Pay Attention to Attention Distribution: A New Local Lipschitz Bound for Transformers

Nikolay Yudin[1], Sergei Kudriashov[1], Alexander Gaponov[1], and Maxim Rakhuba[1]

HSE University
neyudin@edu.hse.ru

Abstract. We introduce a novel upper bound on the local Lipschitz constant of the dot-product self-attention block showing its dependence on the attention map distributions. The proposed bound is not only tighter than the prior art, but for the first time, reveals how the distribution of attention probabilities shapes the local Lipschitz constant of the self-attention block. The theoretical basis of the proposed upper bound lies in the refined closed-form upper bounds on singular values of the Jacobian of softmax function. Leveraging these theoretical insights, we introduce `JaSMin` (**Ja**cobian **S**oftmax norm **Min**imization), a lightweight regularizer that directly controls the local Lipschitz constant of each block and, consequently, the entire model. Additionally, we discuss how the nature of the attention map distribution contributes to the gradient dynamics and, consequently, transformer training stability.

Keywords: Spectral analysis · Softmax Jacobian · Lipschitz constant

1 Introduction

In this work, we analyze the spectral norm of the Jacobian of dot-product self-attention [38], which governs its local Lipschitz properties. While prior works attempted to bound the local Lipschitz constant of self-attention [3, 6, 18, 22], they disregarded the structure of the attention score matrix, proposing bounds that were dependent only on weights' and inputs' norms. We consider this fact to be a significant limitation of these bounds as it naturally omits the powerful component from the analysis. In our work, we pay a particular attention to this problem, emphasizing the dependency of the self-attention local Lipschitz constant on attention map distribution. In particular, we establish a novel theoretical upper bound on the local Lipschitz constant of self-attention that is based on the new refined spectral properties of the softmax Jacobian.

Overall, contributions of this work can be summarized as follows:

– We introduce interlacing bounds for *all* singular values of the softmax Jacobian, improving eigenvalue interlacing theorem [16, 37] (see Theorem 2). For the largest singular value, we prove that our bound surpasses the bound based on the Gershgorin theorem [19, 33].

- We introduce an improved upper bound on the spectral norm of the self-attention Jacobian, which captures the nature of attention map scores (see Theorem 3). We compare our bound with known estimates and show that our bound is tighter than prior art for different attention-based vision models.
- Utilizing the proposed bounds on the softmax Jacobian, we show that the gradient vanishes for both the categorical and uniform attention map distributions. In either extreme, this signal decay hinders optimization for these attention heads and layers, linking these phenomena to the softmax map spectrum. In Section 6, we discuss connection of these findings to prior art.
- Using the proposed bounds, we present `JaSMin` (**Ja**cobian **S**oftmax **Min**imization), a regularization method aimed to control the self-attention Jacobian spectral norm and quantify gradient sensitivity (see Section 7). We show that our regularizer reduces the spectral norm of the whole model Jacobian, supporting our theory. In practice JaSmin leads to the increase of adversarial metrics on CIFAR [24] and Imagenette [17] datasets (see Section 8).

2 Related Work

The dot-product self-attention [38] has become a dominant part of most transformer models, serving as a default option for production-grade large-scale models [20, 44]. Despite its popularity, the theoretical understanding of the fundamental properties of this operation is still incomplete. Recently, it has been observed that dot-product attention is susceptible to fluctuations, which has a negative impact on the model's performance, especially apparent for long sequences, and has motivated researchers to seek for sparse attention alternatives [21, 46]. Another noticeable pattern is an appearance of "attention sinks" [10, 15, 42], when an excessive amount of attention mass allocated to certain tokens, which turn out to serve as "sinks" for the excess attention, when further modifications to the residual stream are no longer necessary. On the other hand, [8] has argued that pure dot-product attention possesses a strong bias towards "token uniformity", which leads to the rank collapse and oversmoothing in deep models [36]. While layer normalization [41] and higher lipschitzness of MLP layers were shown to stabilize the self-attention dynamics and prevent the doubly exponential convergence to rank-1 representations, the theoretical understanding of interconnections between these phenomena remains limited.

In the visual domain, the rapid adoption of Vision Transformers (ViTs) [9] has prompted parallel investigations into their architectural robustness, sensitivity, and efficiency. To address the fundamental limitations of standard dot-product attention, architectures like ShiftViT [39] and LipShiFT [30] proposed abandoning attention entirely in favor of lightweight shifting, while Robust Vision Transformer [29] introduced position-aware attention scaling to improve inherent resistance to corruptions. Beyond the scope of adversarial robustness, foundational modifications for specific tasks and computational regimes were proposed: Swin Transformer [26] localizes attention to shifted windows, improving computational efficiency; CViT [40] switched to encoder-decoder transform-

ers for PDE systems, while EA-ViT [49] adapted models for resource-constrained deployment.

In this work we study attention-based vision models stability problem through the lens of neural network Lipschitz constant. It is known that vulnerability to adversarial attacks is closely linked to Lipschitz continuity and the Lipschitz constant of the neural network [31]. Despite its successful widespread use, the dot-product self-attention is not globally Lipschitz [22]. Such a functional property, may theoretically cause training instability and poor robustness [28, 48].

Recent works that analyze the self-attention's local Lipschitz constant [3, 6, 18, 22] are aimed at bounding the local Lipschitz constant without taking the attention map into account, building bounds dependent on separate norms of inputs and model weights. We argue that, being a core element of self-attention, softmax largely contributes to the local Lipschitz constant bound, motivating us to explore the local Lipschitz constant dependency from attention map scores.

Multiple prior works considered softmax to be 1-Lipschitz [11, 14, 18, 25]. We admit that the concurrent work [33] also proposes that global Lipschitz constant of softmax function can be bounded by $\frac{1}{2}$ under any ℓ_p norm. Our work focuses on bounding all softmax Jacobian singular values, improving eigenvalue interlacing theorem [16, 37] and the result from [33] for the spectral norm.

3 Notation and Background

A neural network can be treated as a function $f\colon \mathcal{X} \subseteq \mathbb{R}^D \to \mathbb{R}^d$, mapping input data to corresponding outputs. A typical approach to quantify its sensitivity to input perturbations is to analyze its Lipschitz continuity. A Lipschitz constant of f over an open set $\mathcal{X} \subseteq \mathbb{R}^D$ with respect to Euclidian norm $\|\cdot\|_2$ can be defined as follows:

$$\texttt{Lip}(f;\mathcal{X}) = \sup_{\substack{x_1,x_2\in\mathcal{X}\\ x_1\neq x_2}} \frac{\|f(x_1)-f(x_2)\|_2}{\|x_1-x_2\|_2}. \tag{1}$$

Rather than considering the global Lipschitz continuity (where $\mathcal{X} = \mathbb{R}^D$), we focus on the analysis of the *local* Lipschitz constants, which is less computationally prohibitive. Specifically, we examine the behavior of f in the neighborhood of balls $B_\varepsilon(x_0)$ of a radius $\varepsilon > 0$ centered at various data points x_0. Assuming differentiability and denoting the Jacobian of $f(x)$ by $\mathcal{J}_f(x) \in \mathbb{R}^{d\times D}$, for each such ball we may write:

$$\texttt{Lip}(f;\mathcal{X}) = \sup_{x\in\mathcal{X}} \|\mathcal{J}_f(x)\|_2, \tag{2}$$

where $\|\cdot\|_2$ for a matrix denotes the spectral norm of a Jacobian matrix, which is equal to its largest singular value. For notation simplicity, we further write all the theoretical results in terms of the spectral norm of the Jacobian matrix, implicitly omitting the supremum across all data points.

In our paper, we focus on controlling the local Lipschitz constant of separate dot product self-attention blocks. The choice of considering local Lipschitz constant instead of global one is based on the fact that dot-product self-attention is not globally Lipschitz [22]. Despite the absence of global lipschitzness, there is an interest in achieving training stability and analyzing key components of attention-based models making them robust to input perturbations.

Let us define dot product self-attention [38] $\texttt{Attn}_h : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times d}$:

$$\texttt{Attn}_h(X) = \texttt{sm}\left(\frac{XW_h^Q(W_h^K)^\top X^\top}{\sqrt{d}}\right)XW_h^V = \texttt{sm}\left(XA_hX^\top\right)XW_h^V, \quad (3)$$

where $X \in \mathbb{R}^{N\times D}$, $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{D\times d}$, $\texttt{sm}$ denotes row-wise softmax and

$$A_h = \frac{W_h^Q(W_h^K)^\top}{\sqrt{d}}. \quad (4)$$

Additionally, let us denote the attention map matrix $P^h \in \mathbb{R}^{N\times N}$ for a head h:

$$P^h = \texttt{sm}(XA_hX^\top), \quad (5)$$

whose rows we further denote as $P^h_{i,:}$. For simplicity, we omit output matrix W_h^O.

In [22] authors firstly derived the closed form for the single-head self-attention Jacobian: using the notation above, we may write the Jacobian as a block matrix:

$$\mathcal{J}_{\texttt{Attn}_h}(X) = \left[\mathcal{J}_{\texttt{Attn}_h}(X)_{ij}\right]_{i,j=1}^{N,N}, \quad (6)$$

so that each $\mathcal{J}_{\texttt{Attn}_h}(X)_{ij} \in \mathbb{R}^{d\times D}$ can be expressed in terms of $W_h^V, X, P^h_{i,:}, A_h$:

$$\mathcal{J}_{\texttt{Attn}_h}(X)_{ij} = (W_h^V)^\top X^\top \mathcal{M}(P^h_{i,:})(E_{ji}XA_h^\top + XA_h\delta_{ij}) + (W_h^V)^\top \cdot P^h_{ij}, \quad (7)$$

where

- $E_{ji} \in \mathbb{R}^{N\times N}$ — a binary matrix with zeros everywhere except (j,i)-th place;
- $\mathcal{M}(P^h_{i,:}) = \text{diag}(P^h_{i,:}) - (P^h_{i,:})^\top P^h_{i,:}$;
- $\delta_{ij} \in \{0,1\}$ is the Kronecker delta.

As a result, the aforementioned Jacobian can be rewritten in a pure matrix form:

$$\begin{aligned}\mathcal{J}_{\texttt{Attn}_h}(X) = (I_N \otimes (W_h^V)^\top X^\top)\,\text{bdiag}\left(\mathcal{M}(P^h_{1,:}), \ldots, \mathcal{M}(P^h_{N,:})\right)\\ (I_N \otimes XA_h^\top + (XA_h \otimes I_N)\mathcal{P}_{N,D}) + P^h \otimes (W_h^V)^\top,\end{aligned} \quad (8)$$

where $\otimes$ denote Kronecker product, $\text{bdiag}(\mathcal{M}(P^h_{1,:}), \ldots, \mathcal{M}(P^h_{N,:}))$ denote a block diagonal matrix with $\mathcal{M}(P^h_{i,:})$ on the diagonal and $\mathcal{P}_{N,D}$ is a "perfect shuffle" permutation matrix [12] satisfying $\text{vec}(X^\top) = \mathcal{P}_{N,D}\text{vec}(X)$ where vec is taken in "c" order.

Utilizing its structure and refined properties of the softmax Jacobian (see Theorem 2), we show both theoretically and empirically that the spectral norm of the matrix from Equation (6) heavily depends on the $\text{bdiag}(\mathcal{M}(P^h_{1,:}), \ldots, \mathcal{M}(P^h_{N,:}))$ (see Theorem 3).

4 Refined spectral properties of the softmax Jacobian

Self-attention mechanism relies on the softmax function, which critically influences the structure of its Jacobian. In this section, we derive new properties of the softmax Jacobian — key to our later analysis of the full self-attention Jacobian. Beyond this immediate application, our results refine previously proposed findings regarding the softmax Jacobian spectral norm [11, 14, 18, 25, 33]. Additionally, our theoretical result improves classic eigenvalue interlacing theorem [16, 37] for this particular structured matrix.

Recall that the softmax function $\texttt{sm}\colon \mathbb{R}^N \to \mathbb{R}^N$ is defined as

$$\texttt{sm}(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}. \tag{9}$$

It is well-known that the softmax Jacobian $\mathcal{J}_{\texttt{sm}}(z) \in \mathbb{R}^{N\times N}$ is a symmetric positive semidefinite matrix of the following form:

$$\mathcal{J}_{\texttt{sm}}(z) = \operatorname{diag}(\texttt{sm}(z)) - \texttt{sm}(z)\texttt{sm}(z)^\top. \tag{10}$$

For simplicity, we further use the following notation:

$$\mathcal{M}(x) = \operatorname{diag}(x) - xx^\top \tag{11}$$

Prior to stating the theorem, let us recall key auxiliary results and define notation central to our analysis. First of all, we formulate eigenvalue interlacing theorem [16, 37] and recall the results introduced in [33].

Theorem 1 (Eigenvalue interlacing theorem [16, 37]). *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices and $\alpha_1 \geqslant \cdots \geqslant \alpha_n$, $\beta_1 \geqslant \cdots \geqslant \beta_n$ be their eigenvalues respectively. If $A = B + \varepsilon pp^\top$, $\|p\|_2 = 1$, $\varepsilon > 0$, then*

$$\alpha_1 \geqslant \beta_1 \geqslant \alpha_2 \geqslant \beta_2 \geqslant \cdots \geqslant \alpha_n \geqslant \beta_n. \tag{12}$$

Directly applying eigenvalue interlacing theorem to the softmax Jacobian matrix $\mathcal{M}(\texttt{sm}(z))$, one may obtain the following interlacing property: for $x = \texttt{sm}(z)$ setting $B = \operatorname{diag}(x) - xx^\top$, $A = \operatorname{diag}(x)$, $p = x/\|x\|_2$, $\varepsilon = \|x\|_2^2$ we obtain

$$x_{(1)} \geqslant \sigma_1(B) \geqslant x_{(2)} \geqslant \sigma_2(B) \geqslant \cdots \geqslant x_{(N)} \geqslant \sigma_N(B), \tag{13}$$

where $x_{(k)}$ here and further is top-k largest vector value. However, even for the case $N = 2$ this bound appears to be loose for the first singular value: in the case of 2-dimensional probability distribution $p = [1-\varepsilon, \varepsilon]$, $0 < \varepsilon < 1$ for a sufficiently small ε eigenvalue interlacing theorem gives the following upper bound:

$$1 - \varepsilon \geqslant \|\mathcal{M}([1-\varepsilon, \varepsilon])\|_2 = \left\| \begin{bmatrix} \varepsilon(1-\varepsilon) & \varepsilon(1-\varepsilon) \\ \varepsilon(1-\varepsilon) & \varepsilon(1-\varepsilon) \end{bmatrix} \right\|_2 = 2\varepsilon(1-\varepsilon). \tag{14}$$

This example indicates that bounds obtainable by eigenvalue interlacing theorem can be improved by utilizing the structure of $\mathcal{M}(x)$. In [33] authors utilize

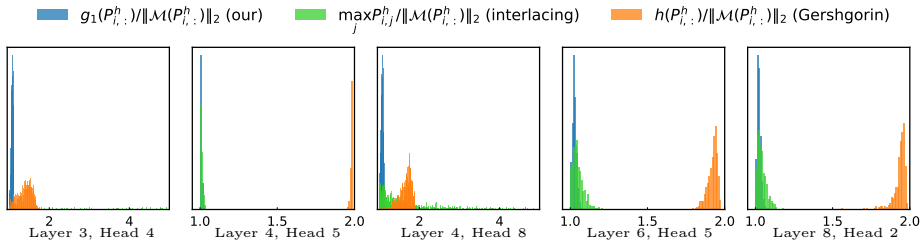

Fig. 1: Comparison of different upper bounds on the spectral norm of the softmax Jacobian. Our bound $g_1(P^h_{i,:})/\|\mathcal{M}(P^h_{i,:})\|_2$ is blue, Gershgorin $h(P^h_{i,:})/\|\mathcal{M}(P^h_{i,:})\|_2$ is orange and interlacing $\max_j P^h_{i,j}$ is green. As anticipated, our bound is tighter than the others. At the same time, the interlacing theorem bound may have large relative approximation error in comparison with the other bound, which is supported by Table 1. Heads, query pixels and ImageNet samples are kept the same for Figure 3.

the structure of $\mathcal{M}(x)$ and prove that $h(x) \coloneqq 2x_{(1)}(1 - x_{(1)})$ upper bounds the spectral norm of the softmax Jacobian using Gershgorin theorem:

$$\|\mathcal{J}_{\mathtt{sm}}(z)\|_2 \leqslant 2 \max_{i=1,\dots,N} (\mathtt{sm}(z)_i)\,(1 - \max_{i=1,\dots,N} \mathtt{sm}(z)_i). \tag{15}$$

We argue that the structure of $\mathcal{M}(x)$ can be utilized even better than it is done in [33]. Below we formulate our main theoretical result about softmax Jacobian singular values.

Definition 1. *Let $x \in \mathbb{R}^N_{\geqslant 0}$ and let $x_{(k)}$ be the k-th largest component of x. For $k = 1, \dots, N-1$ define*

$$g_k(x) \coloneqq x_{(k)}(1 - x_{(k)} + x_{(k+1)}) \tag{16}$$

and, for consistency, $g_N(x) \equiv 0$.

Theorem 2. *Let $x \in \mathbb{R}^N_{\geqslant 0}$ satisfy $\mathbf{1}^\top x = 1$, where $\mathbf{1}$ is a vector of all ones. Then, the singular values $\sigma_i(A)$ of $A = \operatorname{diag}(x) - xx^\top$ admit the following interlacing property with $x_{(i)}$ and $g_i(x)$ from Definition 1:*

$$x_{(1)} \geqslant g_1(x) \geqslant \sigma_1(A) \geqslant x_{(2)} \geqslant g_2(x) \geqslant \sigma_2(A) \geqslant \cdots \geqslant x_{(N)} \geqslant g_N(x) \geqslant \sigma_N(A). \tag{17}$$

Proof. See Appendix A for the proof.

Remark 1. From the theorem above it immediately follows that softmax is 1/2-Lipschitz w.r.t Euclidian norm: $\forall x, y \in \mathbb{R}^n$ we have

$$\|\mathtt{sm}(x) - \mathtt{sm}(y)\|_2 \leqslant 1/2 \cdot \|x - y\|_2. \tag{18}$$

4.1 How accurate is the new upper bound?

We highlight that Theorem 2 for the first singular value is tighter than direct application of both Gershgorin ($h(x)$) and interlacing theorems ($x_{(1)}$):

$$\sigma_1(A) \leqslant g_1(x) \leqslant h(x), \quad \sigma_1(A) \leqslant g_1(x) \leqslant x_{(1)}. \tag{19}$$

Now let us examine what happens on a real-world example. Figure 3 presents various ViT-L attention maps for a certain sequence token i, averaged over 1000 ImageNet samples. We report the average exact value of $\|\mathcal{M}(P^h_{i,:})\|_2$ and the corresponding upper bounds derived via our approach and the Gershgorin theorem.

We observe that our bound consistently outperforms the Gershgorin theorem bound in all the examples. A more detailed view on the distribution of the upper bounds is given in Figure 1. In this figure, we plot the distribution of the ratio $g_1(P^h_{i,:})/\|\mathcal{M}(P^h_{i,:})\|_2$ and compare it with the distribution of the ratio $h(P^h_{i,:})/\|\mathcal{M}(P^h_{i,:})\|_2$ for the same setting as on Figure 3. It can be seen that our bound is noticeably tighter than the other bounds. To additionally show that our bound is sharper than the bound from eigenvalue interlacing theorem, we provide Table 1 which reports mean relative approximation error for different softmax Jacobian spectral norm upper bounding methods. Similarly to setups for Figure 3 and Figure 1, we average the relative approximation error across all token distributions, heads and layers.

Table 1: Mean relative errors of different softmax Jacobian spectral norm estimation methods. Values are averaged across all heads, layers and samples.

Th. 2 (ours)	Th. 1	Eq. 15
0.035	20.19	0.85

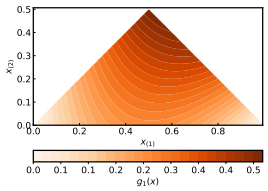

Fig. 2: $g_1(x)$ for $x \in \mathbb{R}^n : \mathbf{1}^\top x = 1$. $\max_x g_1 = 1/2$ is attained at $(x_{(1)}, x_{(2)}) = (1/2, 1/2)$ and $\min_x g_1 = 0$ is attained at $(1, 0)$, $(0, 0)$.

Another noticeable effect is that the smaller the norm values, the closer the distribution becomes to either categorical or uniform distributions. To understand this, we provide draw the behaviour of $g_1(x)$, on Figure 2. We see that these components can be small only at the corners $(1, 0)$, $(0, 0)$.

5 A new upper bound for self-attention Jacobian

Previous works that aimed to bound the self-attention Jacobian spectral norm [3, 6, 18] were primarily focused on the dependency of the self-attention Jacobian

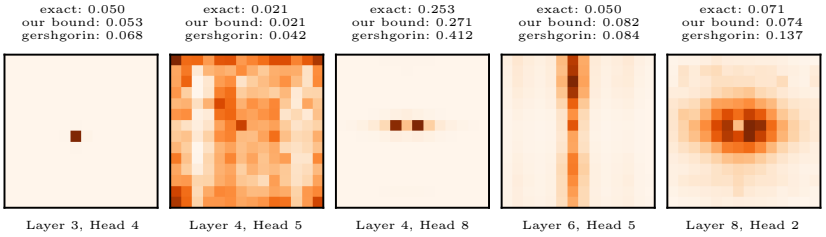

Fig. 3: Attention probabilities for ViT-L model for certain heads. The title of each heatmap describes true average value of $\|\mathcal{M}(P^h_{i,:})\|_2$ together with its our and Gershgorin bounds. Each attention map is averaged across 1000 ImageNet samples.

on the norms of inputs and model weights. Such an approach is rather limiting as it completely disregards all the information about attention map scores, which is a core non-linear mechanism inside self-attention. In prior paper [18] authors omit the term containing softmax by bounding corresponding spectral norm by 1. In paper [3] authors use probabilistic techniques to bound the spectral norm of the self-attention Jacobian, but still omit any softmax scores' dependencies in their main results. In this section, we introduce a novel approach to bound the spectral norm of self-attention Jacobian taking into account attention map scores nature. Below we provide our core theoretical contribution.

Theorem 3. *For the self-attention mechanism defined in Equation* (3)*, the following inequality holds:*

$$\|\mathcal{J}_{\texttt{Attn}_h}(X)\|_2 \leqslant \left\|W_h^V\right\|_2 \cdot \left(\left\|P^h\right\|_2 + 2\,\|X\|_2^2\,\|A_h\|_2 \max_i \left\|\mathcal{M}(P^h_{i,:})\right\|_2 \right), \tag{20}$$

where $\mathcal{M}(P^h_{i,:})$ *is defined as in Equation* (11)*.*

Proof. See Appendix B for the proof.

Remark 2. Note that due to the continuity of the softmax function, we can readily obtain the local Lipschitz bound in a neighborhood of X from Equation (20).

Remark 3. Since multi-head self-attention is simply a linear combination of individual attention heads followed by an output projection, we can bound its norm by the sum of the corresponding spectral norms of each head's transformation:

$$\left\|\left[\mathcal{J}_{\texttt{Attn}_1}(X) \;\dots\; \mathcal{J}_{\texttt{Attn}_H}(X)\right]\right\|_2 \leqslant \sum_{i=1}^{H} \|\mathcal{J}_{\texttt{Attn}_i}(X)\|_2\,. \tag{21}$$

This justifies our use of a summed penalty across heads and layers in the regularization term in Section 7.

Our result provides a conditional interpretation of sink-like behaviour noticed in papers [15, 42] through the lens of self-attention Jacobian matrix spectral properties. The Lipschitz constant decreases when row-wise attention distributions enter a highly concentrated, near-categorical regime. Importantly, the

similarity to attention sinks is that both involve concentration of attention mass, though it is not identical to the usual empirical notion of attention sinks, where the same token or position consistently attracts attention across all inputs. Overall, our upper bound provides a partial theoretical explanation that sufficiently concentrated attention mass produces in a stabilizing effect.

5.1 Comparison with existing bounds

Below we recall existing results for upper bounds of dot-product self-attention local Lipschitz constant in order to compare them with our bound.

Let $B_R(y) \subset \mathbb{R}^D$ represent an open ℓ_2-ball of radius R centered at $y \in \mathbb{R}^D$. Additionally, let $B_R^N(y) \subset \mathbb{R}^{N\times D}$ for $y \in \mathbb{R}^D$ denote a set of matrices $Y \in \mathbb{R}^{N\times D}$ such that $\forall\, i = 1, \ldots, N : \|Y_{i,:}\|_2 \in B_R(y)$ and $\mathrm{vec}(W)$ is a column-wise vectorization of a matrix into a vector.

Theorem 4 ([18]). *The spectral norm of a single-head self-attention admits the following bound:*

$$\|\mathcal{J}_{\mathit{Attn}_h}(X)\|_2 \leqslant N(N+1)\cdot\|X\|_F^2\cdot\Big(\|W_h^V\|_2\|W_h^Q\|_2\|W_h^K\|_2+\|W_h^V\|_2\Big). \quad (22)$$

The upper bound proposed in paper [18] is far from the real spectral norm because it depends quadratically on N. In paper [3] authors provide a tight bound for the local Lipschitz constant of unmasked self-attention and propose to bound the spectral norm of self-attention Jacobian as $\mathcal{O}(\sqrt{N})$ in N.

Theorem 5 ([3]). *Let $X \in \mathbb{R}^{N\times D}$ and there exists such $R > 0$, so that $X \in B_R^N(0) \subset \mathbb{R}^D$. Then,*

$$\|\mathcal{J}_{\mathit{Attn}_h}(X)\|_2 \leqslant \sqrt{3}\|W_h^V\|_2\left(\|A_h\|_2^2R^4(4N+1)+N\right)^{\frac{1}{2}}. \quad (23)$$

One may notice that the bound provided in Theorem 3 has a worse order in N than the bound provided in Theorem 5. Indeed, the proposed bound can be linear in N providing that X has the spectral norm of order $\mathcal{O}(\sqrt{\max(N,D)})$ for both random or Pre-LayerNorm of X [43]. However, in our upper bound we have a term $\max_{i=1,\ldots,N}\|\mathcal{M}(P_{i,:}^h)\|_2$ which in practice makes our bound more precise than the one provided in Theorem 5. To validate it empirically, we take ViT-L and process ImageNet samples showing that our bound on the self-attention Jacobian spectral norm is tighter than prior art for a sequence length $N = 196$ across all heads and layers (see Figure 4). We observe that our bound from Theorem 3 is tighter, outperforming all existing bounds within a considerable margin. This observation supports our theory regarding the influence of attention map distribution on self-attention local Lipschitzness.

Despite the fact that in practice $\max_{i=1,\ldots,N}\|\mathcal{M}(P_{i,:}^h)\|_2$ gives more advantage than $\sqrt{N}$, we are also able to improve the bound from Theorem 5 by a constant factor by slightly altering the end of our proof (see Appendix C).

Proposition 1 *Under the assumptions of Theorem 5:*

$$\|\mathcal{J}_{\mathit{Attn}_h}(X)\|_2 \leqslant \|W_h^V\|_2\big(\sqrt{N}+2\sqrt{N}R^2\|A_h\|_2\big). \quad (24)$$

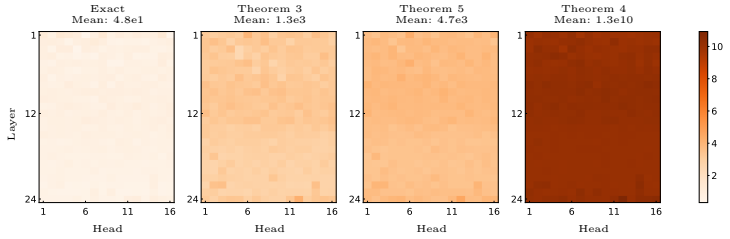

Fig. 4: Self-attention Jacobian spectral norms' comparison for each head and layer of the ViT-L model across 5000 ImageNet samples. The title of each heatmap represents the mean value and standard deviation of the bound across all heads and values.

6 Gradient sensitivity to attention map distributions

1/2-lipschitzness of the softmax mapping already motivates increased stability of transformers to exploding gradients due to the upper bound on each layer's sensitivity, our approach allows for a more fine-grained characterization of gradient dynamics as well.

Although bounding the local lipschitzness was noticed to have a positive effect on the robustness properties of deep learning models [45], it was also shown to tangle generalization and lead to well-known accuracy-robustness trade-off [2]. Unlike previous works, that have mostly focused on generalization properties of locally Lipschitz functions [32], we aim to show that softmax distributions significantly impact gradient dynamics and affect optimization — observations, which has independently been made in studies, which focused on signal propagation, training dynamics and "sink"-like effects in large-scale transformer-based models [15, 34, 47].

To illustrate the point, notice that the gradient of an arbitrary parameter of an attention block X of the shallower layer can be expressed as a matrix-vector multiplication of the gradient w.r.t. its output with the adjoint Jacobian:

$$\left\langle \frac{\partial L}{\partial \texttt{Attn}_h(X)}, \mathcal{J}_{\texttt{Attn}_h}(X)dX \right\rangle = \left\langle \mathcal{J}_{\texttt{Attn}_h}(X)^\top \frac{\partial L}{\partial \texttt{Attn}_h(X)}, dX \right\rangle. \tag{25}$$

Then, with a little abuse of notation, we can bound gradient norms by the product of the Jacobian spectral norm and a gradient obtained from backpropagation.

Thus, for any $W \in \{W_h^Q, W_h^K, W_h^V\}$ from self-attention we have:

$$\left\| \frac{\partial L}{\partial W} \right\|_F = \left\| \text{vec}\left(\frac{\partial L}{\partial W} \right) \right\|_2 \leq \left\| \frac{\partial \texttt{Attn}_h(X)}{\partial W} \right\|_2 \left\| \frac{\partial L}{\partial \texttt{Attn}_h(X)} \right\|_F. \tag{26}$$

It can be seen that gradients w.r.t. the weights depend not only on the sensitivity of deeper layers to attention block's output, but on the spectral norm of the gradient w.r.t. W_h^Q, W_h^K, W_h^V. To show a connection between the attention map distribution nature and the gradient norm, below we provide a distribution-dependent upper bound for $\|\partial \texttt{Attn}_h(X)/\partial W\|_2$, where $W \in \{W_h^Q, W_h^K, W_h^V\}$.

Proposition 2 *For the self-attention mechanism defined in Equation* (3)*, gradient norms can be bounded as follows:*

$$\left\| \frac{\partial \texttt{Attn}_h(X)}{\partial W_h^Q} \right\|_2 \leqslant \frac{1}{\sqrt{d}} \|W_h^V\|_2 \|W_h^K\|_2 \|X\|_2^3 \max_i \left\| \mathcal{M}(P_{i,:}^h) \right\|_2 \tag{27}$$

$$\left\| \frac{\partial \texttt{Attn}_h(X)}{\partial W_h^K} \right\|_2 \leqslant \frac{1}{\sqrt{d}} \|W_h^V\|_2 \|W_h^Q\|_2 \|X\|_2^3 \max_i \left\| \mathcal{M}(P_{i,:}^h) \right\|_2 \tag{28}$$

and for W_h^V*:* $\left\| \frac{\partial \texttt{Attn}_h(X)}{\partial W_h^V} \right\|_2 \leqslant \|P\|_2 \|X\|_2 \leqslant \sqrt{\max_i \sum_{j=1}^N P_{ij}} \; \|X\|_2.$

Proof. See Appendix F for the proof.

To support Proposition 2 empirically, we measure gradient spectral norm w.r.t. Q and K matrices for 1000 ImageNet test samples along with $g_1(P_{i,:}^h)$ — the upper bound on $\max_i \|\mathcal{M}(P_{i,:}^h)\|_2$. Figure 5 shows the correlation between gradient norms divided by $\|X\|_2^3$, which allows to exclude input norm dependency in our bound (following Proposition 2), and $\max_i g_1(P_{i,:}^h)$.

Let us make a few observations regarding the gradient structure. The first thing to be noted is the difference in gradient dynamics for W_h^Q, W_h^K and W_h^V. As P is row-stochastic, and $\forall\, i : X_{i,:}$ of X after LayerNorm [43], $\|X_{i,:}\|_2 = 1$, we can upper bound $\|PX\|_2$ by $\sqrt{N \min(N, D)}$ and notice that $\partial \texttt{Attn}_h(X)/\partial W_h^V$ does not necessarily lead to gradient vanishing. However, we could observe, that uniform attention distribution pushes $\|P\|_2$ towards 1, unlike categorical attention pattens, where the bound approaches its maximal value $\sqrt{N}$.

Contrary, the gradients w.r.t. the W_h^Q and W_h^K may fade, shrinking towards zero in the proximity of *either uniform or categorical distributions for all tokens in the sequence*, an effect overlooked by [34]. It should be also noted — although both lead to the gradient vanishing, the uniform case attains its minima at $1/N$, scaling with sequence length, while the categorical one shrinks exactly to 0.

Similar effects that unify observed regimes were identified in prior literature, but haven't been attributed to the spectral properties of attention maps. When attention distributions inside one attention map all become uniform-like, an update of the attention block becomes close to a rank-1 matrix, leading to the rank collapse phenomenon [8] and training instability [4, 34]. On the other hand, attention spikes have been attributed to "sink-like" effects and entropy collapse, often phrased as an ability to perform "no-op" operation in deeper layers [1]. Similarly to the uniform case, it has been shown to induce training oscillations or divergence [47], with recent "sinkless" architectures displaying superior results on most benchmarks [35]. Importantly, our bound allows for a precise and easy-to-compute quantification of the proximity to these regimes, offering a practical diagnostic tool to use our bound as a proxy for vanishing gradients during the forward pass.

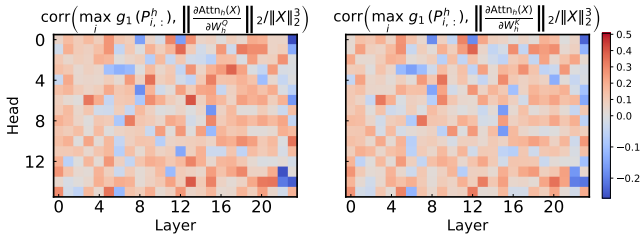

Fig. 5: Correlation between norms for W_Q, W_K and $\max_i g_1(P^h_{i,:})$ averaged across 1000 ImageNet samples for ViT-L model. To discard input norm dependency, we divide gradient norm by $\|X\|_2^3$, following the bound proposed in Proposition 1.

7 `JaSMin` regularizer

From Theorem 3 we observe that the spectral norm of the self-attention depends on three types of components: (1) input-only part $\|X\|_2^2$, (2) parts depending on weight matrices (e.g $\|A_h\|_2$ and $\|W_h^V\|_2$) and (3) mixed terms that are functions of the matrix P^h, where P^h in turn depends on both weights and inputs ($\|P^h\|_2$ and $\|\max_i \mathcal{M}(P^h_{i,:})\|_2$). The regularizer that we present in this section as a method to improve robustness and validate our findings targets the latter terms. In particular, we propose 2 strategies for the choice of a penalty:

$$\mathcal{L} = \mathcal{L}_{\texttt{model}} + \lambda \mathcal{L}_{\texttt{JaSMin}_k}, \tag{29}$$

where $\mathcal{L}_{\texttt{model}}$ is a model loss function, λ is the regularization parameter, and $\texttt{JaSMin}_k$ is the proposed modification. The first way is to penalize $\log(\|\mathcal{M}(P^h_{i,:})\|_2)$ as it naturally constrains the upper bound from Theorem 3:

$$\mathcal{L}_{\texttt{JaSMin}_{k=0}} = \sum_{l,h=1}^{L,H} \max_i \log\left(g_1(P^{l,h}_{i,:})\right). \tag{30}$$

The second strategy is based on the interlacing property of the softmax Jacobian singular values: we aim to penalize the logarithm of the following ratio:

$$\mathcal{L}_{\texttt{JaSMin}_k} = \sum_{l,h=1}^{L,H} \max_i \log\left(\frac{g_1(P^{l,h}_{i,:})}{g_k(P^{l,h}_{i,:})}\right), \quad k > 1. \tag{31}$$

As we show below in Proposition 3, this regularizer relaxes the categorical distribution constraint and yields more flexibility towards a uniform distribution.

Note that the memory overhead of `JaSMin` is only $\mathcal{O}(1)$ due to the associative properties of the summation and maximum operations, while `Specformer` needs to store additional $\mathcal{O}(LD)$ memory for singular vectors at training stage.

Regularization of the upper bound. By minimizing Equation (30) that is based on the upper bound on the first singular value, we implicitly constrain

Table 2: Comparison of `Specformer` and `JaSMin` approaches for ViT-B on CIFAR-100. Number after the name of attack denotes attack budget. We evaluate PGD and AutoAttack with 4 and 10 steps respectively. Each run is averaged across 3 seeds.

Method	CIFAR-100					
	Standard	FGSM2	FGSM4	PGD2	PGD4	AA2
`Baseline`	$91.05_{\pm 0.02}$	$46.64_{\pm 0.71}$	$39.86_{\pm 0.60}$	$27.47_{\pm 0.40}$	$13.11_{\pm 0.33}$	$2.10_{\pm 0.01}$
$\texttt{Specformer}_{(10^{-2},10^{-2},10^{-2})}$	$90.06_{\pm 0.01}$	$44.90_{\pm 1.86}$	$37.65_{\pm 1.20}$	$28.08_{\pm 0.96}$	$12.61_{\pm 0.30}$	$\underline{3.05}_{\pm 0.02}$
$\texttt{Specformer}_{(0,0,10^{-2})}$	$\mathbf{91.14}_{\pm 0.05}$	$46.58_{\pm 0.04}$	$39.97_{\pm 0.07}$	$27.59_{\pm 0.01}$	$12.56_{\pm 0.12}$	$2.10_{\pm 0.01}$
$\texttt{JaSMin}_{k=0,\lambda=1e-2}$	$88.83_{\pm 0.01}$	$\underline{48.28}_{\pm 0.31}$	$\underline{41.38}_{\pm 0.29}$	$\mathbf{32.58}_{\pm 0.35}$	$\mathbf{18.90}_{\pm 0.28}$	$\mathbf{3.62}_{\pm 0.01}$
$\texttt{JaSMin}_{k=10,\lambda=1e-3}$	$\underline{91.08}_{\pm 0.02}$	$\mathbf{48.90}_{\pm 0.05}$	$\mathbf{42.34}_{\pm 0.03}$	$\underline{30.29}_{\pm 0.09}$	$\underline{15.47}_{\pm 0.09}$	$2.45_{\pm 0.01}$

$\|\mathcal{M}(P^{l,h}_{i,:})\|_2$ for every layer l and head h. Let us precisely derive what we can say about $P^{l,h}_{i,:}$ if the g_1 is bounded by a constant $\gamma < 1/4$. By solving the quadratic inequality, we arrive at:

$$(P^{l,h}_{i,:})_{(1)} \leqslant \frac{1-\sqrt{1-4\gamma}}{2} \ \vee \ (P^{l,h}_{i,:})_{(1)} \geqslant \frac{1+\sqrt{1-4\gamma}}{2}. \tag{32}$$

Hence, with this regularization term we enforce attention distributions to become either "more" uniform or "more" categorical, excluding all the intermediate options, according to Figure 2.

Regularization of the ratio. The interlacing property of Theorem 2 allows us to bound the ratio of singular values:

$$\sigma_1(P^{l,h}_{i,:})/\sigma_{k-1}(P^{l,h}_{i,:}) \leqslant g_1\left(P^{l,h}_{i,:}\right)/g_k\left(P^{l,h}_{i,:}\right). \tag{33}$$

An interesting observation that we make is that by bounding the upper bound from Equation (33), we can also constrain the the norm to be small. The following proposition yields an upper bound on the spectral norm in this scenario.

Proposition 3 *Let* $1 \leqslant \gamma \leqslant k/4$ *and* $g_1\left(P^{l,h}_{i,:}\right)/g_k\left(P^{l,h}_{i,:}\right) \leqslant \gamma$. *Then, we have:* $\left\|\mathcal{M}(P^{l,h}_{i,:})\right\|_2 \leqslant \dfrac{1-\sqrt{1-4\gamma/k}}{2} = \mathcal{O}\left(\dfrac{\gamma}{k}\right)$.

Proof. See Appendix D for the proof.

To understand what happens with the distribution, let us consider the corner case of $\gamma = 1$. In this case, due to the interlacing property from Theorem 2:

$$x_{(2)}/x_{(k)} \leqslant g_1/g_k = 1. \tag{34}$$

One can show (see Appendix D) that in this case $x_{(1)} = \cdots = x_{(k)}$. So any uniform distribution for top-p elements, $k \leqslant p \leqslant N$ will fulfill this condition. This yields more flexibility in the uniform distribution as opposed to using $\texttt{JaSMin}_{k=0}$.

It should be noted that our regularizer with $k > 1$ is not immediately applicable to models, utilizing attention masking. For example, for causal attention mask, the regularizer will be unbounded on first $k-1$ tokens for any $k > 1$.

Table 3: Comparison of `Specformer` and `JaSMin` approaches for ViT-B on Imagenette. Number after the name of attack denotes attack budget. We evaluate PGD and AutoAttack with 4 and 10 steps respectively.

Method	Imagenette					
	Standard	FGSM2	FGSM4	PGD2	PGD4	AA2
`Baseline`	<u>98.00</u>	**70.12**	<u>55.51</u>	45.02	15.13	6.86
$\texttt{Specformer}_{(10^{-2},0,0)}$	**98.12**	<u>70.03</u>	**55.95**	45.76	15.98	7.31
$\texttt{JaSMin}_{k=30,\lambda=10^{-2}}$	97.07	68.67	54.94	**48.18**	**20.67**	<u>10.62</u>
$\texttt{JaSMin}_{k=70,\lambda=10^{-2}}$	96.37	66.34	51.66	<u>46.01</u>	<u>18.55</u>	**11.95**

Method efficiency The `JaSMin` method introduces an additional computational cost of $\mathcal{O}\left(BLHN^2\right)$ FLOPs compared to `Specformer`'s $\mathcal{O}\left(LD^2\right)$ where B is the batch size. Although `JaSMin`'s dependence on N^2 might appear to be a limitation, it does not pose a substantial computational cost in ViT experiments, as the sequence length is negligible for most CV models, which results in faster computations than that of `Specformer`. As N increases, the `JaSMin` overhead becomes progressively less significant, being overshadowed by the transformer's $\mathcal{O}\left(BL\left(ND^2 + N^2D\right)\right)$ computational cost.

8 Experiments

We train ViT-B model [9] to classify images using CIFAR-10, CIFAR-100 [24] and Imagenette [17] datasets, and show that our regularizer is able to control the local Lipschitz constant, improving robust metrics without a significant quality drop (see Table 2 and Table 3 for more details). Our models are tuned on the respective datasets starting from checkpoints pretrained on ImageNet dataset [7]. Patch size is set to 4 in this experiment for CIFAR-10 and CIFAR-100, while for Imagenette we utilize patch size equal to 16.

For comparison we use the original ViT training setup with `Specformer` and `JaSMin` method. `Specformer` is parameterized with three regularization coefficients for the largest singular values of the self-attention query, key and value matrices. We reproduced the setup from `Specformer` paper [18] and compare our models to theirs by substituting the regularization parameters only, the distribution prefix length k and a regularization coefficient λ. To show that our method not only reduces the local Lipschitz constant, but also improves the robust metrics, we evaluate our models on several adversarial attacks.

We report FGSM [13] and PGD [27] with 4 steps and AutoAttack (AA) [5] with 10 steps. FGSM and PGD are configured with the attack budgets of $4/255$ and $2/255$, while AutoAttack, being more severe, is used with a smaller budget of $2/255$ to provide more evident results. All attack budgets are measured in terms of the infinity norm. Other training details such as hyperparameters, data augmentation, etc. can be found in Appendix G. Additionally, we report gradient norm dynamics for different model weights in Appendix H for better

understanding of the potential gradient vanishing problem.

Apart from metrics on adversarial attacks, we report the distribution of the local Lipschitz constant. We remove patch embedding layer from the model and compute Jacobian norm for the rest of the model, as `JaSMin` does not regularize positional embeddings and a convolutional layer. As illustrated in Figure 6, our regularizer considerably reduces the local Lipschitz constant of the model, supporting our findings, while `Specformer` regularizer does not considerably change the mean value of the local Lipschitz constant.

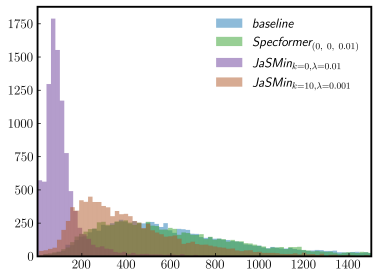

Fig. 6: Comparison of distributions of the spectral norm of the whole model's Jacobian (trained on CIFAR-100) using different regularization methods. The spectral norm is computed via power method at each point of the test dataset.

9 Conclusion

We introduce a new theoretical upper bound for the spectral norm of the self-attention Jacobian. This refinement is based on the improved upper bound for the largest singular value of the softmax Jacobian. Through experiments with vision transformers we observe that the nature of the attention map distribution crucially affects the self-attention local Lipschitz constant. Additionally, we provide a refined interlacing bounds for the softmax Jacobian singular values, provably improving the ones obtainable via an eigenvalue interlacing theorem. To highlight the influence of the attention map distribution on the self-attention local Lipschitz constant, we introduce a lightweight regularizer `JaSMin`. Apart from that, we show that controlling attention map distributions' indeed results in the significant drop of the network's local Lipschitz constant.

Acknowledgements

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4E0002 and the agreement with HSE University № 139-15-2025-009. This research was supported in part through computational resources of HPC facilities at HSE University [23].

Bibliography

[1] Barbero, F., Arroyo, A., Gu, X., Perivolaropoulos, C., Veličković, P., Pascanu, R., Bronstein, M.M.: Why do LLMs attend to the first token? In: Second Conference on Language Modeling (2025), https://openreview.net/forum?id=tu4dFUsW5z

[2] Béthune, L., Boissin, T., Serrurier, M., Mamalet, F., Friedrich, C., González-Sanz, A.: Pay attention to your loss: understanding misconceptions about 1-lipschitz neural networks. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022)

[3] Castin, V., Ablin, P., Peyré, G.: How smooth is attention? In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)

[4] Chen, Z.A., Luo, T.: From condensation to rank collapse: A two-stage analysis of transformer training dynamics. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2026), https://openreview.net/forum?id=gm5mkiTGOy

[5] Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)

[6] Dasoulas, G., Scaman, K., Virmaux, A.: Lipschitz normalization for self-attention layers with application to graph neural networks. In: International Conference on Machine Learning. pp. 2456–2466. PMLR (2021)

[7] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

[8] Dong, Y., Cordonnier, J.B., Loukas, A.: 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)

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy

[10] Feng, W., Wang, H., Wang, J., Zhang, X., Zhao, J., Liang, Y., Chen, X., Han, D.: Edit: Enhancing vision transformers by mitigating attention sink through an encoder-decoder architecture. arXiv preprint arXiv: 2504.06738 (2025)

[11] Gao, B., Pavel, L.: On the properties of the softmax function with application in game theory and reinforcement learning (2018), https://arxiv.org/abs/1704.00805

[12] Golub, G.H., Van Loan, C.F.: Matrix Computations. The Johns Hopkins University Press, Baltimore, MD, 4th edn. (2013)

[13] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2015), https://arxiv.org/abs/1412.6572
[14] Gouk, H., Frank, E., Pfahringer, B., Cree, M.J.: Regularisation of neural networks by enforcing lipschitz continuity. Machine Learning **110**(2), 393–416 (2021). https://doi.org/10.1007/s10994-020-05929-w
[15] Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., Lin, M.: When attention sink emerges in language models: An empirical view. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=78Nn4QJTEN
[16] Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press (1985)
[17] Howard, J.: Imagenette (2019), https://github.com/fastai/imagenette/
[18] Hu, X., Zheng, R., Wang, J., Leung, C.H., Wu, Q., Xie, X.: Specformer: Guarding vision transformer robustness via maximum singular value penalization. In: European Conference on Computer Vision. pp. 345–362. Springer (2024)
[19] Huang, D.A., Farahmand, A.m., Kitani, K., Bagnell, J.: Approximate maxent inverse optimal control and its application for mental simulation of human interactions. Proceedings of the AAAI Conference on Artificial Intelligence **29**(1) (Feb 2015). https://doi.org/10.1609/aaai.v29i1.9605, https://ojs.aaai.org/index.php/AAAI/article/view/9605
[20] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b. arXiv preprint arXiv: 2310.06825 (2023)
[21] Jiang, H., Li, Y., Zhang, C., Wu, Q., Luo, X., Ahn, S., Han, Z., Abdi, A.H., Li, D., Lin, C.Y., et al.: Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. Advances in Neural Information Processing Systems **37**, 52481–52515 (2024)
[22] Kim, H., Papamakarios, G., Mnih, A.: The lipschitz constant of self-attention. In: International Conference on Machine Learning. pp. 5562–5571. PMLR (2021)
[23] Kostenetskiy, P., Chulkevich, R., Kozyrev, V.: HPC resources of the higher school of economics. In: Journal of Physics: Conference Series. vol. 1740, p. 012050. IOP Publishing (2021)
[24] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
[25] Laha, A., Chemmengath, S.A., Agrawal, P., Khapra, M., Sankaranarayanan, K., Ramaswamy, H.G.: On controllable sparse alternatives to softmax. Advances in neural information processing systems **31** (2018)
[26] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV 2021 (October 2021), https://www.microsoft.com/en-us/research/publication/swin-transformer-hierarchical-vision-transformer-using-shifted-windows/

[27] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rJzIBfZAb

[28] Mao, C., Geng, S., Yang, J., Wang, X., Vondrick, C.: Understanding zero-shot adversarial robustness for large-scale models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=P4bXCawRi5J

[29] Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., He, Y., Xue, H.: Towards robust vision transformer. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 12042–12051 (2022)

[30] Menon, R., Franco, N., Günnemann, S.: Lipshift: A certifiably robust shift-based vision transformer. arXiv preprint arXiv:2503.14751 (2025)

[31] Murdock, J.A.: Perturbations: Theory and Methods, Classics in Applied Mathematics, vol. 27. Society for Industrial and Applied Mathematics, Philadelphia, PA (1999). https://doi.org/10.1137/1.9781611971095

[32] Muthukumar, R., Sulam, J.: Sparsity-aware generalization theory for deep neural networks. In: Neu, G., Rosasco, L. (eds.) Proceedings of Thirty Sixth Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 195, pp. 5311–5342. PMLR (12–15 Jul 2023), https://proceedings.mlr.press/v195/muthukumar23a.html

[33] Nair, P.: Softmax is $1/2$-lipschitz: A tight bound across all ℓ_p norms. Transactions on Machine Learning Research (2026), https://openreview.net/forum?id=6dowaHsa6D

[34] Noci, L., Anagnostidis, S., Biggio, L., Orvieto, A., Singh, S.P., Lucchi, A.: Signal propagation in transformers: theoretical perspectives and the role of rank collapse. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022)

[35] Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., Liu, D., Zhou, J., Lin, J.: Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2026), https://openreview.net/forum?id=1b7whO4SfY

[36] Shi, H., GAO, J., Xu, H., Liang, X., Li, Z., Kong, L., Lee, S.M.S., Kwok, J.: Revisiting over-smoothing in BERT from the perspective of graph. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=dUV91uaXm3

[37] Tyrtyshnikov, E.E.: A Brief Introduction to Numerical Analysis. Springer Science & Business Media, Boston, MA (1997)

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/

paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[39] Wang, G., Zhao, Y., Tang, C., Luo, C., Zeng, W.: When shift operation meets vision transformer: An extremely simple alternative to attention mechanism. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2423–2430 (2022)

[40] Wang, S., Seidman, J.H., Sankaran, S., Wang, H., Pappas, G.J., Perdikaris, P.: CViT: Continuous vision transformer for operator learning. arXiv preprint arXiv:2405.13998 (2024)

[41] Wu, X., Ajorlou, A., Wang, Y., Jegelka, S., Jadbabaie, A.: On the role of attention masks and layernorm in transformers. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=lIH6oCdppg

[42] Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=NG7sS51zVF

[43] Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 10524–10533. PMLR (2020), http://proceedings.mlr.press/v119/xiong20b.html

[44] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388

[45] Yang, Y.Y., Rashtchian, C., Zhang, H., Salakhutdinov, R., Chaudhuri, K.: A closer look at accuracy vs. robustness. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)

[46] Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., Wei, F.: Differential transformer. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=OvoCm1gGhN

[47] Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., Susskind, J.: Stabilizing transformer training by preventing attention entropy collapse. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)

[48] Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., Alvarez, J.M.: Understanding the robustness in vision transformers. In: International conference on machine learning. pp. 27378–27394. PMLR (2022)

[49] Zhu, C., Zhao, W., Zhang, H., Zhou, Y., Tang, W., Wang, S., Yuan, Z., Shang, Y., Peng, X., Wang, K., et al.: EA-ViT: Efficient adaptation for elastic vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1038–1047 (2025)