{\mathrm{spec}}]}$ $S^\top = -S \Rightarrow S$ normal [deps : $T0.3, T0.8$]

$\sigma(S) \subset i\mathbb{R}$

$\lambda = i\mu \in \sigma(S) \Rightarrow -\lambda = -i\mu \in \sigma(S)$

(skew-symmetric spectra occur in conjugate-sign pairs)

$\boxed{[T0.17 : J, R_2]}$ $J \stackrel{\varnothing}{=} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, $R_2(\theta) \stackrel{\varnothing}{=} \exp(\theta J)$ [deps : $T0.11$]

$\boxed{[T0.18 : R_2 \in SO(2)]}$ $R_2(\theta)^\top R_2(\theta) = I_2$ [deps : $T0.17, T0.16, T0.11,$

$\det R_2(\theta) = 1$

$\Rightarrow R_2(\theta) \in SO(2)$

(uses $J^\top = -J$ and properties of exp and the determinant)

$\boxed{[T0.19 : \mathrm{Weyl}]}$ $\left|\sigma_{\min}(A + E) - \sigma_{\min}(A)\right| \leq |E|$ [deps : $T0.8, T0.9$]

$\boxed{[T0.20 : (X^{-1})^\top]}$ X invertible $\Rightarrow (X^{-1})^\top = (X^\top)^{-1}$ [deps : $T0.3$]

2 Tier 1: Structured Generators

$\boxed{[T1.1 : L_k]}$ $\forall k \in K : \; L_k \in \mathbb{R}^{d_h \times d_h}, \quad L_k^\top = -L_k, \quad [L_a, L_b] = 0$ [deps : $T0.1, T0.3, T0.10, T0.16$]

(model hypothesis: commuting skew-symmetric generators)

$\boxed{[T1.2 : \text{normal } L_k]}$ $L_k^\top = -L_k \Rightarrow L_k$ normal [deps : $T1.1, T0.16$]

$\boxed{[T1.3 : A(r)]}$ $r \in \mathbb{R}^{d_c} \Rightarrow A(r) \stackrel{\varnothing}{=} \sum_{k=1}^{d_c} r_k L_k$ [deps : $T1.1, T0.1$]

$\boxed{[T1.4 : A(r)^\top]}$ $A(r)^\top = -A(r)$ [deps : $T1.3, T1.1, T0.3$]

$\boxed{[T1.5 : [A(r), A(s)]]}$ $[A(r), A(s)] = 0$ [deps : $T1.3, T1.1, T0.10$]

$\boxed{[T1.6 : R_{\mathrm{STR}}]}$ $R_{\mathrm{STR}}(r) \stackrel{\varnothing}{=} \exp(A(r))$ [deps : $T1.3, T0.11$]

3 Tier 2: Joint Block Structure

$\boxed{[T2.1 : \text{sim-unitary}]}$ $\{L_k\}_{k\in K}$ commute and are normal [deps : $T1.2, T1.1, T0.6, T0.16$]
$\Rightarrow \exists \mathcal{U}_c \in U(d_h) :\ \mathcal{U}_c^* L_k \mathcal{U}_c = \mathrm{diag}(i\lambda_{k,1}, \ldots, i\lambda_{k,d_h})$
$\lambda_{k,j} \in \mathbb{R}$

$\boxed{[T2.2 : \text{real 2D planes}]}$ Fix $j \in H :\ \mathcal{U}_c^* L_k \mathcal{U}_c e_j = i\lambda_{k,j} e_j$ [deps : $T2.1, T0.17, T0.16, T0.3$]
$\Rightarrow L_k(\Re e_j) = \lambda_{k,j} J(\Re e_j, \Im e_j)$
$L_k(\Im e_j) = \lambda_{k,j} J(\Re e_j, \Im e_j)$
$\Rightarrow \mathrm{span}\{\Re e_j, \Im e_j\}$ is a real, 2-D, L_k-invariant plane

$\boxed{[T2.3 : \text{JointBlockDiag}]}$ $\exists m, d_{\text{null}} \in \mathbb{N} :\ 2m + d_{\text{null}} = d_h$ [deps : $T2.2, T0.6, T0.16, T0.5$]
$$\exists U \in O(d_h) :\ U^\top L_k U = \left(\bigoplus_{u=1}^{m} \lambda_{k,u} J \right) \oplus 0_{d_{\text{null}} \times d_{\text{null}}}$$

$\boxed{[T2.4 : \beta, \theta]}$ $\beta_u \stackrel{\varnothing}{=} (\lambda_{1,u}, \ldots, \lambda_{d_c,u})^\top \in \mathbb{R}^{d_c}$ [deps : $T2.3, T1.1, T0.1$]
$\theta_u(r) \stackrel{\varnothing}{=} \beta_u^\top r \in \mathbb{R}$

$\boxed{[T2.5 : U^\top A(r) U]}$ $U^\top A(r) U = \left(\bigoplus_{u=1}^{m} \theta_u(r) J \right) \oplus 0_{d_{\text{null}} \times d_{\text{null}}}$ [deps : $T1.3, T2.3, T2.4, T0.17, T0.5$]

$\boxed{[T2.6 : \exp \text{blk}]}$ $\exp\big(U^\top A(r) U\big) = \left(\bigoplus_{u=1}^{m} R_2(\theta_u(r)) \right) \oplus I_{d_{\text{null}}}$ [deps : $T2.5, T0.17, T0.11, T0.5$]

$\boxed{[T2.7 : R_{\text{STR}} \text{ form}]}$ $R_{\text{STR}}(r) = U \left[\left(\bigoplus_{u=1}^{m} R_2(\theta_u(r)) \right) \oplus I_{d_{\text{null}}} \right] U^\top$ [deps : $T1.6, T2.6, T0.11, T2.3$]

$\boxed{[T2.8 : R_{\text{STR}} \in SO]}$ $R_{\text{STR}}(r) \in SO(d_h)$ [deps : $T2.7, T$
(uses $R_2(\theta) \in SO(2)$, block determinant multiplicativity, and $U \in O(d_h)$)

$\boxed{[T2.9 : \text{rel } R_{\text{STR}}]}$ $R_{\text{STR}}(r_i)^\top R_{\text{STR}}(r_j) = R_{\text{STR}}(r_j - r_i)$ [deps : $T1.4, T1.5, T1.6, T0.11$]

$\boxed{[T2.10 : \Pi_{\text{act}}, \Pi_{\text{null}}, d_{\text{act}}]}$ $\Pi_{\text{act}} \stackrel{\varnothing}{=} U\, \mathrm{diag}(I_{2m}, 0)\, U^\top$ [deps : $T2.3, T0.15, T0.2, T0.14, T0.6$]
$\Pi_{\text{null}} \stackrel{\varnothing}{=} U\, \mathrm{diag}(0, I_{d_{\text{null}}})\, U^\top$
$\Pi_{\text{act}}^\top = \Pi_{\text{act}},\ \Pi_{\text{act}}^2 = \Pi_{\text{act}}$
$\Pi_{\text{null}}^\top = \Pi_{\text{null}},\ \Pi_{\text{null}}^2 = \Pi_{\text{null}}$
$\Pi_{\text{act}} \Pi_{\text{null}} = 0 \Rightarrow \mathrm{im}(\Pi_{\text{act}}) \perp \mathrm{im}(\Pi_{\text{null}})$
$d_{\text{act}} \stackrel{\varnothing}{=} \mathrm{rank}(\Pi_{\text{act}}) = 2m$

$\boxed{[T2.11 : \Pi_{\text{act}} \text{ comm}]}$ $R_{\text{STR}}(r) \Pi_{\text{act}} = \Pi_{\text{act}} R_{\text{STR}}(r)$ [deps : $T2.7, T2.10, T0.5, T0.6$]

4 Tier 3: Cayley Post-Rotations

$\boxed{[T3.1 : \text{Cayley}]}$ $S \in \mathbb{R}^{d_h \times d_h},\ S^\top = -S \Rightarrow (I+S)$ invertible [deps : $T0.16, T0.3, T0.2$]

$$P_{\text{sp}} \overset{\varnothing}{=} (I-S)(I+S)^{-1}$$
$$(I-S)(I+S) = (I+S)(I-S) = I - S^2$$
(uses $\sigma(S) \subset i\mathbb{R}$ so $1 + i\mu \neq 0$)

$\boxed{[T3.2 : P_{\text{sp}} \in SO, \mathcal{B}]}$ $P_{\text{sp}}^\top = ((I+S)^{-1})^\top (I-S)^\top = I$ [deps : $T3.1, T0.16, T0.20, T0.7, T0.6, T0.4$

$$\det(P_{\text{sp}}) = \prod_u \frac{1 - i\mu_u}{1 + i\mu_u} = 1$$
$$\mathcal{B} \overset{\varnothing}{=} \{(I-S)(I+S)^{-1} \mid S^\top = -S\} \subseteq SO(d_h)$$

$\boxed{[T3.3 : \text{Cayley surj}]}$ $Q \in SO(d_h),\ -1 \notin \sigma(Q) \Rightarrow S_Q \overset{\varnothing}{=} (I-Q)(I+Q)^{-1}$ [deps : $T0.16, T0.20, T0.$

$$S_Q^\top = -S_Q, \quad (I - S_Q)(I + S_Q)^{-1} = Q$$
(Cayley transform bijection on $SO(d_h)$ with no -1 eigenvalues)

$\boxed{[T3.4 : \mathcal{A}]}$ $\mathcal{A} \overset{\varnothing}{=} \left\{ U \operatorname{diag}(I_{2m}, R_{\text{null}})\, U^\top \;\middle|\; R_{\text{null}} \in SO(d_{\text{null}}) \right\} \subseteq SO(d_h)$ [deps : $T2.3, T0.6, T0.7, T0.5, T0.4, T$

$$R_{\text{null}} \in SO(d_{\text{null}}),\ -1 \notin \sigma(R_{\text{null}}) \Rightarrow U \operatorname{diag}(I_{2m}, R_{\text{null}})\, U^\top \in \mathcal{B}$$

$\boxed{[T3.5 : R_{\text{sp}}]}$ $R_{\text{sp}}(r) \overset{\varnothing}{=} R_{\text{STR}}(r) P_{\text{sp}}$ [deps : $T2.8, T3.2, T0.6, T0.4$]

$$R_{\text{STR}}(r) \in SO(d_h),\ P_{\text{sp}} \in SO(d_h) \Rightarrow R_{\text{sp}}(r) \in SO(d_h)$$

$\boxed{[T3.6 : \Pi_{\text{act}} R_{\text{sp}} \Pi_{\text{act}}]}$ $P_{\text{sp}} \in \mathcal{A} \Rightarrow \Pi_{\text{act}} P_{\text{sp}} = \Pi_{\text{act}} = P_{\text{sp}} \Pi_{\text{act}}$ [deps : $T3.5, T2.10, T2.11, T3.4, T0.5, T0.6$]

$$R_{\text{STR}}(r) \Pi_{\text{act}} = \Pi_{\text{act}} R_{\text{STR}}(r)$$
$$\Rightarrow \Pi_{\text{act}} R_{\text{sp}}(r) \Pi_{\text{act}} = \Pi_{\text{act}} R_{\text{STR}}(r) \Pi_{\text{act}}$$

$\boxed{[T3.7 : \Pi\text{-rel } R_{\text{sp}}]}$ $P_{\text{sp}} \in \mathcal{A} \Rightarrow \Pi_{\text{act}} R_{\text{sp}}(r_i)^\top R_{\text{sp}}(r_j) \Pi_{\text{act}} = \Pi_{\text{act}} R_{\text{STR}}(r_j - r_i) \Pi_{\text{act}}$ [deps : $T3.6, T2.9, T3.5, T3.2,$

5 Tier 4: Attention Queries, Keys, and Scores

$\boxed{[T4.1 : W, q, k, v]}$ $W_Q, W_K, W_V \in \mathbb{R}^{d_h \times D}$ [deps : $T0.1, T0.3$]

$$x_i \in \mathbb{R}^D,\ r_i \in \mathbb{R}^{d_c}$$
$$q_i = W_Q x_i, \quad k_j = W_K x_j, \quad v_j = W_V x_j$$

$\boxed{[T4.2 : q^{(\text{act})}, k^{(\text{act})}, d_{\text{act}}]}$ $q_i^{(\text{act})} \overset{\varnothing}{=} \Pi_{\text{act}} q_i, \quad k_j^{(\text{act})} \overset{\varnothing}{=} \Pi_{\text{act}} k_j$ [deps : $T2.10, T4.1$]

$$d_{\text{act}} \overset{\varnothing}{=} 2m$$

$$\boxed{[T4.3 : \tilde{q}, \tilde{k}]}\ \tilde{q}_i = \Pi_{\text{act}} R_{\text{sp}}(r_i) q_i^{(\text{act})} \qquad [\text{deps} : T3.5, T4.2, T2.10]$$
$$\tilde{k}_j = \Pi_{\text{act}} R_{\text{sp}}(r_j) k_j^{(\text{act})}$$

$$\boxed{[T4.4 : \alpha_{ij} \text{ def}]} \qquad \alpha_{ij} \triangleq \frac{\tilde{q}_i^\top \tilde{k}_j}{\sqrt{d_{\text{act}}}} \qquad [\text{deps} : T4.3, T4.2, T2.10, T0.3]$$
$$= \frac{1}{\sqrt{d_{\text{act}}}} (q_i^{(\text{act})})^\top \big(\Pi_{\text{act}} R_{\text{sp}}(r_i) \Pi_{\text{act}}\big)^\top \big(\Pi_{\text{act}} R_{\text{sp}}(r_j) \Pi_{\text{act}}\big) k_j^{(\text{act})}$$
$$(\text{uses } \Pi_{\text{act}}^\top = \Pi_{\text{act}} = \Pi_{\text{act}}^2)$$

$$\boxed{[T4.5 : \alpha_{ij} \text{ rel}]}\ P_{\text{sp}} \in \mathcal{A} \Rightarrow \alpha_{ij} = \frac{1}{\sqrt{2m}} (q_i^{(\text{act})})^\top \Big(\Pi_{\text{act}} R_{\text{STR}}(r_j - r_i) \Pi_{\text{act}}\Big) k_j^{(\text{act})} \qquad [\text{deps} : T4.4, T3.6, T3.7, T4.2, T$$

6 Tier 5: Stability Estimates

$$\boxed{[T5.1 : \varepsilon \text{ comm}]} \qquad \varepsilon_{ab} \triangleq |[L_a, L_b]|, \quad \varepsilon \triangleq \max_{a,b} \varepsilon_{ab} \qquad [\text{deps} : T1.3, T1.1, T0.10, T0.9]$$
$$\big|[A(r), A(s)]\big| \le \sum_{a,b} |r_a||s_b| \varepsilon_{ab} \le |r|_1 |s|_1 \, \varepsilon$$

$$\boxed{[T5.2 : \text{BCH err}]}\ \big|\log\big(\exp(A(r)) \exp(A(s))\big) - (A(r) + A(s))\big| \le \tfrac{1}{2} |[A(r), A(s)]| + C |[A(r), A(s)]|^2 \qquad [\text{deps} : T0.1$$

$$\boxed{[T5.3 : R_{\text{STR}} \text{ rel approx}]}\ \big|R_{\text{STR}}(r)^\top R_{\text{STR}}(s) - R_{\text{STR}}(s - r)\big| \le C \, \varepsilon \, |r|_1 |s|_1 + O(\varepsilon^2) \qquad [\text{deps} : T1.6, T1.4, T1.5, T5$$

$$\boxed{[T5.4 : S = S_- + E]}\ S = S_- + E, \quad S_-^\top = -S_-, \quad |E| \le \eta, \quad 0 \le \eta < 1 \qquad [\text{deps} : T0.16, T0.9]$$

$$\boxed{[T5.5 : \sigma_{\min}(I + S_-)]} \qquad S_-^\top = -S_- \Rightarrow \sigma(S_-) \subset i\mathbb{R} \qquad [\text{deps} : T5.4, T0.16, T0.8]$$
$$\Rightarrow |1 + i\mu| \ge 1 \text{ for } \lambda = i\mu \in \sigma(S_-)$$
$$\sigma_{\min}(I + S_-) \ge 1$$

$$\boxed{[T5.6 : \sigma_{\min}(I + S)]}\ \sigma_{\min}(I + S) \ge \sigma_{\min}(I + S_-) - |E| \ge 1 - \eta > 0 \qquad [\text{deps} : T5.4, T5.5, T0.19, T0.9]$$
$$\Rightarrow I + S \text{ invertible}$$

$$\boxed{[T5.7 : \text{Cayley Lip}]}\ P_{\text{sp}}(S) \triangleq (I - S)(I + S)^{-1}, \quad P_{\text{sp}}(S_-) \triangleq (I - S_-)(I + S_-)^{-1} \qquad [\text{deps} : T5.6, T5.4, T0.20, T0$$
$$|P_{\text{sp}}(S) - P_{\text{sp}}(S_-)| \le C' \eta$$

$\boxed{[T5.8 : \eta_{\text{mix}}]}\ P_{\text{sp}} = U \begin{bmatrix} A & B \\ C & D \end{bmatrix} U^\top, \quad \eta_{\text{mix}} \overset{\varnothing}{=} |B| + |C| \qquad [\text{deps} : T3.2, T2.3, T2.10, T0.9, T0.6, T0.5, T0.7]$

$$P_{\text{sp}} \in SO(d_h) \Rightarrow \begin{bmatrix} A & B \\ C & D \end{bmatrix}^\top \begin{bmatrix} A & B \\ C & D \end{bmatrix} = I$$
$$\Rightarrow |A - I_{2m}| \le C'' \eta_{\text{mix}}$$
$$\Rightarrow |\Pi_{\text{act}} P_{\text{sp}} \Pi_{\text{act}} - \Pi_{\text{act}}| \le C'' \eta_{\text{mix}}$$

$\boxed{[T5.9 : \Pi R_{\text{sp}} \Pi \text{ approx}]}\ \left|\Pi_{\text{act}} R_{\text{sp}}(r) \Pi_{\text{act}} - \Pi_{\text{act}} R_{\text{STR}}(r) \Pi_{\text{act}}\right| \le C'' \eta_{\text{mix}} \qquad [\text{deps} : T3.5, T2.11, T5.8, T2.10, T0.9, T$

$\boxed{[T5.10 : \alpha_{ij} \text{ stab}]}\ \Delta_{ij} \overset{\varnothing}{=} C\big(\varepsilon |r_i|_1 |r_j|_1 + \eta_{\text{mix}}\big) |q_i^{(\text{act})}|\, |k_j^{(\text{act})}| \qquad [\text{deps} : T4.4, T5.9, T5.3, T4.2, T5$

$$\left| \alpha_{ij} - \frac{1}{\sqrt{2m}} (q_i^{(\text{act})})^\top \Big(\Pi_{\text{act}} R_{\text{STR}}(r_j - r_i) \Pi_{\text{act}} \Big) k_j^{(\text{act})} \right| \le \Delta_{ij}$$

7 Tier 6: Conclusions

$\boxed{[T6.1 : \text{GOAL}]}\ \big(T1.1 \wedge T2.3 \wedge P_{\text{sp}} \in \mathcal{A}\big) \Rightarrow \alpha_{ij} = \frac{1}{\sqrt{2m}} (q_i^{(\text{act})})^\top \Big(\Pi_{\text{act}} R_{\text{STR}}(r_j - r_i) \Pi_{\text{act}} \Big) k_j^{(\text{act})} \qquad [\text{deps} : T4.5, T$

$\boxed{[T6.2 : \text{GOAL}_{\text{rob}}]}\ \Big(|[L_a, L_b]| \le \varepsilon\ \forall a, b, \quad P_{\text{sp}} \in SO(d_h) \text{ with mixing } \eta_{\text{mix}} \Big) \Rightarrow \alpha_{ij} \approx_{\Delta_{ij}} \frac{1}{\sqrt{2m}} (q_i^{(\text{act})})^\top \Big(\Pi_{\text{act}} R_{\text{STR}}($

$$\Delta_{ij} \le C\big(\varepsilon |r_i|_1 |r_j|_1 + \eta$$

8 Tier 6′: Quantified Robustness

$\boxed{[T6'.0 : \varepsilon_{\text{comm}}, \Lambda, R, S, S_-, E, \rho, \eta, \eta_{\text{mix}}]}\ \varepsilon_{\text{comm}} \overset{\varnothing}{=} \max_{a,b \in K} \big|[L_a, L_b]\big| \qquad [\text{deps} : T1.1, T2.3, T2.4,$

$$\Lambda \overset{\varnothing}{=} \max_{u \in \{1,\dots,m\}} \sum_{k=1}^{d_c} |\lambda_{k,u}|$$
$$R \overset{\varnothing}{=} \max_i |r_i|_1$$
$$S \overset{\varnothing}{=} S_- + E, \quad S_-^\top = -S_-, \quad E^\top = -E$$
$$\rho \overset{\varnothing}{=} |S|, \quad \eta \overset{\varnothing}{=} |E|$$
$$\eta_{\text{mix}} \overset{\varnothing}{=} \max \Big\{ \big|\Pi_{\text{null}} P_{\text{sp}} \Pi_{\text{act}}\big|,\ \big|\Pi_{\text{act}} P_{\text{sp}} \Pi_{\text{null}}\big| \Big\}$$

$\boxed{[T6'.1 : |A(r)|, \text{BCH const}]}\ |A(r)| \overset{T2.5}{\le} \sum_{u=1}^{m} |\theta_u(r)| \overset{T2.4}{=} \sum_{u=1}^{m} \Big| \sum_{k=1}^{d_c} r_k \lambda_{k,u} \Big| \le \Lambda\, |r|_1 \qquad [\text{deps} : T2.5, T2.4, T0.9, T0.12]$

$$C_{\text{BCH}}(r, s) \overset{\varnothing}{=} C_\star \exp\big(2\Lambda(|r|_1 + |s|_1)\big)$$

$\boxed{[T6'.2 : \text{BCH quant}]}$ $\left|\log\big(e^{A(r)}e^{A(s)}\big) - (A(r)+A(s))\right| \overset{T0.13}{\leq} \frac{1}{2}|[A(r),A(s)]| + C|[A(r),A(s)]|^2$ [deps : $T0.13, T5$

$\overset{T5.1}{\leq} \frac{1}{2}\,\varepsilon_{\text{comm}}|r|_1|s|_1 + C_{\text{BCH}}(r,s)\,\varepsilon_{\text{comm}}^2|r|_1^2|s|_1^2$

$\boxed{[T6'.3 : \text{Rel approx } R_{\text{STR}}]}$ $R_{\text{STR}}(r)^\top R_{\text{STR}}(s) \overset{T1.6}{=} e^{-A(r)}e^{A(s)}$ [deps :

$\log\big(R_{\text{STR}}(r)^\top R_{\text{STR}}(s)\big) \overset{T0.11}{=} \log\big(e^{-A(r)}e^{A(s)}\big) \overset{T6'.2}{=} A(s) - A(r) + \Delta_{rs}$

$A(s) - A(r) \overset{T1.3}{=} A(s-r)$

$\Rightarrow \left|R_{\text{STR}}(r)^\top R_{\text{STR}}(s) - R_{\text{STR}}(s-r)\right| \leq C_1\,\varepsilon_{\text{comm}}|r|_1|s|_1 + C_2\,\varepsilon_{\text{comm}}^2|r|_1^2|s|_1^2$

$\boxed{[T6'.4 : \sigma_{\min}(I+S), |(I+S)^{-1}|]}$ $\sigma_{\min}(I+S) \overset{T0.19}{\geq} \sigma_{\min}(I) - |S| = 1-\rho$ [deps : $T0.19, T0.8, T0.9$]

$|(I+S)^{-1}| = \sigma_{\min}(I+S)^{-1} \leq (1-\rho)^{-1}$

$\boxed{[T6'.5 : \text{Cayley Lip}]}$ $P_{\text{sp}}(S) \overset{\varnothing}{=} (I-S)(I+S)^{-1}$ [deps : $T3$

$P_{\text{sp}}(S_1) - P_{\text{sp}}(S_2) = (S_2-S_1)(I+S_1)^{-1} + (I-S_2)\big[(I+S_2)^{-1} - (I+S_1)^{-1}\big]$

$(I+S_2)^{-1} - (I+S_1)^{-1} = (I+S_2)^{-1}(S_1-S_2)(I+S_1)^{-1}$

$\Rightarrow |P_{\text{sp}}(S_1) - P_{\text{sp}}(S_2)| \leq \dfrac{2}{\sigma_{\min}(I+S_1)\,\sigma_{\min}(I+S_2)}\,|S_1-S_2|$

$\Rightarrow |P_{\text{sp}}(S) - P_{\text{sp}}(S_-)| \leq \dfrac{2}{(1-\rho)^2}\,\eta$

$\boxed{[T6'.6 : \text{Neumann}]}$ $|S| = \rho < 1 \Rightarrow (I+S)^{-1} = \sum_{t=0}^{\infty}(-S)^t$ [deps : $T0.9, T0.12$]

$P_{\text{sp}}(S) = (I-S)\sum_{t=0}^{\infty}(-S)^t = I - 2\sum_{u=1}^{\infty}(-1)^{u-1}S^u$

$\boxed{[T6'.7 : \text{QuadMix } \Pi_{\text{act}}]}$ $\Pi_{\text{act}}S\Pi_{\text{act}} = 0, \quad \Pi_{\text{null}}S\Pi_{\text{null}} = S_{\text{null}}$ [deps

$\Pi_{\text{act}}S^2\Pi_{\text{act}} = (\Pi_{\text{act}}S\Pi_{\text{null}})(\Pi_{\text{null}}S\Pi_{\text{act}}) \Rightarrow |\Pi_{\text{act}}S^2\Pi_{\text{act}}| \leq \eta^2$

$\Pi_{\text{act}}P_{\text{sp}}(S)\Pi_{\text{act}} - \Pi_{\text{act}} = -2\sum_{u=1}^{\infty}(-1)^{u-1}\Pi_{\text{act}}S^u\Pi_{\text{act}} = -2\sum_{u=2}^{\infty}(-1)^{u-1}\Pi_{\text{act}}S^u\Pi_{\text{act}}$

$\Rightarrow \left|\Pi_{\text{act}}P_{\text{sp}}(S)\Pi_{\text{act}} - \Pi_{\text{act}}\right| \leq 2\sum_{u=2}^{\infty}\left|\Pi_{\text{act}}S^u\Pi_{\text{act}}\right| \leq 2\sum_{u=2}^{\infty}\rho^{u-2}\,|\Pi_{\text{act}}S^2\Pi_{\text{act}}|$

$\leq \dfrac{2}{1-\rho}\,\eta^2 \leq \dfrac{2}{(1-\rho)^3}\,\eta^2$

$\boxed{[T6'.8 : \text{Mix model-free}]}$ $P_{\text{sp}} = U\begin{bmatrix}A & B\\ C & D\end{bmatrix}U^\top, \quad A^\top A + C^\top C = I, \quad AA^\top + BB^\top = I$ [deps : $T3.2, T2.3, T2.$

$|A - I| \leq 2\max\{|B|,|C|\} = 2\,\eta_{\text{mix}}$

$\Rightarrow \left|\Pi_{\text{act}}P_{\text{sp}}\Pi_{\text{act}} - \Pi_{\text{act}}\right| \leq 2\,\eta_{\text{mix}}$

$\boxed{[T6'.9 : \text{Rob } R_{\text{sp}} \text{ on act}]}$ $\big|\Pi_{\text{act}} R_{\text{sp}}(r)\Pi_{\text{act}} - \Pi_{\text{act}} R_{\text{STR}}(r)\Pi_{\text{act}}\big| = \big|\Pi_{\text{act}} R_{\text{STR}}(r)\big(P_{\text{sp}} - \Pi_{\text{act}} - \Pi_{\text{null}}\big)\Pi_{\text{act}}\big|$ [deps :

$$\overset{T0.9}{\leq} \big|P_{\text{sp}} - \Pi_{\text{act}} - \Pi_{\text{null}}\big| \leq \begin{cases} \dfrac{2}{(1-\rho)^3}\,\eta^2, & \text{(ESPR/Cayley/pure mix)}, \\ 2\,\eta_{\text{mix}}, & \text{(model-free)}. \end{cases} =: \delta_{\text{mix}}$$

$\boxed{[T6'.10 : \text{Logit err}]}$ $\alpha_{ij} \overset{T4.4}{=} \frac{1}{\sqrt{2m}} (q_i^{(\text{act})})^\top \Big(\Pi_{\text{act}} R_{\text{sp}}(r_i)^\top R_{\text{sp}}(r_j)\Pi_{\text{act}}\Big) k_j^{(\text{act})}$ [de

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

$$|\alpha_{ij} - \alpha_{ij}^\star| \leq \tfrac{1}{\sqrt{2m}} |q_i^{(\text{act})}|\,|k_j^{(\text{act})}|$$

$$\cdot \Big(\big|\Pi_{\text{act}} \big(R_{\text{sp}}(r_i)^\top R_{\text{sp}}(r_j) - R_{\text{STR}}(r_j - r_i)\big)\Pi_{\text{act}}\big| + \big|\Pi_{\text{act}} R_{\text{sp}}(\cdot)\Pi_{\text{act}} - \Pi_{\text{act}} R_{\text{STR}}(\cdot)\Pi_{\text{act}}\big| \Big)$$

$$\leq |q_i^{(\text{act})}|\,|k_j^{(\text{act})}| \Big(C_1\,\varepsilon_{\text{comm}} |r_i|_1 |r_j|_1 + C_2\,\varepsilon_{\text{comm}}^2 |r_i|_1^2 |r_j|_1^2 + \delta_{\text{mix}} \Big)$$

$\boxed{[T6'.11 : \text{Rob theorem}]}$ $\big|\alpha_{ij} - \alpha_{ij}^\star\big| \leq \Big(\frac{1}{2} + C_{\text{BCH}}(r_i, r_j)\Big) \varepsilon_{\text{comm}} |r_i|_1 |r_j|_1\, |q_i^{(\text{act})}|\,|k_j^{(\text{act})}|$ [deps : $T6'.1, T6'.3, T6'.9$

$$+ C_{\text{mix}} \delta_{\text{mix}} |q_i^{(\text{act})}|\,|k_j^{(\text{act})}|$$

$$\begin{cases} C_{\text{mix}} = 1,\ \delta_{\text{mix}} = \eta_{\text{mix}}, & \text{(model-free)}, \\ C_{\text{mix}} = \dfrac{2}{(1-\rho)^3},\ \delta_{\text{mix}} = \eta^2, & \text{(ESPR/Cayley/pure mix)}. \end{cases}$$

$\boxed{[T6'.12 : \text{Replacements}]}$ $\{T5.2,\ T5.3\} \rightsquigarrow \{T6'.2,\ T6'.3\}$ [deps : $T5.2, T5.3, T5.7, T5.8, T5.9, T6.2, T6'.2, T6'.3,$

$$\{T5.7\} \rightsquigarrow \{T6'.5\}$$

$$\{T5.8,\ T5.9\} \rightsquigarrow \{T6'.7,\ T6'.9\}$$

$$\{T6.2\} \rightsquigarrow \{T6'.11\}$$

9 Tier 7: Positive Random Features and Variance Bounds

$\boxed{[T7.0 : \mathcal{N}(0, I_n), \mathbb{E}, \text{Var}]}$ $\mathcal{N}(0, I_n) \overset{\varnothing}{=}$ the mean-zero Gaussian law on $\mathbb{R}^n$ with density $(2\pi)^{-n/2} \exp\big(-\frac{1}{2}|w|^2\big)$ [

$$w \sim \mathcal{N}(0, I_n) \Rightarrow \mathbb{E}[w] = 0, \quad \mathbb{E}[w w^\top] = I_n$$

$$\mathbb{E}[\cdot] \overset{\varnothing}{=} \text{expectation with respect to the (joint) law of all Gaussian draws } w$$

$$\text{Var}[Z] \overset{\varnothing}{=} \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \quad \text{for any scalar random variable } Z$$

[$T7.1$: RF setup]

$d_{\text{act}} = 2m$ from T2.10

For probabilistic estimates we draw $w \sim \mathcal{N}(0, I_{d_{\text{act}}})$ in the sense of T7.0

$$\phi_w(x) \overset{\varnothing}{=} \exp\Big(w^\top x - \tfrac{1}{2}|x|^2\Big) \quad (x \in \mathbb{R}^{d_{\text{act}}})$$

(positive softmax random feature map as in FAVOR + estimators)

$$Z(x, y; w) \overset{\varnothing}{=} \phi_w(x)\,\phi_w(y) = \exp\Big(w^\top (x+y) - \tfrac{1}{2}(|x|^2 + |y|^2)\Big) \quad (x, y \in \mathbb{R}^{d_{\text{act}}})$$

$$\widehat{\text{SM}}_{n_{\text{rf}}}(x, y) \overset{\varnothing}{=} \frac{1}{n_{\text{rf}}} \sum_{t=1}^{n_{\text{rf}}} Z(x, y; w_t) \quad \text{for } n_{\text{rf}} \in \mathbb{N},\ w_t \text{ i.i.d. } \mathcal{N}(0, I_{d_{\text{act}}})$$

Note: $\widehat{\text{SM}}_{n_{\text{rf}}}(x, y)$ is our Monte Carlo estimator of $\exp(x^\top y)$ (unnormalized softmax kernel)

[$T7.2$: $\widehat{q}_i, \widehat{k}_j, \alpha_{ij}$]

$\tilde{q}_i, \tilde{k}_j$ from T4.3,

$$\alpha_{ij} \overset{\varnothing}{=} \frac{1}{\sqrt{d_{\text{act}}}} \tilde{q}_i^\top \tilde{k}_j$$

$$\widehat{q}_i \overset{\varnothing}{=} d_{\text{a}}^{-}$$

$$\widehat{q}_i^\top \widehat{k}_j = (d_{\text{act}}^{-1/4} \tilde{q}_i)^\top (d_{\text{act}}^{-1/4}$$

(The attention logit α_{ij} is *exactly* a Euclidean dot product between $\widehat{q}_i, \widehat{k}_j \in \mathbb{R}^{d_{\text{act}}}$, with no ren

[$T7.3$: $\mathbb{E}Z = \exp(x^\top y)$, unbiasedness]

Note: $\widehat{\text{SM}}_{n_{\text{rf}}}(\widehat{q}_i, \widehat{k}_j)$ is an *unbiased* positive estimator of $\exp(\alpha_{ij})$, the *unn*

$[T7.4 : \mathbb{E}Z^2, \mathrm{Var}Z, \text{closed form}]$

Continue with fixed $x, y \in \mathbb{R}$

$$Z(x,y;w) = \exp\Big(w^\top(x+y) - \tfrac{1}{2}(|x|^2 + |y$$

$$Z(x,y;w)^2 = \exp\Big(2w^\top(x+y) - (|x|^2 + |y$$

$$\mathbb{E}[Z(x,y;w)^2] = \exp\Big(\tfrac{1}{2}|2(x+y)|^2 - (|x|^2+|y|^2)\Big) = \exp\Big(2|x+y|^2 - (|x|^2 + |y$$

$$|x+y|^2 = |x|^2 + |y|^2 + 2x^\top y \text{ (as used in T}$$

$$\Rightarrow 2|x+y|^2 - (|x|^2+|y|^2) = 2(|x|^2+|y|^2+2x^\top y) - (|x|^2+|y|^2) = |x|^2 + |y|^2 + 4$$

$$\Rightarrow \mathbb{E}[Z(x,y;w)^2] = \exp\Big(|x|^2 + |y|^2 + 4x$$

$$\mathbb{E}[Z(x,y;w)] = \exp(x^\top y) \quad \text{from}$$

$$(\mathbb{E}[Z(x,y;w)])^2 = \exp(2x$$

$$\mathrm{Var}[Z(x,y;w)] = \mathbb{E}[Z(x,y;w)^2] - (\mathbb{E}[Z(x,y;w$$

$$= \exp\Big(|x|^2 + |y|^2 + 4x^\top y\Big) - \exp\Big(2x$$

$$\mathrm{Var}[Z(x,y;w)] = \exp\Big(2x^\top y\Big)\Big[\exp\Big(|x|^2 + |y|^2 + 2x^\top y\Big)$$

Note: The single-sample variance is finite and given in closed form for all finite

$[T7.5 : \text{Var } \widehat{\text{SM}}_{n_{\text{rf}}}, \text{geom link}]$ The draws w_t in T7.1 are i.i.d., so the $Z(\cdot; w_t)$ ar

$$\widehat{\text{SM}}_{n_{\text{rf}}}(x, y) = \frac{1}{n_{\text{rf}}} \sum_{t=1}^{n_{\text{rf}}} Z(x,$$

$$\text{Var}\Big[\widehat{\text{SM}}_{n_{\text{rf}}}(x, y)\Big] = \frac{1}{n_{\text{rf}}^2} \sum_{t=1}^{n_{\text{rf}}} \text{Var}[Z(x, y; w_t)] = \frac{1}{n_{\text{rf}}} \text{Var}[Z(x, y; w)] \quad \text{(i.i.d. variance sc}$$

Insert the expression from

$$\text{Var}\Big[\widehat{\text{SM}}_{n_{\text{rf}}}(x, y)\Big] = \frac{1}{n_{\text{rf}}} \exp\Big(2x^\top y\Big) \Big[\exp\Big(|x|^2 + |y|^2 + 2x^\top y$$

Now specialise $x = \widehat{q}_i,\ y = \widehat{k}_j$ from T7.2. Then $x^\top y$

$$\Rightarrow \text{Var}\Big[\widehat{\text{SM}}_{n_{\text{rf}}}(\widehat{q}_i, \widehat{k}_j)\Big] = \frac{1}{n_{\text{rf}}} \exp\Big(2\alpha_{ij}\Big) \Big[\exp\Big(|\widehat{q}_i|^2 + |\widehat{k}_j|^2 + 2\alpha_{ij}$$

Relate $|\widehat{q}_i|, |\widehat{k}_j|$ back to geometric quantities already controlled in

$$\widehat{q}_i = d_{\text{act}}^{-1/4} \tilde{q}_i, \quad \widehat{k}_j = d_{\text{act}}^{-1/4} \tilde{k}_j \text{ from T7.2} \Rightarrow |\widehat{q}_i|^2 + |\widehat{k}_j|^2 = d_{\text{act}}^{-1/2}(|\tilde{q}_i|^2 +$$

$$\tilde{q}_i = \Pi_{\text{act}} R_{\text{sp}}(r_i) q_i^{(\text{act})} \text{ and } \tilde{k}_j = \Pi_{\text{act}} R_{\text{sp}}(r_j) k_j^{(\text{act})} \text{ from}$$

$R_{\text{sp}}(r) \in SO(d_h)$ (T3.5) so $|R_{\text{sp}}(r)z| = |z|$ for all z Π_{act} is an orthogonal projector (T2.10) so $|\Pi_{\text{act}} z| \le |z|$ for

$$\Rightarrow |\tilde{q}_i| \le |q_i^{(\text{act})}|, \quad |\tilde{k}_j| \le |$$

$$\Rightarrow |\tilde{q}_i|^2 + |\tilde{k}_j|^2 \le |q_i^{(\text{act})}|^2 + |k$$

(1) The estimator $\widehat{\text{SM}}_{n_{\text{rf}}}(\widehat{q}_i, \widehat{k}_j)$ is unbiased for $\exp(\alpha_{ij})$

10 Tier 8′: Tight Assumptions and OOD Dominance

$[T8'.0 : |\cdot|_2, |\cdot|_F,\ \mathbb{S}^{d_h-1}]$ $|M|_2 \overset{\varnothing}{=} |M|$ (operator norm from T0.9), $\quad |M|_F \overset{\varnothing}{=} \sqrt{\text{tr}(M^\top M)}$ $\quad$ [deps : $T0.9, T0.3$]

$$\mathbb{S}^{d_h-1} \overset{\varnothing}{=} \{z \in \mathbb{R}^{d_h} \mid |z| = 1\}$$

$[T8'.1 : (\mathcal{X}, \mathcal{Y}),\ \mathcal{D},\ \mathcal{T},\ \Phi_{\#}\mathcal{D}]$ $\mathcal{X}, \mathcal{Y}$ measurable, $(X, Y) \sim \mathcal{D}$

$\mathcal{T} \subseteq \{\Phi : \mathcal{X} \to \mathcal{X}$ bijective, measurable$\}$

$\Phi_{\#}\mathcal{D}$ (pushforward) satisfies $(X', Y') \sim \Phi_{\#}\mathcal{D} \Leftrightarrow (X', Y') = (\Phi \cdot X, Y),\ (X, Y) \sim \mathcal{D}$

$$(\Phi \cdot x)(u) \overset{\varnothing}{=} x(\Phi^{-1} u)$$

$[T8'.1bis : \text{Parameterization map}]$ $\exists\ \Pi_\Delta : \mathcal{T} \to \text{supp}(\nu)$ measurable s.t. $\Pi_{\text{act}} R_{\text{STR}}(\Pi_\Delta(\Phi)) \Pi_{\text{act}}$ matches the acti

$\boxed{[T8'.2 : \text{Label invariance}]}\ \forall \Phi \in \mathcal{T} : \ \Pr_{(X,Y)\sim\mathcal{D}}\big[Y' = Y \text{ where } (X',Y') \sim \Phi_{\#}\mathcal{D}\big] = 1 \qquad [\text{deps} : T8'.1]$

$\boxed{[T8'.3 : \psi,\ \ell,\ \text{bounded logits},\ L]}\ \psi : \mathbb{R}^{d_h} \to \mathbb{R}^C, \quad \ell : \mathbb{R}^C \times \mathcal{Y} \to \mathbb{R}_+ \qquad [\text{dep}$

$\exists B_\psi > 0 : \ \sup_{x\in\mathcal{X}} |\psi(z_f(x))|_2 \le B_\psi \ \text{(spectral control or fixed LayerNorm)}$

$\exists L > 0 : \ \forall y,\ \ell(\cdot, y) \text{ is } L\text{-Lipschitz on } \{u \in \mathbb{R}^C : |u|_2 \le B_\psi\}$

$\boxed{[T8'.4 : \ z_f,\ \text{measurable},\ B_z]}\ z_f : \mathcal{X} \to \mathbb{R}^{d_h} \text{ measurable}, \quad \sup_{x\in\mathcal{X}} |z_f(x)| \le B_z \qquad [\text{deps} : T0.9]$

$\boxed{[T8'.5 : \ \nu,\ \Delta,\ R_\Delta]}\ \nu : \text{ prob. law on } \mathbb{R}^{d_c}, \quad \Delta \sim \nu, \quad |\Delta|_1 \le R_\Delta \text{ a.s.} \qquad [\text{deps} : T0.1, T0.9]$

$\boxed{[T8'.5bis : \ r\text{-radius}]}\ r \in \mathbb{R}^{d_c} \text{ ranges over } \{r : \ |r|_1 \le R\},\ R > 0 \text{ fixed} \qquad [\text{deps} : T0.9]$

$\boxed{[T8'.6 : \ \mathcal{H}_{\text{STRING}},\ \mathcal{H}_{\text{ESPR}}]}\ \mathcal{H}_{\text{STRING}} \overset{\varnothing}{=} \{f : \ P_{\text{sp}} \in \mathcal{A},\ [L_a, L_b] = 0\} \qquad [\text{deps} : T3.4, T$

$\mathcal{H}_{\text{ESPR}} \overset{\varnothing}{=} \{f : \ P_{\text{sp}} \in SO(d_h),\ |[L_a, L_b]| \le \varepsilon_{\text{comm}},\ \eta_{\text{mix}} \text{ as in T6'.0}\}$

$\mathcal{H}_{\text{STRING}} \subset \mathcal{H}_{\text{ESPR}}$

$\boxed{[T8'.7 : \ \text{IR}_{\text{spec}}(f)]}\ \ \Pi_{\text{act}} \in \mathbb{R}^{d_h \times d_h} \text{ is the orthogonal projector onto the first } 2m \text{ coordinates (T2.10)},\ \Pi_{\text{null}} = I$

$R_{\text{STR}}(\cdot)$ acts as block 2×2 rotations on m planes with angles from β (identity on null block;

$R_{\text{sp}}(r) \overset{\varnothing}{=} P_{\text{sp}} R_{\text{STR}}(r) P_{\text{sp}}^\top \text{ for } P_{\text{sp}} \in SO(d_h)$

$\text{IR}_{\text{spec}}(f) \overset{\varnothing}{=} \mathbb{E}_{\Delta\sim\nu} \ \sup_{|r|_1 \le R} \Big\| \Pi_{\text{act}} R_{\text{sp}}(r)^\top R_{\text{sp}}(r+\Delta) \Pi_{\text{act}} - \Pi_{\text{act}} R_{\text{STR}}(\Delta)$

$\boxed{[T8'.8 : \ C_{\text{comm}}, C_{\text{mix}}, C_{\mathcal{T}},\ \rho]}\ C_{\text{comm}} \overset{\varnothing}{=} \tfrac{1}{2}\,\varepsilon_{\text{comm}}(R + R_\Delta)^2 + C_{\text{BCH}}(2R + R_\Delta)\,\varepsilon_{\text{comm}}^2(R + R_\Delta)^4 \qquad [\text{deps} : T6'.1, T$

$\rho \overset{\varnothing}{=} |S|_2 < 1 \text{ for } P_{\text{sp}} = \text{cayley}(S) \text{ (Neumann regime, T6'.6)}$

$$C_{\text{mix}} \overset{\varnothing}{=} \begin{cases} \dfrac{2}{(1-\rho)^3}\,\eta^2, & \text{(ESPR/Cayley/pure mix)}, \\ \eta_{\text{mix}}, & \text{(model-free)} \end{cases}$$

$C_{\mathcal{T}} \overset{\varnothing}{=} 1$

$\boxed{[T8'.8bis : \ \text{Small-angle/BCH regime}]}\ \varepsilon_{\text{comm}},\ \eta_{\text{mix}},\ R,\ R_\Delta \text{ are small so that higher-order BCH terms are absorb}$

$\boxed{[T8'.9 : \ \Delta_{\mathcal{T}}(f)]}\ \Delta_{\mathcal{T}}(f) \overset{\varnothing}{=} \mathbb{E}_{\Phi\sim\mu}\ \mathbb{E}_{(X,Y)\sim\mathcal{D}}\ \big|z_f(\Phi \cdot X) - z_f(X)\big| \qquad [\text{deps} : T8'.1, T8'.4]$

$\boxed{[T8'.10:\ \text{Geom} \Rightarrow \text{IR link (tight } C_z)]}$ $z' = U'z,\ z = Uz,\ |z| \le B_z \Rightarrow |z'-z| = |(U'-U)z| \le |U'-U|_2\, B_z$ [d

$C_z \overset{\varnothing}{=} B_z \Rightarrow \Delta_{\mathcal{T}}(f) \le C_z\, \mathrm{IR}_{\mathrm{spec}}(f)$

$\boxed{[T8'.11:\ R_{\mathrm{train}},\ R_{\mathcal{T}}]}$ $R_{\mathrm{train}}(f) \overset{\varnothing}{=} \mathbb{E}_{(X,Y)\sim\mathcal{D}}\ell(\psi(z_f(X)), Y)$ $[\text{deps} : T8'.3, T8'.4, T8'.1]$

$R_{\mathcal{T}}(f) \overset{\varnothing}{=} \mathbb{E}_{\Phi\sim\mu}\, \mathbb{E}_{(X,Y)\sim\mathcal{D}}\ell(\psi(z_f(\Phi\cdot X)), Y)$

$\boxed{[T8'.12:\ \text{Shift-Lipschitz}]}$ $R_{\mathcal{T}}(f) - R_{\mathrm{train}}(f) \le L\,\Delta_{\mathcal{T}}(f) \le LC_z\,\mathrm{IR}_{\mathrm{spec}}(f)$ $[\text{deps} : T8'.3, T8'.10, T8'.11]$

$\boxed{[T8'.13:\ \text{Upper control of IR}_{\mathrm{spec}}]}$ $\mathrm{IR}_{\mathrm{spec}}(f) \le C_{\mathrm{comm}} + C_{\mathrm{mix}}$ $[\text{deps} : T6'.11, T8'.7, T8'.8, T8'.8bis]$

$\boxed{[T8'.14:\ c_{\mathcal{T}}\ (\text{noncommuting lower bound on active block})]}$ $m < d_h/2\ \wedge\ \exists a \neq b :\ [L_a, L_b] \neq 0$

$\exists \kappa_{\mathcal{T}} > 0 :\ \mathbb{E}_{\Delta\sim\nu} \inf_{|r|_1\le R} \big\| R_{\mathrm{STR}}(r)^\top R_{\mathrm{STR}}(r+\Delta) - R$

$\Rightarrow \inf_{f\in\mathcal{H}_{\mathrm{STRING}}} \mathrm{IR}_{\mathrm{spec}}(f) \ge$

$\boxed{[T8'.15:\ \text{Reachability in ESPR}]}$ $\forall \delta > 0\ \exists f_\delta \in \mathcal{H}_{\mathrm{ESPR}} :\ \varepsilon_{\mathrm{comm}} \le \delta,\ \eta_{\mathrm{mix}} \le \delta$ $[\text{deps} : T6'.11, T8'.6, T8'.7, T8'.$

$\Rightarrow\ \mathrm{IR}_{\mathrm{spec}}(f_\delta) \le C_{\mathrm{comm}}(\delta) + C_{\mathrm{mix}}(\delta) \xrightarrow{\delta\to 0} 0$

$\boxed{[T8'.16:\ \text{Infimum gap}]}$ $\inf_{f\in\mathcal{H}_{\mathrm{ESPR}}} R_{\mathcal{T}}(f) \le \inf_{f\in\mathcal{H}_{\mathrm{ESPR}}} \big(R_{\mathrm{train}}(f) + LC_z\,\mathrm{IR}_{\mathrm{spec}}(f)\big)$ $[\text{deps} : T8'.12, T8'.14, T8'.15, T$

$\le \inf_{f\in\mathcal{H}_{\mathrm{STRING}}} \big(R_{\mathrm{train}}(f) + LC_z\,\mathrm{IR}_{\mathrm{spec}}(f)\big) - LC_z\, c_{\mathcal{T}}$

$\le \inf_{f\in\mathcal{H}_{\mathrm{STRING}}} R_{\mathcal{T}}(f) - LC_z\, c_{\mathcal{T}}$

$\boxed{[T8'.17:\ \text{Optimization: Polyak–Łojasiewicz (PL)}]}$ $\exists \mu > 0,\ L_g > 0 :\ \forall H \in \{\mathcal{H}_{\mathrm{STRING}}, \mathcal{H}_{\mathrm{ESPR}}\},\ \mathcal{L}_H(f) \overset{\varnothing}{=} \widehat{R}_{\mathrm{train}}($

satisfies PL: $\frac{1}{2}\|\nabla\mathcal{L}_H(f)\|^2 \ge \mu\big(\mathcal{L}_H(f$

$\nabla\mathcal{L}_H$ is L_g-Lipschitz $\Rightarrow$ GD finds global minimizer $\hat{f}_{\lambda,H} \in$

$\boxed{[T8'.18:\ \mathfrak{C}_n(H,\delta)\ (\text{Rademacher; explicit scaling})]}$ $\exists S_0, \tilde{c} > 0 :\ \mathfrak{C}_n(H,\delta) \overset{\varnothing}{=} c\Big(\frac{S(H)}{\sqrt{n}} +$

$S(\mathcal{H}_{\mathrm{STRING}}) \le S_0, \quad S(\mathcal{H}_{\mathrm{ESPR}}) \le S_0 + \tilde{c}\cdot\big(\mathrm{rank}(E) + \varepsilon_{\mathrm{comm}}\big)$

$\boxed{[T8'.19:\ \text{Uniform generalization}]}$ $\Pr\Big[\ \forall f \in H :\ R_{\mathrm{train}}(f) \le \widehat{R}_{\mathrm{train}}(f) + \mathfrak{C}_n(H,\delta)\ \Big]\ \ge\ 1-\delta$ $[\text{deps} : T8'.18]$

$\boxed{[T8'.20:\ \text{Shift risk bound (uniform)}]}\ \Pr\Big[\ \forall f \in H:\ R_{\mathcal{T}}(f) \le \widehat{R}_{\text{train}}(f) + \mathfrak{C}_n(H,\delta) + LC_z\,\text{IR}_{\text{spec}}(f)\Big]\ \ge\ 1-\delta$

$\boxed{[T8'.21:\ \text{Achieved OOD gap}]}\quad \epsilon_0 \overset{\varnothing}{=} LC_z c_{\mathcal{T}},\quad \Delta_{\text{gen}} \overset{\varnothing}{=} \mathfrak{C}_n(\mathcal{H}_{\text{ESPR}},\delta) - \mathfrak{C}_n(\mathcal{H}_{\text{STRING}},\delta)\qquad [\text{deps}: T8'.16, T8'.1$

$$\Pr\Big[R_{\mathcal{T}}(\hat{f}_{\lambda,\mathcal{H}_{\text{ESPR}}}) \le R_{\mathcal{T}}(\hat{f}_{\lambda,\mathcal{H}_{\text{STRING}}}) - \epsilon_0 + \Delta_{\text{gen}}\Big] \ge 1-\delta$$

$\boxed{[T8'.22:\ \text{Worst-case corollary}]}\qquad R_{\sup}(f) \overset{\varnothing}{=} \sup_{\Phi\in\mathcal{T}} \mathbb{E}_{(X,Y)\sim\mathcal{D}}\ell(\psi(z_f$

$$R_{\sup}(f) - R_{\text{train}}(f) \le LC_z \sup_{\Delta\in\text{supp}(\nu)} \sup_{|r|_1\le R} \Big\|\Pi_{\text{act}} R_{\text{sp}}(r)^\top R_{\text{sp}}(r+\Delta)\Pi_{\text{act}} - \Pi_{\text{act}} R_{\text{S}}$$

$$\text{and}\ \sup_{\Phi\in\mathcal{T}} = \sup_{\Delta\in\text{supp}(\nu)}$$

$\boxed{[T8'.23:\ \text{No benefit when commuting (sharpness)}]}\ \text{If}\ \forall\Delta\in\text{supp}(\nu):\ [L_a,L_b]=0\ \wedge\ P_{\text{sp}}\in\mathcal{A} \Rightarrow \text{IR}_{\text{spec}}(f)=0$

$$\Rightarrow\ c_{\mathcal{T}}=0,\ \epsilon_0=0$$