Mean-Field Transformer Dynamics: Well-Posedness, Gaussian Clustering

Valérie Castin
Ecole Normale Supérieure PSL, Paris

DESY, Hamburg, 26 September 2025


José A. Carrillo
University of Oxford


Gabriel Peyré
ENS PSL


Pierre Ablin
Apple Paris

A Unified Perspective on the Dynamics of Deep Transformers, Castin, Ablin, Carrillo, Peyré, *preprint 2025*

How Smooth Is Attention?, Castin, Ablin and Peyré, *in ICML 2024*

Outline

I - Discrete and mean-field model for Transformers

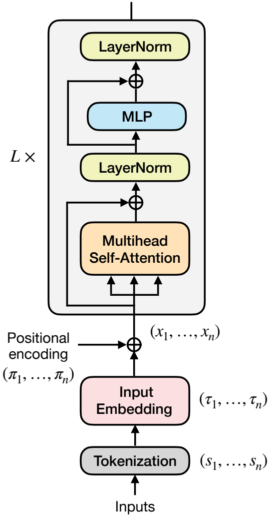
II - Well-posedness of the Transformer PDE for compactly supported initial data

III - Clustering in the Gaussian case

IV - Mean-field can be overpessimistic: refined estimates on the attention map

I - Discrete and mean-field model for Transformers

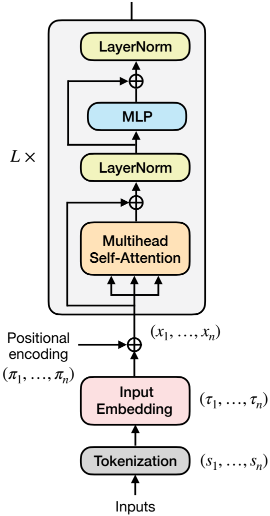
The Transformer architecture



Transformers process tuples:

- Traditional neural network: $x \in \mathbb{R}^{d_{\mathrm{in}}} \mapsto x' \in \mathbb{R}^{d_{\mathrm{out}}}$

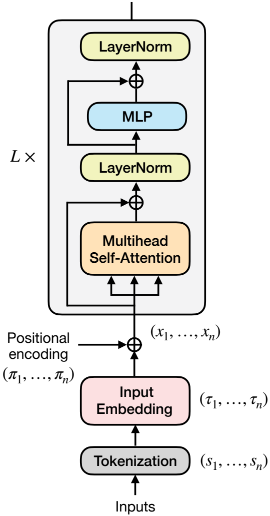
The Transformer architecture



Transformers process tuples:

- Traditional neural network: $x \in \mathbb{R}^{d_{\text{in}}} \mapsto x' \in \mathbb{R}^{d_{\text{out}}}$
- Transformer:
  $(x_1, \ldots, x_n) \in (\mathbb{R}^d_{\text{in}})^n \mapsto (x'_1, \ldots, x'_n) \in (\mathbb{R}^d_{\text{out}})^n$

The Transformer architecture

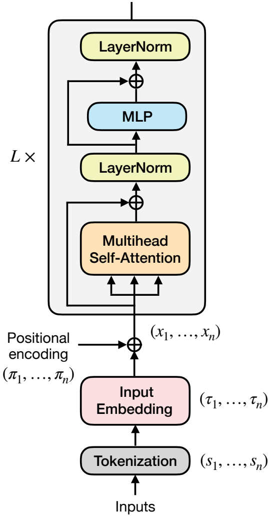


Transformers process tuples:

- Traditional neural network: $x \in \mathbb{R}^{d_{\text{in}}} \mapsto x' \in \mathbb{R}^{d_{\text{out}}}$
- Transformer: $(x_1, \dots, x_n) \in (\mathbb{R}^d_{\text{in}})^n \mapsto (x'_1, \dots, x'_n) \in (\mathbb{R}^d_{\text{out}})^n$

Tokens $X = (x_1, \dots, x_n)$ are obtained through **tokenization**:

The Transformer architecture



Transformers process tuples:

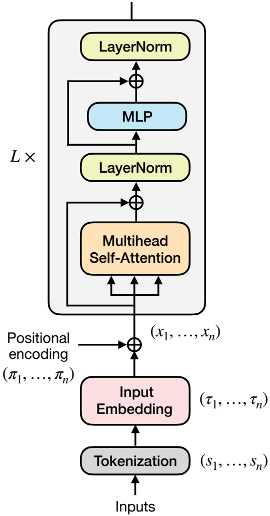
- Traditional neural network: $x \in \mathbb{R}^{d_{\text{in}}} \mapsto x' \in \mathbb{R}^{d_{\text{out}}}$
- Transformer: $(x_1, \dots, x_n) \in (\mathbb{R}^{d}_{\text{in}})^n \mapsto (x'_1, \dots, x'_n) \in (\mathbb{R}^{d}_{\text{out}})^n$

Tokens $X = (x_1, \dots, x_n)$ are obtained through **tokenization**:

- NLP: tokens $=$ words or subwords

This is how GPT-3 tokenizes this sentence.

The Transformer architecture



Transformers process tuples:

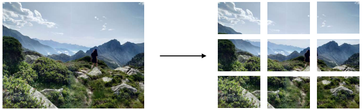
- Traditional neural network: $x \in \mathbb{R}^{d_{\text{in}}} \mapsto x' \in \mathbb{R}^{d_{\text{out}}}$
- Transformer:
  $(x_1, \ldots, x_n) \in (\mathbb{R}^d_{\text{in}})^n \mapsto (x'_1, \ldots, x'_n) \in (\mathbb{R}^d_{\text{out}})^n$

Tokens $X = (x_1, \ldots, x_n)$ are obtained through **tokenization**:

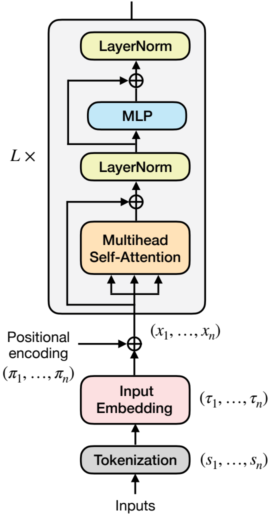
- NLP: tokens $=$ words or subwords

  This is how GPT-3 tokenizes this sentence.

- Vision: tokens $=$ image patches

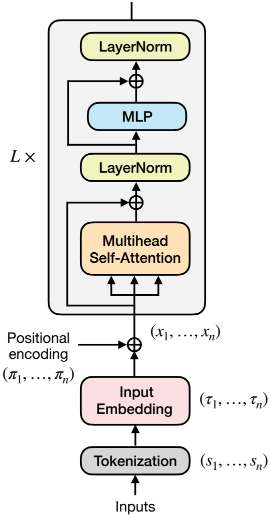


The Transformer architecture



- One input $\to$ n tokens $X := (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$

The Transformer architecture

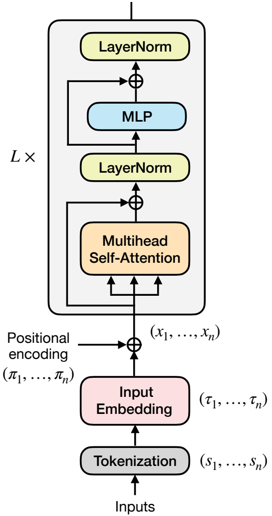


- One input $\to$ n tokens $X := (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$
- Self-attention with params A, V:

$$f(x_1, \dots, x_n) := (\Gamma_X(x_1), \dots, \Gamma_X(x_n)) \text{ where}$$

$$\Gamma_X(x_i) := \sum_{j=1}^{n} p_j(x_i) V x_j \quad \text{with} \quad p_j(x_i) := e^{\langle A x_i, x_j \rangle} \Big/ \sum_{\ell=1}^{n} e^{\langle A x_i, x_\ell \rangle}$$

The Transformer architecture



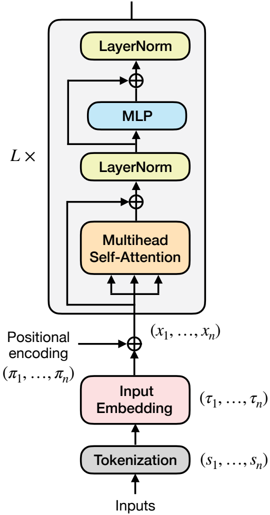
- One input $\to$ n tokens $X := (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$
- Self-attention with params A, V:

$$f(x_1, \ldots, x_n) := (\Gamma_X(x_1), \ldots, \Gamma_X(x_n)) \text{ where}$$

$$\Gamma_X(x_i) := \sum_{j=1}^{n} p_j(x_i) V x_j \quad \text{with} \quad p_j(x_i) := e^{\langle Ax_i, x_j \rangle} / \sum_{\ell=1}^{n} e^{\langle Ax_i, x_\ell \rangle}$$

- Layer normalization $\mathrm{LN}: x \mapsto \beta \odot \frac{x}{|x|}\sqrt{d}$ (applied token-wise)

The Transformer architecture



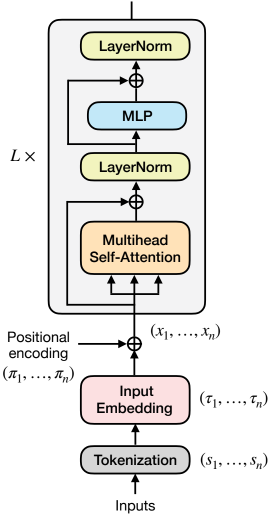
- One input $\to$ n tokens $X := (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$
- Self-attention with params A, V:

$$f(x_1, \dots, x_n) := (\Gamma_X(x_1), \dots, \Gamma_X(x_n)) \text{ where}$$

$$\Gamma_X(x_i) := \sum_{j=1}^{n} p_j(x_i) V x_j \quad \text{with} \quad p_j(x_i) := e^{\langle A x_i, x_j \rangle} / \sum_{\ell=1}^{n} e^{\langle A x_i, x_\ell \rangle}$$

- Layer normalization $\mathrm{LN}\colon x \mapsto \beta \odot \frac{x}{|x|}\sqrt{d}$ (applied token-wise)
- Multilayer perceptron $g\colon \mathbb{R}^d \to \mathbb{R}^d, x \mapsto W\sigma(Ux + b)$ (applied token-wise)

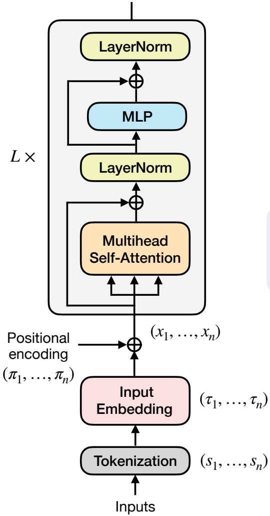
Studying the dynamics of tokens across layers



Without MLP and LN:

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \qquad 1 \le i \le n$$

Studying the dynamics of tokens across layers

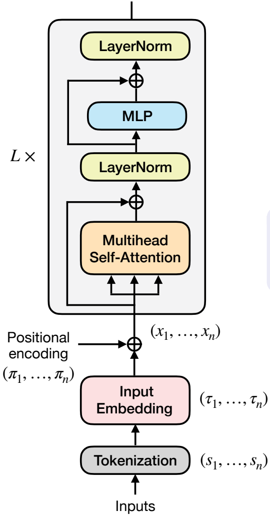


Without MLP and LN:

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \qquad 1 \leq i \leq n$$

What are the dynamics of tokens going through the Transformer? The geometry of learned representations?

Studying the dynamics of tokens across layers



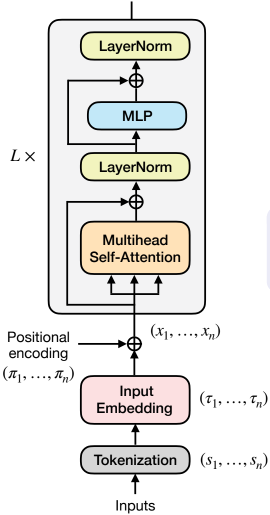
Without MLP and LN:

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \qquad 1 \leq i \leq n$$

What are the dynamics of tokens going through the Transformer? The geometry of learned representations?

- understand clustering effect

Studying the dynamics of tokens across layers



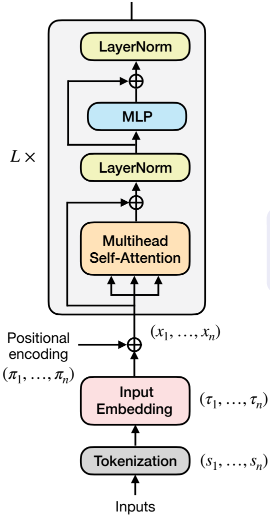
Without MLP and LN:

$$x_i(\ell + 1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \qquad 1 \leq i \leq n$$

What are the dynamics of tokens going through the Transformer? The geometry of learned representations?

- understand clustering effect
- impact of parameters A_ℓ, , V_ℓ

Studying the dynamics of tokens across layers



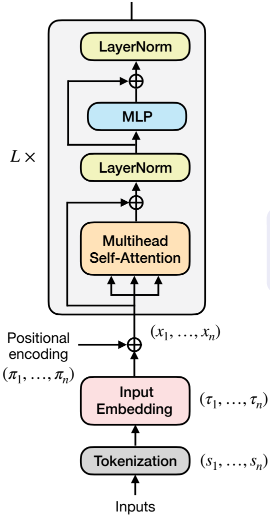
Without MLP and LN:

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \qquad 1 \leq i \leq n$$

What are the dynamics of tokens going through the Transformer? The geometry of learned representations?

- understand clustering effect
- impact of parameters A_ℓ, , V_ℓ
- impact of the initial support and the sequence length

Studying the dynamics of tokens across layers



Without MLP and LN:

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \qquad 1 \leq i \leq n$$

What are the dynamics of tokens going through the Transformer? The geometry of learned representations?

- understand clustering effect
- impact of parameters A_ℓ, , V_ℓ
- impact of the initial support and the sequence length
- compare attention variants

Clustering for the time-continuous interacting particle system

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \quad \longleftrightarrow \quad \dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \quad \text{(similar to Neural ODEs)}$$

Clustering for the time-continuous interacting particle system

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \quad \longleftrightarrow \quad \dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \quad \text{(similar to Neural ODEs)}$$

Clustering result [Geshkovski et al., 2024]

- Assume $A \succ 0$ constant and $V = I_d$
- Rescale $z_i(t) := e^{-tV} x_i(t)$

Clustering for the time-continuous interacting particle system

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \quad \longleftrightarrow \quad \dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \quad \text{(similar to Neural ODEs)}$$

Clustering result [Geshkovski et al., 2024]

- Assume $A \succ 0$ constant and $V = I_d$
- Rescale $z_i(t) := e^{-tV} x_i(t)$

Then $\forall (z_1(0), \ldots, z_n(0))$ there exists a convex polytope $K \subset \mathbb{R}^d$ such that

$$\text{all } z_i(t) \text{ converge in } \{0\} \cup \partial K.$$

Clustering for the time-continuous interacting particle system

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \quad \longleftrightarrow \quad \dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \quad \text{(similar to Neural ODEs)}$$

Clustering result [Geshkovski et al., 2024]

- Assume $A \succ 0$ constant and $V = I_d$
- Rescale $z_i(t) := e^{-tV} x_i(t)$

Then $\forall (z_1(0), \dots, z_n(0))$ there exists a convex polytope $K \subset \mathbb{R}^d$ such that

$$\text{all } z_i(t) \text{ converge in } \{0\} \cup \partial K.$$

- rescaling mimics layer normalization

Clustering for the time-continuous interacting particle system

$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \quad \longleftrightarrow \quad \dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \quad \text{(similar to Neural ODEs)}$$

Clustering result [Geshkovski et al., 2024]

- Assume $A \succ 0$ constant and $V = I_d$
- Rescale $z_i(t) := e^{-tV} x_i(t)$

Then $\forall (z_1(0), \dots, z_n(0))$ there exists a convex polytope $K \subset \mathbb{R}^d$ such that

$$\text{all } z_i(t) \text{ converge in } \{0\} \cup \partial K.$$

- rescaling mimics layer normalization
- typically $\sharp\text{vertices}(K) \ll n$

Clustering for the time-continuous interacting particle system

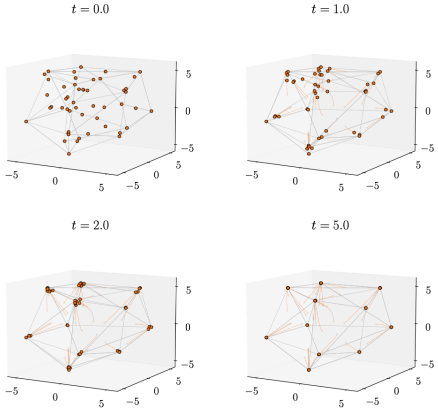
$$x_i(\ell+1) = x_i(\ell) + \Gamma_{X_\ell}(x_i(\ell)) \quad \longleftrightarrow \quad \dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \quad \text{(similar to Neural ODEs)}$$

Clustering result [Geshkovski et al., 2024]

- Assume $A \succ 0$ constant and $V = I_d$
- Rescale $z_i(t) := e^{-tV} x_i(t)$

Then $\forall (z_1(0), \ldots, z_n(0))$ there exists a convex polytope $K \subset \mathbb{R}^d$ such that

$$\text{all } z_i(t) \text{ converge in } \{0\} \cup \partial K.$$



- rescaling mimics layer normalization
- typically $\sharp$vertices$(K) \ll n$

Mean-field attention and the Transformer PDE

Self-attention $f(X) = (\Gamma_X(x_1), \ldots, \Gamma_X(x_n))$ where

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

Mean-field attention and the Transformer PDE

Self-attention $f(X) = (\Gamma_X(x_1), \dots, \Gamma_X(x_n))$ where

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

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

Mean-field attention and the Transformer PDE

Self-attention $f(X) = (\Gamma_X(x_1), \dots, \Gamma_X(x_n))$ where

$$\Gamma_X(z) = \sum_{j=1}^{n} p_j(z) \color{red}{V} x_j \quad \text{with} \quad p_j(z) = \exp(\langle \color{red}{A} z, x_j \rangle) / \sum_{\ell=1}^{n} \exp(\langle \color{red}{A} z, x_\ell \rangle)$$

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

Mean-field self-attention [Sander et al., 2022, Vuckovic et al., 2020]

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

$$\Gamma_\mu : z \in \mathbb{R}^d \mapsto \int k(z,y) \color{red}{V} y \mathrm{d}\mu(y) \quad \text{with} \quad k(z,y) := e^{\langle \color{red}{A} z, y \rangle} / \int e^{\langle \color{red}{A} z, y' \rangle} \mathrm{d}\mu(y')$$

Mean-field attention and the Transformer PDE

Self-attention $f(X) = (\Gamma_X(x_1), \dots, \Gamma_X(x_n))$ where

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

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

Mean-field self-attention [Sander et al., 2022, Vuckovic et al., 2020]

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

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

- Suited for handling very long sequence lengths
- Well-defined for μ compactly supported, Gaussian...
- Covers discrete case: view X as $\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$

Mean-field attention and the Transformer PDE

Self-attention $f(X) = (\Gamma_X(x_1), \ldots, \Gamma_X(x_n))$ where

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

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

Mean-field self-attention [Sander et al., 2022, Vuckovic et al., 2020]

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

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

$$1 \leq i \leq n, \quad \dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \quad \longrightarrow \quad \partial_t \mu + \mathrm{div}(\mu \Gamma_\mu) = 0 \tag{1}$$

II - Well-posedness of the Transformer PDE for compactly supported initial data

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

The Transformer PDE in the compactly supported case

$$\mu_0 \text{ compactly supported}, \qquad \partial_t \mu + \mathrm{div}(\mu \Gamma_\mu) = 0 \tag{1}$$

The Transformer PDE in the compactly supported case

$$\mu_0 \text{ compactly supported}, \quad \partial_t \mu + \mathrm{div}(\mu \Gamma_\mu) = 0 \tag{1}$$

Well-posedness [Geshkovski et al., 2024, Castin et al., 2025]

- Assume $A(t), V(t)$ continuous
- Assume $\mathrm{supp}\, \mu_0 \subset B(0, R_0)$

Then (1) has a unique global weak solution μ, such that

$$\mathrm{supp}\, \mu(t) \subset B(0, e^{\int_0^t \|V(s)\|_2 \mathrm{d}s} R_0).$$

The Transformer PDE in the compactly supported case

$$\mu_0 \text{ compactly supported}, \quad \partial_t \mu + \operatorname{div}(\mu \Gamma_\mu) = 0 \tag{1}$$

Well-posedness [Geshkovski et al., 2024, Castin et al., 2025]

- Assume $A(t), V(t)$ continuous
- Assume $\operatorname{supp} \mu_0 \subset B(0, R_0)$

Then (1) has a unique global weak solution μ, such that

$$\operatorname{supp} \mu(t) \subset B(0, e^{\int_0^t \|V(s)\|_2 \mathrm{d}s} R_0).$$

If $\operatorname{supp} \nu_0 \subset B(0, R_0)$ then

$$W_p(\mu(t), \nu(t)) \leq C(t, R_0) W_p(\mu_0, \nu_0) \quad \forall p \geq 1$$

with $C(t, R_0) \propto e^{tR(t)^2}$

The Transformer PDE in the compactly supported case

$$\Gamma_\mu(z) = \int k(z,y) V y \,\mathrm{d}\mu(y) \quad \text{with} \quad k(z,y) := e^{\langle Az, y\rangle} / \int e^{\langle Az, y'\rangle} \mathrm{d}\mu(y')$$

Central estimates for proof

1. $\sup_{x \in \mathbb{R}^d} |\Gamma_\mu(x)| \leq \|V\|_2 R$,
2. $\sup_{x \in \mathbb{R}^d} \|D_x \Gamma_\mu(x)\|_2 \leq \|V\|_2 \|A\|_2 R^2$,
3. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x, R) W_p(\mu, \nu)$

The Transformer PDE in the compactly supported case

$$\Gamma_\mu(z) = \int k(z,y) V y \,\mathrm{d}\mu(y) \quad \text{with} \quad k(z,y) := e^{\langle Az, y\rangle} \Big/ \int e^{\langle Az, y'\rangle} \mathrm{d}\mu(y')$$

Central estimates for proof

1. $\sup_{x\in\mathbb{R}^d} |\Gamma_\mu(x)| \leq \|V\|_2 R$,
2. $\sup_{x\in\mathbb{R}^d} \|D_x \Gamma_\mu(x)\|_2 \leq \|V\|_2 \|A\|_2 R^2$,
3. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x,R) W_p(\mu,\nu)$

- Eq. 1 controls radius growth: $\operatorname{supp}\mu(t) \subset B(0, e^{\int_0^t \|V(s)\|_2 \mathrm{d}s} R_0)$

The Transformer PDE in the compactly supported case

$$\Gamma_\mu(z) = \int k(z,y) V y \mathrm{d}\mu(y) \quad \text{with} \quad k(z,y) := e^{\langle Az, y\rangle} / \int e^{\langle Az, y'\rangle} \mathrm{d}\mu(y')$$

Central estimates for proof

1. $\sup_{x\in\mathbb{R}^d} |\Gamma_\mu(x)| \leq \|V\|_2 R$,
2. $\sup_{x\in\mathbb{R}^d} \|D_x \Gamma_\mu(x)\|_2 \leq \|V\|_2 \|A\|_2 R^2$,
3. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x,R) W_p(\mu,\nu)$

- Eq. 1 controls radius growth: $\operatorname{supp} \mu(t) \subset B(0, e^{\int_0^t \|V(s)\|_2 \mathrm{d}s} R_0)$
- Modular framework $\to$ extends to attention variants!

Extending to attention variants

Attention map: $\Gamma_\mu(z) = \int k(z, y) V y \mathrm{d}\mu(y)$

✓ Softmax attention: $k(z, y) = e^{\langle Az, y \rangle} / \int e^{\langle Az, y' \rangle} \mathrm{d}\mu(y')$

Extending to attention variants

Attention map: $\Gamma_\mu(z) = \int k(z,y) V y \mathrm{d}\mu(y)$

- ✓ Softmax attention: $k(z,y) = e^{\langle Az, y\rangle} / \int e^{\langle Az, y'\rangle} \mathrm{d}\mu(y')$
- ✓ L2 attention: $k(z,y) = e^{-|Qz - Ky|^2} / \int e^{-|Qz - Ky'|^2} \mathrm{d}\mu(y')$

Extending to attention variants

Attention map: $\Gamma_\mu(z) = \int k(z,y) V y \mathrm{d}\mu(y)$

- ✓ Softmax attention: $k(z,y) = e^{\langle Az,y\rangle} / \int e^{\langle Az,y'\rangle} \mathrm{d}\mu(y')$
- ✓ L2 attention: $k(z,y) = e^{-|Qz-Ky|^2} / \int e^{-|Qz-Ky'|^2} \mathrm{d}\mu(y')$
- ✓ Sinkhorn attention: $k(z,y)$ is the limit $j \to +\infty$ of

$$\kappa^0(z,y) = e^{\langle Az,y\rangle}, \quad \kappa^{j+1}(z,y) = \begin{cases} \frac{\kappa^j(z,y)}{\int \kappa^j(z,y')\mathrm{d}\mu(y')} & \text{if } j \text{ is even,} \\ \frac{\kappa^j(z,y)}{\int \kappa^j(z',y)\mathrm{d}\mu(z')} & \text{if } j \text{ is odd} \end{cases}$$

Extending to attention variants

Attention map: $\Gamma_\mu(z) = \int k(z,y) V y \mathrm{d}\mu(y)$

- ✓ Softmax attention: $k(z,y) = e^{\langle Az,y\rangle} / \int e^{\langle Az,y'\rangle} \mathrm{d}\mu(y')$
- ✓ L2 attention: $k(z,y) = e^{-|Qz-Ky|^2} / \int e^{-|Qz-Ky'|^2} \mathrm{d}\mu(y')$
- ✓ Sinkhorn attention: $k(z,y)$ is the limit $j \to +\infty$ of

$$\kappa^0(z,y) = e^{\langle Az,y\rangle}, \quad \kappa^{j+1}(z,y) = \begin{cases} \frac{\kappa^j(z,y)}{\int \kappa^j(z,y')\mathrm{d}\mu(y')} & \text{if } j \text{ is even,} \\ \frac{\kappa^j(z,y)}{\int \kappa^j(z',y)\mathrm{d}\mu(z')} & \text{if } j \text{ is odd} \end{cases}$$

- ✓ Masked attention

Extending to attention variants

Attention map: $\Gamma_\mu(z) = \int k(z,y) V y \mathrm{d}\mu(y)$

- ✓ Softmax attention: $k(z,y) = e^{\langle Az,y\rangle} / \int e^{\langle Az,y'\rangle}\mathrm{d}\mu(y')$
- ✓ L2 attention: $k(z,y) = e^{-|Qz-Ky|^2} / \int e^{-|Qz-Ky'|^2}\mathrm{d}\mu(y')$
- ✓ Sinkhorn attention: $k(z,y)$ is the limit $j \to +\infty$ of

$$\kappa^0(z,y) = e^{\langle Az,y\rangle}, \quad \kappa^{j+1}(z,y) = \begin{cases} \frac{\kappa^j(z,y)}{\int \kappa^j(z,y')\mathrm{d}\mu(y')} & \text{if } j \text{ is even,} \\ \frac{\kappa^j(z,y)}{\int \kappa^j(z',y)\mathrm{d}\mu(z')} & \text{if } j \text{ is odd} \end{cases}$$

- ✓ Masked attention
- ✓ Multihead attention: $\Gamma_\mu = \sum_{h=1}^{H} \Gamma_\mu^{(h)}$

Extending to attention variants

Attention map: $\Gamma_\mu(z) = \int k(z,y) V y \mathrm{d}\mu(y)$

- ✗ Unnormalized attention: $k(z,y) = e^{\langle Az, y\rangle}$
- ✗ Linear attention: $k(z,y) = \langle Az, y\rangle$
- ✗ ReLU attention: $k(z,y) = \mathrm{ReLU}(\langle Az, y\rangle)$
- ✗ Sigmoid attention: $k(z,y) = \sigma(\langle Az, y\rangle)$

Extending to attention variants

Attention map: $\Gamma_\mu(z) = \int k(z,y) V y \mathrm{d}\mu(y)$

- ✗ Unnormalized attention: $k(z,y) = e^{\langle Az, y\rangle}$
- ✗ Linear attention: $k(z,y) = \langle Az, y\rangle$
- ✗ ReLU attention: $k(z,y) = \mathrm{ReLU}(\langle Az, y\rangle)$
- ✗ Sigmoid attention: $k(z,y) = \sigma(\langle Az, y\rangle)$

- Estimate 1

$$\sup_{x\in\mathbb{R}^d} |\Gamma_\mu(x)| \leq \|V\|_2 R$$

is *not* satisfied $\to$ no global solution

Extending to attention variants

Attention map: $\Gamma_\mu(z) = \int k(z,y) V y \mathrm{d}\mu(y)$

- ✗ Unnormalized attention: $k(z,y) = e^{\langle Az,y\rangle}$
- ✗ Linear attention: $k(z,y) = \langle Az, y\rangle$
- ✗ ReLU attention: $k(z,y) = \mathrm{ReLU}(\langle Az, y\rangle)$
- ✗ Sigmoid attention: $k(z,y) = \sigma(\langle Az, y\rangle)$

- Estimate 1

$$\sup_{x\in\mathbb{R}^d} |\Gamma_\mu(x)| \leq \|V\|_2 R$$

  is *not* satisfied $\to$ no global solution
- But LayerNorm solves the problem

III - Clustering in the Gaussian case

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

The Transformer PDE in the Gaussian case

Lemma: attention map on Gaussians

If $\mu = \mathcal{N}(\alpha, \Sigma)$ then

$$\Gamma_\mu(x) = V(\alpha + \Sigma A x)$$

Similar for L2 and Sinkhorn attention!

The Transformer PDE in the Gaussian case

Lemma: attention map on Gaussians

If $\mu = \mathcal{N}(\alpha, \Sigma)$ then

$$\Gamma_\mu(x) = V(\alpha + \Sigma A x)$$

Similar for L2 and Sinkhorn attention!

Proposition: evolution of Gaussian initial data [Castin et al., 2025]

Consider

$$\partial_t \mu + \mathrm{div}(\mu \Gamma_\mu) = 0 \tag{1}$$

Assume A, V continuous and $\mu_0 = \mathcal{N}(\alpha_0, \Sigma_0)$. Then (1) has a unique maximal solution on $[0, t_{\max})$, Gaussian for all t.

The Transformer PDE in the Gaussian case

Lemma: attention map on Gaussians

If $\mu = \mathcal{N}(\alpha, \Sigma)$ then

$$\Gamma_\mu(x) = V(\alpha + \Sigma A x)$$

Similar for L2 and Sinkhorn attention!

Proposition: evolution of Gaussian initial data [Castin et al., 2025]

Consider

$$\partial_t \mu + \mathrm{div}(\mu \Gamma_\mu) = 0 \tag{1}$$

Assume A, V continuous and $\mu_0 = \mathcal{N}(\alpha_0, \Sigma_0)$. Then (1) has a unique maximal solution on $[0, t_{\max})$, Gaussian for all t. Denoting $\mu(t) = \mathcal{N}(\alpha(t), \Sigma(t))$:

$$\begin{cases} \dot{\alpha} = V(I_d + \Sigma A)\alpha \\ \dot{\Sigma} = V\Sigma A\Sigma + \Sigma A^\top \Sigma V^\top \end{cases}$$

Clustering emerges in the Gaussian case

Proposition: closed-form analysis

Consider

$$\dot{\Sigma} = V\Sigma A\Sigma + \Sigma A^\top \Sigma V^\top.$$

Assume

- A, V constant
- V commutes with $VA + A^\top V^\top$ and Σ_0

Then

$$\Sigma(t) = (\Sigma_0^{-1} - t(VA + A^\top V^\top))^{-1}$$

Clustering emerges in the Gaussian case

Proposition: closed-form analysis

Consider

$$\dot{\Sigma} = V\Sigma A\Sigma + \Sigma A^\top \Sigma V^\top.$$

Assume

- A, V constant
- V commutes with $VA + A^\top V^\top$ and Σ_0

Then

$$\Sigma(t) = (\Sigma_0^{-1} - t(VA + A^\top V^\top))^{-1}$$

- If $VA + A^\top V^\top \preceq 0$ the solution is global and converges to Σ^* such that

$$\lambda_i(VA + A^\top V^\top) < 0 \Rightarrow \lambda_i(\Sigma^*) = 0 \qquad \text{"clustering"}$$

Clustering emerges in the Gaussian case

Proposition: closed-form analysis

Consider

$$\dot{\Sigma} = V\Sigma A\Sigma + \Sigma A^\top \Sigma V^\top.$$

Assume

- A, V constant
- V commutes with $VA + A^\top V^\top$ and Σ_0

Then

$$\Sigma(t) = (\Sigma_0^{-1} - t(VA + A^\top V^\top))^{-1}$$

- If $VA + A^\top V^\top \preceq 0$ the solution is global and converges to Σ^* such that

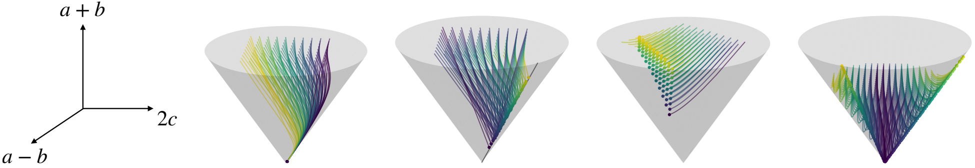
$$\lambda_i(VA + A^\top V^\top) < 0 \Rightarrow \lambda_i(\Sigma^*) = 0 \qquad \text{"clustering"}$$

- Otherwise $\lambda_1(\Sigma(t)) \to +\infty$ in finite time

Clustering emerges in the Gaussian case

Plot the covariance evolution for $d = 2$:

(a) Conv. to zero (b) Conv. to a line (c) Conv. to a plane (d) Conv. to two lines



$$\begin{pmatrix} a & c \\ c & b \end{pmatrix} \in \mathcal{S}_2^+ \mapsto (x, y, z) := (a - b, 2c, a + b)$$

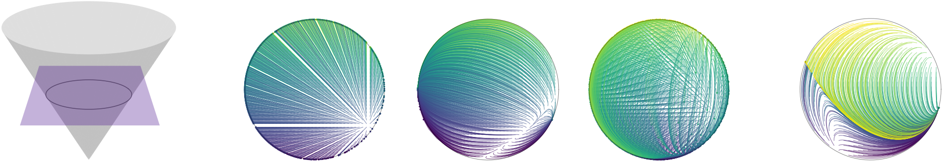
Comparing attention variants

- Softmax, L2, Sinkhorn attention $\rightarrow$ similar behavior when converge
- L2 does not blow-up in finite time $\rightarrow$ more regular

Comparing attention variants

- Softmax, L2, Sinkhorn attention $\rightarrow$ similar behavior when converge
- L2 does not blow-up in finite time $\rightarrow$ more regular

$\Sigma \mapsto \Sigma / \mathrm{Tr}\Sigma$ (a) Blow-up (b) Blow-up (c) Div. (d) Conv./blow-up



SM att. MH att. L2 att. SM att.

Conclusion of parts II and III

- Mean-field attention and the Transformer PDE generalize dynamics to infinitely many tokens
- Compactly supported data: well-posed PDE, very sensitive to initial condition
- Gaussian data: clustering, possible finite-time blow-up $\rightarrow$ LayerNorm changes a lot the dynamics

- Beyond Gaussian case?
- What about next-token prediction? (Masked attention dynamics)
- From discrete to continuous time, what changes?

IV - Mean-field can be overpessimistic: refined estimates on the attention map

Lipschitz constant of mean-field attention

Estimates on Γ_μ

1. $\sup_{x\in\mathbb{R}^d}|\Gamma_\mu(x)| \leq \|V\|_2 R$,
2. $\sup_{x\in\mathbb{R}^d}\|D_x\Gamma_\mu(x)\|_2 \leq \|V\|_2 \|A\|_2 R^2$,
3. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x, R) W_p(\mu, \nu)$

2. and 3. imply the Wasserstein Lipschitz bound

$$W_p(F(\mu), F(\nu)) \leq c(A, V) R^2 e^{\|A\|_2 R^2} W_p(\mu, \nu)$$

Lipschitz constant of mean-field attention

Estimates on Γ_μ

1. $\sup_{x\in\mathbb{R}^d} |\Gamma_\mu(x)| \leq \|V\|_2 R$,
2. $\sup_{x\in\mathbb{R}^d} \|D_x\Gamma_\mu(x)\|_2 \leq \|V\|_2 \|A\|_2 R^2$,
3. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x, R) W_p(\mu, \nu)$

2. and 3. imply the Wasserstein Lipschitz bound

$$W_p(F(\mu), F(\nu)) \leq c(A, V) R^2 e^{\|A\|_2 R^2} W_p(\mu, \nu)$$

that translates to a discrete bound

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

Lipschitz constant of mean-field attention

Estimates on Γ_μ

1. $\sup_{x\in\mathbb{R}^d}|\Gamma_\mu(x)| \leq \|V\|_2 R$,
2. $\sup_{x\in\mathbb{R}^d}\|D_x\Gamma_\mu(x)\|_2 \leq \|V\|_2 \|A\|_2 R^2$,
3. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x,R)W_p(\mu,\nu)$

2. and 3. imply the Wasserstein Lipschitz bound

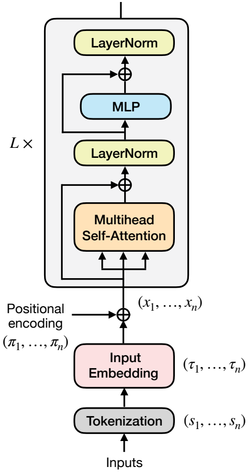
$$W_p(F(\mu),F(\nu)) \leq c(A,V)R^2 e^{\|A\|_2 R^2} W_p(\mu,\nu)$$

that translates to a discrete bound

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

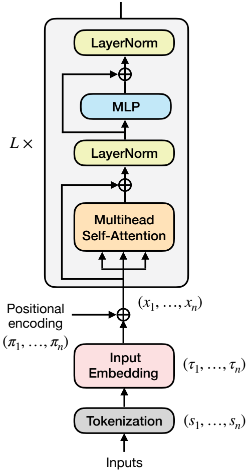
$\rightarrow$ Refine the bound with n?

Studying the robustness of Transformers



How much can the output of a Transformer change when slightly perturbing the input?

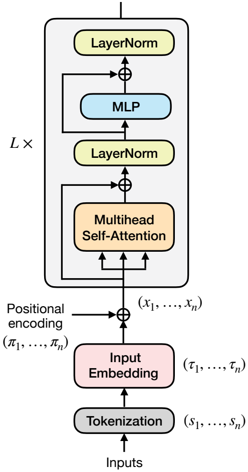
Studying the robustness of Transformers



How much can the output of a Transformer change when slightly perturbing the input?

- controls robustness and expressive power

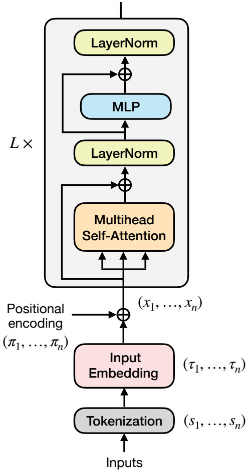
Studying the robustness of Transformers



How much can the output of a Transformer change when slightly perturbing the input?

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

Studying the robustness of Transformers



How much can the output of a Transformer change when slightly perturbing the input?

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

Does the robustness of an input depend on the sequence length?

Measuring regularity with the local Lipschitz constant

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

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:

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

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]

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

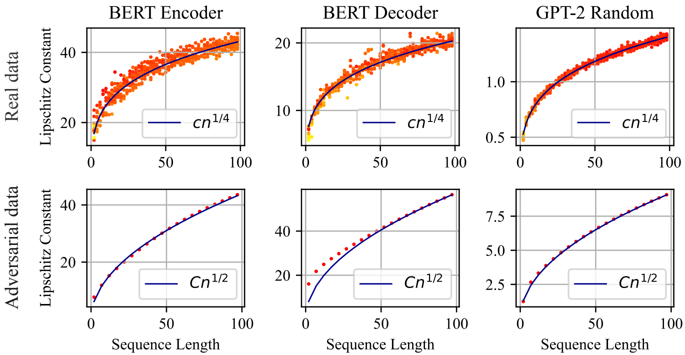
and if $V = I_d$,

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

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 $C\sqrt{n}$
- mean-field bound: $\mathrm{Lip}(f_{|B_R^n}) \leq c(A, V)R^2 e^{\|A\|_2 R^2}$

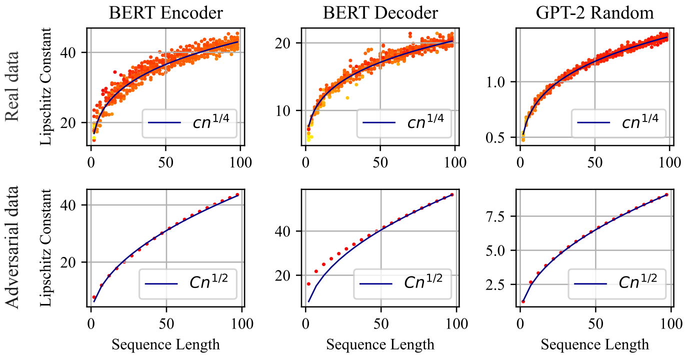
Experiments: typical case and worst case



Local Lipschitz constant of real vs. adversarial data

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

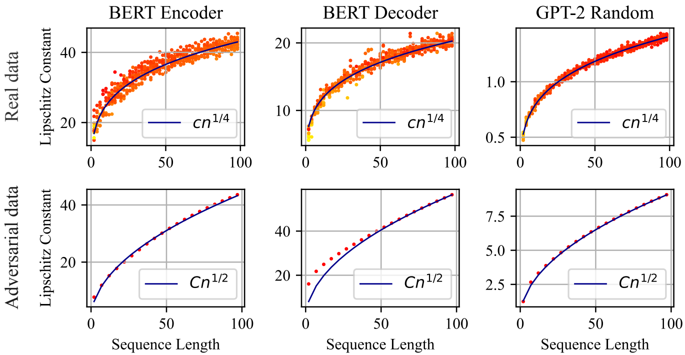
Experiments: typical case and worst case



Local Lipschitz constant of real vs. adversarial data

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

Conclusion of part IV

- Mean-field estimates are over-pessimistic for practical sequence lenghts
- Yet, smoothness of attention at an input depends on the sequence length

Open question: find a statistical model for the data that explains the observed growth rate?

Thank you!

References

Castin, V., Ablin, P., Carrillo, J. A., and Peyré, G. (2025).
A unified perspective on the dynamics of deep transformers.
In *arXiv preprint arXiv:2501.18322*.

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.

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.

Vuckovic, J., Baratin, A., and Combes, R. T. d. (2020).
A mathematical theory of attention.
arXiv preprint arXiv:2007.02876.

V - 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, $V \in \mathbb{R}^{d \times d}$

Mean-field regime? $\rightarrow$ 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

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

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

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

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

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

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

with

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