BEYOND THE ATTENTION LOTTERY: EXPRESSIVE POWER AND TOPOLOGICAL COMPLEXITY VIA ALGEBRAIC GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study algebraic and geometric properties of simplified attention models. For exact computation of all pairwise inner products between two independently varying collections of N vectors in $\mathbb{R}^d$, we prove that a single-hidden-layer network with square activations requires at least $N^2 + 1$ hidden units and give an explicit construction using $2N^2d$ units. Thus the width is $\Theta(N^2)$ for fixed d, while the optimal dependence on a growing d is left open. For a fixed latent representation and explicitly specified admissible query and key maps, allowing a nonlinear key map strictly enlarges the linear span of attention score functions. When $d \geq 2$ and each summand may instead use an unrestricted preceding map, the standard, single-sided, and dual-sided spans coincide. In a bias-free, softmax-free residual attention model, we derive that every output coordinate is a polynomial of odd degree at most 3^L after L layers; the occurrence of particular lower degrees requires nondegeneracy conditions. We then discuss Milnor–Thom bounds only for the corresponding idealized polynomial decision boundaries, and analyze actual input scaling in softmax attention: small scales produce nearly uniform averaging, whereas large scales concentrate on maximal scores under a positive score gap. The final discussion presents the "lottery of attention" as a conjectural geometric interpretation of parameter sparsity and initialization; it is not a theorem guaranteeing optimization feasibility or homotopic convergence.

1 INTRODUCTION

Deep learning has achieved unprecedented success across a broad range of domains, primarily driven by the Transformer architecture (Vaswani et al., 2017). Despite its empirical dominance, a rigorous, first-principles mathematical understanding of why Transformers are so effective remains an active area of investigation. While many theoretical works treat Transformers as black-box universal approximators or focus on their optimization landscapes, deeper structural understanding of their representation capacity can be obtained by analyzing them through the lens of algebraic geometry.

In this paper, we study the algebraic and geometric structures of multi-layer Transformers. Our analysis focuses on three main directions:

1. **The Expressive Hierarchy of Attention Variants:** For a fixed latent representation, we prove strict inclusions between the score spans of projection-free, standard, single-sided nonlinear, and dual-sided nonlinear attention under explicit assumptions. We also identify how these relations change when each summand may use an unrestricted preceding representation.
2. **Width Bounds for Quadratic Simulation:** We prove an $\Omega(N^2)$ lower bound and give an $O(N^2d)$ constructive upper bound for exact simulation of all pairwise inner products with a single-hidden-layer square-activation network. The bounds match in their dependence on N for fixed d; their dependence on a growing feature dimension remains unresolved by this analysis.
3. **Polynomial Filtrations, Topological Complexity, and Long-Context Boundaries:** For a simplified residual attention model, we study a filtration of polynomial function spaces and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

use token-logit differences to describe decision boundaries. We bound the topology of the associated algebraic hypersurfaces and discuss a conjectural *topological lottery hypothesis* motivated by parameter and data sparsity.

2 Background and Related Work

2.1 Attention and Transformer expressivity

The Transformer combines attention, token-wise feed-forward networks, and residual paths (Vaswani et al., 2017; He et al., 2016). A head first computes query–key scores, applies row-wise softmax, and then aggregates value vectors. We distinguish the function spaces spanned by these scores from the functions realized by a complete network. This distinction matters because architecture changes can alter the behavior of a stack: Dong et al. (2021) establish rapid rank collapse for pure attention under their assumptions, while Strobl et al. (2024) survey how Transformer expressivity depends on architectural and computational assumptions. Our single-layer comparison fixes the input representation and stops at the score matrix; the later polynomial model studies a separately specified simplified stack.

2.2 Geometric and statistical measures of complexity

Piecewise-linear networks can be studied through their affine response regions (Pascanu et al., 2013). VC-dimension and pseudodimension quantify other aspects of capacity, with bounds for piecewise-linear networks developed by Bartlett et al. (2019). For polynomial models, the classical work of Milnor (1964) and Thom (1965) instead relates polynomial degree to the topology of real zero sets. These measures answer different questions: a large degree bound allows complicated boundaries but does not establish that a particular trained network attains that complexity.

2.3 The Lottery Ticket Hypothesis

Frankle & Carbin (2018) propose that randomly initialized dense networks contain sparse subnetworks which, when trained in isolation from their original initialization, can match the accuracy of the original network. Their hypothesis concerns subnetworks and initialization. Our later topological lottery discussion draws an analogy with useful local portions of decision boundaries; the original hypothesis does not establish high Betti numbers or guarantee a topology-preserving training trajectory.

3 Expressive Power of Attention Variants

We compare standard attention with three variants through the function spaces spanned by their pre-softmax scores. Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be fixed, with image $\mathcal{U}$ containing a nonempty open set U. For $N \geq 2$, let $x_i = f(u_i)$ and $x_j = f(u_j)$ be token representations in the same input matrix $X \in \mathbb{R}^{N\times d}$. For distinct positions $i \neq j$, we allow these two rows to vary independently over $\mathcal{U}$; sharing X does not impose $x_i = x_j$. A span means finite real linear combinations of score functions, with equality on all of $\mathcal{U} \times \mathcal{U}$.

3.1 Standard Attention and Three Variants

Fix a head dimension $r \geq 1$. Query and key maps act on each token separately, and all linear projections below are bias-free. The score at position (i, j) uses x_i as its query input and x_j as its key input; $x_{i,a}$ denotes coordinate a of token i.

Standard attention. Independent query and key projections give $A(X) = XW_QW_K^TX^T$ and the score family

$$\mathcal{S}_T = \{(x_i, x_j) \mapsto \langle x_iW_Q, x_jW_K\rangle : W_Q, W_K \in \mathbb{R}^{d\times r}\}. \tag{1}$$

Each score is bilinear, with coefficient matrix of rank at most r. Neither $W_QW_K^T$ nor the score matrix $A(X)$ is required to be symmetric.

Variant I: Removing the projections. Using the latent vectors directly gives

$$\mathcal{S}_0 = \{(x_i, x_j) \mapsto \langle x_i, x_j \rangle\}. \tag{2}$$

With f fixed, this family contains just one score. Removing the projections is distinct from tying trainable projections: scores $\langle x_i W, x_j W \rangle$ have coefficient matrices WW^T, whose real linear span is the space of symmetric matrices of dimension $d(d+1)/2$.

Variant II: A nonlinear query map. Let $\mathcal{F}$ be a class of maps $\mathcal{U} \to \mathbb{R}^r$ containing every linear map. Assume also that $x_i \mapsto h(x_i)e_1$ belongs to $\mathcal{F}$ for some scalar function h satisfying

$$h|_U \notin \mathcal{L}(U), \qquad \mathcal{L}(U) = \mathrm{Span}\{x_j \mapsto x_{j,a} : 1 \le a \le d\},$$

where e_1 is the first coordinate vector in $\mathbb{R}^r$. Replacing only the query map gives

$$\mathcal{S}_1 = \{(x_i, x_j) \mapsto \langle q(x_i), x_j W \rangle : q \in \mathcal{F},\ W \in \mathbb{R}^{d \times r}\}. \tag{3}$$

The key dependence remains linear. The assumptions concern exact representation. For example, sufficiently wide ReLU networks with biases and a linear output layer represent linear maps using $u = \mathrm{ReLU}(u) - \mathrm{ReLU}(-u)$ and admit $h(x_i) = \mathrm{ReLU}(x_{i,1} - c_1)$ for a ball $B(c, \rho) \subseteq U$; this h is not linear on U.

Variant III: Nonlinear query and key maps. Allowing both maps to vary in $\mathcal{F}$ gives

$$\mathcal{S}_2 = \{(x_i, x_j) \mapsto \langle q(x_i), k(x_j) \rangle : q, k \in \mathcal{F}\}. \tag{4}$$

Unlike $\mathcal{S}_1$, this family permits nonlinear dependence on the key input as well.

3.2 Inclusion Relations

Proposition 3.1 (Strict containment for a fixed latent representation)**.** *Under the assumptions above, for $d \ge 2$,*

$$\mathrm{Span}(\mathcal{S}_0) \subsetneq \mathrm{Span}(\mathcal{S}_T) \subsetneq \mathrm{Span}(\mathcal{S}_1) \subsetneq \mathrm{Span}(\mathcal{S}_2). \tag{5}$$

For $d = 1$, the first inclusion is an equality and the other two remain strict.

The proof in Appendix A.1 uses the dimensions 1 and d^2 for the first inclusion, and the witnesses $h(x_i)x_{j,1}$ and $x_{i,1}h(x_j)$ for the two nonlinear strict inclusions.

The four spans can be described more explicitly. On $\mathcal{U}$, let $\mathcal{L} = \mathrm{Span}\{x_i \mapsto x_{i,a} : 1 \le a \le d\}$ and $\mathcal{H} = \mathrm{Span}\{q_b : q \in \mathcal{F},\ 1 \le b \le r\}$. Write $\mathcal{A} \otimes \mathcal{B}$ for finite sums $\sum_\ell a_\ell(x_i) b_\ell(x_j)$ with $a_\ell \in \mathcal{A}$ and $b_\ell \in \mathcal{B}$. Then

$$\begin{aligned}
\mathrm{Span}(\mathcal{S}_0) &= \{(x_i, x_j) \mapsto c\langle x_i, x_j \rangle : c \in \mathbb{R}\},\\
\mathrm{Span}(\mathcal{S}_T) &= \mathcal{L} \otimes \mathcal{L},\\
\mathrm{Span}(\mathcal{S}_1) &= \mathcal{H} \otimes \mathcal{L},\\
\mathrm{Span}(\mathcal{S}_2) &\subseteq \mathcal{H} \otimes \mathcal{H}.
\end{aligned}$$

If $\mathcal{F}$ is closed under linear transformations of its output, as for MLPs with an unrestricted linear output layer, the last inclusion is also an equality. Under this additional condition, if $\dim \mathcal{H} = m < \infty$, the four dimensions are $1, d^2, md, m^2$. Appendix A.2 proves these characterizations.

These are statements about spans. For one standard head the rank bound r still applies; when $r < d$, even $\langle x_i, x_j \rangle$ is in $\mathrm{Span}(\mathcal{S}_T)$ but not in $\mathcal{S}_T$. Composing with $(u, u') \mapsto (f(u), f(u'))$ preserves the strict inclusions because f maps onto $\mathcal{U}$.

Scope of the fixed-representation assumption. For comparison, let $\widetilde{\mathcal{S}}_\nu$ allow an arbitrary preceding map $g : \mathcal{X} \to \mathbb{R}^d$, where $\mathcal{X} = \mathbb{R}^d$, in each score, for $\nu \in \{0, T, 1, 2\}$. The admissible query and key maps are now defined on $\mathbb{R}^d$ and contain all linear maps. No regularity or architecture constraint is imposed on g, and taking a span permits a different g in each summand. In this separate comparison, x_i, x_j denote token vectors before g, and attention receives $g(x_i), g(x_j)$. Define

$$\mathcal{R} = \left\{ (x_i, x_j) \mapsto \sum_{\ell=1}^{m} a_\ell(x_i) b_\ell(x_j) : m < \infty,\ a_\ell, b_\ell : \mathcal{X} \to \mathbb{R} \right\},$$

Here $\mathcal{R}$ includes asymmetric functions; $\mathcal{R}_{\mathrm{sym}}$ denotes only its symmetric subspace, used for the projection-free variant.

Proposition 3.2 (Spans under unrestricted adaptation). *For $d \geq 2$,*

$$\mathrm{Span}(\widetilde{\mathcal{S}}_0) = \mathcal{R}_{\mathrm{sym}} \subsetneq \mathcal{R},$$
$$\mathrm{Span}(\widetilde{\mathcal{S}}_T) = \mathrm{Span}(\widetilde{\mathcal{S}}_1) = \mathrm{Span}(\widetilde{\mathcal{S}}_2) = \mathcal{R}.$$

Appendix A.3 gives the construction and proof.

For $d = 1$, the standard and projection-free spans both equal $\mathcal{R}_{\mathrm{sym}}$, and no equality with the nonlinear spans is asserted. The unrestricted result concerns finite separable sums with separately chosen representations; it does not establish equivalence for one shared representation, one head, or a fixed parameter budget.

4 Complexity and Parameter Efficiency: Attention vs. MLP

We study the width required to simulate the pairwise inner-product operation underlying attention using a single-hidden-layer network with square activations. We establish a quadratic lower bound in the sequence length and a constructive upper bound that also depends linearly on the feature dimension. The comparison distinguishes hidden-unit counts from the number of trainable parameters.

4.1 Simulating Attention with Quadratic Activation Networks

Let $N, d \geq 1$ be integers and let $x_i, y_j \in \mathbb{R}^d$ for $i, j \in \{1, \ldots, N\}$. Both collections vary independently over all of $\mathbb{R}^{N \times d}$. We wish to compute all pairwise inner products $S_{ij} = \sum_{\alpha=1}^{d} x_{i,\alpha} y_{j,\alpha}$ simultaneously using a bias-free network with a single hidden layer, activation function $\sigma(z) = z^2$, and a linear output layer.

Proposition 4.1. *The attention bilinear form $S_{ij} = x_i \cdot y_j^T$ can be exactly simulated by a two-layer neural network with $2N^2 d$ hidden units and quadratic activation functions.*

The construction uses the polarization identity

$$xy = \tfrac{1}{4}\big((x+y)^2 - (x-y)^2\big),$$

with two hidden units for each token pair and feature coordinate. The full network construction and verification appear in Appendix B.1.

The construction uses $2N^2 d$ hidden units. If both weight matrices are stored densely, their total number of entries is $(2Nd + N^2)\, 2N^2 d = 4N^3 d^2 + 2N^4 d$. However, this particular construction uses only two nonzero input weights and one nonzero output weight per hidden unit, for a total of $6N^2 d$ nonzero entries. These are implementation counts for the construction, not lower bounds on the number of trainable parameters.

4.2 Lower and Upper Bounds on Hidden Neurons

We now establish a lower bound from the dimension of the target quadratic-form space. Combining it with Proposition 4.1 gives bounds between quadratic order in N and order $N^2 d$.

Theorem 4.2. *Let $N, d \geq 1$ and let $Z = [x_{\mathit{flat}}^T, y_{\mathit{flat}}^T]^T \in \mathbb{R}^{2Nd}$ represent the concatenated input of two independently varying collections of token vectors. Let $H_{\min}(N, d)$ be the minimum hidden width among networks of the form*

$$T_{ij} = \sum_{k=1}^{H} W_{2,(ij),k} (\mathbf{w}_k^T Z)^2 \tag{6}$$

that exactly represent $S_{ij} = \sum_{\alpha=1}^{d} x_{i,\alpha} y_{j,\alpha}$ for every $Z \in \mathbb{R}^{2Nd}$ and every i, j. Then

$$N^2 + 1 \leq H_{\min}(N, d) \leq 2N^2 d. \tag{7}$$

In particular, $H_{\min}(N, d) = \Omega(N^2)$ and $H_{\min}(N, d) = O(N^2 d)$. For each fixed d, these bounds imply $H_{\min}(N, d) = \Theta(N^2)$ as $N \to \infty$.

The lower bound follows because the N^2 target quadratic forms are linearly independent. A width of exactly N^2 is impossible: their coefficient matrices have zero trace, whereas each nonzero hidden square has a rank-one positive-semidefinite coefficient matrix with positive trace. The upper bound follows from Proposition 4.1. Appendix B.2 gives the detailed proof.

Remark 4.3. *Theorem 4.2 concerns hidden width for exact computation of all pairwise inner products on two independent inputs. It does not establish the optimal dependence on d when d grows with N, or a lower bound on the number of distinct trainable parameters. In attention, the shared query and key projections contain $2dd_k$ parameters, independent of N, while the pairwise inner products are computed by a fixed operation. These width bounds therefore quantify a cost of implementing that operation with a square-activation network; they do not establish parameter optimality of attention among general architectures. Applying a lower bound to self-attention on a shared input X, with queries XW_Q and keys XW_K, would additionally require accounting for the constraints relating those queries and keys.*

4.3 Geometric Interpretation of ReLU Activation Regions

While we analyzed quadratic activations above, standard MLPs typically use rectified linear units (ReLU), $\sigma(z) = \max(0, z)$. The geometric action of a ReLU layer is fundamentally different. Let $h(x) = \text{ReLU}(Wx + b)$, where $W \in \mathbb{R}^{n\times d}$.

The linear map $z(x) = Wx + b$ maps the d-dimensional input space to an n-dimensional affine subspace. The ReLU activation function projects this affine subspace onto the positive orthant $\mathbb{R}^n_+$. Specifically, for any component where $z_i(x) < 0$, the activation sets $h_i(x) = 0$, which geometrically corresponds to projecting the point onto the boundary hyperplane $h_i = 0$.

This mapping partitions the input space $\mathbb{R}^d$ into a set of polyhedral regions (cells), within each of which the network behaves as a purely affine transformation. As shown in the classical work *On the Number of Linear Regions of Deep Neural Networks*, a multi-layer FFN with L layers, hidden width n, and input dimension n_0 can partition the input space into:

$$\Omega\left(\left(\frac{n}{n_0}\right)^{(L-1)n_0} n^{n_0}\right) \tag{8}$$

distinct linear regions (Pascanu et al., 2013). Thus, while FFNs approximate non-linear functions by gluing together a massive number of local linear patches, the attention mechanism generates smooth, globally non-linear algebraic varieties directly.

5 Multi-Layer Transformers: Polynomial Filtrations and Real Algebraic Geometry

We study a simplified residual attention stack with bias-free projections and without attention softmax, feed-forward sublayers, or normalization. Its polynomial structure lets us connect continuous input representations to the regions in which a fixed model predicts different vocabulary tokens.

5.1 Homogeneous Cubic Vector Fields

For $X \in \mathbb{R}^{N\times d}$, a single head with compatible projection dimensions has attention output

$$\mathcal{A}(X) = (XW_QW_K^TX^T)XW_V, \qquad \mathcal{A}(\lambda X) = \lambda^3\mathcal{A}(X). \tag{9}$$

Thus the attention branch is a homogeneous cubic map; fixed scale factors can be absorbed into its weights. Multiplying the scores entrywise by a fixed zero–one mask also preserves this degree bound.

5.2 Polynomial Filtrations and Residual Connections

Residual paths, introduced in deep residual networks (He et al., 2016) and used in Transformers (Vaswani et al., 2017), give

$$X_l = X_{l-1} + \mathcal{A}_l(X_{l-1}), \qquad l = 1, \ldots, L. \tag{10}$$

For the unmasked model above, let $\mathcal{V}_l$ be the real linear span of all scalar output-coordinate functions at depth l, as the weights vary, and let $\mathcal{D}_l$ collect the monomial degrees occurring in this space.

Theorem 5.1 (Polynomial Filtration). *For the specified residual polynomial model with unrestricted compatible weights,*

$$\mathcal{V}_0 \subsetneq \mathcal{V}_1 \subsetneq \cdots \subsetneq \mathcal{V}_L, \qquad \mathcal{D}_l = \{1, 3, 5, \ldots, 3^l\}. \tag{11}$$

Every output coordinate contains only odd-degree monomials of degree at most 3^l. The equality for $\mathcal{D}_l$ describes degrees attainable across the family, not degrees present for every fixed choice of weights.

Cubic composition gives the degree bound, and setting the final attention branch to zero preserves the preceding function space. A scalar restriction $p_l(t) = p_{l-1}(t) + p_{l-1}(t)^3$ realizes the highest degree and all intervening odd degrees. Appendix C gives the detailed induction and construction. Without residual paths, nonzero outputs are homogeneous of degree 3^l; retaining lower-degree terms does not alone guarantee dynamical stability.

5.3 Token Predictions and Decision Boundaries

Fix the model parameters, sequence length, and prediction position s (for next-token prediction, usually $s = N$). Regard the input embedding matrix $X = X_0 \in \mathbb{R}^{N\times d}$ as a continuous variable. The final hidden vector $h_s(X) = (X_L(X))_{s,:}$ is mapped to one logit for each vocabulary token $a \in \mathcal{T}$:

$$z_a(X) = h_s(X)w_a + b_a, \qquad \widehat{\tau}(X) = \operatorname*{argmax}_{a\in\mathcal{T}} z_a(X). \tag{12}$$

Here w_a, b_a are the output-head parameters and ties are resolved by a fixed rule. The final vocabulary softmax preserves the ordering of logits, so it does not change this greedy prediction rule. Sampling from the resulting distribution is a different decoding procedure.

Each token has a region in input space where its logit is maximal. A change from predicting token a to predicting token b must cross a location where their logits tie. Define

$$\mathcal{C}_a = \{X : z_a(X) \geq z_c(X) \text{ for all } c \in \mathcal{T}\}, \tag{13}$$

$$F_{ab}(X) = z_a(X) - z_b(X), \qquad H_{ab} = \{X : F_{ab}(X) = 0\}. \tag{14}$$

The closed regions $\mathcal{C}_a$ can overlap at ties; the tie-breaking rule assigns the actual prediction. The relevant part of H_{ab} is where $z_a = z_b \geq z_c$ for every other token c. Equality of two logits below a third token's score does not define a prediction boundary. At a tie where a, b strictly dominate the other tokens and $\nabla F_{ab} \neq 0$, this set is locally a hypersurface separating the two token predictions. Degenerate ties need not form a smooth separating surface.

For example, a boundary between predicting `cat` and `dog` is the part of $z_{\text{cat}}(X) = z_{\text{dog}}(X)$ where these tokens have the largest logits; on its two regular sides the greedy prediction differs. In the final hidden space this equality is affine, while its preimage in the input space can be curved and disconnected. Actual embedded token sequences occupy a discrete subset of this continuous space, so the surfaces describe an extension of the model to continuous embeddings.

In our polynomial model, F_{ab} has degree at most 3^L; an output bias can add a constant term. Thus H_{ab} is an algebraic zero set, and imposing the maximal-logit inequalities gives a semi-algebraic set. This is the precise connection between the polynomial stack and token decision surfaces.

5.4 Topological Complexity and the Milnor–Thom Theorem

The Betti numbers of a zero set count connected components and higher-dimensional holes. The classical bound below applies to a full pairwise equality surface H_{ab}.

Theorem 5.2 (Milnor–Thom). *For a real polynomial P in m variables with degree at most $D \geq 1$, the zero set $H = \{x \in \mathbb{R}^m : P(x) = 0\}$ satisfies (Milnor, 1964; Thom, 1965)*

$$\sum_{k=0}^{m} \beta_k(H) \leq D(2D-1)^{m-1}. \tag{15}$$

Corollary 5.3. *For the polynomial stack with an affine vocabulary head, each pairwise logit-equality set obeys*

$$\sum_{k=0}^{Nd} \beta_k(H_{ab}) \leq 3^L(2\cdot 3^L - 1)^{Nd-1} = O\big((2\cdot 3^L)^{Nd}\big). \tag{16}$$

This is an upper bound on possible topology, not a lower bound on attained complexity. The active token boundary additionally satisfies dominance inequalities, and the full prediction boundary involves many token pairs. Bounding either requires a semi-algebraic analysis accounting for those inequalities; the displayed bound cannot simply be transferred to an arbitrary subset or union of the H_{ab}.

5.5 Effects of Architectural Components on the Polynomial Model

Biases in query, key, or value maps introduce additional degrees. ReLU applied to polynomial outputs gives piecewise-polynomial maps, while normalization and other nonlinearities generally change the polynomial description. Internal attention softmax makes the hidden map non-polynomial, so the degree and Milnor–Thom bounds above do not directly apply to a standard Transformer. This differs from the final vocabulary softmax, which leaves logit comparisons unchanged. Attention weights sum to one but do not bound value norms; Section 6 analyzes actual input scaling.

6 Asymptotic Analysis of Softmax Normalization

We examine how the magnitude of the input changes attention, keeping the temperature fixed at one. Consider a single unmasked attention layer with fixed bias-free linear projections and no normalization layers. For a fixed baseline input X, write

$$S_{ij} = \frac{x_i W_Q W_K^T x_j^T}{\sqrt{d_k}}, \qquad v_j = x_j W_V. \tag{17}$$

Scaling the actual input to tX, with $t > 0$, multiplies the scores by t^2 and the values by t. Thus

$$P_{ij}(t) = \frac{\exp(t^2 S_{ij})}{\sum_{k=1}^N \exp(t^2 S_{ik})}, \qquad \widetilde{Y}_i(t) = t \sum_{j=1}^N P_{ij}(t) v_j. \tag{18}$$

The regimes below depend on the scaled differences between scores, not on their common offset.

6.1 Small Input Scale: Nearly Uniform Averaging

When $t^2 \max_j |S_{ij} - \bar{S}_i| \ll 1$, where $\bar{S}_i = N^{-1} \sum_j S_{ij}$, the weights in row i are nearly uniform. Consequently, the attention branch weakly distinguishes among tokens and its leading output is the same average of the scaled values for every query. More precisely, for fixed X and weights, as $t \to 0^+$,

$$P_{ij}(t) = \frac{1}{N} + \frac{t^2}{N}(S_{ij} - \bar{S}_i) + O(t^4), \tag{19}$$

$$\widetilde{Y}_i(t) = t\bar{v} + \frac{t^3}{N} \sum_{j=1}^N (S_{ij} - \bar{S}_i) v_j + O(t^5), \qquad \bar{v} = \frac{1}{N} \sum_{j=1}^N v_j. \tag{20}$$

To obtain these expansions, subtract the row mean from the logits and use $\exp(t^2(S_{ij} - \bar{S}_i)) = 1 + t^2(S_{ij} - \bar{S}_i) + O(t^4)$; the centered scores sum to zero. The first possible query-dependent correction is therefore cubic in the input scale. The unnormalized output tends to zero, while $\widetilde{Y}_i(t)/t \to \bar{v}$. If $\bar{v}$ or the cubic coefficient vanishes, the corresponding leading term is absent. This averaging statement concerns the attention branch; a residual path still carries the scaled input.

6.2 Large Input Scale: Concentration on Maximal Scores

At large scales, attention increasingly favors the tokens with maximal scores and suppresses contributions from the others. Suppose $N \geq 2$ and row i has a unique maximizing index j^* with gap

$$\Delta_i = S_{ij^*} - \max_{j \neq j^*} S_{ij} > 0. \tag{21}$$

Dividing the softmax denominator by $\exp(t^2 S_{ij^*})$ gives

$$1 - P_{ij^*}(t) \leq (N-1) e^{-t^2 \Delta_i}, \qquad P_{ij}(t) \leq e^{-t^2 \Delta_i} \quad (j \neq j^*). \tag{22}$$

Hence concentration is strong when $(N-1)e^{-t^2\Delta_i} \ll 1$, and

$$\|\widetilde{Y}_i(t) - tv_{j^*}\|_2 \le t(N-1)e^{-t^2\Delta_i} \max_{j\neq j^*} \|v_j - v_{j^*}\|_2 \longrightarrow 0. \tag{23}$$

Thus the layer approaches hard token selection: its normalized output tends to v_{j^*}, while its output norm grows linearly in t if $v_{j^*} \neq 0$. Bounded attention weights do not bound the magnitude of the scaled values or the resulting output.

If several tokens tie for the maximum, let $J_i = \operatorname{argmax}_j S_{ij}$. The limiting weights are uniform on J_i and zero elsewhere, so

$$\frac{\widetilde{Y}_i(t)}{t} \longrightarrow \frac{1}{|J_i|} \sum_{j\in J_i} v_j. \tag{24}$$

This also covers $N = 1$. A small gap can delay concentration even for a large input norm; the relevant quantity is $t^2\Delta_i$ relative to the number of competing tokens.

7 Discussion: Parameter Bottleneck, Cylinder Degeneracy, and Long-Context Capacity Boundary

Corollary 5.3 bounds the topology permitted by polynomial degree and input dimension. The family actually generated by a model is further constrained by its shared parameters, which need not increase with context length.

7.1 Sparsity of Parameter Image in the Polynomial Ring Space

For fixed depth L and degree bound $D = 3^L$, the space of input polynomials has dimension

$$\dim \mathcal{P}_{\le D}(\mathbb{R}^{Nd}) = \binom{Nd+D}{D} = O((Nd)^D). \tag{25}$$

Let P count the trainable parameters, including the output head, with architecture and vocabulary size held fixed as N grows. The map $\bar{\Phi}$ from parameters to the coefficients of a token-logit difference is polynomial in this idealized model. Its image therefore has dimension at most P, giving

$$\frac{\dim \operatorname{Im}(\bar{\Phi})}{\dim \mathcal{P}_{\le D}} \le \frac{P}{\binom{Nd+D}{D}} \longrightarrow 0 \qquad (N \to \infty). \tag{26}$$

This image is a restricted, generally nonlinear subset of coefficient space; when $P < \binom{Nd+D}{D}$, it has measure zero there. This dimension argument does not by itself supply a VC-dimension bound. The results of Bartlett et al. (2019) concern piecewise-linear networks and cannot be transferred directly to the present polynomial architecture.

A cylinder interpretation requires an additional condition: if a logit difference depends only on a lower-dimensional linear projection of the input, its zero set extends unchanged along the projection's nullspace. A small parameter count alone does not establish this invariance. These observations motivate studying how the reachable boundary family changes with context length without claiming a universal context-length threshold.

7.2 Topological Lottery Hypothesis and Sparsity-Induced Decoupling

The Lottery Ticket Hypothesis concerns sparse subnetworks and their initial weights (Frankle & Carbin, 2018). Inspired by that perspective, we conjecture that a randomly initialized attention model may already contain local portions of token decision surfaces that training can move toward useful class boundaries. This geometric interpretation requires evidence about the initialization distribution, architecture, and data; neither the lottery-ticket result nor the degree bound guarantees large initial Betti numbers or alignment with a target boundary.

Shared parameters couple changes to different regions of input space. If the data occupy only a small part of that space, some collateral boundary changes may occur away from observed inputs and incur no additional empirical loss. This possibility motivates a dual-sparsity interpretation: the model

explores a restricted family of polynomials, while the data constrain only part of the ambient input domain. Data sparsity alone, however, does not ensure that parameter updates can satisfy all sample constraints or avoid optimization conflicts.

A testable version of this hypothesis would track pairwise logit margins and boundary changes near data during training, and compare the effects of pruning or resetting initial weights. Smooth changes in weights can still pass through singular boundaries where topology changes. We therefore present the topological lottery as a conjectural account of training behavior, rather than a guarantee of optimization feasibility or homotopic convergence.

8 Conclusion

We compare the function spaces spanned by single-layer attention scores under explicit assumptions on shared representations, and prove $N^2 + 1 \leq H_{\min}(N, d) \leq 2N^2 d$ for exact square-network simulation of all pairwise inner products. In a simplified residual polynomial stack, we derive a filtration of attainable function spaces and describe token prediction regions through logit comparisons. The resulting algebraic equality surfaces admit degree-based topological upper bounds; full token boundaries also involve dominance inequalities. Input-scaling analysis separately treats attention softmax, while the topological lottery remains a hypothesis connecting initialization, parameter constraints, and data geometry.

Ethics Statement and AI Use Disclosure

Large language models and AI-assisted editing tools were used to polish the language, organize the manuscript, draft and revise portions of the text, and assist with the presentation of related work. AI assistance also supported the exploration of mathematical formulations and the development and examination of proof arguments, including attention-score span comparisons, their assumptions, and the interpretation of token decision boundaries. These uses extended beyond proofreading. AI-generated suggestions are not treated as independent evidence or as formal verification of mathematical claims. The author retains responsibility for the manuscript, the accuracy and attribution of its references, and the validity and limitations of its mathematical arguments; AI tools are not credited as authors.

References

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pp. 2793–2803. PMLR, 2021.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

John Milnor. On the betti numbers of real varieties. *Proceedings of the American Mathematical Society*, 15(2):275–280, 1964.

Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*, 2013.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages can transformers express? a survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 2024.

René Thom. Sur l'homologie des variétés algébriques réelles, differential and combinatorial topology (a symposium in honor of marston morse), 1965.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A Detailed Proofs of the Attention-Variant Span Relations

This appendix proves the span relations stated in Section 3. The distinction between a score realized by one head and a finite linear combination of scores is essential throughout. Because analyzing multiple Transformer layers jointly is complex, we compare only the function spaces spanned by the pre-softmax score functions of a single attention layer, without considering preceding MLPs or other layers.

A.1 Fixed latent representation

Let $N \geq 2$ and $d, r \geq 1$. As in Section 3, fix $f : \mathbb{R}^d \to \mathbb{R}^d$ with image $\mathcal{U}$ containing a nonempty open set U. Fix distinct token positions $i \neq j$ and allow the corresponding rows x_i, x_j of the same input matrix X to vary independently over $\mathcal{U}$. Let $\mathcal{F}$ contain every linear map $\mathcal{U} \to \mathbb{R}^r$ and contain $x_i \mapsto h(x_i)e_1$, where

$$h|_U \notin \mathcal{L}(U), \qquad \mathcal{L}(U) = \mathrm{Span}\{x_j \mapsto x_{j,a} : 1 \leq a \leq d\}.$$

Here e_1 is the first coordinate vector in $\mathbb{R}^r$. All linear projections are bias-free, and all spans below are finite real linear spans.

Proof of Proposition 3.1. We prove the three successive containments.

Step 1: Projection-free and standard spans. Every member of $\mathcal{S}_T$ is bilinear:

$$\langle x_i W_Q, x_j W_K \rangle = x_i W_Q W_K^T x_j^T.$$

Conversely, for every pair (a, b), choose $q(x_i) = x_{i,a}e_1$ and $k(x_j) = x_{j,b}e_1$. Their score is $x_{i,a}x_{j,b}$. Therefore all d^2 monomials $x_{i,a}x_{j,b}$ belong to $\mathcal{S}_T$, and

$$\mathrm{Span}(\mathcal{S}_T) = \{(x_i, x_j) \mapsto x_i M x_j^T : M \in \mathbb{R}^{d\times d}\}.$$

The monomials are linearly independent on the nonempty open set $U \times U$, so this space has dimension d^2.

The family $\mathcal{S}_0$ contains only the inner product. Hence

$$\mathrm{Span}(\mathcal{S}_0) = \{(x_i, x_j) \mapsto c\langle x_i, x_j \rangle : c \in \mathbb{R}\},$$

which has dimension one. Since the inner product is a nonzero bilinear form, it belongs to the standard span. Thus

$$\mathrm{Span}(\mathcal{S}_0) \subseteq \mathrm{Span}(\mathcal{S}_T),$$

and the inclusion is strict exactly when $d^2 > 1$, namely when $d \geq 2$.

Step 2: Strict enlargement by nonlinear queries. Because $\mathcal{F}$ contains all linear maps, every standard score is a single-sided score, so

$$\mathrm{Span}(\mathcal{S}_T) \subseteq \mathrm{Span}(\mathcal{S}_1).$$

The function

$$K_1(x_i, x_j) = h(x_i)x_{j,1}$$

belongs to $\mathcal{S}_1$: use $q(x_i) = h(x_i)e_1$ and a linear Key map with first output coordinate $x_{j,1}$. Suppose that K_1 belonged to $\mathrm{Span}(\mathcal{S}_T)$. Then there would be a matrix M such that

$$h(x_i)x_{j,1} = x_i M x_j^T \qquad \text{for all } (x_i, x_j) \in U \times U.$$

Since U is open, it contains a point x_j° with $x_{j,1}^\circ \neq 0$. Fixing $x_j = x_j^\circ$ gives

$$h(x_i)x_{j,1}^\circ = x_i M (x_j^\circ)^T,$$

whose right-hand side is a linear function of x_i. Division by $x_{j,1}^\circ$ implies $h|_U \in \mathcal{L}(U)$, contradicting the assumption on h. Therefore

$$\mathrm{Span}(\mathcal{S}_T) \subsetneq \mathrm{Span}(\mathcal{S}_1).$$

Step 3: Strict enlargement by nonlinear keys. Again, $\mathcal{S}_1 \subseteq \mathcal{S}_2$ because every linear Key map is admissible. More generally, if

$$K \in \mathrm{Span}(\mathcal{S}_1),$$

then for some finite collection of queries q_ℓ and matrices W_ℓ,

$$K(x_i, x_j) = \sum_{\ell=1}^{m} c_\ell \langle q_\ell(x_i), x_j W_\ell \rangle = \sum_{a=1}^{d} A_a(x_i) x_{j,a},$$

where

$$A_a(x_i) = \sum_{\ell=1}^{m} c_\ell \sum_{b=1}^{r} (q_\ell(x_i))_b (W_\ell)_{ab}.$$

Thus every fixed-x_i slice $K(x_i, \cdot)|_U$ lies in the linear space $\mathcal{L}(U)$. Here $i \neq j$: we hold row i of X fixed while varying row j, so the query coefficients remain fixed even though query and key use the same input matrix.

Now consider

$$K_2(x_i, x_j) = x_{i,1} h(x_j).$$

It belongs to $\mathcal{S}_2$ by choosing $q(x_i) = x_{i,1} e_1$ and $k(x_j) = h(x_j) e_1$. If it belonged to $\mathrm{Span}(\mathcal{S}_1)$, choose $x_i^\circ \in U$ with $x_{i,1}^\circ \neq 0$. The slice

$$K_2(x_i^\circ, x_j) = x_{i,1}^\circ h(x_j)$$

would have to lie in $\mathcal{L}(U)$. Dividing by $x_{i,1}^\circ$ again gives $h|_U \in \mathcal{L}(U)$, a contradiction. Hence

$$\mathrm{Span}(\mathcal{S}_1) \subsetneq \mathrm{Span}(\mathcal{S}_2).$$

Combining the three steps proves, for $d \geq 2$,

$$\mathrm{Span}(\mathcal{S}_0) \subsetneq \mathrm{Span}(\mathcal{S}_T) \subsetneq \mathrm{Span}(\mathcal{S}_1) \subsetneq \mathrm{Span}(\mathcal{S}_2).$$

When $d = 1$, the first two spaces are both one-dimensional and therefore equal; the two nonlinear strict containments use the same slice arguments. □

For the full score matrix, each entry is $A_{ij}(X) = K(x_i, x_j)$. Since any pair in $\mathcal{U} \times \mathcal{U}$ can occupy two distinct rows of X, equality of score-matrix functions for every such X implies equality of their two-input score functions. The strict span relations therefore also hold for these matrix-valued functions. The slice argument does not apply to diagonal entries alone: when $i = j$, the witness becomes $x_{i,1} h(x_i)$, which a single-sided score also realizes using $q(x_i) = h(x_i) e_1$ and a linear key selecting $x_{i,1}$. Independence here is an assumption on the allowed inputs to this single layer; constraints imposed by preceding layers are outside this comparison.

Composition with the fixed preceding map preserves these conclusions. To see this, suppose

$$K(f(u), f(u')) = 0 \qquad \text{for all } u, u' \in \mathbb{R}^d.$$

Surjectivity onto $\mathcal{U}$ implies $K(x_i, x_j) = 0$ for all $x_i, x_j \in \mathcal{U}$. Thus composition defines an injective linear map on the score spaces and preserves strict inclusions.

A.2 Exact tensor-product descriptions

Let

$$\mathcal{L} = \mathrm{Span}\{x_i \mapsto x_{i,a} : 1 \leq a \leq d\}, \qquad \mathcal{H} = \mathrm{Span}\{q_b : q \in \mathcal{F},\ 1 \leq b \leq r\}.$$

Both are spaces of scalar functions on $\mathcal{U}$. For function spaces $\mathcal{A}, \mathcal{B}$, write $\mathcal{A} \otimes \mathcal{B}$ for finite sums $\sum_\ell a_\ell(x_i) b_\ell(x_j)$, with $a_\ell \in \mathcal{A}$ and $b_\ell \in \mathcal{B}$. Appendix A.1 already establishes

$$\mathrm{Span}(\mathcal{S}_0) = \{(x_i, x_j) \mapsto c\langle x_i, x_j \rangle : c \in \mathbb{R}\}, \qquad \mathrm{Span}(\mathcal{S}_T) = \mathcal{L} \otimes \mathcal{L}.$$

Expanding a single-sided score gives

$$\langle q(x_i), x_j W \rangle = \sum_{b=1}^{r} \sum_{a=1}^{d} W_{ab} q_b(x_i) x_{j,a},$$

so $\mathrm{Span}(\mathcal{S}_1) \subseteq \mathcal{H} \otimes \mathcal{L}$. For the reverse inclusion, fix $q \in \mathcal{F}$, $1 \le b \le r$, and $1 \le a \le d$. Choose W whose only nonzero entry is $W_{ab} = 1$. The resulting score is $q_b(x_i)x_{j,a}$. Since these coordinates generate $\mathcal{H}$ and $\mathcal{L}$, their finite linear combinations generate every element of $\mathcal{H} \otimes \mathcal{L}$. Therefore

$$\mathrm{Span}(\mathcal{S}_1) = \mathcal{H} \otimes \mathcal{L}.$$

Similarly, the expansion $\langle q(x_i), k(x_j)\rangle = \sum_{b=1}^{r} q_b(x_i)k_b(x_j)$ proves

$$\mathrm{Span}(\mathcal{S}_2) \subseteq \mathcal{H} \otimes \mathcal{H}.$$

Suppose additionally that $\mathcal{F}$ is closed under linear transformations of its output coordinates. For any $q, k \in \mathcal{F}$ and any coordinates b, c, this closure gives admissible maps $\widehat{q}(x_i) = q_b(x_i)e_1$ and $\widehat{k}(x_j) = k_c(x_j)e_1$. Both use the same output coordinate, so

$$\langle \widehat{q}(x_i), \widehat{k}(x_j)\rangle = q_b(x_i)k_c(x_j).$$

These products generate $\mathcal{H} \otimes \mathcal{H}$, proving equality under the additional closure assumption.

For completeness, if $\{a_s\}_{s=1}^{p}$ and $\{b_t\}_{t=1}^{q}$ are bases of two finite-dimensional function spaces on $\mathcal{U}$, then the pq products $a_s(x_i)b_t(x_j)$ are linearly independent on $\mathcal{U} \times \mathcal{U}$. Indeed, in an identity $\sum_{s,t} c_{st}a_s(x_i)b_t(x_j) = 0$, fix x_i and use independence of the b_t to obtain $\sum_s c_{st}a_s(x_i) = 0$ for each t and every x_i. Independence of the a_s then gives $c_{st} = 0$. Consequently $\dim(\mathcal{A} \otimes \mathcal{B}) = pq$. Since $\dim \mathcal{L} = d$, when $\dim \mathcal{H} = m < \infty$ and the output closure condition holds, the four dimensions are $1, d^2, md, m^2$.

A.3 Unrestricted preceding representations

In this separate comparison, let $\mathcal{X} = \mathbb{R}^d$ be the domain of token vectors x_i, x_j before the map g; attention receives $g(x_i)$ and $g(x_j)$. Allow an arbitrary map $g : \mathcal{X} \to \mathbb{R}^d$ to be chosen separately in each summand of a finite linear combination. Within each summand, the same g is applied to both token positions. No regularity, parameter-budget, or open-image condition is imposed on g. The query and key maps are defined on $\mathbb{R}^d$, take values in $\mathbb{R}^r$ for $r \ge 1$, and their admissible class contains every linear map. Define

$$\mathcal{R} = \left\{ (x_i, x_j) \mapsto \sum_{\ell=1}^{m} a_\ell(x_i)b_\ell(x_j) : m < \infty,\ a_\ell, b_\ell : \mathcal{X} \to \mathbb{R} \right\},$$

Here a_ℓ and b_ℓ are arbitrary real-valued functions of one token vector, and m is the finite number of product terms. Each $a_\ell(x_i)b_\ell(x_j)$ multiplies one scalar feature of token i by one scalar feature of token j. Throughout this appendix, K denotes a scalar score function, with no symmetry or positive-definiteness assumption. The subspace $\mathcal{R}_{\mathrm{sym}}$ consists only of those $K \in \mathcal{R}$ satisfying $K(x_i, x_j) = K(x_j, x_i)$ for all inputs, and is used below for the projection-free variant.

Proof of Proposition 3.2. **Standard and nonlinear variants: no symmetry restriction.** Every score in the four unrestricted families is a finite sum of products of a function of x_i and a function of x_j, so all four spans are contained in $\mathcal{R}$. For one arbitrary pair of scalar functions $a, b : \mathcal{X} \to \mathbb{R}$ and $d \ge 2$, choose

$$g(v) = (a(v), b(v), 0, \ldots, 0) \qquad (v \in \mathcal{X}).$$

For a vector $v \in \mathbb{R}^d$, choose the linear maps $q(v) = v_1 e_1$ and $k(v) = v_2 e_1$. They select the first and second coordinates of g and place them in the same head coordinate, giving

$$\langle q(g(x_i)), k(g(x_j))\rangle = a(x_i)b(x_j).$$

Thus every element of $\mathcal{R}$ is a finite sum of standard scores, where the preceding map can be chosen separately for each product. Since the admissible maps contain all linear maps, $\widetilde{\mathcal{S}}_T \subseteq \widetilde{\mathcal{S}}_1 \subseteq \widetilde{\mathcal{S}}_2$. Therefore

$$\mathrm{Span}(\widetilde{\mathcal{S}}_T) = \mathrm{Span}(\widetilde{\mathcal{S}}_1) = \mathrm{Span}(\widetilde{\mathcal{S}}_2) = \mathcal{R}.$$

These functions need not be symmetric: $a(x_i)b(x_j)$ generally differs from $a(x_j)b(x_i)$. For example, with g the identity and the coordinate-selecting query and key maps above, $K(x_i, x_j) = x_{i,1}x_{j,2}$.

Taking $x_i = (1, 0, \ldots, 0)$ and $x_j = (0, 1, 0, \ldots, 0)$ gives scores 1 and 0 in the two directions, respectively, even though both tokens belong to the same input matrix.

Projection-free variant: symmetry from the shared inner product. For this variant alone, the score is

$$K_0(x_i, x_j) = \langle g(x_i), g(x_j) \rangle = \langle g(x_j), g(x_i) \rangle = K_0(x_j, x_i).$$

Each such score, and every finite real linear combination of these scores, is symmetric. Hence

$$\mathrm{Span}(\widetilde{\mathcal{S}}_0) \subseteq \mathcal{R}_{\mathrm{sym}}.$$

Conversely, $a(x_i)a(x_j)$ is realized by $g(v) = (a(v), 0, \ldots, 0)$, and polarization gives

$$a(x_i)b(x_j) + b(x_i)a(x_j) = (a+b)(x_i)(a+b)(x_j) - a(x_i)a(x_j) - b(x_i)b(x_j).$$

To prove the reverse inclusion for this variant, take $K \in \mathcal{R}_{\mathrm{sym}}$ with a finite separable decomposition $K(x_i, x_j) = \sum_\ell a_\ell(x_i) b_\ell(x_j)$. Its defining symmetry gives

$$K(x_i, x_j) = \frac{1}{2} \sum_\ell \big(a_\ell(x_i) b_\ell(x_j) + b_\ell(x_i) a_\ell(x_j) \big).$$

Each summand is in the projection-free span by polarization, which proves

$$\mathrm{Span}(\widetilde{\mathcal{S}}_0) = \mathcal{R}_{\mathrm{sym}}.$$

This is a proper subspace of $\mathcal{R}$, for example because $K(x_i, x_j) = x_{j,1}$ is not symmetric. □

All symmetry statements above concern scores before masking and softmax; even a symmetric score matrix can become asymmetric after row-wise softmax normalization.

For $d = 1$, every standard score has the form $c\, g(x_i) g(x_j)$ for some $c \in \mathbb{R}$. Hence its span equals the projection-free span, which is $\mathcal{R}_{\mathrm{sym}}$ by the same polarization argument. The two-coordinate construction above is unavailable, so no general equality with the nonlinear spans is asserted in that case. In every dimension the comparison concerns finite sums with separately chosen preceding maps; it does not establish equivalence for one shared representation, one head, or a fixed parameter budget.

B Detailed Proofs for Complexity and Parameter Efficiency: Attention vs. MLP

This appendix supplies the detailed proofs of Proposition 4.1 and Theorem 4.2 from Section 4, under the same assumptions on independent inputs and bias-free square-activation networks.

B.1 Explicit quadratic network construction

Proof of Proposition 4.1. For each token pair index (i, j) and each feature index $\alpha \in \{1, \dots, d\}$, we allocate two hidden units, $h^{(1)}_{ij,\alpha}$ and $h^{(2)}_{ij,\alpha}$. We construct the network as follows:

1. **Input to Hidden Layer:** For $h^{(1)}_{ij,\alpha}$, we set the weight from $x_{i,\alpha}$ to 1, the weight from $y_{j,\alpha}$ to 1, and all other weights to 0. The activation is:

$$h^{(1)}_{ij,\alpha} = \sigma(x_{i,\alpha} + y_{j,\alpha}) = (x_{i,\alpha} + y_{j,\alpha})^2 \tag{B.1}$$

   For $h^{(2)}_{ij,\alpha}$, we set the weight from $x_{i,\alpha}$ to 1, the weight from $y_{j,\alpha}$ to -1, and all other weights to 0. The activation is:

$$h^{(2)}_{ij,\alpha} = \sigma(x_{i,\alpha} - y_{j,\alpha}) = (x_{i,\alpha} - y_{j,\alpha})^2 \tag{B.2}$$

2. **Hidden to Output Layer:** We define the output for the pair (i, j) as:

$$S_{ij} = \sum_{\alpha=1}^{d} \left[\frac{1}{4} h^{(1)}_{ij,\alpha} - \frac{1}{4} h^{(2)}_{ij,\alpha} \right] \tag{B.3}$$

   Substituting the expressions for the hidden units:

$$S_{ij} = \sum_{\alpha=1}^{d} \left[\frac{1}{4}(x_{i,\alpha} + y_{j,\alpha})^2 - \frac{1}{4}(x_{i,\alpha} - y_{j,\alpha})^2 \right] = \sum_{\alpha=1}^{d} x_{i,\alpha} y_{j,\alpha} \tag{B.4}$$

This exactly reproduces the bilinear attention matrix. □

B.2 Lower and upper bounds on hidden width

Proof of Theorem 4.2. Let $R_k = \mathbf{w}_k \mathbf{w}_k^T \in \mathbb{R}^{2Nd \times 2Nd}$ be the outer product of the weight vector $\mathbf{w}_k$ for the k-th hidden unit. It is symmetric and positive semidefinite, and has rank one whenever $\mathbf{w}_k \neq 0$. The output of the k-th hidden unit can be written as:

$$h_k = (\mathbf{w}_k^T Z)^2 = Z^T R_k Z \tag{B.5}$$

The network's output for the entry (i, j) is a linear combination of these hidden units:

$$T_{ij} = \sum_{k=1}^{H} W_{2,(ij),k} Z^T R_k Z = Z^T \left(\sum_{k=1}^{H} W_{2,(ij),k} R_k \right) Z \tag{B.6}$$

Let $M_{ij} = \sum_{k=1}^{H} W_{2,(ij),k} R_k$. For exact representation, we must have $T_{ij}(Z) = S_{ij}(Z)$ for every Z and all i, j. The target bilinear form $S_{ij} = \sum_{\alpha=1}^{d} x_{i,\alpha} y_{j,\alpha}$ can be written as a quadratic form in Z:

$$S_{ij} = Z^T M^*_{ij} Z \tag{B.7}$$

where $M^*_{ij} \in \mathbb{R}^{2Nd \times 2Nd}$ is the unique symmetric matrix corresponding to the quadratic form S_{ij}. We can write M^*_{ij} in block form as:

$$M^*_{ij} = \frac{1}{2} \begin{pmatrix} 0 & B_{ij} \\ B_{ij}^T & 0 \end{pmatrix} \tag{B.8}$$

where $B_{ij} \in \mathbb{R}^{Nd \times Nd}$ has the identity matrix I_d at the block index (i, j) and zeros elsewhere.

Let $\mathcal{M} = \mathrm{Span}\{M_{ij}^* : i, j = 1, \ldots, N\}$. A linear combination with coefficient matrix $C = (c_{ij}) \in \mathbb{R}^{N\times N}$ has the form

$$\sum_{i,j=1}^{N} c_{ij} M_{ij}^* = \frac{1}{2} \begin{pmatrix} 0 & C \otimes I_d \\ C^T \otimes I_d & 0 \end{pmatrix}. \tag{B.9}$$

Since $d \geq 1$, this matrix vanishes if and only if $C = 0$. Hence the N^2 target matrices are linearly independent, and

$$\dim \mathcal{M} = N^2. \tag{B.10}$$

Although each M_{ij}^* is a sum of d components with disjoint feature supports, the same coefficient c_{ij} multiplies every feature component. Thus, $\mathcal{M}$ is a diagonal subspace of the direct sum of the feature-wise spaces; it is not their full direct sum when $d > 1$.

Equality of the quadratic forms for every Z, together with symmetry of their coefficient matrices, implies $M_{ij} = M_{ij}^*$. Consequently,

$$\mathcal{M} \subseteq \mathrm{Span}(\{R_k\}_{k=1}^{H}) \tag{B.11}$$

Since the dimension of the span of H matrices is at most H,

$$H \geq \dim \mathcal{M} = N^2. \tag{B.12}$$

Moreover, equality $H = N^2$ would force $\mathrm{Span}\{R_k\}_{k=1}^{H} = \mathcal{M}$. Every matrix in $\mathcal{M}$ has zero trace, whereas $\mathrm{tr}(R_k) = \|\mathbf{w}_k\|^2$. Thus all R_k would have to be zero, contradicting $\dim \mathcal{M} = N^2 > 0$. Therefore $H \geq N^2 + 1$. Proposition 4.1 supplies the upper bound $H_{\min}(N, d) \leq 2N^2 d$, completing the proof. □

C Detailed Proof of the Polynomial Filtration

The spaces in Theorem 5.1 are linear spans over varying parameters. A particular fixed model can omit degrees through zero weights or cancellation.

Proof of Theorem 5.1. We compute the degrees attainable as the weights vary, rather than those present for every fixed parameter choice. We first analyze the possible degrees by induction on l, and then give a construction that ensures their simultaneous occurrence. For $l = 0$, the output is simply the input X_0. Thus, $\mathcal{D}_0 = \{1\}$, which holds. For $l = 1$, the output is $X_1 = X_0 + \mathcal{A}_1(X_0)$. Since $\mathcal{A}_1$ is homogeneous of degree 3, only degrees 1 and 3 are possible. The construction below realizes both, so $\mathcal{D}_1 = \{1, 3\}$.

Assume the induction hypothesis holds for $l - 1$, meaning $\mathcal{D}_{l-1} = \{1, 3, 5, \ldots, 3^{l-1}\}$. For layer l, the output is:

$$X_l = X_{l-1} + \mathcal{A}_l(X_{l-1}) \tag{C.1}$$

The first term X_{l-1} contributes the polynomial degrees already present in $\mathcal{D}_{l-1}$. The second term $\mathcal{A}_l(X_{l-1})$ comes from a trilinear expression in X_{l-1}. Any scalar monomial contribution is of the form:

$$P(X_0) = c \cdot P_1(X_0)P_2(X_0)P_3(X_0) \tag{C.2}$$

where P_1, P_2, P_3 are monomial terms from X_{l-1}, and c is a scalar coefficient. The degree of this term is the sum of the degrees of its constituents:

$$\deg(P) = d_1 + d_2 + d_3, \quad \text{where } d_1, d_2, d_3 \in \mathcal{D}_{l-1} \tag{C.3}$$

Since d_1, d_2, d_3 are odd integers, their sum $d_1 + d_2 + d_3$ must also be an odd integer. The minimum possible degree is $1 + 1 + 1 = 3$. The maximum possible degree is $3^{l-1} + 3^{l-1} + 3^{l-1} = 3^l$. By choosing appropriate combinations of $d_1, d_2, d_3 \in \{1, 3, \ldots, 3^{l-1}\}$, we can generate any odd integer in the range $[3, 3^l]$. For example, to generate any odd integer $2k + 1 \leq 3^l$, we can partition it as $d_1 + d_2 + d_3$ using odd integers from the previous layer. Thus, the set of degrees at layer l is:

$$\mathcal{D}_l = \mathcal{D}_{l-1} \cup \{3, 5, \ldots, 3^l\} = \{1, 3, 5, \ldots, 3^l\} \tag{C.4}$$

To show the strict containment $\mathcal{V}_{l-1} \subsetneq \mathcal{V}_l$, we observe that by setting $W_V^{(l)} = 0$, the cubic term vanishes and we recover $X_l = X_{l-1}$, which implies $\mathcal{V}_{l-1} \subseteq \mathcal{V}_l$. The containment is strict because $\mathcal{V}_l$ contains terms of degree up to 3^l, which cannot be represented in $\mathcal{V}_{l-1}$ (since the maximum degree in $\mathcal{V}_{l-1}$ is $3^{l-1} < 3^l$). To verify attainment without coefficient cancellation, restrict the input to $X_0 = tE_{11}$, where E_{11} has a single nonzero entry equal to one. In every layer choose query and key projections that select the first feature into the first head coordinate, and a value projection that preserves only that feature. Then $X_l = p_l(t)E_{11}$, with $p_0(t) = t$ and $p_l(t) = p_{l-1}(t) + p_{l-1}(t)^3$. All coefficients are nonnegative. If p_{l-1} contains every odd degree from 1 to 3^{l-1} with positive coefficient, its cube contains every odd degree from 3 to 3^l, and the residual preserves degree 1. Induction therefore gives all claimed degrees, including a nonzero term of degree 3^l. Such a coordinate function cannot belong to $\mathcal{V}_{l-1}$, whose elements have degree at most 3^{l-1}. This completes the proof. □