Self-Attention as a Multi-Agent System: a Dynamical Perspective on Transformers

Jan Peszek

Institute of Applied Mathematics and Mechanics, University of Warsaw, Warszawa, Poland
J.PESZEK@UW.EDU.PL

Abstract

Transformers [1], the architecture underlying modern large language models, (specifically its self-attention component) was recast in [2] as the multi-agent system

$$\dot{x}_k(t) = \mathrm{P}^{\perp}_{x_k(t)} \left(\sum_{j=1}^{n} \exp\left(\langle Q x_k(t), K x_j(t) \rangle\right) V x_j(t) \right), \quad k = 1, \dots, n, \tag{1}$$

where $x_k(t) \in \mathbb{S}^{d-1}$ represents the state of the k-th token after the t-th attention layer and $Q, K, V \in \mathbb{R}^{d \times d}$ are the trained query, key and value matrices and $\mathrm{P}^{\perp}_{x_k(t)}$ is the projection onto $\mathbb{S}^{d-1}$. The output of the model is determined by the long-time state of the system, making its asymptotic behavior a central object of study. The goal of this talk is to illustrate how methods from collective dynamics can help analyze large language models.

In this talk I present a dynamical systems perspective on a 2D linearized transformer architecture, where the evolution of tokens across attention layers can be reformulated as a Kuramoto-type interacting particle system. Using the Ott–Antonsen ansatz, the high-dimensional dynamics admits a low-dimensional reduction describing the evolution in a reduced vocabulary regime. I will show that this reduced system exhibits a rich range of behaviors depending on the trained parameters, including gradient-flow regimes leading to steady states and Hamiltonian regimes generating oscillatory dynamics. I will also discuss how these phenomena persist beyond the Ott–Antonsen manifold using the Watanabe–Strogatz transformation to analyze the stability of the full particle system.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[2] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, A mathematical perspective on transformers, *arXiv preprint arXiv:2312.10794*, 2023.