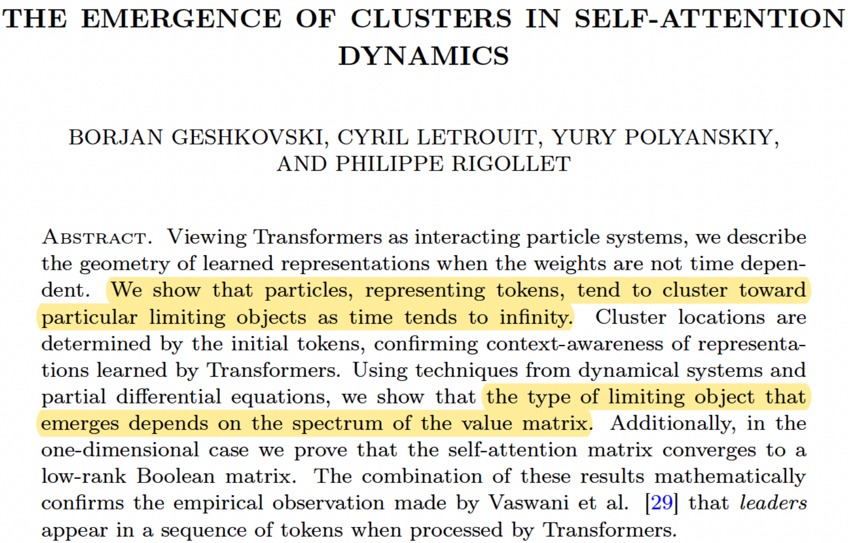
# Emergence of Clustering in Self-Attention

Anton Sugolov and Murdock Aubry

MAT1510: Theory and (or) Data Science

November 27, 2023

# Table of Contents



# THE EMERGENCE OF CLUSTERS IN SELF-ATTENTION DYNAMICS

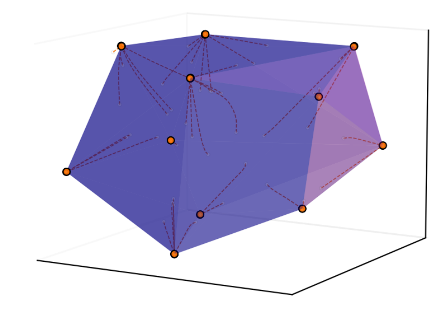
BORJAN GESHKOVSKI, CYRIL LETROUIT, YURY POLYANSKIY,
AND PHILIPPE RIGOLLET

Abstract. Viewing Transformers as interacting particle systems, we describe the geometry of learned representations when the weights are not time dependent. We show that particles, representing tokens, tend to cluster toward particular limiting objects as time tends to infinity. Cluster locations are determined by the initial tokens, confirming context-awareness of representations learned by Transformers. Using techniques from dynamical systems and partial differential equations, we show that the type of limiting object that emerges depends on the spectrum of the value matrix. Additionally, in the one-dimensional case we prove that the self-attention matrix converges to a low-rank Boolean matrix. The combination of these results mathematically confirms the empirical observation made by Vaswani et al. [29] that *leaders* appear in a sequence of tokens when processed by Transformers.

# Geshkovski et al.

The emergence of clusters in self-attention dynamics. B. Geshkovski, C. Letrouit, Y. Polyanskiy, and Philippe Rigollet. (2023).

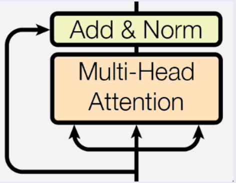
## Token flows



- View token embeddings as particles
- Self-attention defines ‘dynamics’ on particles
- Study clustering of geometric representations after repetition of dynamics

# Setting

## Dynamical framework



$$
\begin{aligned}
x(t+1) &= x(t) + f_\theta(x(t)) \\
&= x(t) + \dot{x}(t)
\end{aligned}
$$

$$
\dot{x}_i(t) = \sum_{j=1}^{n} \underbrace{P_{ij}(t)}_{\text{attention mat.}} \quad Vx_j(t)
$$

- Consider reptitions of self-attention
- Study change in embeddings (particles) as time variable
- Residual connection modifies input to self-attention matrix
- Defines ‘dynamics’ on particles

# Setting

## Self-attention matrix

$$P_{ij}(t) = \frac{e^{\langle Qx_i(t), Kx_j(t)\rangle}}{\sum_{\ell=1}^{n} e^{\langle Qx_i(t), Kx_\ell(t)\rangle}} \quad (i,j) \in [n]^2$$

- $P(t) = \operatorname{softmax}\left(Qx(t)(Kx(t))^T\right)$
- $x(t) = (x_1(t), \ldots, x_n(t)) \in \mathbb{R}^{n\times d}$ tokens
- $Q, K$ are usually denoted $W_Q, W_K$

- Iterated dynamics of self-attention matrix
- Under what conditions on $Q, K, V$ can we describe dynamics as $t \to \infty$?

# Geskovski et al. Low Rank Convergence

## Theorem

*For $x_i(t) \in \mathbb{R}$ and as $t \to \infty$, $P(t) \to P^*$ where $P^*$ is a low-rank boolean matrix.*

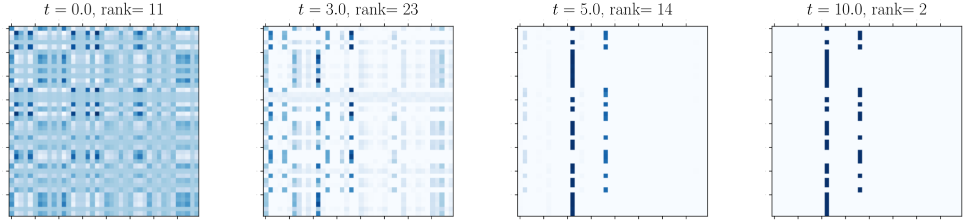


Figure: Example of Theorem 1 result when $Q = K = V = I$ with $n = 40$ tokens.

# Geskovski et al. Convex Polytopes

## Theorem

*When $V = I$ and $Q^T K > 0$ (positive matrix) then points flow to corners of convex polytope.*

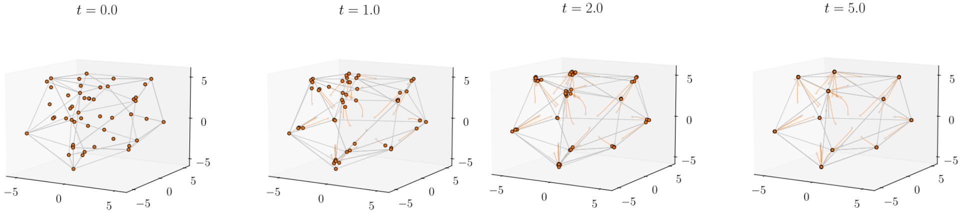


Figure: Example of Theorem 2 result when $Q = K = V = I \in \mathbb{R}^{3\times 3}$ with $n = 40$ tokens.

# Questions

- Similar dynamics occur for trained weights from real transformers?
- Effect of multihead self-attention?
- How does number of heads affect dynamics?
- How does token initialization affect dynamics?
- Does the number of tokens affect dynamics?

# Table of Contents

1. Summary

2. Experiments

# ALBERT Transformer Weights by Lan et. al (2020)

Repeated weight sharing between multi-head layers. Trained value matrix eigenvalue for head 5 is positive and real.

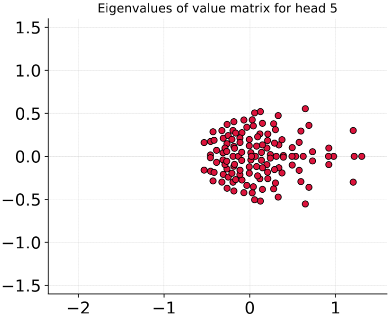


Figure: Top eigenvalue of $V$ is real.

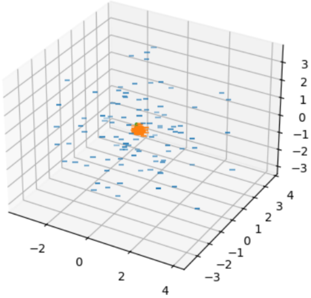
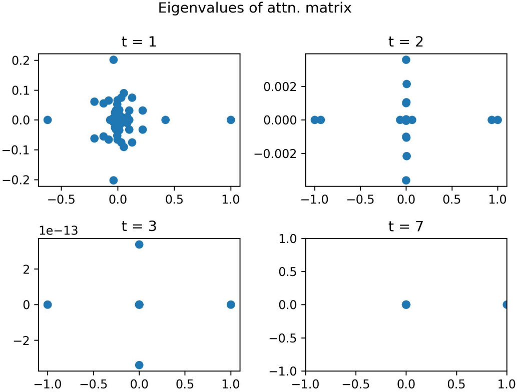
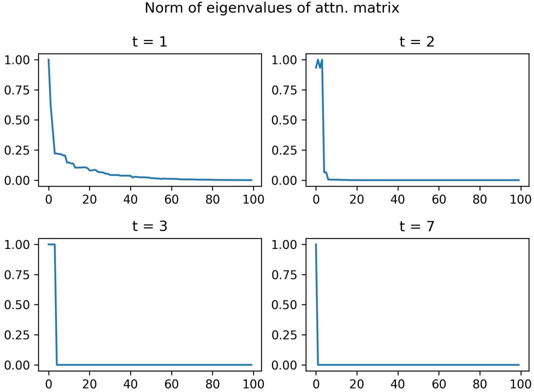


Figure: PCA of flows with head 5 weights shows clustering.

# ALBERT Transformer Weights by Lan et. al (2020)

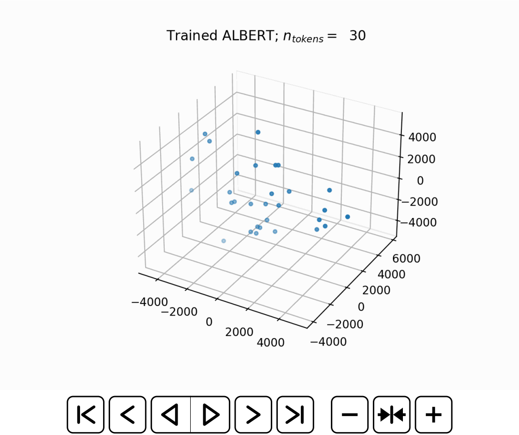


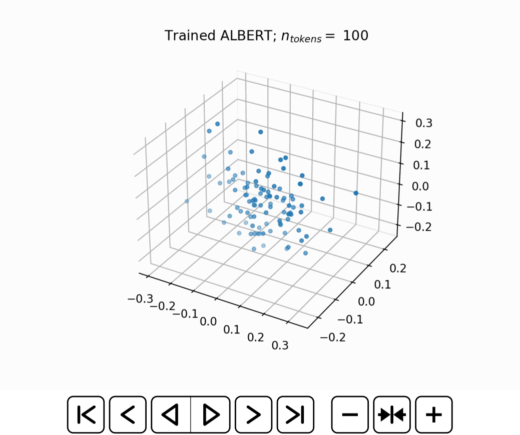
Eigenvalues of attn. matrix



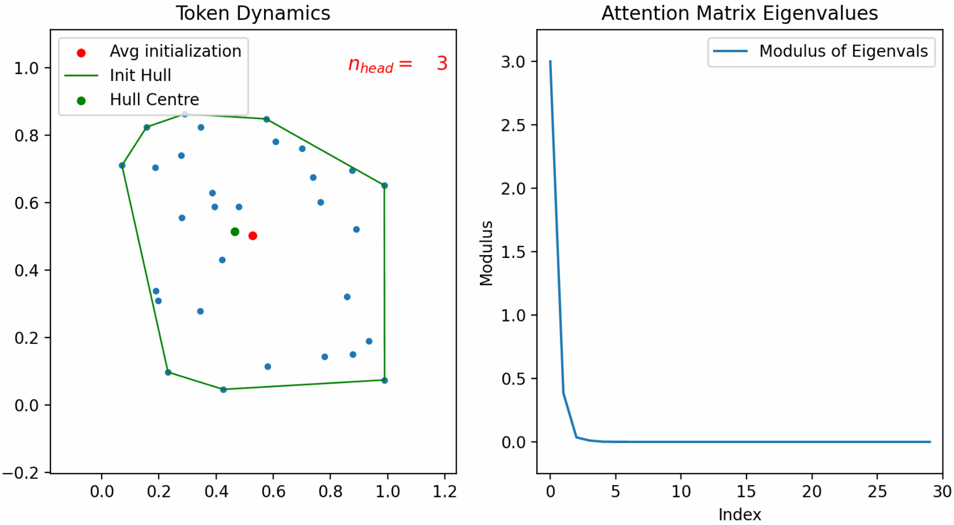
Norm of eigenvalues of attn. matrix

# ALBERT Transformer Weights by Lan et. al (2020). Multihead Implementation.

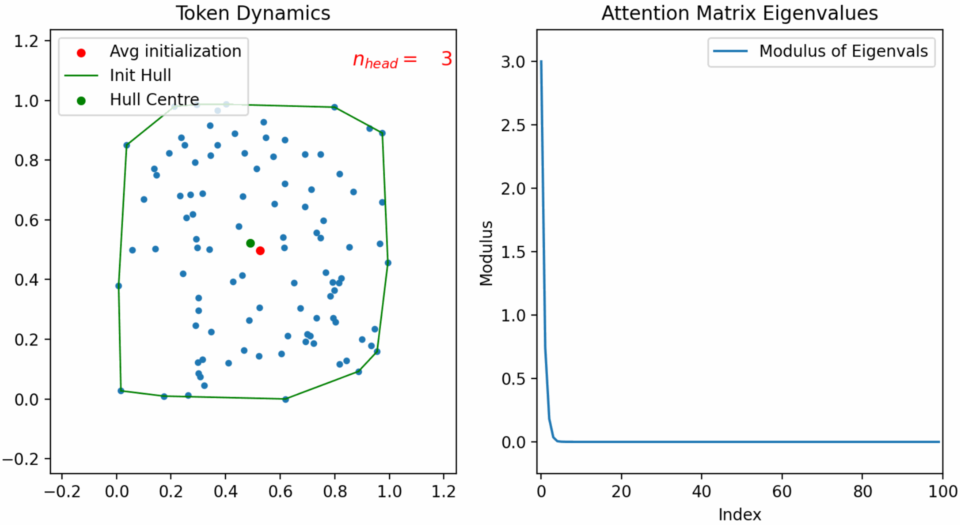




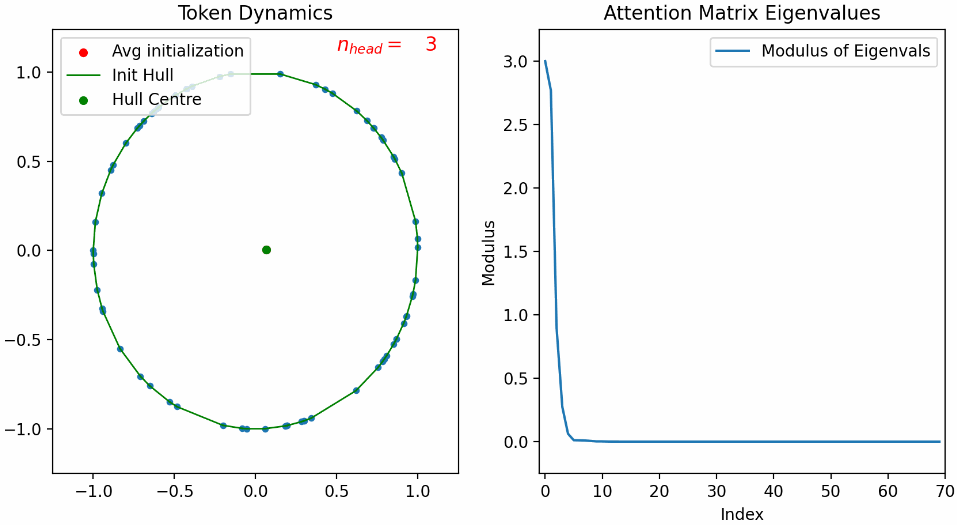
# Multihead Dynamics - Eigenvalue Analysis



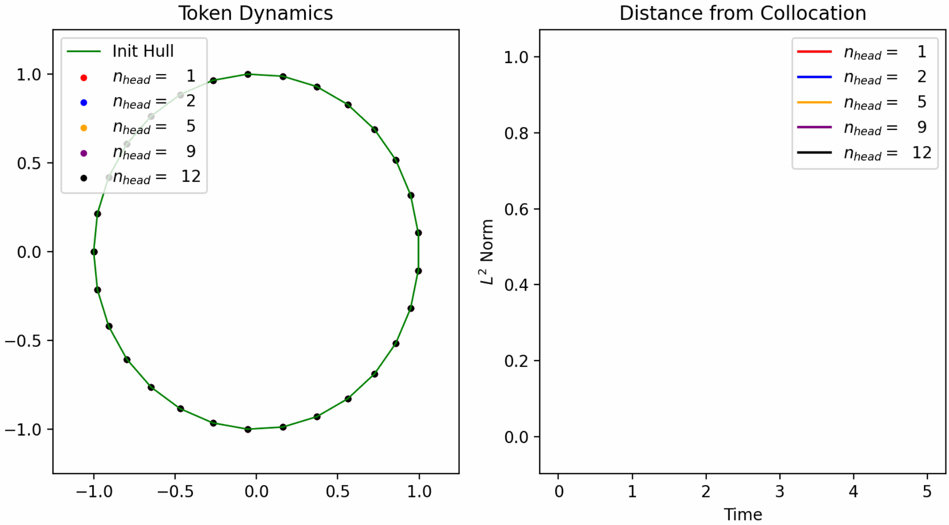
# Multihead Dynamics - Eigenvalue Analysis: More Tokens



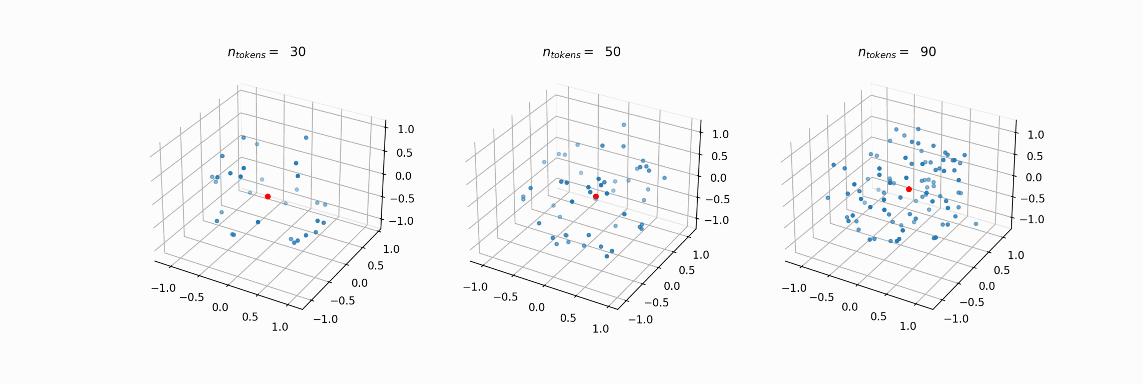
# Multihead Dynamics - Unit Circle Initialization



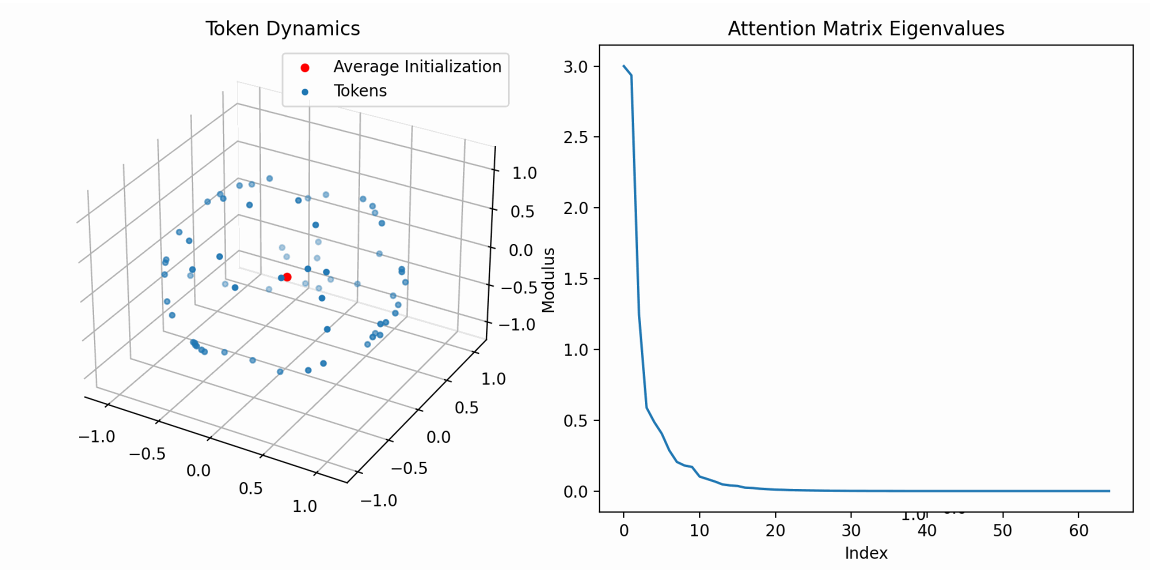
# Multihead Dynamics - Rate of Convergence



# Multihead Dynamics - 3 Dimensions



# Multihead Dynamics - Eigenvalue Analysis



# Next Experiments

- Test dynamics for on the weights of more trained transformer models. Iterpret dependence of dynamics on the model architecture.
- Observe dynamics when a real tokenized sentence is passed.
- Explore relationship between Neural collapse.
- Quantify the relationship between limiting structure and number of tokens and token initialization.
- Comparison between dynamics predicted by the Master equation.

# The Master Equation

- The dynamics of the tokens are governed by the discrete-time versions of

  $$\dot{X}(t) = f_\theta(X(t)) = P(X(t))X(t)$$

  where $P(t)$ is the attention matrix.
- Analogy with the time-dependent Master equation:

  $$\frac{d\vec{P}}{dt} = A(t)\vec{P}(t)$$

- If $A$ is approximately constant, then the solutions are given by

  $$\vec{P}(t) = \sum_{i=1}^{n} c_i e^{\lambda_i t}\vec{v}_t(t)$$

- This can act as a measure of the effect that the initialization of the tokens and weight matrices on the clustering patters and location of collocation points.