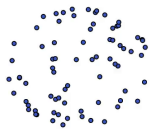
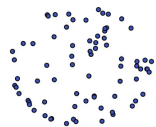
Mode collapse and metastability in Transformers





Berlin-Leipzig hybrid Seminar: Mathematics of Machine Learning

Viktor Stein, 15.09.2025

Outline

I. What are transformers?

II. ODE and PDE description of transformers

III. Long-time behavior - emergence of clusters

Table of Contents

LLMs, GPTs, etc

- GPT = generative pretrained transformer, a type of LLM = large language model.
- ChatGPT receives "question" (text input sequence) and *generates* "answer" (text output sequence) left-to-right.
- Before transformers: sequence-to-sequence (Seq2Seq) models use two particular RNNs (called Long-Short-Term-Memory, LSTM) in an *encoder-decoder architecture*. CNNs alike struggle to capture long-range dependencies However: quadratic complexity of transformers!
- text is not only sequential (order matters), but also structured: there is *context*!

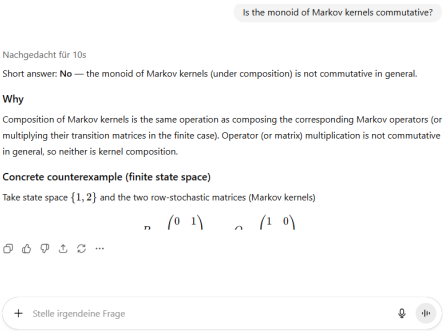


Fig. 1: ChatGPT5' UI.

The preprocessing step of *tokenization* uses a *vocabulary*, an *embedding* and positional encoding.

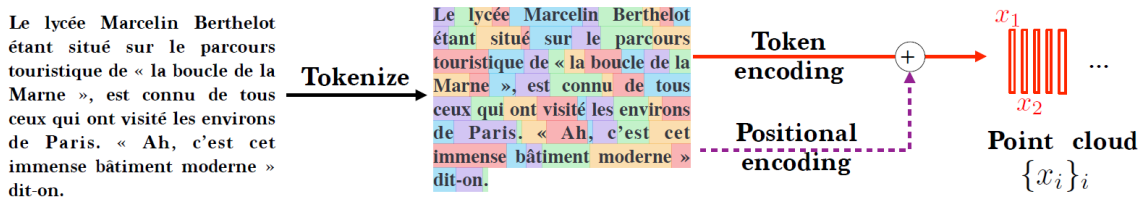


Fig. 2: Text is encoded into a point cloud. © G. Peyré

The points x_i are called (context) *tokens*.

Transformer architecture

The transformer architecture consists of stacked decoder-like *sublayers*, made up of (masked multi-head) *self-attention* + (token-wise) feed-forward neural networks aka MLP with residual aka skip connections + layer normalization

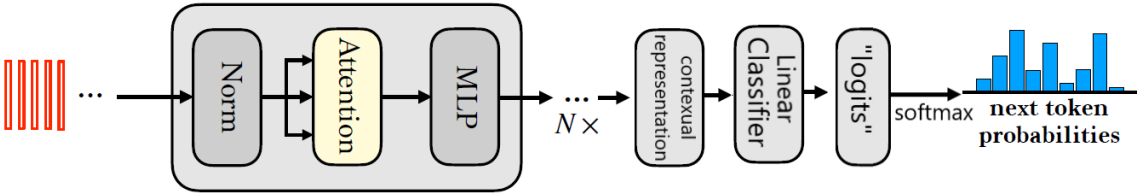


Fig. 3: Autoregressive decoder-only transformer architecture (GPT). Figure modified from Peyré 2024.

(Variant of encoder-decoder transformer "T5" [Vaswani et al. 2017])

Training via backpropagation: loss = predict next token given the previous ones

Generation: predict next token, add to rest ("context"), repeat ("autoregression")

$\text{Attention}(Q, K, V) \coloneqq \text{softmax}\left(\frac{QK^\mathsf{T}}{\sqrt{d_k}}\right) V$, where the "soft version" of argmax is

$$\text{softmax}\colon \mathbb{R}^d \to \text{int}(\Delta_{d-1}), \qquad x \mapsto \left(\frac{\exp(x_j)}{\sum_{\ell=1}^d \exp(x_\ell)}\right).$$

learned during training: query, key, value matrices Q, K, V

Per i-th token in the k-th layer with step siz $\Delta > 0$:

$$x_i^{(k+1)} = x_i^k + \Delta \sum_{j=1}^n \frac{\exp\left(\langle Qx_i^{(k)}, Kx_j^{(k)}\rangle\right)}{\sum_{\ell=1}^n \exp\left(\langle Qx_i^{(k)}, Kx_\ell^{(k)}\rangle\right)} V x_j^{(k)}.$$

$\langle Qx_i, Kx_j\rangle$ is the (non-symmetric!) *alignment score* between x_i and x_j. If alignment is high, then x_i is relevant for x_j. (Other alignment: $a(x, y) \coloneqq v^\mathsf{T} \tanh(Wx + Uy)$)

Table of Contents

From now on: ignore normalization & MLP.

Each transformer sublayer = one discrete time step. Letting the step size $\Delta \to 0$ (like in neural ODEs) we obtain (unmasked single-head) *self-attention*

$$\dot{x}_i(t) = \sum_{j=1}^{n} \underbrace{\frac{\exp\left(\langle Qx_i(t), Kx_j(t)\rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\langle Qx_i(t), Kx_\ell(t)\rangle\right)}}_{=:P_{i,j}(t)} Vx_j(t), \qquad i \in [n],\ t > 0. \tag{1}$$

$x_i(t)$ - tokens or *representations* at time t, $P_{i,j}$ is called the (stochastic) *attention matrix*

(1) is a simplified version of forward pass through the *trained* transformer with the same Q, K, V in all layers.

Mean field limit of infinitely many tokens:

$$\{x_i\}_{i=1}^N \longleftrightarrow \frac{1}{N}\sum_{i=1}^N \delta_{x_i} \xrightarrow{N\to\infty} \mu \in \mathcal{P}(\mathbb{R}^d).$$

On probability measures $\mathcal{P}(\mathbb{R}^d)$, the transformer ODE becomes the *transformer PDE*

$$\dot{\mu}_t = -\nabla \cdot \big(\mu_t \Gamma(\mu_t)\big), \quad t > 0, \qquad [\Gamma(\mu)](x) := \int_{\mathbb{R}^d} Vy \frac{\exp\left(\langle Qx, Ky\rangle\right)}{\int_{\mathbb{R}^d} \exp\left(\langle Qx, Kz\rangle\right) \mathrm{d}\mu(z)} \,\mathrm{d}\mu(y)$$

Γ is the *softmax attention mapping*. There are other forms of attention (Sinkhorn, L2, ...)

Table of Contents

Key results from [Geshkovski et al. 2023]

Time rescaling: $z_i(t) \coloneqq e^{-tV} x_i(t) \rightsquigarrow$ controls $\|z_i(t)\|$ for $t \to \infty$

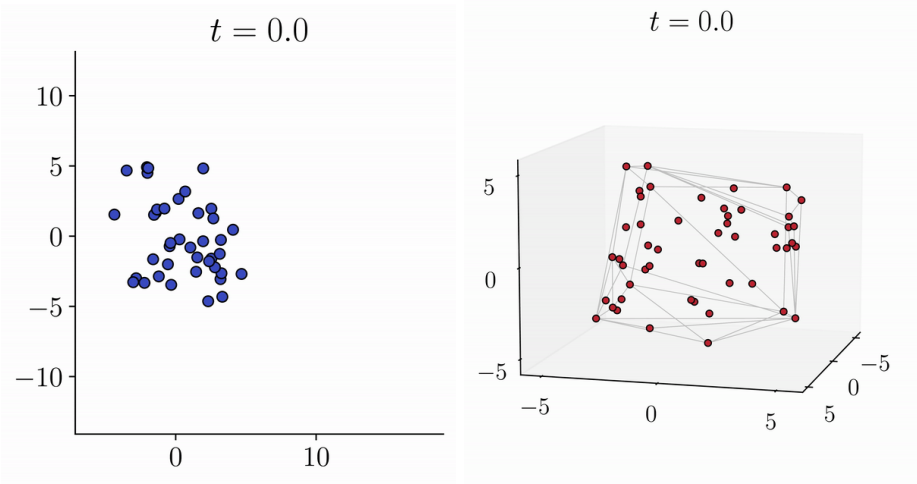
Mathematical surrogate for normalization, since usually $\|x_i(t)\| \in O(e^t)$ for $t \to \infty$.

Value	Key and query	Limit geometry
$V = \mathrm{I}_d$	$Q^\mathsf{T} K \succ 0$	vertices of convex polytope
$\lambda_{\max}(V) > 0$ simple	$\langle Q\varphi_1, K\varphi_1 \rangle > 0$	union of 3 parallel hyperplanes
V paranormal	$Q^\mathsf{T} K \succ 0$	polytope $\times$ subspaces
$V = -\mathrm{I}_d$	$Q^\mathsf{T} K = \mathrm{I}$	single cluster at origin

Table 1: Clustering taxonomy for rescaled dynamics (except last row).

last row $\leftrightarrow$ heat equation (for Sinkhorn attention) [Agarwal et al. 2024].

V paranormal $\iff \exists F, G \subset \mathbb{R}^d$ with $F \oplus G = \mathbb{R}^d$, $VF = F$, $VG = G$, $V|_F = \lambda \mathrm{I}$, $\rho(V|_G) < \lambda$ ($\rho =$ spectral radius). Also, $\varphi_1 \in \ker(V - \lambda_{\max}(V)\,\mathrm{I})$.



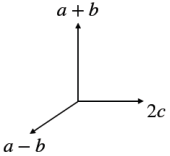
Leaders $\triangleq$ leading tokens capture the attention of all tokens (except one(?)) & carry the largest amount of information

Let $A \coloneqq K^\mathsf{T} Q$. For $\mu_0 \sim \mathcal{N}(\alpha_0, \Sigma_0)$ we have $\mu_t \sim \mathcal{N}(\alpha_t, \Sigma_t)$ with

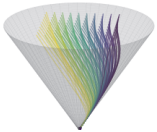
$$\dot{\Sigma}_t = 2\,\mathrm{Sym}(V\Sigma A\Sigma), \qquad \dot{\alpha} = V(\mathrm{I} + \Sigma A)\alpha_t$$

For single-head attention and Q, K, V constant in time:

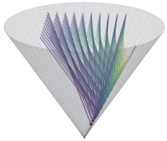
- Limiting points have low rank (under commutativity assumptions)
- Rank 1 is preserved
- Stationary points have rank 1 if $V = \mathrm{I}$ and $A = A^\mathsf{T}$.



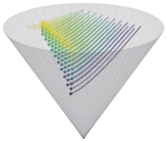
(a) Conv. to zero



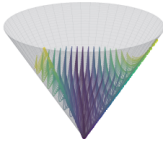
(b) Conv. to a line



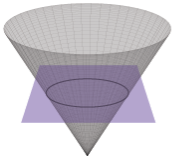
(c) Conv. to a plane



(d) Conv. to two lines



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



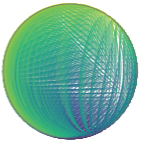
(a) Blow-up

SM att.

(b) Blow-up

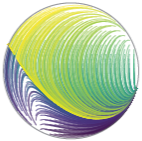
MH att.

(c) Div.



L2 att.

(d) Conv./blow-up



SM att.

Thank you for your *attention*!

References I

Agarwal, Medha et al. [2024]. "Iterated Schrödinger bridge approximation to Wasserstein Gradient Flows". arXiv preprint arXiv:2406.10823.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio [May 2015]. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *International Conference on Learning Representations (ICLR)*. Published as a conference paper. San Diego, CA, USA. URL: https://arxiv.org/abs/1409.0473.

Burger, Martin et al. [2025]. "Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization". In: *Philosophical Transactions A* 383.2298, p. 20240233.

Castin, Valérie et al. [2025]. "A unified perspective on the dynamics of deep transformers". arXiv preprint arXiv:2501.18322.

References II

Geshkovski, Borjan et al. [2023]. "The Emergence of Clusters in Self-Attention Dynamics". In: *Advances in Neural Information Processing Systems 36*. Ed. by H. Larochelle et al., pp. 57026–57037. DOI: 10.5555/3666122.3668615. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/b2b3e1d9840eba17ad9bbf073e009afe-Abstract-Conference.html.

Lu, Yiping et al. [2019]. "Understanding and improving transformer from a multi-particle dynamic system point of view". In: *Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.*

References III

Michel, Paul, Omer Levy, and Graham Neubig [2019]. “Are Sixteen Heads Really Better than One?” In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf.

Peyré, Gabriel [2024]. *Transformers are universal in context leaners*. Slides of a talk given at the conference "Learning and Optimization in Luminy".

Shalova, Anna and Mark Peletier [n.d.]. “Porous medium is the message: variational analysis of toy transformers”. Manuscript in preparation.

References IV

Vaswani, Ashish et al. [2017]. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vuckovic, James, Aristide Baratin, and Remi Tachet des Combes [2020]. “A mathematical theory of attention”. arXiv preprint arXiv:2007.02876.

The monoid structure of Markov kernels

Definition (Markov kernel)

Let E be a metric space such that $(E, \mathcal{E})$ is a measurable space, and $\mathcal{P}(E)$ the set of Borel probability measures on $(E, \mathcal{E})$. A *Markov kernel* is $(M(x, \cdot))_{x \in E} \subset \mathcal{P}(E)$ such that $x \mapsto M(x, A)$ is measurable for all $A \in \mathbb{E}$.

Example. By the disintegration theorem, one can "disintegrate" any coupling $\pi \in \mathcal{P}(E \times E)$ between $\mu, \nu \in \mathcal{P}(E)$ "into" a Markov kernel K with $\pi(A \times B) = \int_A K_\pi(x, B) \, \mathrm{d}\mu(x)$ for all $A, B \in \mathcal{E}$.

The set $\mathrm{MK}(E) \subset \mathcal{P}(E)^E$ of Markov kernels is a *(non-Abelian) monoid* (group without inverses) with identity $(\delta_x)_{x \in E}$ and the operation

$$\mathrm{MK}(E)^2 \to \mathrm{MK}(E), \quad (M, N) \mapsto MN(x, A) \coloneqq \int_E N(y, A) \, \mathrm{d}[M(x, \cdot)](y).$$

Action of $\mathrm{MK}(E)$ on $\mathcal{P}(E)$ and $\mathrm{Meas}(E)$ via duality

$\mathrm{MK}(E)$ acts (from the right) on $\mathcal{P}(E)$ by

$$\mathrm{MK}(E) \times \mathcal{P}(E) \to \mathcal{P}(E), \quad (\mu, M) \mapsto \mu M(A) \coloneqq \int_E M(x, A) \, \mathrm{d}\mu(x).$$

Example. We have $\delta_{x_0} M = M(x_0, \cdot)$.

Let $\langle \mu, f \rangle \coloneqq \int_E f(x) \, \mathrm{d}\mu(x)$ for $\mu \in \mathcal{P}(E)$ and any real-valued measurable function $f \in \mathrm{Meas}(E)$ on E.

The dual – via $\langle \mu M, f \rangle = \langle \mu, M(f) \rangle$ – action is

$$\mathrm{MK}(E) \times \mathrm{Meas}(E) \to \mathrm{Meas}(E), \quad (M, f) \mapsto Mf(x) \coloneqq \int_E f(y) \, \mathrm{d}[M(x, \cdot)](y).$$

Attention as Markov kernel

[Vuckovic, Baratin, and Combes 2020]

TODO

Multiple heads - Competing wisdom

Replace

$$\sum_{j=1}^{n} P_{i,j}(t) V x_j(t) \longleftrightarrow \sum_{h=1}^{H} \sum_{j=1}^{n} P_{i,j}^{(h)}(t) V^{(h)} x_j(t),$$

where $H =$ number of heads.

[Shalova and Peletier n.d.]: behavior of the mean field ODE vastly differs between one and two heads, if one injects noise and considers the dynamics on $\mathcal{P}(\mathbb{S}^d)$ (thus modelling LayerNorm)

Some empirical evidence however suggests that in practice one can often dramatically prune the number of heads [Michel, Levy, and Neubig 2019]

TODO: what is the rationale behind multiple heads?