How Smooth Is Attention?

Valérie Castin [1] Pierre Ablin [2] Gabriel Peyré [1,3]

[1]ENS PSL, Paris [2]Apple [3]CNRS

WHAT'S IN THIS POSTER?

We **investigate smoothness of the self-attention map**, by providing sharp bounds on its Lipschitz constant as a function of the **sequence length** n and the **magnitude of tokens** R.

- The local Lipschitz constant with real data grows like $Cn^{1/4}$ → **More tokens mean less robustness!**
- The **worst-case rate** is $Cn^{1/2}$ for n small, and $CR^2e^{CR^2}$ for n very large ($n \sim e^{cR^2}$).
- **Masked** self-attention can be **generalized to probability measures** by adding a position coordinate.

The Transformer Architecture

Transformers represent each data point by a **sequence of tokens** $X = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$

This is how GPT-3 tokenizes this sentence.

Figure 1. Tokenization of text (GPT2 tokenizer)

Figure 2. Tokenization of images

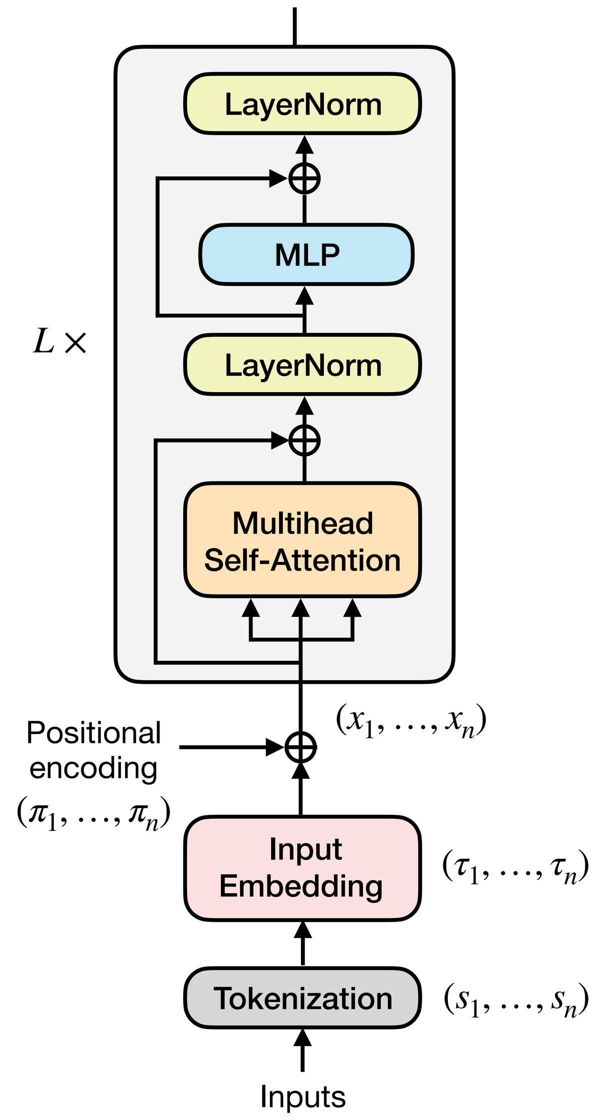


Figure 3. Architecture of a Transformer's Encoder [3]

Main building blocks:

- **Self-attention** with $Q, K, V \in \mathbb{R}^{d\times d}$:

$$f: \begin{cases} (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n \\ (x_1, \dots, x_n) \mapsto \left(V \sum_{j=1}^n P_{ij} x_j\right)_{1\le i\le n} \end{cases}$$

with

$$P_{ij} := \exp(\langle Qx_i, Kx_j\rangle/\sqrt{d}) / \sum_{k=1}^n \exp(\langle Qx_i, Kx_k\rangle/\sqrt{d}).$$

Denote $A := K^\top Q/\sqrt{d}$.

- **Multi-head self-attention:**

$$f^{MH} := \sum_{h=1}^H W_h f_{A_h, V_h}$$

- **Masked self-attention:**

$$f^m(X)_i := f(x_1, \dots, x_i)_i$$

- **Layer normalization:** "projects" each x_i on an ellipsis

$$\mathrm{LayerNorm}\colon x \in \mathbb{R}^d \mapsto \alpha \odot \frac{x - \mathrm{mean}(x)}{\mathrm{std}(x)} + \beta \in \mathbb{R}^d$$

$$\mathrm{RMSNorm}\colon x \in \mathbb{R}^d \mapsto \alpha \odot \frac{x}{|x|}\sqrt{d} \in \mathbb{R}^d$$

Definition – Lipschitz constant

$$\mathrm{Lip}(f_{|B_R^n}) := \sup_{X\neq Y \in B_R^n} \frac{\|f(X) - f(Y)\|}{\|X - Y\|} = \sup_{X\in B_R^n} \|D_X f\|_2 \qquad B_R := \{x \in \mathbb{R}^d : |x| \le R\}$$

State of the art

Kim et al. [2]

$$\mathrm{Lip}(f_{|B_R^n}) \ge c(A, V) R^2$$

Geshkovski et al. [1]

$$\mathrm{Lip}(f_{|B_R^n}) \le \|V\|_2 (1 + 3\|A\|_2 R^2) e^{2\|A\|_2 R^2}$$

Big discrepancy! Which bound is tighter? Dependency in n?

Denote $\gamma_1 \ge \dots \ge \gamma_\delta$ the real eigenvalues of A, and $\gamma := \max(-\gamma_\delta, \gamma_1/8)$.

Contribution 1 – Discrete bound

$$\mathrm{Lip}(f_{|B_R^n}) \le \sqrt{3}\|V\|_2 \left(\|A\|_2^2 R^4 (4n+1) + n\right)^{1/2} \approx R^2\sqrt{n}$$

and if $V = I_d$,

$$\mathrm{Lip}(f_{|B_R^n}) \ge \frac{1}{1 + (n-1)e^{-2R^2\gamma}}\sqrt{n-1}$$

where $R^2\gamma \approx 10^{2-3}$ in practical Transformers.

Numerical experiments

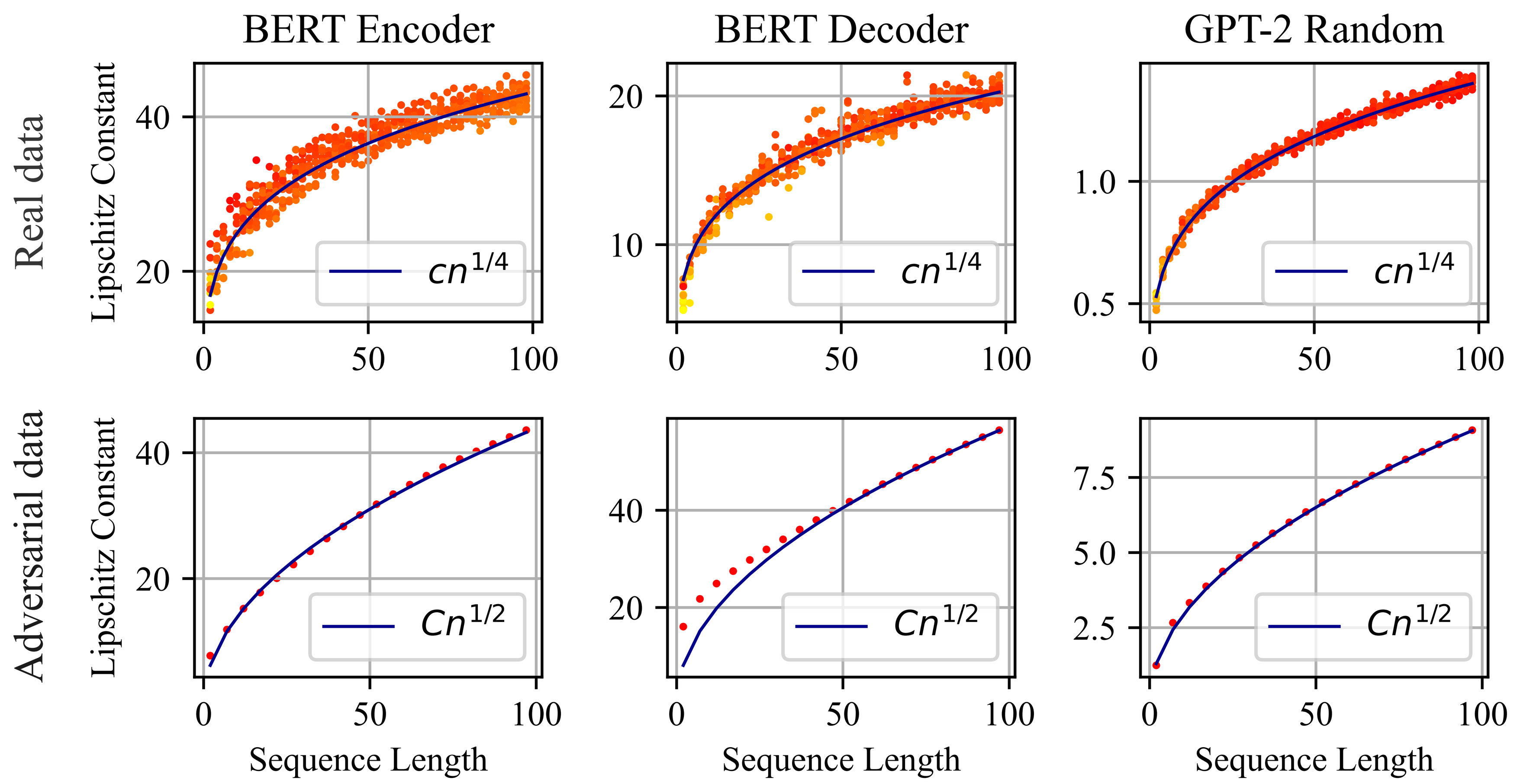


Figure 4. Local Lipschitz constant of self-attention and masked self-attention as a function of the sequence length.

Multi-head attention

From single-head to multi-head:

$$\mathrm{Lip}(f^{MH}_{|B_R^n}) \le \sum_{h=1}^H \|W_h\|_2 \mathrm{Lip}(f_{h|B_R^n}).$$

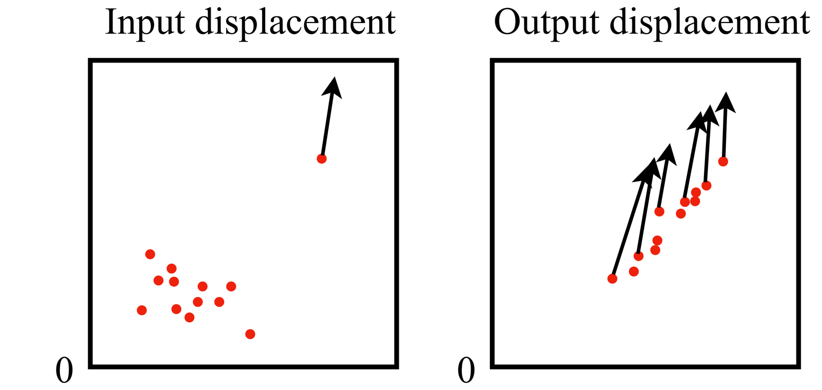
Adversarial configurations also work for multi-head!

What are the adversarial configurations?

One token x_j far away from the others and such that for all i:

$$\langle Ax_i, x_j\rangle \approx \max_k \langle Ax_i, x_k\rangle$$

→ local Lipschitz constant **proportional to** $\sqrt{n}$



Mean-field framework

In-context mapping: $f(X) = (\Gamma_X(x_1), \dots, \Gamma_X(x_n))$ with

$$\Gamma_X\colon x \in \mathbb{R}^d \mapsto \frac{\sum_{j=1}^n e^{\langle Ax, x_j\rangle} V x_j}{\sum_{j=1}^n e^{\langle Ax, x_j\rangle}}.$$

Generalization to probability measures: $F\colon \mu \mapsto (\Gamma_\mu)_\sharp \mu$ with

$$\Gamma_\mu\colon x \in \mathbb{R}^d \mapsto \frac{\int V y e^{\langle Ax, y\rangle} \mathrm{d}\mu(y)}{\int e^{\langle Ax, y\rangle} \mathrm{d}\mu(y)}.$$

Wasserstein distance: $W_2(\mu, \nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int |x - y|^2 \mathrm{d}\pi(x, y)\right)^{1/2}$.

Mean-field Lipschitz constant: $\mathrm{Lip}(F_{|\mathcal{P}(B_R)}) := \sup_{\mu\neq\nu\in\mathcal{P}(B_R)} \frac{W_2(F(\mu), F(\nu))}{W_2(\mu,\nu)}$.

Contribution 2 – Mean-field masked self-attention

For $\bar\mu \in \mathcal{P}_c([0,1]\times\mathbb{R}^d)$, denote $\mu(\mathcal{A}) := \int_{s=0}^1 \int_{x\in\mathcal{A}} \mathrm{d}\bar\mu(s, x)$. We define

$$F^m\colon \bar\mu \mapsto (\Gamma_{\bar\mu})_\sharp \bar\mu \quad \text{where}$$

$$\Gamma_{\bar\mu}(s, x) := \left(s, \frac{\int_{[0,1]\times\mathbb{R}^d} V y e^{\langle Ax, y\rangle} \mathbf{1}_{\tau\le s} \mathrm{d}\bar\mu(\tau, y)}{\int_{[0,1]\times\mathbb{R}^d} e^{\langle Ax, y\rangle} \mathbf{1}_{\tau\le s} \mathrm{d}\bar\mu(\tau, y)}\right).$$

Same upper bound as unmasked mean-field self-attention!

Contribution 3 – Mean-field lower bound

It holds [1]:

$$\mathrm{Lip}(F_{|\mathcal{P}(B_R)}) \le \|V\|_2 (1 + 3\|A\|_2 R^2) e^{2\|A\|_2 R^2}.$$

We show that if $V = I_d$ and $n \sim_{R\to+\infty} e^{2\gamma R^2}$, then

$$\mathrm{Lip}(F_{|\mathcal{P}(B_R)}) \ge \mathrm{Lip}(f_{|B_R^n}) \gtrsim \frac{\gamma}{2} R^2 e^{\gamma R^2}.$$

References

SCAN ME!

[1] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

[2] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pages 5562–5571. PMLR, 2021.

[3] 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.

ICML 2024

valerie.castin@ens.psl.eu