A Numerical Study of Robustness Verification for Lightning Self-Attention

Anonymous Authors[1]

Abstract

We study the algebraic complexity of robustness verification for lightning self-attention. Since lightning self-attention is polynomial, a binary linear readout defines a cubic decision boundary, and exact Euclidean robustness certification reduces to a nearest-boundary problem. We use the Euclidean distance degree to measure boundary complexity and the ED discriminant to study how the real nearest-boundary problem varies with the input. Our computations show that lightning self-attention boundaries have ED degree far below that of generic cubic hypersurfaces, while still varying systematically with sequence length, embedding dimension, output dimension, and attention rank. We also compute two-dimensional slices of the ED discriminant for a small architecture, showing chambers with different numbers of real critical points.

1. Introduction

Robustness verification asks whether the prediction of a neural network is stable under perturbations of an input. For classifiers with polynomial outputs, exact Euclidean certification is a nearest-boundary problem for an algebraic decision boundary. This perspective was developed theoretically for polynomial neural networks in the work of Alexandr et al. (2026). Here we specialize it to linear self-attention; the corresponding computational study is contained in the aforementioned work.

We study this problem for lightning self-attention, also called linear self-attention, obtained from softmax self-attention by replacing the softmax nonlinearity with the identity map. The resulting attention block is cubic in the input tokens, and after a binary linear readout its decision boundary is a cubic hypersurface.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Our main complexity measure is the *Euclidean distance degree* (ED degree), the generic number of complex critical points of the squared Euclidean distance function to an algebraic variety (Draisma et al., 2016). A generic cubic hypersurface in $\mathbb{R}^N$ has ED degree $3(2^N - 1)$, but lightning self-attention boundaries are highly non-generic because their coefficients come from the structured attention parametrization.

Contribution. We give a focused numerical study of ED degrees and ED discriminants for lightning self-attention decision boundaries. The experiments show that these boundaries have ED degree far below the generic cubic benchmark. At the same time, the degree depends strongly on the factorization $N = td$ into sequence length and embedding dimension, on the output dimension d', and on the rank bound a for the attention matrix. We also approximate ED discriminant chambers for a small attention architecture, showing that the number of real critical points varies across the data space. These computations point to a systematic algebraic structure behind attention decision boundaries and motivate a theoretical study of their complexity.

2. Nearest-boundary certification

Let $f_\theta : \mathbb{R}^N \to \mathbb{R}^k$ be a classifier. For two classes $c \neq c'$, write

$$g_{cc'}(x) = f_{\theta,c}(x) - f_{\theta,c'}(x).$$

The pairwise decision boundary is the hypersurface $\mathcal{V}_{cc'} = \{x : g_{cc'}(x) = 0\}$. For a test input u, exact Euclidean certification requires computing the distance from u to the relevant decision boundary. When $g = g_{cc'}$ is polynomial, the critical equations for this nearest-boundary problem are

$$g(x) = 0, \qquad x - u = \lambda \nabla g(x). \tag{1}$$

The ED degree of $\{g = 0\}$ is the number of complex solutions of (1) for generic u.

For a generic degree-D hypersurface in $\mathbb{R}^N$, the ED degree is $D \sum_{i=0}^{N-1} (D-1)^i$. Hence the generic cubic benchmark is $G_N = 3(2^N - 1)$. We use G_N only as a coarse dense comparison: attention boundaries are structured and sparse, so sharper comparisons should eventually account for the monomial support of the corresponding Lagrange system

(Breiding et al., 2022; Lindberg et al., 2023). The number of monomials that appear in lightning self-attention is $\frac{1}{6}dt(d+1)(3dt-2d+2)$, as proved in (Alexandr et al., 2026, Proposition 5).

3. Lightning self-attention boundaries

We study the ED degree of the decision boundary induced by lightning self-attention followed by a binary linear readout. Fix integers t, d, d', a, and let

$$\varphi_W(X) = VX(X^\top AX) \in \mathbb{R}^{d' \times t}, \tag{2}$$

where $X \in \mathbb{R}^{d\times t}$ is the input, $V \in \mathbb{R}^{d'\times d}$ is the value matrix, and $A \in \mathbb{R}^{d\times d}$ is an attention matrix satisfying $\operatorname{rank}(A) \le a$. The entries of (2) are cubic polynomials in the $N = dt$ entries of X.

Vectorizing $\varphi_W(X)$ and composing with an affine map $L : \mathbb{R}^{d't} \to \mathbb{R}^2$ gives two polynomial scores $f(X) = (f_1(X), f_2(X))$. The binary decision boundary is

$$f_1(X) - f_2(X) = 0.$$

This is the natural boundary associated with the classifier, since the predicted label changes when the two scores are equal. The *ED degree* of this hypersurface measures the number of complex critical points of the squared Euclidean distance from a generic input $u \in \mathbb{R}^{d\times t}$ to the boundary.

4. Experimental ED degrees

Table 1 reports ED degrees for a range of lightning self-attention architectures. Here $G_N = 3(2^N - 1)$ is the ED degree of a generic cubic hypersurface in the same ambient dimension. The computations use parameter homotopy and monodromy in `HomotopyContinuation.jl` (Breiding and Timme, 2018). A specialization from a generic cubic start system is numerically unstable for these structured boundaries, so we instead treat the attention weights and the target data point as parameters of the critical system.

The ED degrees are substantially smaller than the generic cubic benchmark. Thus, although the decision boundary is cubic in $N = td$ variables, it is not a generic cubic hypersurface. The gap reflects the sparse and constrained algebraic structure imposed by the attention parametrization.

The experiments show three main trends. First, for fixed N, the ED degree depends strongly on the factorization $N = td$. The cases $t = N, d = 1$ and $t = 1, d = N$ stabilize at ED degree 7 and 12, respectively, while balanced or higher-dimensional embedding cases can be much larger. Second, for fixed t and d, increasing the rank bound a and output dimension d' usually increases the ED degree. Third, the growth can saturate with d'. For example, when $(t, d, a) = (2, 4, 2)$, the ED degree is 359 for $d' = 2, 3, 4$. This is consistent with the fact that $\varphi_W(X)$ is a $d' \times t$ matrix, so its rank is at most $\min(d', t)$. Once $d' \ge t$, increasing d' no longer changes this rank ceiling.

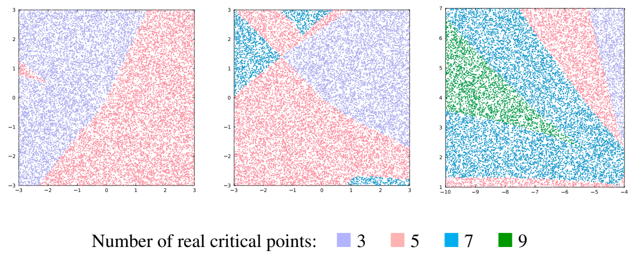

Figure 1. Two-dimensional slices of the data space for a fixed lightning self-attention architecture. Colors indicate the number of real critical points; the visible interfaces approximate the intersections with the ED discriminant.

5. ED discriminant chambers

The ED degree counts complex critical points for generic data, but robustness verification over real inputs also depends on how many of these critical points are real. This number can vary with the test input. The *ED discriminant* is the real hypersurface in data space where the number of real critical points changes; it partitions the data space into chambers with constant real critical-point count (Draisma et al., 2016).

We study this phenomenon for the architecture $(t, d, d', a) = (2, 2, 1, 1)$. Here the input space is $\mathbb{R}^{2\times 2}$ and the ED degree is 17. Since non-real critical points occur in complex conjugate pairs, a generic real data point has an odd number of real critical points. Computing the ED discriminant symbolically is infeasible even in this small attention case, so we approximate the chamber decomposition numerically.

We fix random network parameters, choose a generic base point $u_0 \in \mathbb{R}^{2\times 2}$ and two orthogonal unit directions v_1, v_2, and sample the affine plane $u(s, c) = u_0 + sv_1 + cv_2$. For each sampled point, we use parameter homotopy to track the 17 complex critical points from u_0 to $u(s, c)$ and then count the real solutions.

In these slices we observe chambers with 3, 5, 7, and 9 real critical points. Thus, even when the complex ED degree is fixed, the real nearest-boundary problem has data-dependent structure. This gives an instance-level refinement of the ED degree: the architecture determines the complex critical-point count, while the location of the input relative to the ED discriminant controls the number of real candidates that can contribute to the actual robustness radius.

Table 1. ED degrees for lightning self-attention decision boundaries. The table includes representative computations from the larger experimental sweep. Here $N = td$ and $G_N = 3(2^N - 1)$.

t	d	d'	a	ED	G_N	t	d	d'	a	ED	G_N
4	1	1	1	7	45	1	4	1	1	12	45
2	2	1	1	17	45	2	2	2	2	39	45
6	1	1	1	7	189	1	6	1	1	12	189
3	2	1	1	17	189	2	3	1	1	20	189
3	2	2	1	46	189	2	3	2	1	49	189
3	2	2	2	99	189	2	3	2	3	143	189
8	1	1	1	7	765	1	8	1	1	12	765
4	2	1	1	17	765	2	4	1	1	20	765
4	2	2	1	50	765	2	4	2	1	56	765
2	4	1	2	85	765	2	4	1	4	157	765
2	4	2	2	359	765	2	4	4	2	359	765
9	1	1	1	7	1533	1	9	1	1	12	1533
3	3	1	1	20	1533	3	3	1	2	79	1533
3	3	2	1	85	1533	3	3	3	1	106	1533
3	3	1	3	113	1533	3	3	3	3	895	1533
10	1	1	1	7	3069	1	10	1	1	12	3069
5	2	1	1	17	3069	2	5	1	1	20	3069
5	2	1	2	39	3069	5	2	2	1	50	3069
2	5	2	2	481	3069	2	5	2	5	1155	3069

6. Discussion

Lightning self-attention gives a structured family of cubic decision boundaries whose ED degrees are far below those of generic cubic hypersurfaces. The experiments show that this reduction is not arbitrary: the degree changes systematically with the sequence length, embedding dimension, rank bound, and output dimension. This makes lightning self-attention a useful test case for studying algebraic complexity in neural network verification.

These computations are meant as a starting point for a theoretical study. The repeated patterns in the table suggest that the ED degrees should be explained by the sparsity, rank constraints, and tensor structure of the attention map. The ED discriminant computations point to a complementary real question, namely how the number of real critical points changes across input space. Together, these observations suggest that lightning self-attention provides a tractable family in which to study how architectural structure controls the algebraic geometry of robustness verification.

Impact statement

This work aims to improve the mathematical understanding of robustness verification for structured neural network architectures. The work is theoretical and computational in nature and does not introduce deployed systems or datasets involving human subjects.

References

Yulia Alexandr, Hao Duan, and Guido Montúfar. Algebraic Invariants of Lightning Self-Attention arXiv:2604.15632, 2026.

Yulia Alexandr, Hao Duan, and Guido Montúfar. Robustness Verification of Polynomial Neural Networks arXiv:2602.06105, 2026.

Paul Breiding and Sascha Timme. HomotopyContinuation.jl: A package for homotopy continuation in Julia. In *Mathematical Software – ICMS 2018*, pp. 458–465, 2018.

Paul Breiding, Frank Sottile, and James Woodcock. Euclidean distance degree and mixed volume. *Foundations of Computational Mathematics*, 22:1743–1775, 2022.

Jan Draisma, Emil Horobeţ, Giorgio Ottaviani, Bernd Sturmfels, and Rekha Thomas. The Euclidean distance degree of an algebraic variety. *Foundations of Computational Mathematics*, 16:99–149, 2016.

Julia Lindberg, Leonid Monin, and Kemal Rose. The algebraic degree of sparse polynomial optimization. arXiv:2308.07765, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

A. Computational details

We use two separate homotopy-continuation computations.

ED degree computations. For each architecture (t, d, d', a), let $g(X) = f_1(X) - f_2(X)$ be the cubic decision-boundary polynomial and write $x = \mathrm{vec}(X)$. The ED critical equations are $g(x) = 0$ and $x - u = \lambda \nabla g(x)$.

A direct computation with fixed random network weights is numerically unstable in these examples. One would solve the resulting ED system by specializing from a generic cubic start system, which has $G = 3(2^N - 1)$ paths. Since the attention boundary is a structured cubic with ED degree much smaller than G, many of these paths diverge or become ill-conditioned under specialization. This makes the computation unreliable.

Instead, we keep the network parameters as parameters of the polynomial system. In the rank-constrained case, we write $A = KQ^\top$ with $K, Q \in \mathbb{C}^{d\times a}$ and treat K, Q, V, the readout parameters, and the target point u as parameters:

```
@var X[1:d, 1:t]; x = vec(X)
@var K[1:d, 1:a]; @var Q[1:d, 1:a]
@var V[1:d_prime, 1:d]
@var W_out[1:2, 1:(d_prime*t)]
@var b_out[1:2]
@var u[1:(d*t)]; @var lambda

A = K * transpose(Q)
Y = V * X * (transpose(X) * A * X)
Y_flat = vec(Y)
scores = W_out * Y_flat + b_out
g = scores[1] - scores[2]
grad_g = differentiate(g, x)

params = [vec(K); vec(Q); vec(V); vec(W_out);
          vec(b_out); u]
sys = System([g; x - u - lambda * grad_g],
             [x; lambda], params)
```

To start monodromy, we construct one explicit solution of the ED system. We choose a random complex input x_0, multiplier λ_0, and network weights. We then choose the readout bias so that $g(x_0) = 0$ and set $u_0 = x_0 - \lambda_0 \nabla g(x_0)$. Thus (x_0, λ_0) satisfies the ED critical equations for the corresponding parameter values.

```
# Choose b_out so that g(x0) = 0
b0_2 = randn(ComplexF64)
b0_1 = b0_2 - sum((W0[1, :] .- W0[2, :]) .* Y0_flat)
b0 = [b0_1, b0_2]

# Choose u0 so that x0 - u0 = lambda0 * grad_g(x0)
u0 = x0 - lambda0 * grad_g_eval
```

Starting from this seed solution, `monodromy_solve` follows loops in the parameter space of the attention family and recovers the remaining critical points. The resulting number of complex solutions is the ED degree for generic parameters in the chosen architecture.

ED discriminant slice computations. The ED discriminant computation is separate. Here we fix the architecture $(t, d, d', a) = (2, 2, 1, 1)$ and fix one real random choice of network parameters. This gives a single cubic decision boundary $g(X) = 0$ with ED degree 17. We then define the ED critical system with u as the only parameter:

$$x - u - \lambda \nabla g(x) = 0, \qquad g(x) = 0.$$

Schematically, this is implemented as follows.

```
@var X[1:d, 1:t] u[1:N] lambda
x = vec(X)

# K_val, Q_val, V_val, W_val, b_val are fixed
A_val = K_val * transpose(Q_val)
Y = V_val * X * (transpose(X) * A_val * X)
f = W_val * vec(Y) + b_val
g = f[1] - f[2]

grad_g = differentiate(g, x)
sys = System([x - u - lambda * grad_g; g],
             [x; lambda], u)
```

We choose a generic base point u_0 and compute the 17 ED critical points over this base point. Then we choose two orthonormal directions $v_1, v_2 \in \mathbb{R}^N$ and sample points in the affine plane $u(s, c) = u_0 + sv_1 + cv_2$. For each sampled point, we track the base solutions from u_0 to $u(s, c)$ by parameter homotopy.

```
base_sols = solutions(solve(sys;
            target_parameters = u0))

u_target = u0 + s * v1 + c * v2

res = solve(sys, base_sols;
            start_parameters = u0,
            target_parameters = u_target)
```

At each sampled point, we count the real tracked solutions. Away from the ED discriminant, this count should be odd, since the total ED degree is 17 and non-real solutions occur in conjugate pairs. Even counts indicate samples on or near the ED discriminant, where the real/non-real classification is numerically unstable; we discard these samples when plotting the chambers.