Limitations of Normalization in Attention Mechanism

Timur Mudarisov[1] Mikhail Burtsev[2] Tatiana Petrova[1] Radu State[1]

October 20, 2025

University of Luxembourg
London Institute for Mathematical Sciences

Outline

Motivation

Why Normalization in Attention Matters

1. **Attention is a selector**: it must prioritize informative tokens among many.
2. **Softmax causes vanishing attention** as context length L grows:

$$\alpha_i = O\left(\frac{1}{L}\right)$$

3. **Consequences**
   - Loss of discriminative power between tokens
   - Noise dominating relevant context
   - Unstable gradients when aggressively sharpening

Problem Setting

We study attention as a **general normalized selection mechanism**.

Input: sequence embeddings $X = \{x_i\}_{i=1}^{L}$, $x_i \in \mathbb{R}^d$, and

$$q_m = f_q(x_m), \quad k_n = f_k(x_n), \quad v_n = f_v(x_n).$$

General attention normalisation:

$$a_{m,n} = \frac{F(q_m^\top k_n; \theta)}{\sum_{j=1}^{L} F(q_m^\top k_j; \theta)}, \tag{1}$$

where $F : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a positive **scoring function**.

Goal: Examine following points:

- capacity to **separate informative vs. non-informative tokens**,
- **geometric structure** of selected tokens,
- **gradient stability** during training.

Contributions

Contributions

We provide a quantitative analysis of attention as a **capacity-limited selector**.

1. **Distance bound**:
   - Derive non-asymptotic upper bounds on representation distance
   - Show collapse when the active set size N grows proportionally to context length L
2. **Geometric separability bound**:
   - Analyze attention in embedding space using metric geometry
   - Prove that no more than $\sim 80\%$ of selected tokens can be simultaneously separated
3. **Gradient sensitivity bound**:
   - General Jacobian bound for any normalization function F
   - Recovers the $\frac{1}{4T}$ instability of softmax as a special case
4. **Empirical validation on GPT-2**:
   - Confirm distance collapse, separability saturation, and sharpness–stability trade-off

Key message: normalization fundamentally limits attention capacity.

Distance Bound

Top-N Selection and Representation Distance

Why study top-N selection?

- In attention, most weights α_i are small — only a few tokens matter.
- We model attention as a **token selector**: it highlights the N most relevant tokens.

Formal setup:

- Let $I_N = \{i_1, \dots, i_N\} \subset \{1, \dots, L\}$ - indices of largest attention weights. Aggregated context:
$$s = \sum_{i \in I_N} \alpha_i x_i.$$
- Distance to non-selected tokens (loss of separation):
$$\tilde{d} = \sum_{i \in I \setminus I_N} \|\alpha_i x_i - s\|_2.$$

Goal: measure how well attention separates informative from non-informative tokens.

Theorem: Distance Analysis

Theorem 1 (Non-asymptotic Distance Upper Bound)
Let I_N be the indices of the top-N attention weights $\{\alpha_i\}_{i=1}^{L}$ and $\bar{\alpha}_N = \sum_{i \in I_N} \alpha_i$. Then the representation distance satisfies:

$$\tilde{d} \leq (1 - \bar{\alpha}_N)\, d_1 + \max_{j \in I_N} \|x_j\|_2^2 \Big[\bar{\alpha}_N (L - N) - (1 - \bar{\alpha}_N)\Big],$$

where $d_1 = \max_{i \notin I_N,\, j \in I_N} \|x_i - x_j\|_2$.

If I_N is selected uniformly at random among subsets of size N, then

$$\mathbb{E}[\tilde{d}] = \frac{L - N}{L} \sum_{i=1}^{L} \Big\|(\alpha_i + \tfrac{N}{L-1}) x_i - \bar{x}\Big\|_2^2 + \varepsilon, \qquad \bar{x} = \sum_{i=1}^{L} \alpha_i x_i.$$

Corollary
If N grows proportionally to sequence length L (i.e. $N = \Theta(L)$), then

$$\tilde{d} \to 0,$$

Geometric Separability

Geometric Separability: Definition of N_s

Assumptions.

- Token embeddings lie on a sphere of radius M; minimum pairwise separation $\delta > 0$.
- Let $I_N = \{i_1, \dots, i_N\}$ be indices of the top-N tokens, and

$$s \;=\; \sum_{i \in I_N} \alpha_i x_i \quad \text{(context from selected tokens).}$$

Radius. Choose a tolerance radius so that all non–selected tokens are outside the ball around s:

$$r \;:=\; \min_{j \notin I_N} \left\| \alpha_j x_j - s \right\|_2. \tag{2}$$

Definition (what we count).

$$N_s \;:=\; \#\Big\{\, i \in I_N \;:\; \left\| \alpha_i x_i - s \right\|_2 \le r \,\Big\}. \tag{3}$$

Interpretation: among the N selected tokens, N_s are *geometrically distinguishable* — their (weighted) embeddings stay within the selective ball $B_r(s)$, while every non–selected lies outside.

Interpretation figure

To understand the previous definition better, consider the following figure.

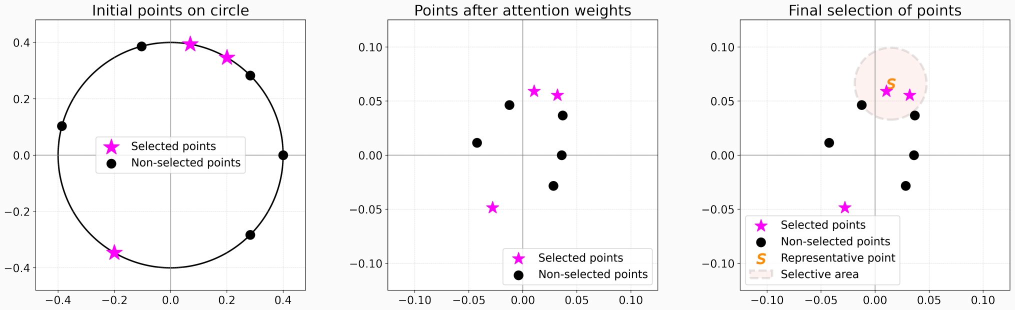

Figure 1: Illustrative example of the geometric separation. **Left:** Token embeddings lie on a circle. **Middle:** After scaling by their attention weights α_i, both attended (magenta stars) and non-attended (black dots) points move toward the origin. **Right:** Only the selected tokens that remain inside the ball $B_r(s)$ (shaded) are deemed distinguishable.

Theorem 2: Bounds on Fraction of Distinguishable Tokens

Theorem (Geometric separability)
Assume embeddings $\{x_i\}_{i=1}^{L}$ lie on a sphere of radius M with minimum pairwise separation $\delta > 0$. Let I_N be the top-N indices, $s = \sum_{i \in I_N} \alpha_i x_i$, and define

$$r := \min_{j \notin I_N} \left\| \alpha_j x_j - s \right\|_2, \qquad N_s := \#\{\, i \in I_N : \|\alpha_i x_i - s\|_2 \le r \,\}.$$

For each $i \in I_N$ set

$$\xi_i^2 = M^2 \sum_{\substack{j \in I_N \\ j \neq i}} \alpha_j^2 \;+\; \left(M^2 - \frac{\delta^2}{2} \right) \sum_{\substack{j,k \in I_N \\ j \neq k,\; j \neq i}} \alpha_j \alpha_k .$$

Then the expected fraction of distinguishable selected tokens satisfies

$$1 - \frac{1}{rN} \sum_{i \in I_N} \xi_i \;\le\; \mathbb{E}\left[\frac{N_s}{N} \right] \;\le\; \frac{1}{N} \sum_{i \in I_N} \exp\left(-\frac{(r - \xi_i)^2}{16 M^2} \right).$$

Gradient Sensitivity

Gradient Sensitivity of Attention: Why It Matters

The attention mechanism must be selective to distinguish informative tokens. However, making attention sharper during training exposes a second difficulty: **gradient sensitivity**.

Consider two nearly identical logit vectors:

$$\ell^{(1)} = (0, \dots, 0, a, a + \varepsilon), \qquad \ell^{(2)} = (0, \dots, 0, a + 2\varepsilon, a),$$

with

$$\|\ell^{(1)} - \ell^{(2)}\|_2 = \sqrt{5}\,\varepsilon.$$

Let $\alpha^{(1)}, \alpha^{(2)}$ be the corresponding attention weight vectors. A first-order expansion gives:

$$\|\alpha^{(1)} - \alpha^{(2)}\|_2 \;\approx\; \|\nabla_\ell \alpha^{(1)}\,(\ell^{(1)} - \ell^{(2)})\|_2 \;\sim\; \sqrt{2}\,\frac{\varepsilon}{T}.$$

Observation: even a tiny change in logits can cause large changes in attention weights when T is small. This makes the gradient step **highly unstable** during training.

Theorem: Gradient Sensitivity of Normalization Functions

Lemma 2 (Jacobian Bound for General Normalizers)
For the attention weights

$$\alpha_i = \frac{F(\ell_i, \theta)}{\sum_{j=1}^{L} F(\ell_j, \theta)},$$

the Jacobian w.r.t. logits satisfies

$$\|\nabla_\ell \alpha\|_2 \;\leq\; \min\left\{ \frac{\|F'\|_2}{L \, \min_j F(\ell_j, \theta)} + \frac{\|F\|_2 \|F'\|_2}{L^2 \, \min_j F^2(\ell_j, \theta)},\; \sqrt{2} \right\}.$$

Corollary (Softmax Instability)
For the softmax normalization $F(z) = \exp(z/T)$,

$$\|\nabla_\ell \alpha\|_2 \leq \min\left\{ \frac{1}{4T}, \sqrt{2} \right\}.$$

Empirical Validation

Experimental Setup

Model.

- GPT-2 (124M, 12 layers, 12 heads/layer); full attention matrices extracted.
- Hidden size: 768, context length $L = 1024$.

Data.

- Consecutive segments from *War and Peace* (public domain).
- BPE tokenization (HuggingFace), no truncation beyond context window.
- 1024-token sequences sampled sequentially (no shuffling).

Metrics.

- **Distance:** $\tilde{d}$ from Theorem 1 (collapse analysis).
- **Separability:** N_s/N from Theorem 2 (geometric capacity).
- **Sensitivity:** finite-difference Jacobian $\|\nabla_\ell \alpha\|_2$.

Results: Distance vs. Sequence Length

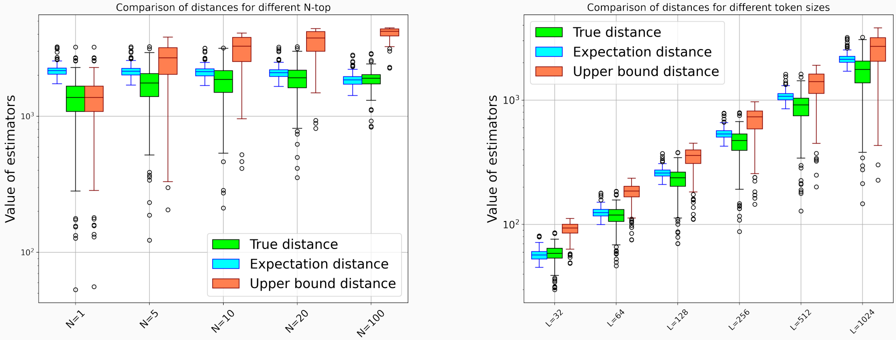

Figure 2: Distance statistics validate Theorem 1. (a) With L = 1024, increasing N beyond 20 yields diminishing returns: the distance plateaus while the bound tightens. (b) With N = 5, both the true distance (green) and its expectation (blue) grow roughly linearly in L; the red upper bound is safe but conservative.

Results: Top-N Plateau

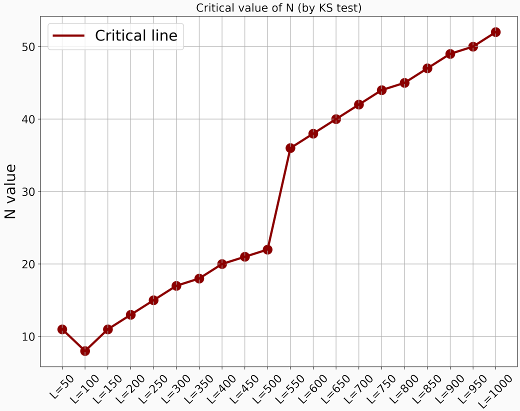

Figure 3: Critical top-N obtained by a KS test ($\alpha = 0.01$); fewer than 6 % of the tokens need to be selected before the empirical and expected distances become statistically indistinguishable.

Results: Geometric Separability Saturation

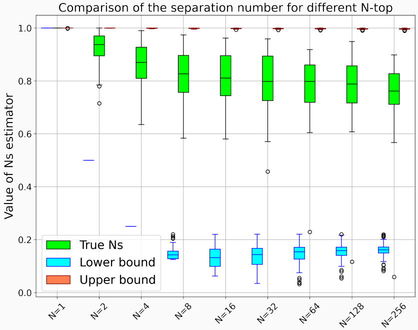

Figure 4: Geometric separability saturates at 70–85%. For increasing top-N, the empirical fraction of distinguishable embeddings N_s/N (green boxes) quickly plateaus; roughly one-fifth of selected tokens remain outside $B_r(s)$. The red line shows the exponential upper bound from Theorem 2, while the blue line shows the conservative lower bound.

Results: Gradient Sensitivity vs. Temperature

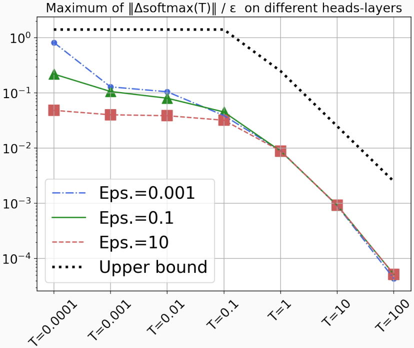

Figure 5: Gradient sensitivity decays as 1/T. Maximum finite-difference Jacobian norm $g(T, \varepsilon)$ for three perturbation magnitudes (coloured curves, log–log scale). The dashed black curve is the theoretical bound $\min\{1/(4T), \sqrt{2}\}$ from gradient's corollary.

Implications

Selectivity vs. Stability Trade-off

From theory:

- From Theorem 1: increasing N (active set size) leads to **distance collapse**.
- From Theorem 2: geometric capacity is **bounded** ($N_s/N \leq 0.8$ even ideally).
- From Gradient Lemma: sharp attention ($T \to 0$) **explodes sensitivity**:

$$\|\nabla_\ell \alpha\|_2 \leq \frac{1}{4T}.$$

Conclusion: attention cannot be simultaneously

- highly selective (*small* N, *sharp* distribution),
- stable during optimization (*bounded gradients*),
- and robust to long context (*large* L)

$\Rightarrow$ **Every normalization rule must trade off selectivity vs. stability.**

Design Guidelines Derived from Theory

The following guidelines follow directly from our theoretical results.

- **Control active set size (Theorem 1)**:

  $$N \ll L \quad \Rightarrow \quad \text{avoid distance collapse and loss of token discrimination.}$$

  Use small top-k selection or sparsity constraints.

- **Monitor geometric capacity (Theorem 2)**:

  $$\frac{N_s}{N} \to 0.7\text{–}0.85 \quad \Rightarrow \quad \text{head is saturated.}$$

  Use N_s/N or attention entropy to detect when a head stops being selective.

- **Avoid sharp softmax (Gradient Lemma)**:

  $$\|\nabla_\ell \alpha\|_2 \propto \frac{1}{T} \quad \Rightarrow \quad T \lesssim 0.1 \text{ leads to unstable gradients.}$$

- **Use adaptive normalization**: Length-aware (Scalable-Softmax), sparse (Sparsemax/Entmax), or gradient-controlled (SA-Softmax) normalizers mitigate the selectivity–stability trade-off.

Conclusion

Conclusion

Normalization fundamentally limits the capacity of attention.

- **Capacity limit:** Any F independent of L forces $\alpha_i = O(1/L)$ — attention mass vanishes as context grows.
- **Distance collapse (Theorem 1):** once $N = \Theta(L)$, attention loses separation power between informative and non-informative tokens.
- **Geometric bound (Theorem 2):** at most 70–85% of selected tokens remain geometrically distinguishable — heads have finite resolution.
- **Gradient instability:** sharper distributions amplify sensitivity as $\|\nabla_\ell \alpha\|_2 \propto 1/T$.

Implication:

No normalization rule can be simultaneously sharp, stable, and long-context scalable.

Questions?