Modeling Transformers with PDEs: Well-Posedness and Asymptotic Behavior

Valérie Castin [1] Gabriel Peyré [1,3] Pierre Ablin [2] José Antonio Carrillo [4]

[1]ENS PSL, Paris [2]Apple [3]CNRS [4]University of Oxford

The Transformer Architecture

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

This is how GPT-3 tokenizes this sentence.

Figure 1. Tokenization of text (GPT2 tokenizer)

Figure 2. Tokenization of images

Self-attention grasps dependencies between tokens (e.g. semantic dependencies) and is coupled with 2-layer **multi-layer perceptron** and **layer normalization**. All layers are **residual**.

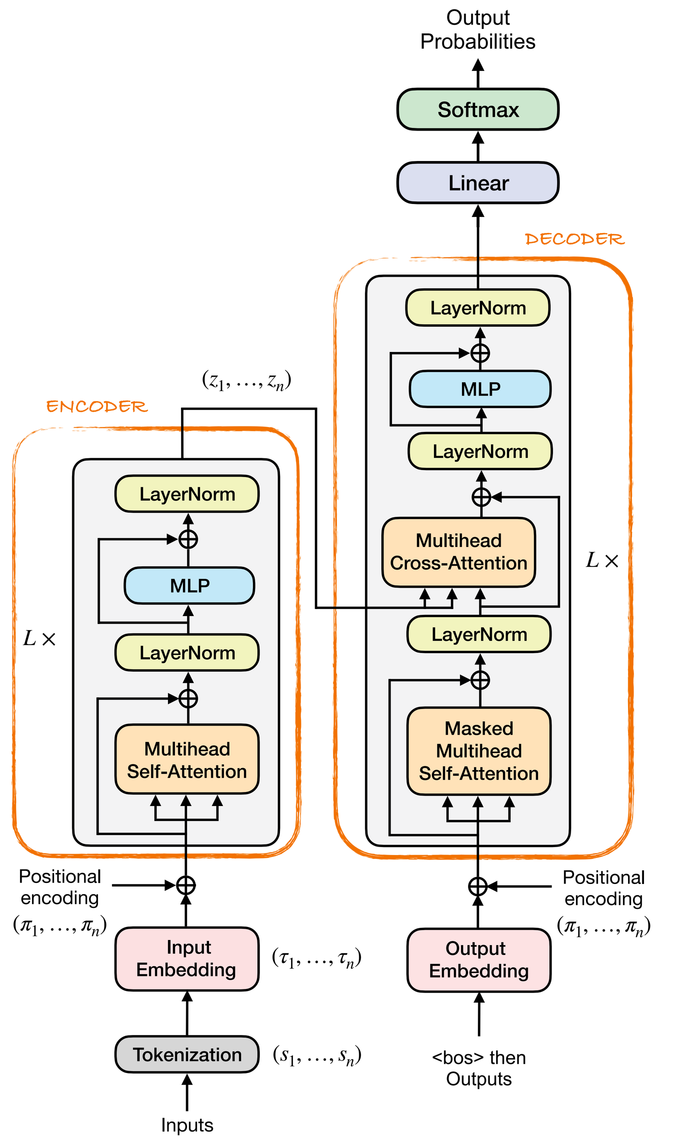


Figure 3. The original Transformer architecture, by Vaswani et al. [4]

Tokens can be seen as **interacting particles** in the Encoder.

Setup – Viewing a Transformer as a PDE

We consider a simplified Transformer with only residual self-attention blocks:

$$f := (\mathrm{id} + f^L) \circ \cdots \circ (\mathrm{id} + f^1) \tag{1}$$

with

$$f^\ell : X := (x_1, \dots, x_n) \mapsto (\Gamma^\ell_X(x_1), \dots, \Gamma^\ell_X(x_n))$$

for some function $\Gamma^\ell_X : \mathbb{R}^d \to \mathbb{R}^d$. Equation (1) can then be seen as the discretization of

$$\dot{x}_i = \Gamma(t, X(t))_i \qquad 1 \le i \le n.$$

The mean-field limit of this system of equations is then of the form

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

Variants of self-attention

Self-attention has three parameters $Q, K, V \in \mathbb{R}^{d \times d}$. Denote $A := K^\top Q$.

- **Traditional self-attention** by Vaswani et al. [4]

$$\Gamma^{(\mathrm{trad})}_\mu : x \in \mathbb{R}^d \mapsto \frac{\int V y\, e^{Ax\cdot y} \mathrm{d}\mu(y)}{\int e^{Ax \cdot y} \mathrm{d}\mu(y)}.$$

- **L2 self-attention** by Kim et al. [2]

$$\Gamma^{(\mathrm{L2})}_\mu : x \in \mathbb{R}^d \mapsto \frac{\int V y\, e^{-|Qx - Ky|^2} \mathrm{d}\mu(y)}{\int e^{-|Qx-Ky|^2} \mathrm{d}\mu(y)}.$$

- **Sinkformer self-attention** by Sander et al. [3]

$$\Gamma^{(\mathrm{sink})}_\mu : x \in \mathbb{R}^d \mapsto \int V y\, k^\infty(x,y) \mathrm{d}\mu(y)$$

where k^∞ is obtained by performing the Sinkhorn-Knopp algorithm on $k^0(x,y) := e^{-|Qx-Ky|^2}$, i.e. $k^\infty(x,y)$ is the limit of the following sequence:

$$k^{j+1}(x,y) = \begin{cases} \frac{k^j(x,y)}{\int k^j(x,y')\mathrm{d}\mu(y')} & \text{if } j \text{ is even,} \\ \frac{k^j(x,y)}{\int k^j(x',y)\mathrm{d}\mu(x')} & \text{if } j \text{ is odd.} \end{cases}$$

Contribution 1 – Well-posedness for compactly supported initial data

Equip $\mathcal{P}_p(\mathbb{R}^d)$ with the p-Wasserstein distance W_p. If $Q, K, V : [0, +\infty) \to \mathbb{R}^{d\times d}$ are **continuous** and the initial data μ_0 is **compactly supported**, then for all considered types of self-attention, the evolution

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

has a unique global weak solution $\mu \in \mathcal{C}([0,+\infty), \mathcal{P}_p(\mathbb{R}^d))$. Moreover, the radius $R(t)$ of the support of $\mu(t)$ satisfies

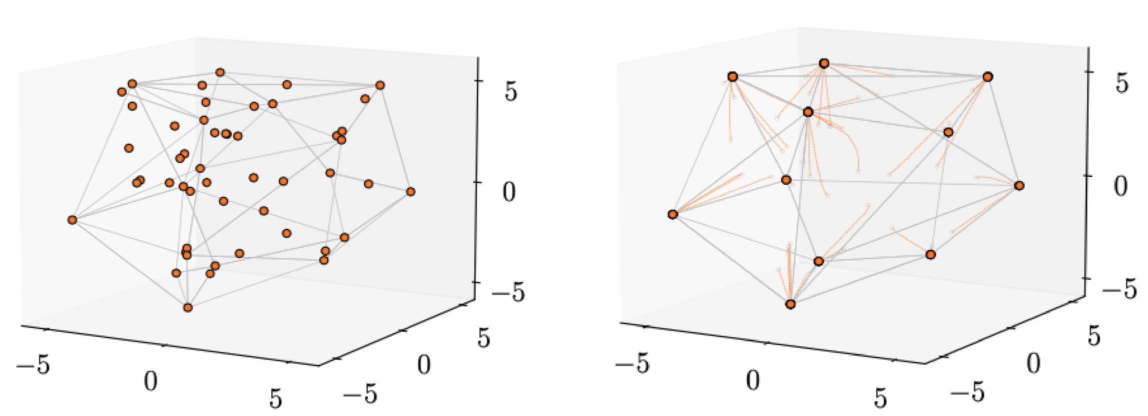
$$R(t) \le e^{\int_0^t \|V(s)\|_2 \mathrm{d}s} R_0$$

and we have a stability estimate

$$W_p(\mu(t), \nu(t)) \le C(T, R_0) W_p(\mu_0, \nu_0).$$

State of the art – Behavior for n tokens

Geshkovski et al. [1] show the emergence of clusters when μ_0 is an empirical measure, after the rescaling $z_i := e^{-tV} x_i$.



(a) Tokens at $t = 0$ (b) Tokens at $t = 5$

Figure 4. Clustering dynamics evidentiated by Geshkovski et al. [1] after the rescaling $z_i := e^{-tV} x_i$, for $Q = K = V = I_3$.

Contribution 2 – Behavior for a Gaussian initial condition

When $\mu_0 \sim \mathcal{N}(\alpha, \Sigma)$, the solution $\mu(t)$ of

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

stays Gaussian over time for all considered types of self-attention. We derive an ODE on the covariance matrix Σ of $\mu(t)$.

- **Traditional self-attention**

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

Two cases: finite-time blow-up or convergence to a low-rank matrix → clustering effect.

- **L2 self-attention**

$$\dot{\Sigma} = 2V(\Sigma^{-1} + 2K^\top K)^{-1} A \Sigma + 2\Sigma A^\top (\Sigma^{-1} + 2K^\top K)^{-1} V^\top.$$

We always have global existence. Two cases: divergence of at least one eigenvalue or convergence to a low-rank matrix → clustering effect.

- **Sinkformer self-attention**

$$\dot{\Sigma} = V C_\Sigma \Sigma^{-1} (A^\top)^{-1} \Sigma + \Sigma A^{-1} \Sigma^{-1} C_\Sigma^\top V^\top,$$

with $C_\Sigma := \frac{1}{2}\left(\Sigma^{1/2}(4\Sigma^{1/2} A \Sigma A^\top \Sigma^{1/2} + I_d)^{1/2} \Sigma^{-1/2} - I_d\right)$.

Contribution 3 – Handling masked self-attention

Masked self-attention is defined as

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

with f traditional self-attention. How to extend it to probability measures?

Mean-field masked self-attention:

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

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

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

Then the evolution

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

with compactly supported initial data is well-posed.

References

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

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

[3] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

 valerie.castin@orange.fr