How Smooth Is Attention?

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

[1]Ecole Normale Supérieure PSL, Paris
[2]Apple

ENS de Lyon, 21 January 2025

Transformers process tuples

Transformers are state-of-the-art in several fields: NLP, computer vision, multimodal learning, generative modeling...

Transformers process tuples

Transformers are state-of-the-art in several fields: NLP, computer vision, multimodal learning, generative modeling...

- Important feature: data represented as tuples $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ of tokens

Transformers process tuples

Transformers are state-of-the-art in several fields: NLP, computer vision, multimodal learning, generative modeling...

- Important feature: data represented as tuples $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ of tokens

This is how GPT-3 tokenizes this sentence.

Tokenization of text

Tokenization of images

Transformers process tuples

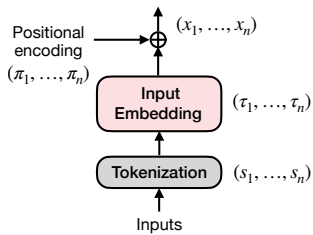
Transformers are state-of-the-art in several fields: NLP, computer vision, multimodal learning, generative modeling...

- Important feature: data represented as tuples $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ of tokens

This is how GPT-3 tokenizes this sentence.

Tokenization of text

Tokenization of images



- Positional encoding encodes the order of tokens

Transformers process tuples

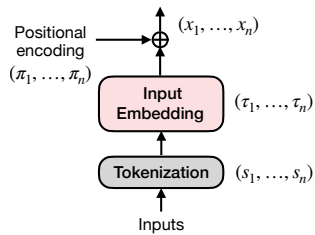
Transformers are state-of-the-art in several fields: NLP, computer vision, multimodal learning, generative modeling...

- Important feature: data represented as tuples $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ of tokens

This is how GPT-3 tokenizes this sentence.

Tokenization of text

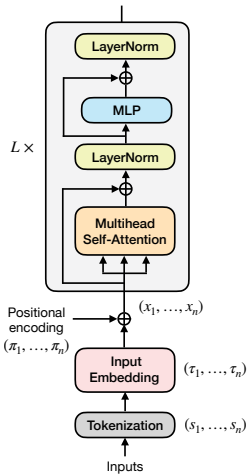
Tokenization of images



- Positional encoding encodes the order of tokens

Allows to learn local dependencies!

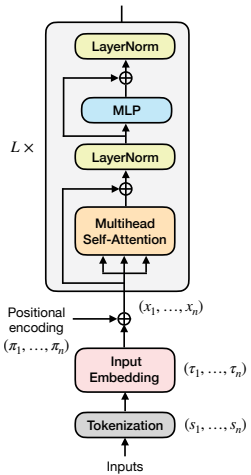
Attention is the core component of Transformers



Architecture of Transformers:

- Attention layers: $\cup_n(\mathbb{R}^d)^n \to \cup_n(\mathbb{R}^d)^n$ learn dependencies between tokens
- Multilayer perceptrons: $\mathbb{R}^d \to \mathbb{R}^d$ (applied token-wise)
- Layer normalization: project each token on an ellipsis

Attention is the core component of Transformers

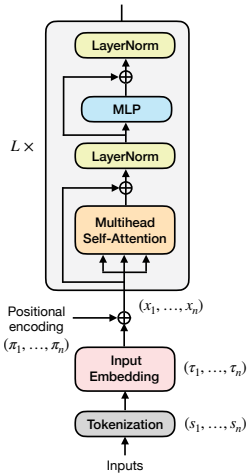


Architecture of Transformers:

- Attention layers: $\cup_n(\mathbb{R}^d)^n \to \cup_n(\mathbb{R}^d)^n$ learn dependencies between tokens
- Multilayer perceptrons: $\mathbb{R}^d \to \mathbb{R}^d$ (applied token-wise)
- Layer normalization: project each token on an ellipsis

How much can the output change when slightly perturbing the input?

Attention is the core component of Transformers



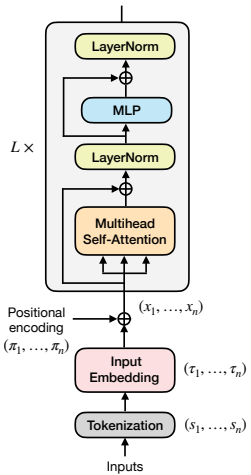
Architecture of Transformers:

- Attention layers: $\cup_n(\mathbb{R}^d)^n \to \cup_n(\mathbb{R}^d)^n$ learn dependencies between tokens
- Multilayer perceptrons: $\mathbb{R}^d \to \mathbb{R}^d$ (applied token-wise)
- Layer normalization: project each token on an ellipsis

How much can the output change when slightly perturbing the input?

- controls robustness and expressive power

Attention is the core component of Transformers



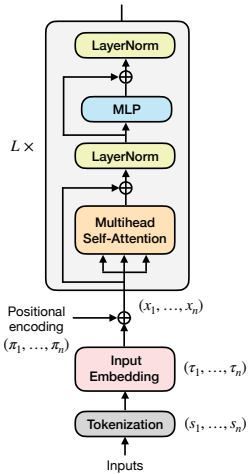
Architecture of Transformers:

- Attention layers: $\cup_n(\mathbb{R}^d)^n \to \cup_n(\mathbb{R}^d)^n$ learn dependencies between tokens
- Multilayer perceptrons: $\mathbb{R}^d \to \mathbb{R}^d$ (applied token-wise)
- Layer normalization: project each token on an ellipsis

How much can the output change when slightly perturbing the input?

- controls robustness and expressive power
- parameters are fixed

Attention is the core component of Transformers



Architecture of Transformers:

- Attention layers: $\cup_n(\mathbb{R}^d)^n \to \cup_n(\mathbb{R}^d)^n$ learn dependencies between tokens
- Multilayer perceptrons: $\mathbb{R}^d \to \mathbb{R}^d$ (applied token-wise)
- Layer normalization: project each token on an ellipsis

How much can the output change when slightly perturbing the input?

- controls robustness and expressive power
- parameters are fixed
- we analyze only one attention layer

The self-attention map

- **Parameters**: $Q, K \in \mathbb{R}^{k \times d}$ and $V \in \mathbb{R}^{d \times d}$

The self-attention map

- **Parameters**: $Q, K \in \mathbb{R}^{k \times d}$ and $V \in \mathbb{R}^{d \times d}$
- Attention of $X = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ w.r.t. $z \in \mathbb{R}^d$:

$$\Gamma_X(z) := \sum_{j=1}^{n} p_j V x_j \quad \text{with} \quad p_j := \exp(\langle Qz, Kx_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle Qz, Kx_\ell \rangle)$$

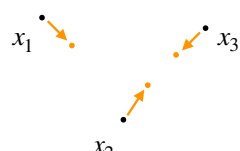
The self-attention map

- **Parameters**: $Q, K \in \mathbb{R}^{k \times d}$ and $V \in \mathbb{R}^{d \times d}$
- Attention of $X = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ w.r.t. $z \in \mathbb{R}^d$:

$$\Gamma_X(z) \coloneqq \sum_{j=1}^{n} p_j V x_j \quad \text{with} \quad p_j \coloneqq \exp(\langle Qz, Kx_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle Qz, Kx_\ell \rangle)$$

- **Self-attention** of $X = (x_1, \ldots, x_n)$:

$$f \colon X \in (\mathbb{R}^d)^n \mapsto (\Gamma_X(x_1), \ldots, \Gamma_X(x_n)) \in (\mathbb{R}^d)^n$$



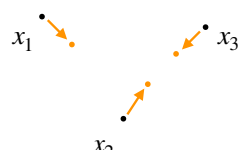
The self-attention map

- **Parameters**: $Q, K \in \mathbb{R}^{k \times d}$ and $V \in \mathbb{R}^{d \times d}$
- Attention of $X = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ w.r.t. $z \in \mathbb{R}^d$:

$$\Gamma_X(z) \coloneqq \sum_{j=1}^{n} p_j V x_j \quad \text{with} \quad p_j \coloneqq \exp(\langle Qz, Kx_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle Qz, Kx_\ell \rangle)$$

- **Self-attention** of $X = (x_1, \ldots, x_n)$:

$$f \colon X \in (\mathbb{R}^d)^n \mapsto (\Gamma_X(x_1), \ldots, \Gamma_X(x_n)) \in (\mathbb{R}^d)^n$$



- Depends only on $A \coloneqq K^\top Q$

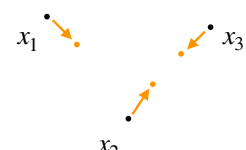
The self-attention map

- **Parameters**: $Q, K \in \mathbb{R}^{k \times d}$ and $V \in \mathbb{R}^{d \times d}$
- Attention of $X = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ w.r.t. $z \in \mathbb{R}^d$:

$$\Gamma_X(z) \coloneqq \sum_{j=1}^{n} p_j V x_j \quad \text{with} \quad p_j \coloneqq \exp(\langle Qz, Kx_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle Qz, Kx_\ell \rangle)$$

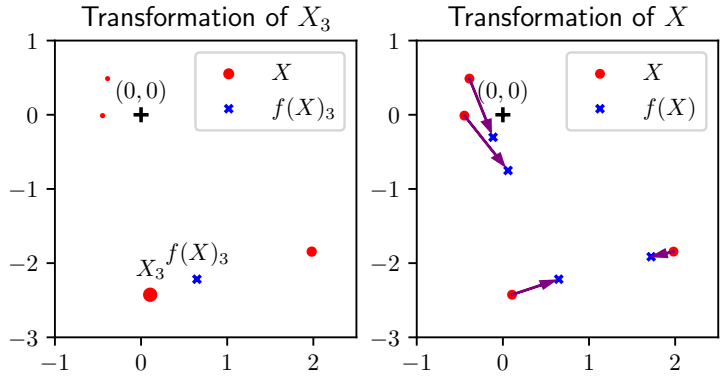
- **Self-attention** of $X = (x_1, \ldots, x_n)$:

$$f \colon X \in (\mathbb{R}^d)^n \mapsto (\Gamma_X(x_1), \ldots, \Gamma_X(x_n)) \in (\mathbb{R}^d)^n$$



- Depends only on $A \coloneqq K^\top Q$
- **Multi-head attention**: linear combination $\sum_{h=1}^{H} W^{(h)} f^{(h)}$

The self-attention map



$$f(X)_i := \sum_{j=1}^{n} P_{ij} V x_j \quad \text{with} \quad P_{ij} := \exp(\langle Q x_i, K x_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle Q x_i, K x_\ell \rangle)$$

Measuring regularity with the local Lipschitz constant

$f\colon (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ self-attention

Local Lipschitz constant

Norm on $(\mathbb{R}^d)^n$: $\|X\|^2 := \sum_{i=1}^n |x_i|^2$

Measuring regularity with the local Lipschitz constant

$f\colon (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ self-attention

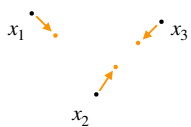
Local Lipschitz constant

Norm on $(\mathbb{R}^d)^n$: $\quad \|X\|^2 := \sum_{i=1}^n |x_i|^2$

Local Lipschitz constant of f at X:

$$\mathrm{Lip}_X(f) := \|D_X f\|_2 = \sup_{\|\varepsilon\|=1} \|D_X f(\varepsilon)\|$$

where $D_X f\colon (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ Jacobian of f



Measuring regularity with the local Lipschitz constant

$f\colon (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ self-attention

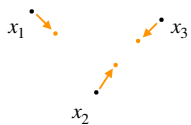
Local Lipschitz constant

Norm on $(\mathbb{R}^d)^n$: $\quad \|X\|^2 := \sum_{i=1}^n |x_i|^2$

Local Lipschitz constant of f at X:

$$\mathrm{Lip}_X(f) := \|D_X f\|_2 = \sup_{\|\varepsilon\|=1} \|D_X f(\varepsilon)\|$$

where $D_X f\colon (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ Jacobian of f



- Gives global guarantees:

$$\sup_{X \neq Y \in B_R^n} \frac{\|f(X) - f(Y)\|}{\|X - Y\|} = \sup_{X \in B_R^n} \mathrm{Lip}_X(f)$$

Measuring regularity with the local Lipschitz constant

$f\colon (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ self-attention

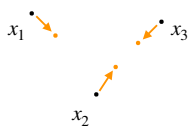
Local Lipschitz constant

Norm on $(\mathbb{R}^d)^n$: $\quad \|X\|^2 := \sum_{i=1}^n |x_i|^2$

Local Lipschitz constant of f at X:

$$\mathrm{Lip}_X(f) := \|D_X f\|_2 = \sup_{\|\varepsilon\|=1} \|D_X f(\varepsilon)\|$$

where $D_X f\colon (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ Jacobian of f



- Gives global guarantees:

$$\sup_{X \neq Y \in B_R^n} \frac{\|f(X) - f(Y)\|}{\|X - Y\|} = \sup_{X \in B_R^n} \mathrm{Lip}_X(f)$$

- Direction of maximal change: $U = (u_1, \ldots, u_n)$ first singular vector of $D_X f$

$$\frac{\|f(X + \eta U) - f(X)\|}{\|U\|} \to_{\eta \to 0} \mathrm{Lip}_X(f)$$

Presentation of the problem

- Self-attention is not globally Lipschitz continuous [Kim et al., 2021]

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

$\rightarrow$ set one particle at zero + spread the others

Presentation of the problem

- Self-attention is not globally Lipschitz continuous [Kim et al., 2021]

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

  $\rightarrow$ set one particle at zero + spread the others

- Bound independent of n [Geshkovski et al., 2024]

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

Presentation of the problem

- Self-attention is not globally Lipschitz continuous [Kim et al., 2021]

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

  $\rightarrow$ set one particle at zero + spread the others

- Bound independent of n [Geshkovski et al., 2024]

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

Questions

- Big discrepancy! Which bound is tighter? Dependency on n?
- Characterize adversarial configurations?
- Local Lipschitz constant of real data?

I - Dependency on n of the Lipschitz constant of self-attention

Theoretical dependency on the sequence length n

Theorem 1 [Castin et al., 2024]

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

Theoretical dependency on the sequence length n

Theorem 1 [Castin et al., 2024]

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

and if $V = I_d$,

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

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

Theoretical dependency on the sequence length n

Theorem 1 [Castin et al., 2024]

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

and if $V = I_d$,

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

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

- R fixed by layer normalization

Theoretical dependency on the sequence length n

Theorem 1 [Castin et al., 2024]

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

and if $V = I_d$,

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

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

- R fixed by layer normalization
- n not too large: $\mathrm{Lip}(f_{|B_R^n})$ grows like $C\sqrt{n}$

Theoretical dependency on the sequence length n

Theorem 1 [Castin et al., 2024]

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

and if $V = I_d$,

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

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

- R fixed by layer normalization
- n not too large: $\mathrm{Lip}(f_{|B_R^n})$ grows like $C\sqrt{n}$
- n exponentially large: analyzed separately (mean-field regime)

Theoretical dependency on the sequence length n

Theorem 1 [Castin et al., 2024]

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

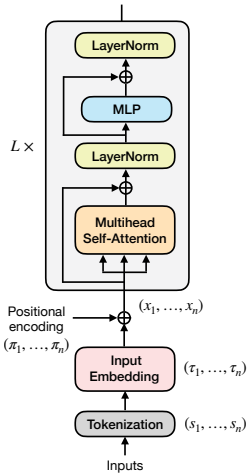
and if $V = I_d$,

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

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

- Configuration for lower bound: $(Ru, -Ru, \ldots, -Ru)$ or $(Ru, Ru/2, \ldots, Ru/2)$ with u eigenvector of A
- Similar bound for multi-head

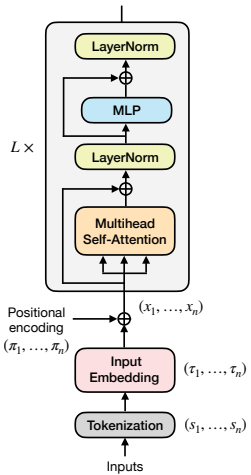
Experiments: typical case and worst case



Plot dependency of $\mathrm{Lip}_X(f)$ in number of tokens n:

- for X obtained from real text (Alice in Wonderland, AG_NEWS)

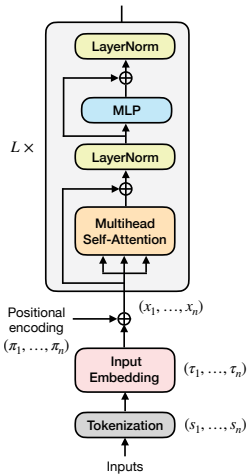
Experiments: typical case and worst case



Plot dependency of $\mathrm{Lip}_X(f)$ in number of tokens n:

- for X obtained from real text (Alice in Wonderland, AG_NEWS)
- for adversarial data of the lower bound

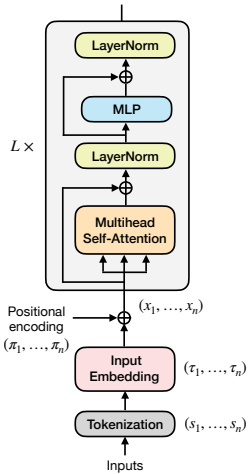
Experiments: typical case and worst case



Plot dependency of $\mathrm{Lip}_X(f)$ in number of tokens n:

- for X obtained from real text (Alice in Wonderland, AG_NEWS)
- for adversarial data of the lower bound
- $\|D_X f\|_2$ computed with power method ($d = 768$)

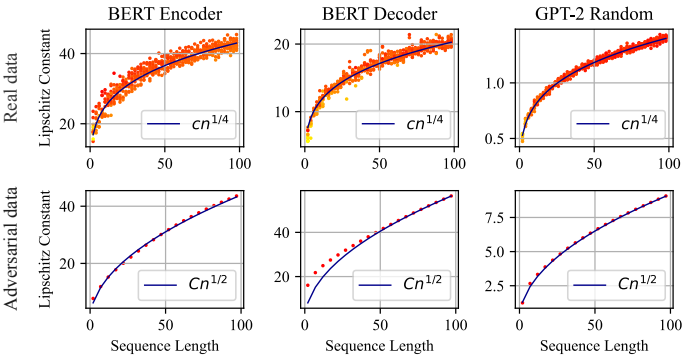
Experiments: typical case and worst case



Plot dependency of $\mathrm{Lip}_X(f)$ in number of tokens n:

- for X obtained from real text (Alice in Wonderland, AG_NEWS)
- for adversarial data of the lower bound
- $\|D_X f\|_2$ computed with power method ($d = 768$)
- Same theory for masked self-attention

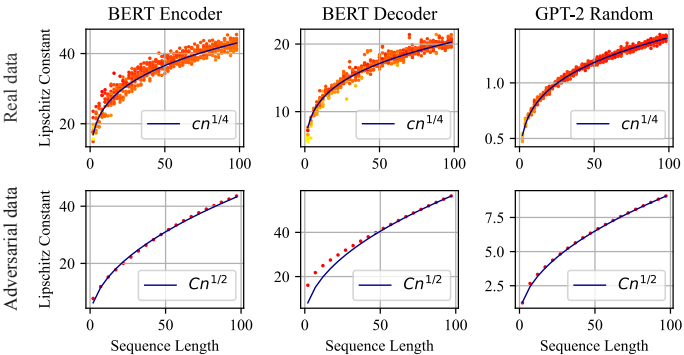
Experiments: typical case and worst case



Local Lipschitz constant of real vs. adversarial data

- Growth in $Cn^{1/4}$ for real data

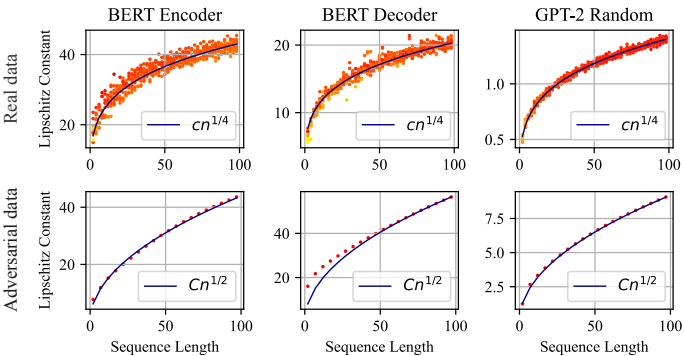
Experiments: typical case and worst case



Local Lipschitz constant of real vs. adversarial data

- Growth in $Cn^{1/4}$ for real data
- Growth in $C\sqrt{n}$ for adv. data $\rightarrow$ matches lower bound

Experiments: typical case and worst case

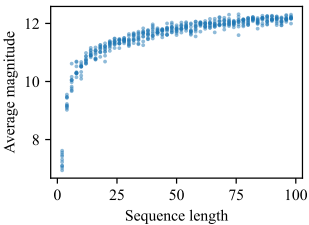


Local Lipschitz constant of real vs. adversarial data

- Growth in $Cn^{1/4}$ for real data
- Growth in $C\sqrt{n}$ for adv. data $\rightarrow$ matches lower bound

Obstacle to Lipschitz attention (GPT-4o context window: 128k)

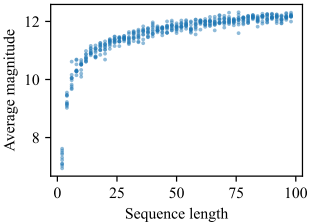
The special case of learnt positional encoding



GPT-2 pretrained: magnitude of tokens R grows with number of tokens n

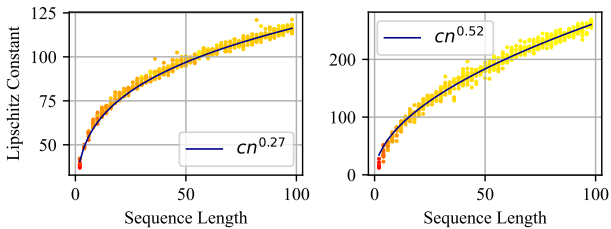
$\mathrm{Lip}(f_{|B_R^n}) \leq R^2\sqrt{n}$

The special case of learnt positional encoding



GPT-2 pretrained: magnitude of tokens R grows with number of tokens n

$\mathrm{Lip}(f_{|B_R^n}) \leq R^2\sqrt{n}$



Gaining intuition with the large radius regime

Large radius regime

$R \to +\infty$ while n fixed

Denote $m_i = \operatorname{argmax}_{1 \leq j \leq n} \langle A x_i, x_j \rangle$ (a.s. unique)

$$(D_{RX} f)(\varepsilon) \to_{R \to +\infty} (V \varepsilon_{m_1}, \dots, V \varepsilon_{m_n}) \qquad \varepsilon \in (\mathbb{R}^d)^n$$

Proposition [Castin et al., 2024]

In the large radius regime:

$$\operatorname{Lip}(f_{|B_R^n}) \lesssim_{R \to +\infty} \|V\|_2 \sqrt{n}$$

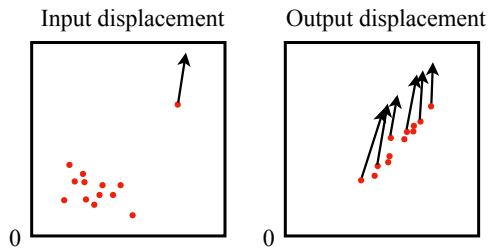
reached if $m_1 = \cdots = m_n$

Adversarial configurations in the large radius regime

One token x_j far away from the others s.t.

$$\forall i = 1, \dots, n, \quad \langle Ax_i, x_j \rangle \approx \max_k \langle Ax_i, x_k \rangle$$

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



II - The mean-field regime (n exponentially large)

Generalizing self-attention to probability measures

Self-attention $f(X) \coloneqq (\Gamma_X(x_1), \dots, \Gamma_X(x_n))$ with

$$\Gamma_X : x \in \mathbb{R}^d \mapsto \sum_{j=1}^{n} p_j V x_j \quad \text{with} \quad p_j \coloneqq \exp(\langle Ax, x_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle Ax, x_\ell \rangle)$$

Generalizing self-attention to probability measures

Self-attention $f(X) \coloneqq (\Gamma_X(x_1), \dots, \Gamma_X(x_n))$ with

$$\Gamma_X : x \in \mathbb{R}^d \mapsto \sum_{j=1}^{n} p_j V x_j \quad \text{with} \quad p_j \coloneqq \exp(\langle Ax, x_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle Ax, x_\ell \rangle)$$

Permutation equivariant: $x_i \leftrightarrow x_j \Rightarrow f(X)_i \leftrightarrow f(X)_j$

Generalizing self-attention to probability measures

Self-attention $f(X) \coloneqq (\Gamma_X(x_1), \dots, \Gamma_X(x_n))$ with

$$\Gamma_X : x \in \mathbb{R}^d \mapsto \sum_{j=1}^{n} p_j V x_j \quad \text{with} \quad p_j \coloneqq \exp(\langle Ax, x_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle Ax, x_\ell \rangle)$$

Permutation equivariant: $x_i \leftrightarrow x_j \Rightarrow f(X)_i \leftrightarrow f(X)_j$

Mean-field self-attention [Sander et al., 2022]

$F : \mu \in \mathcal{P}(\mathbb{R}^d) \mapsto (\Gamma_\mu)_\sharp \mu$ with

$$\Gamma_\mu : x \in \mathbb{R}^d \mapsto \int k(x, y) V y \, \mathrm{d}\mu(y) \quad \text{with} \quad k(x, y) \coloneqq e^{\langle Ax, y \rangle} / \int e^{\langle Ax, z \rangle} \mathrm{d}\mu(z)$$

Generalizing self-attention to probability measures

Self-attention $f(X) \coloneqq (\Gamma_X(x_1), \dots, \Gamma_X(x_n))$ with

$$\Gamma_X : x \in \mathbb{R}^d \mapsto \sum_{j=1}^{n} p_j V x_j \quad \text{with} \quad p_j \coloneqq \exp(\langle Ax, x_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle Ax, x_\ell \rangle)$$

Permutation equivariant: $x_i \leftrightarrow x_j \Rightarrow f(X)_i \leftrightarrow f(X)_j$

Mean-field self-attention [Sander et al., 2022]

$F : \mu \in \mathcal{P}(\mathbb{R}^d) \mapsto (\Gamma_\mu)_\sharp \mu$ with

$$\Gamma_\mu : x \in \mathbb{R}^d \mapsto \int k(x, y) V y \, \mathrm{d}\mu(y) \quad \text{with} \quad k(x, y) \coloneqq e^{\langle Ax, y \rangle} / \int e^{\langle Ax, z \rangle} \mathrm{d}\mu(z)$$

Covers discrete case representing X as $\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$

Generalizing self-attention to probability measures

Mean-field self-attention [Sander et al., 2022]

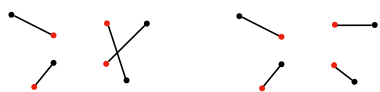
$F\colon \mu \in \mathcal{P}_c(\mathbb{R}^d) \mapsto (\Gamma_\mu)_\sharp \mu$ with

$$\Gamma_\mu\colon x \in \mathbb{R}^d \mapsto \int k(x,y) V y \,\mathrm{d}\mu(y) \quad \text{with} \quad k(x,y) \coloneqq e^{\langle Ax, y\rangle} \Big/ \int e^{\langle Ax, z\rangle} \mathrm{d}\mu(z)$$

Lipschitz constant measured with **Wasserstein distance**

$$W_2(\mu,\nu) \coloneqq \left(\inf_\pi \int |x-y|^2 \,\mathrm{d}\pi(x,y)\right)^{1/2}$$

where $\int \mathrm{d}\pi(x,\cdot) = \mathrm{d}\mu(x)$ and $\int \mathrm{d}\pi(\cdot,y) = \mathrm{d}\nu(y)$



Generalizing self-attention to probability measures

Mean-field self-attention [Sander et al., 2022]

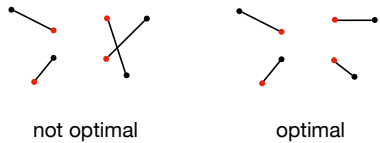
$F\colon \mu \in \mathcal{P}_c(\mathbb{R}^d) \mapsto (\Gamma_\mu)_\sharp \mu$ with

$$\Gamma_\mu\colon x \in \mathbb{R}^d \mapsto \int k(x,y) V y \,\mathrm{d}\mu(y) \quad \text{with} \quad k(x,y) \coloneqq e^{\langle Ax, y\rangle} \Big/ \int e^{\langle Ax, z\rangle} \mathrm{d}\mu(z)$$

Lipschitz constant measured with **Wasserstein distance**

$$W_2(\mu,\nu) \coloneqq \left(\inf_\pi \int |x-y|^2 \,\mathrm{d}\pi(x,y)\right)^{1/2}$$

where $\int \mathrm{d}\pi(x,\cdot) = \mathrm{d}\mu(x)$ and $\int \mathrm{d}\pi(\cdot,y) = \mathrm{d}\nu(y)$



$$\mathrm{Lip}(F_{|\mathcal{P}(B_R)}) \coloneqq \sup_{\mu \neq \nu \in \mathcal{P}(B_R)} \frac{W_2(F(\mu), F(\nu))}{W_2(\mu,\nu)}$$

Generalizing self-attention to probability measures

Mean-field self-attention [Sander et al., 2022]

$F\colon \mu \in \mathcal{P}_c(\mathbb{R}^d) \mapsto (\Gamma_\mu)_\sharp \mu$ with

$$\Gamma_\mu\colon x \in \mathbb{R}^d \mapsto \int k(x,y) V y \,\mathrm{d}\mu(y) \quad \text{with} \quad k(x,y) \coloneqq e^{\langle Ax, y\rangle} \Big/ \int e^{\langle Ax, z\rangle} \mathrm{d}\mu(z)$$

Lipschitz constant measured with **Wasserstein distance**

$$W_2(\mu,\nu) \coloneqq \left(\inf_\pi \int |x-y|^2 \,\mathrm{d}\pi(x,y)\right)^{1/2}$$

where $\int \mathrm{d}\pi(x,\cdot) = \mathrm{d}\mu(x)$ and $\int \mathrm{d}\pi(\cdot,y) = \mathrm{d}\nu(y)$

$$\mathrm{Lip}(F_{|\mathcal{P}(B_R)}) \coloneqq \sup_{\mu\neq\nu\in\mathcal{P}(B_R)} \frac{W_2(F(\mu),F(\nu))}{W_2(\mu,\nu)}$$

$\to$ upp. bound $R^2 e^{R^2}$ [Geshkovski et al., 2024]

Lower bound in the mean-field regime

Upper bound [Geshkovski et al., 2024]

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

Lower bound in the mean-field regime

Upper bound [Geshkovski et al., 2024]

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

Proposition 2 [Castin et al., 2024]

If $V = I_d$ and $n \sim e^{2\gamma R^2}$:

$$\mathrm{Lip}(f_{|B_R^n}) \gtrsim \frac{\gamma}{2} R^2 e^{\gamma R^2}$$

Lower bound in the mean-field regime

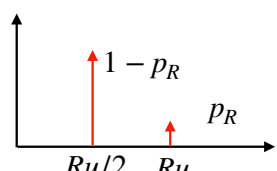
Upper bound [Geshkovski et al., 2024]

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

Proposition 2 [Castin et al., 2024]

If $V = I_d$ and $n \sim e^{2\gamma R^2}$:

$$\mathrm{Lip}(f_{|B_R^n}) \gtrsim \frac{\gamma}{2} R^2 e^{\gamma R^2}$$



Lower bound in the mean-field regime

Upper bound [Geshkovski et al., 2024]

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

Proposition 2 [Castin et al., 2024]

If $V = I_d$ and $n \sim e^{2\gamma R^2}$:

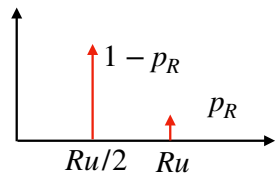
$$\mathrm{Lip}(f_{|B_R^n}) \gtrsim \frac{\gamma}{2} R^2 e^{\gamma R^2}$$

Probability measures for lower bound:

$$p_R \delta_{Ru} + (1 - p_R)\delta_{Ru/2} \quad \text{or} \quad p_R \delta_{Ru} + (1 - p_R)\delta_{-Ru}$$

with

$$p_R = e^{-2\gamma R^2} \to_{R \to +\infty} 0$$



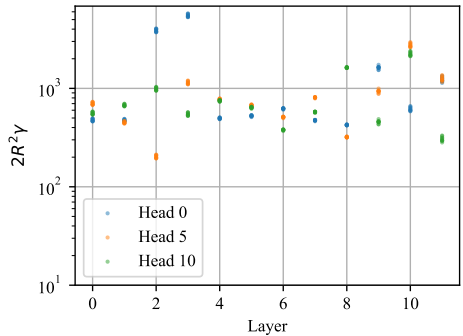
The mean-field regime is not practically relevant

Proposition 2 [Castin et al., 2024]

If $V = I_d$ and $n \sim e^{2\gamma R^2}$:

$$\mathrm{Lip}(f_{|B_R^n}) \gtrsim \frac{\gamma}{2} R^2 e^{\gamma R^2}$$

In practice $2\gamma R^2 \approx 10^3 \rightarrow$ not realistic



III - Masked self-attention

Masked self-attention processes tokens sequentially

Masked self-attention $f^m \colon (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ such that

$$f^m(X)_i \coloneqq f(x_1, \ldots, x_i)_i$$

with params $A, V \in \mathbb{R}^{d \times d}$

Theorem 2 [Castin et al., 2024]

The upper and lower bounds of Theorem 1 on $\mathrm{Lip}(f_{|B_R^n})$ also hold for $\mathrm{Lip}(f^m_{|B_R^n})$

Mean-field regime? $\to$ not permutation equivariant!

Generalizing masked self-attention to measures

Mean-field self-attention $F\colon \mu \in \mathcal{P}_c(\mathbb{R}^d) \mapsto (\Gamma_\mu)_\sharp \mu$ where

$$\Gamma_\mu\colon x \in \mathbb{R}^d \mapsto \int_{\mathbb{R}^d} k(x,y) V y \mathrm{d}\mu(y) \quad \text{with} \quad k(x,y) \coloneqq e^{\langle Ax, y\rangle} / \int e^{\langle Ax, z\rangle} \mathrm{d}\mu(z)$$

Generalizing masked self-attention to measures

Mean-field self-attention $F\colon \mu \in \mathcal{P}_c(\mathbb{R}^d) \mapsto (\Gamma_\mu)_\sharp \mu$ where

$$\Gamma_\mu\colon x \in \mathbb{R}^d \mapsto \int_{\mathbb{R}^d} k(x,y) V y \mathrm{d}\mu(y) \quad \text{with} \quad k(x,y) \coloneqq e^{\langle Ax, y\rangle} / \int e^{\langle Ax, z\rangle} \mathrm{d}\mu(z)$$

Mean-field masked self-attention [Castin et al., 2024]

Replace $\mu \in \mathcal{P}_c(\mathbb{R}^d)$ by $\bar{\mu} \in \mathcal{P}_c([0,1] \times \mathbb{R}^d)$:

$$F^m\colon \bar{\mu} \mapsto (\Gamma_{\bar{\mu}})_\sharp \bar{\mu} \quad \text{where} \quad \Gamma_{\bar{\mu}}(s,x) \coloneqq \left(s, \int_{[0,1]\times\mathbb{R}^d} V y k_s(x,y) \mathrm{d}\bar{\mu}(\tau, y) \right)$$

with

$$k_s(x,y) \coloneqq e^{\langle Ax, y\rangle} 1_{\tau \le s} / \int_{[0,1]\times\mathbb{R}^d} e^{\langle Ax, y\rangle} 1_{\tau \le s} \mathrm{d}\bar{\mu}(\tau, y)$$

Generalizing masked self-attention to measures

Mean-field self-attention $F\colon \mu \in \mathcal{P}_c(\mathbb{R}^d) \mapsto (\Gamma_\mu)_\sharp \mu$ where

$$\Gamma_\mu\colon x \in \mathbb{R}^d \mapsto \int_{\mathbb{R}^d} k(x,y) V y \mathrm{d}\mu(y) \quad \text{with} \quad k(x,y) \coloneqq e^{\langle Ax, y\rangle} / \int e^{\langle Ax, z\rangle} \mathrm{d}\mu(z)$$

Mean-field masked self-attention [Castin et al., 2024]

Replace $\mu \in \mathcal{P}_c(\mathbb{R}^d)$ by $\bar{\mu} \in \mathcal{P}_c([0,1] \times \mathbb{R}^d)$:

$$F^m\colon \bar{\mu} \mapsto (\Gamma_{\bar{\mu}})_\sharp \bar{\mu} \quad \text{where} \quad \Gamma_{\bar{\mu}}(s,x) \coloneqq \left(s, \int_{[0,1]\times\mathbb{R}^d} V y k_s(x,y) \mathrm{d}\bar{\mu}(\tau, y) \right)$$

with

$$k_s(x,y) \coloneqq e^{\langle Ax, y\rangle} 1_{\tau \le s} / \int_{[0,1]\times\mathbb{R}^d} e^{\langle Ax, y\rangle} 1_{\tau \le s} \mathrm{d}\bar{\mu}(\tau, y)$$

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

IV - Application of the mean-field framework: modeling Transformers as PDEs

Modeling an infinitely deep Transformer as a PDE

- Simplified Transformer with only attention layers:

$$f = f^L \circ \cdots \circ f^1 \quad \text{with} \quad f^\ell(X) \coloneqq X + \frac{1}{L}(\Gamma^\ell_X(x_1), \ldots, \Gamma^\ell_X(x_n))$$

Modeling an infinitely deep Transformer as a PDE

- Simplified Transformer with only attention layers:

$$f = f^L \circ \cdots \circ f^1 \quad \text{with} \quad f^\ell(X) \coloneqq X + \frac{1}{L}(\Gamma^\ell_X(x_1), \ldots, \Gamma^\ell_X(x_n))$$

- Discretizes

$$\dot{x}_i(t) = \Gamma_{\mu(t)}(x_i(t)), \qquad 1 \le i \le n$$

with $\mu(t) \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$ and $\Gamma_{\mu(t)} \colon \mathbb{R}^d \to \mathbb{R}^d$ velocity field

Modeling an infinitely deep Transformer as a PDE

- Simplified Transformer with only attention layers:

$$f = f^L \circ \cdots \circ f^1 \quad \text{with} \quad f^\ell(X) \coloneqq X + \frac{1}{L}(\Gamma^\ell_X(x_1), \ldots, \Gamma^\ell_X(x_n))$$

- Discretizes

$$\dot{x}_i(t) = \Gamma_{\mu(t)}(x_i(t)), \qquad 1 \le i \le n$$

with $\mu(t) \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$ and $\Gamma_{\mu(t)} \colon \mathbb{R}^d \to \mathbb{R}^d$ velocity field

Corresponding PDE:

$$\partial_t \mu + \operatorname{div}(\mu \Gamma_\mu) = 0$$

also on continuous measures

Tokens cluster after renormalization

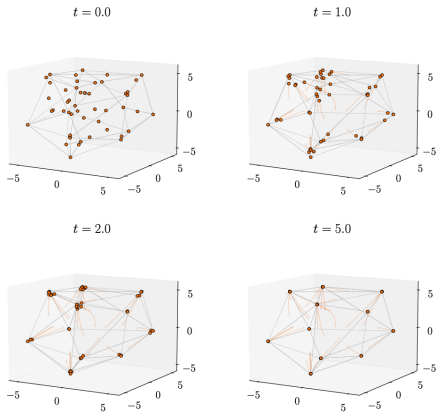
For discrete tokens $x_1, \ldots, x_n$ following

$$\dot{x}_i(t) = \Gamma_{X(t)}(x_i(t))$$

and renormalizing

$$y_i(t) := e^{-tV} x_i(t)$$

$\rightarrow$ **clusters emerge** [Geshkovski et al., 2024]



Tokens cluster after renormalization

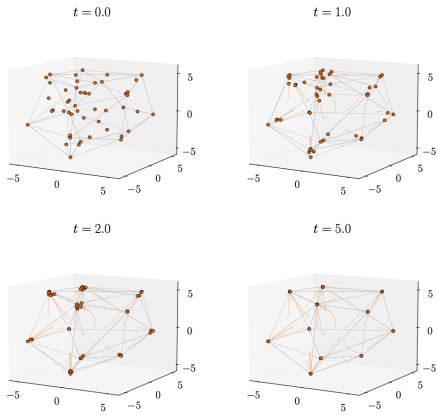
For discrete tokens $x_1, \dots, x_n$ following

$$\dot{x}_i(t) = \Gamma_{X(t)}(x_i(t))$$

and renormalizing

$$y_i(t) := e^{-tV} x_i(t)$$

$\rightarrow$ **clusters emerge** [Geshkovski et al., 2024]



Similar phenomenon for Gaussian inputs! (Castin, Carrillo, Peyré, Ablin, in preparation)

References


Castin, V., Ablin, P., and Peyré, G. (2024).
How smooth is attention?
In *ICML 2024*.


Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. (2024).
The emergence of clusters in self-attention dynamics.
Advances in Neural Information Processing Systems, 36.


Kim, H., Papamakarios, G., and Mnih, A. (2021).
The lipschitz constant of self-attention.
In *International Conference on Machine Learning*, pages 5562–5571. PMLR.


Sander, M. E., Ablin, P., Blondel, M., and Peyré, G. (2022).
Sinkformers: Transformers with doubly stochastic attention.
In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR.