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

Valérie Castin
Ecole Normale Supérieure PSL, Paris

RSME, Universidad de Alicante, 21 January 2026

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*

Outline

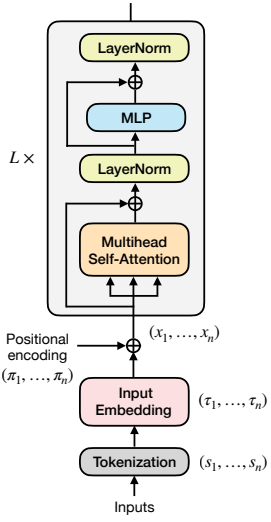
I - Discrete and mean-field model for Transformers

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

III - Clustering in the Gaussian case

I - Discrete and mean-field model for Transformers

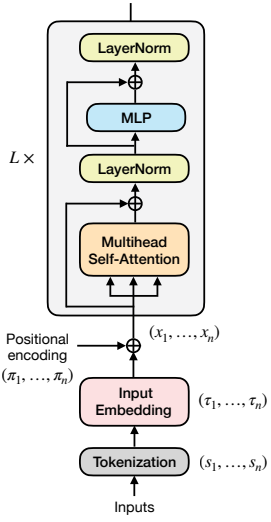
The Transformer architecture



Transformers process tuples:

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

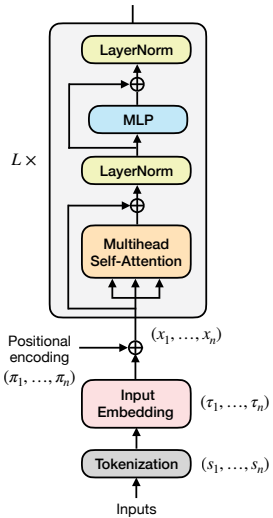
The Transformer architecture



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

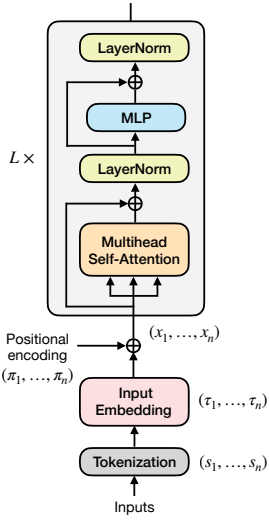


Transformers process tuples:

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

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

The Transformer architecture



Transformers process tuples:

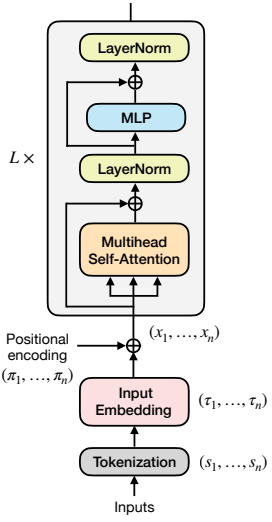
- Traditional neural network: $x \in \mathbb{R}^{d_{\text{in}}} \mapsto x' \in \mathbb{R}^{d_{\text{out}}}$
- Transformer:
  $(x_1, \ldots, x_n) \in (\mathbb{R}^d_{\text{in}})^n \mapsto (x'_1, \ldots, x'_n) \in (\mathbb{R}^d_{\text{out}})^n$

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

- NLP: tokens = words or subwords

This is how GPT-3 tokenizes this sentence.

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

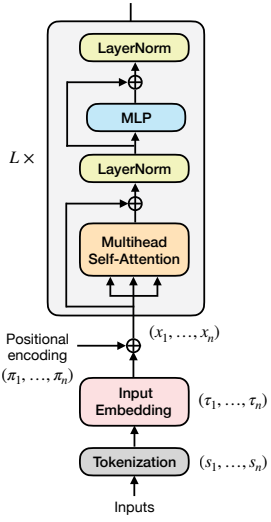
Tokens $X = (x_1, \ldots, x_n)$ are obtained through **tokenization**:

- NLP: tokens = words or subwords

  This is how GPT-3 tokenizes this sentence.

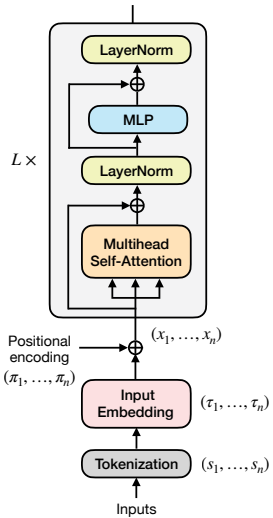
- Vision: tokens = image patches

The Transformer architecture



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

The Transformer architecture

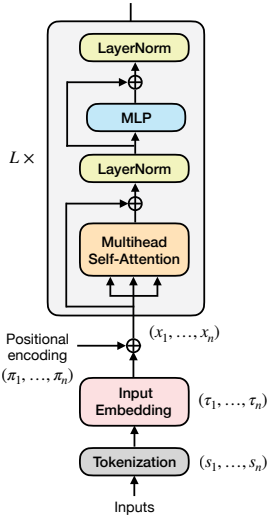


- One input $\to$ n tokens $X \coloneqq (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$
- Self-attention with params A, V:

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

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

The Transformer architecture



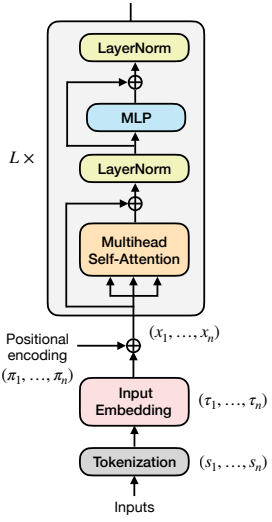
- One input $\to$ n tokens $X \coloneqq (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$
- Self-attention with params A, V:

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

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

- Multilayer perceptron $g\colon \mathbb{R}^d \to \mathbb{R}^d, x \mapsto W\sigma(Ux + b)$ (applied token-wise)

The Transformer architecture



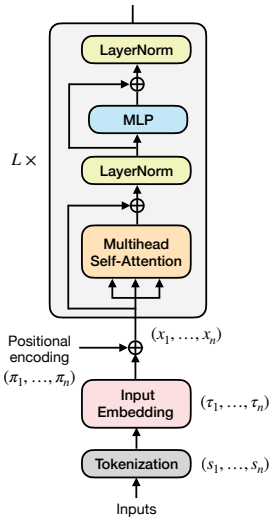
- One input $\to$ n tokens $X \coloneqq (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$
- Self-attention with params A, V:

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

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

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

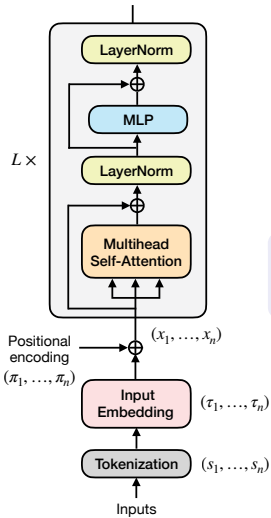
Studying the dynamics of tokens across layers



Without MLP and LN:

$$x_i(t+1) = x_i(t) + \Gamma_{X_t}(x_i(t)) \qquad 1 \leq i \leq n$$

Studying the dynamics of tokens across layers

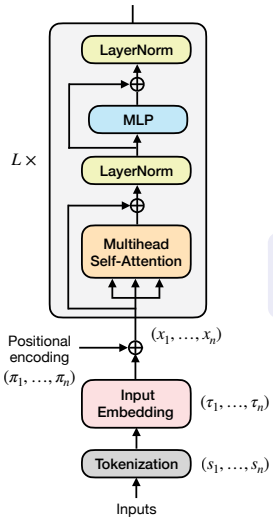


Without MLP and LN:

$$x_i(t+1) = x_i(t) + \Gamma_{X_t}(x_i(t)) \qquad 1 \leq i \leq n$$

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

Studying the dynamics of tokens across layers



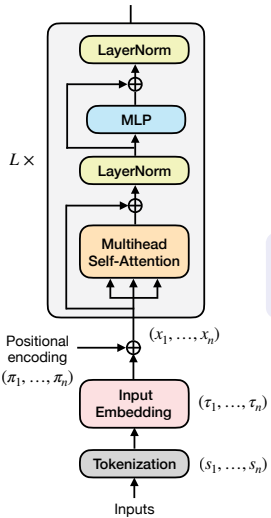
Without MLP and LN:

$$x_i(t+1) = x_i(t) + \Gamma_{X_t}(x_i(t)) \qquad 1 \leq i \leq n$$

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



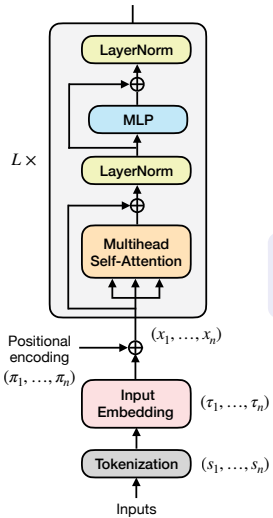
Without MLP and LN:

$$x_i(t+1) = x_i(t) + \Gamma_{X_t}(x_i(t)) \qquad 1 \leq i \leq n$$

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:

$$x_i(t+1) = x_i(t) + \Gamma_{X_t}(x_i(t)) \qquad 1 \leq i \leq n$$

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

Modeling the Transformer dynamics

Infinite-depth limit [Sander et al., 2022, Geshkovski et al., 2023]

$$\dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \qquad 1 \leq i \leq n \qquad \text{(similar to Neural ODEs)}$$

➔ well-posed

➔ clustering of tokens when $t \to +\infty$

Modeling the Transformer dynamics

Infinite-depth limit [Sander et al., 2022, Geshkovski et al., 2023]

$$\dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \qquad 1 \leq i \leq n \qquad \text{(similar to Neural ODEs)}$$

➔ well-posed

➔ clustering of tokens when $t \to +\infty$

Infinite-depth & infinite-length limit [Sander et al., 2022, Geshkovski et al., 2023, Castin et al., 2025]

Mean-field attention map ("infinite sequence length"):

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

Modeling the Transformer dynamics

Infinite-depth limit [Sander et al., 2022, Geshkovski et al., 2023]

$$\dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \qquad 1 \leq i \leq n \qquad \text{(similar to Neural ODEs)}$$

➔ well-posed

➔ clustering of tokens when $t \to +\infty$

Infinite-depth & infinite-length limit [Sander et al., 2022, Geshkovski et al., 2023, Castin et al., 2025]

Mean-field attention map ("infinite sequence length"):

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

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

Modeling the Transformer dynamics

Infinite-depth limit [Sander et al., 2022, Geshkovski et al., 2023]

$$\dot{x}_i(t) = \Gamma_{X(t)}(x_i(t)) \qquad 1 \leq i \leq n \qquad \text{(similar to Neural ODEs)}$$

➔ well-posed

➔ clustering of tokens when $t \to +\infty$

Infinite-depth & infinite-length limit [Sander et al., 2022, Geshkovski et al., 2023, Castin et al., 2025]

Mean-field attention map ("infinite sequence length"):

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

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

➔ well-posed for compactly supported μ_0

➔ behavior for measures with density?

II - Well-posedness and stability estimates 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

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

The Transformer PDE in the compactly supported case

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

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

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

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

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

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. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x,R) W_p(\mu,\nu)$ with $c(x,R)$ exp in R^2

The Transformer PDE in the compactly supported case

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

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. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x,R) W_p(\mu,\nu)$ with $c(x,R)$ exp in R^2

- **Refine estimate 3** if μ empirical measure with n diracs [Castin et al., 2024]:

$$c(x,R) \leq \|A\|_2 R^2$$

The Transformer PDE in the compactly supported case

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

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. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x,R) W_p(\mu,\nu)$ with $c(x,R)$ exp in R^2

- **Refine estimate 3** if μ empirical measure with n diracs [Castin et al., 2024]:

$$c(x,R) \leq \|A\|_2 R^2$$

- **LayerNorm** prevents exp growth of $R(t)$

The Transformer PDE in the compactly supported case

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

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. $|\Gamma_\mu(x) - \Gamma_\nu(x)| \leq c(x,R) W_p(\mu,\nu)$ with $c(x,R)$ exp in R^2

- **Refine estimate 3** if μ empirical measure with n diracs [Castin et al., 2024]:

$$c(x,R) \leq \|A\|_2 \, R^2$$

- **LayerNorm** prevents exp growth of $R(t)$
- Modular framework $\rightarrow$ 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

III - Clustering in the Gaussian case

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

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

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

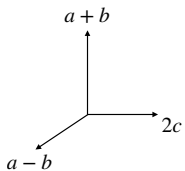
- If $VA + A^\top V^\top \preceq 0$ the solution is global and converges to Σ^* such that

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

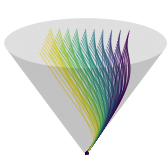
- Otherwise $\lambda_1(\Sigma(t)) \to +\infty$ in finite time

Clustering emerges in the Gaussian case

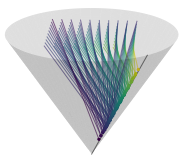
Plot the covariance evolution for $d = 2$:



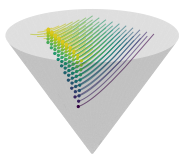
(a) Conv. to zero



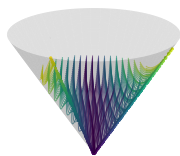
(b) Conv. to a line



(c) Conv. to a plane



(d) Conv. to two lines



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

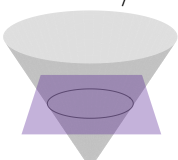
Comparing attention variants

- 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

$\Sigma \mapsto \Sigma/\mathrm{Tr}\Sigma$



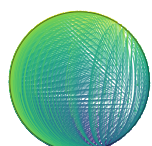
(a) Blow-up

SM att.

(b) Blow-up

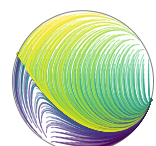
MH att.

(c) Div.



L2 att.

(d) Conv./blow-up



SM att.

Conclusion

- 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?

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. (2023).
A mathematical perspective on transformers.
arXiv preprint arXiv:2312.10794.


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.