Diverse Dynamical Behaviors in Deep Linear Transformers

Thomas Maranzatto, University of Maryland, College Park
Jan Peszek, University of Warsaw

Mathematics of Transformers — 2nd edition, Oxford

In dimension two, we uncover a hidden low-dimensional structure in linear self-attention by reducing the dynamics to a second-harmonic Kuramoto-type model and applying Watanabe–Strogatz theory. This reduction reveals a rich range of behaviors: clustering, oscillations, and bifurcations.

Key references:

1. On the Diverse Dynamical Behaviors Arising in Deep Linear Transformers Garcia Trillos and Li (Wisconsin Madison), Riedl (Oxford), Teolis (Rice), TM, Akkoc, Ulukus (UMD, College Park) and JP (U. Warsaw).
2. *Attention is all you need*, Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin. The seminal work, Cited by 277191 (Google Scholar).
3. *A mathematical perspective on Transformers*, Geshkovski, Letrouit, Polyanskiy, Rigollet, Bulletin of the American Mathematical Society. Mathematical basics.
4. *Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention*, Katharopoulos, Vyas, Pappas, Fleuret, Proceedings of Machine Learning Research. Key from our perspective.

Self-Attention and Feed Forwad as a Dynamical System

Let $\mathbf{x}(k)$ be token representations at layer k. Then $\mathbf{x}(k+1)$ is derived as follows.

$$y_i = x_i(k) + \varepsilon \frac{1}{Z_i(k)} \sum_{j=1}^{N} \exp\left(\langle Qx_i(k), Kx_j(k)\rangle\right) Vx_j(k), \quad Z_i(k) := \sum_{l=1}^{N} \exp\left(\langle Qx_i(k), Kx_l(k)\rangle\right),$$

$$z_i = y_i + \varepsilon(W_2\sigma(W_1 y_i + b_1) + b_2) =: y_i + \varepsilon F(y_i),$$

$$x_i(k+1) = \frac{z_i}{\|z_i\|}$$

The matrices Q, K, V, W_1, W_2 and the vectors b_1, b_2 contain trained parameters. In modern large language models, such matrices and vectors may contain billions, or even trillions, of trained parameters.

A mathematician looks at it and sees:

$$\dot{x}_i = \mathbf{P}^{\perp}_{x_i}\left(\frac{1}{Z_i}\sum_{j=1}^{N}\exp\left(\langle Qx_i, Kx_j\rangle\right) Vx_j + F(x_i)\right), \qquad k = 1, \ldots, N.$$

A Mathematical Perspective on Transformers.

The Goal

$$\dot{x}_i = \mathbf{P}^{\perp}_{x_i}\left(\frac{1}{Z_i}\sum_{j=1}^{N}\exp\left(\langle Qx_i, Kx_j\rangle\right)Vx_j\right), \qquad k = 1,\dots,N, \quad \mathbf{P}^{\perp}_{x}y = y - \langle x, y\rangle x$$

The goal is to study large-time behavior of Self-Attention dynamics because it is what dictates the output of the model.

- Matrices Q, K, V contain billions of parameters and it is difficult to justify reducing the problem to mathematically meaningful cases.
- Instead, we rely on reducing the admissible initial data to certain invariant manifolds to discover possible dynamics for diverse parameters.
- This can be interpreted as follows: we aim to keep the parameters general, but we reduce the admissible vocabulary, while still obtaining extraordinary diversity of dynamics.
- Classification of large-time behavior of the reduced dynamics.

Self-Attention and Kuramoto Oscillators

Take $d = 2$ and $K = 0$, $V = I$. Then

$$Z_i = N, \qquad \frac{1}{Z_i}\sum_{j=1}^{N} \exp\left(\langle Qx_i, Kx_j\rangle\right) Vx_j = \frac{1}{N}\sum_{j=1}^{N} x_j.$$

Thus for $x_i = e^{i\theta_i}$

$$i\dot{\theta}_i e^{i\theta_i} = \dot{x}_i = \mathbf{P}^{\perp}_{x_i}\left(\frac{1}{N}\sum_{j=1}^{N} x_j\right) = \frac{1}{N}\sum_{j=1}^{N}(x_j - \langle x_i, x_j\rangle x_i)$$

$$= \frac{1}{N}\sum_{j=1}^{N}(e^{i\theta_j} - \cos(\theta_i - \theta_j)e^{i\theta_i}) = e^{i\theta_i}\frac{1}{N}\sum_{j=1}^{N}\underbrace{(e^{i(\theta_j-\theta_i)} - \cos(\theta_i - \theta_j))}_{i\sin(\theta_j-\theta_i)}.$$

Dividing both sides by $ie^{i\theta_i}$ we obtain the Kuramoto model

$$\dot{\theta}_i = \frac{1}{N}\sum_{j=1}^{N}\sin(\theta_j - \theta_i), \qquad i = 1, \ldots, N.$$

Which techniques known for the Kuramoto model may be of use?

Watanabe–Strogatz Transform

Consider a generic ℓ-th harmonic coupling

$$\dot{\theta}_i = \omega(t) + \mathrm{Im}\left(H(t)e^{-i\ell\theta_i}\right), \qquad i = 1, \ldots, N. \tag{1}$$

Kuramoto: $\ell = 1$, $H(t) = \frac{1}{N}\sum_{j=1}^{N} e^{i\theta_j}$, $\omega(t)$ is the common natural frequency. Applying the Möbius transform

$$e^{i\ell\theta_i} = \frac{\zeta + e^{i(\psi_i+\Phi)}}{1+\overline{\zeta}e^{i(\psi_i+\Phi)}}, \qquad \zeta \in \mathbb{D}, \quad \Phi \in \mathbb{R}, \quad \psi_i = \mathrm{const}.$$

Eq. (1) is equivalent to

$$\dot{\zeta} = i\ell\omega(t)\zeta + \frac{\ell}{2}\left(H(t) - \overline{H(t)}\zeta^2\right),$$

$$\dot{\Phi} = \ell\omega(t) + \ell\,\mathrm{Im}\left(\overline{H(t)}\zeta\right),$$

$$\dot{\psi}_i = 0, \qquad i = 1, \ldots, N.$$

For Kuramoto we get

$$\dot{\zeta} = \frac{1}{2}\left(R_1 - \overline{R_1}\zeta^2\right), \quad \dot{\Phi} = \mathrm{Im}\left(\overline{R_1}\zeta\right),$$

$$R_1 := \frac{1}{N}\sum_{j=1}^{N} e^{i\theta_j} = \frac{1}{N}\sum_{j=1}^{N} \frac{\zeta + e^{i(\psi_j+\Phi)}}{1+\overline{\zeta}e^{i(\psi_j+\Phi)}}.$$

Watanabe–Strogatz Transform

Consider a generic ℓ-th harmonic coupling

$$\dot{\theta}_i = \omega(t) + \operatorname{Im}\left(H(t)e^{-i\ell\theta_i}\right), \qquad i = 1, \ldots, N. \tag{1}$$

Kuramoto: $\ell = 1$, $H(t) = \frac{1}{N}\sum_{j=1}^{N} e^{i\theta_j}$, $\omega(t)$ is the common natural frequency. Applying the Möbius transform

$$e^{i\ell\theta_i} = \frac{\zeta + e^{i(\psi_i+\Phi)}}{1+\overline{\zeta}e^{i(\psi_i+\Phi)}}, \qquad \zeta \in \mathbb{D}, \quad \Phi \in \mathbb{R}, \quad \psi_i = \mathrm{const}.$$

We reduce a system of N ODEs to 3 ODEs with N constants of motion. However, this gives little computational gain, since the N constants still have to be accounted for at each time. The reduction becomes genuinely useful only on special invariant manifolds.

For Kuramoto we get

$$\dot{\zeta} = \frac{1}{2}\left(R_1 - \overline{R_1}\zeta^2\right), \quad \dot{\Phi} = \operatorname{Im}\left(\overline{R_1}\zeta\right),$$

$$R_1 := \frac{1}{N}\sum_{j=1}^{N} e^{i\theta_j} = \frac{1}{N}\sum_{j=1}^{N} \frac{\zeta + e^{i(\psi_j+\Phi)}}{1+\overline{\zeta}e^{i(\psi_j+\Phi)}}.$$

Linear Self-Attention Model

$$\dot{x}_i = \mathbf{P}^{\perp}_{x_i}\left(\sum_{j=1}^{N} h\left(\langle Qx_i, Kx_j\rangle\right) Vx_j\right), \qquad k = 1, \ldots, N, \quad \mathbf{P}^{\perp}_{x} y = y - \langle x, y\rangle x$$

Goal: Can we find functions h, for which SA dynamics is a harmonic coupling for all Q, K, V?

Problems:

- Minor: Large dimension of the ambient space is a problem: let's stick to $d = 2$ although $d = 1000$ can be pursued via existing extensions of the WS framework.
- Major: WS works only when the harmonic coupling does not mix harmonics.

Linear Self-Attention Model

$$\dot{x}_i = \mathbf{P}^{\perp}_{x_i}\left(\sum_{j=1}^{N} h\left(\langle Qx_i, Kx_j\rangle\right) Vx_j\right), \qquad k = 1,\ldots,N, \quad \mathbf{P}^{\perp}_{x} y = y - \langle x, y\rangle x$$

Goal: Can we find functions h, for which SA dynamics is a harmonic coupling for all Q, K, V?

Linear Self-Attention

$$\dot{x}_i = \mathbf{P}^{\perp}_{x_i}\left(\sum_{j=1}^{N}\langle Qx_i, Kx_j\rangle Vx_j\right), \quad \dot{\theta}_i(t) = C(R_2(t)) + \operatorname{Im}\left(B(R_2(t))e^{-i2\theta_i}\right), \quad x_i = e^{i\theta_i}$$

- We have found one – linear h. Miraculously, this is a second harmonic coupling with

$$R_2 = \frac{1}{N}\sum_{j=1}^{N} e^{i2\theta_j}.$$

- The model has been established as a fast variant of transformer (LSA is $O(N)$ compared to SA's $O(N^2)$) (*Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention*).

LSA Model in WS terms

Watanabe-Strogatz transform let's us change the system

$$\dot{x}_i = \mathbf{P}^{\perp}_{x_i}\left(\sum_{j=1}^{N}\langle Qx_i, Kx_j\rangle Vx_j\right), \quad x_i = e^{i\theta},\ \theta \sim f_t,\ \psi \sim \nu, \quad R_2(t) = \int_0^{2\pi} e^{i2\theta}\, df_t(\theta) = \int_0^{2\pi} \frac{\zeta + e^{i(\psi+\Phi)}}{1+\overline{\zeta}e^{i(\psi+\Phi)}}\, d\nu(\psi)$$

into

$$\underbrace{\frac{d}{dt}(\zeta,\Phi) = \mathrm{F}(\zeta, R_2)}_{\text{Full (LSA) system}} = \mathrm{F}(\zeta,\zeta) + (\mathrm{F}(\zeta,R_2) - \mathrm{F}(\zeta,\zeta)),$$

$$\underbrace{\frac{d}{dt}(\zeta,\Phi) = \mathrm{F}(\zeta,\zeta)}_{\text{Autonomus system of 3 ODEs}} \qquad + \qquad \underbrace{R}_{\text{Remainder to control}} \qquad .$$

Low-dimensional Invariant Manifold

The system in WS terms reads

$$\underbrace{\frac{d}{dt}(\zeta, \Phi) = \mathrm{F}(\zeta, R_2)}_{\text{Full (LSA) system}} = \mathrm{F}(\zeta, \zeta) + \underbrace{(\mathrm{F}(\zeta, R_2) - \mathrm{F}(\zeta, \zeta))}_{R},$$

- What is the difference between ζ and R_2? What if $\zeta = R_2$?

Let $g_t(\theta) = \frac{1}{2}(f_t(\frac{\theta}{2}) + f_t(\frac{\theta}{2} + \pi))$. The following are equivalent:

$\zeta(t) = R_2(t)$	$\theta \sim g_t(\theta) = \frac{1}{2\pi} \frac{1-\lvert\zeta(t)\rvert^2}{\lvert e^{i(\theta - \Phi(t))} - \zeta(t)\rvert^2}$	Fourier modes of g_t are $(R_2(t))^n$	$\psi \sim \mathrm{Unif}([0, 2\pi])$

If we accept that they are equivalent then they are clearly independent of t since the last one is. This is the so called Ott-Antonsen ansatz that establishes the invariant manifold. The key point is that for all t we have $\zeta(t) = R_2(t)$.

Low-dimensional Invariant Manifold – Stability

$$\frac{d}{dt}(\zeta,\Phi) = \mathrm{F}(\zeta,\zeta) + R, \quad R_2(t) = \int_0^{2\pi} e^{i2\theta} df_t(\theta) = \int_0^{2\pi} \frac{\zeta + e^{i(\psi+\Phi)}}{1+\bar{\zeta} e^{i(\psi+\Phi)}} d\nu(\psi) \tag{2}$$

Let $\zeta(t) = |\zeta(t)|e^{i\eta(t)}$ solve (2), and let $\psi \sim \nu$ be a probability measure on $[0, 2\pi)$. Then

$$R \lesssim C_\nu(|\zeta(t)|)(1-|\zeta(t)|), \quad C_\nu(|\zeta|) := \sup_{\chi\in[0,2\pi)} \int_0^{2\pi} \frac{2}{|1+|\zeta|e^{i(\psi+\chi)}|} \nu(d\varphi), \tag{3}$$

$$R \lesssim \left(1 + \frac{2}{1-|\zeta(t)|}\right) d_{BL}(\nu, \mathrm{Uniff}), \quad \text{Unifform } \nu \text{ yields the invariant manifoold}. \tag{4}$$

Conclusion: It is relatively easy to grasp the dynamics on the OA manifold. To control the general solutions we do as follows:

- In the low synchonization regime $|\zeta| < 1$ we use (4) with $d_{BL}(\nu, \mathrm{Unif}) < \varepsilon$, where ε depends on the system parameters.
- As ζ and R_2 get close to 1, the general solution automatically begins to behave like the OA solution by (3).

Case Studies: Prologue

Recall that on the OA manifold we have $\zeta(t) = R_2(t)$ and hence,

$$\frac{d}{dt}\zeta(t) = \frac{d}{dt}R_2(t) = F(R_2, R_2)$$

Thus, understanding the long-time limit of dynamics on the OA manifold reduces to understanding the behavior of the complex ODE $\dot{R}_2 = F(R_2, R_2)$.

Case Studies: Prologue

Recall that on the OA manifold we have $\zeta(t) = R_2(t)$ and hence,

$$\frac{d}{dt}\zeta(t) = \frac{d}{dt}R_2(t) = F(R_2, R_2)$$

Thus, understanding the long-time limit of dynamics on the OA manifold reduces to understanding the behavior of the complex ODE $\dot{R}_2 = F(R_2, R_2)$.

Challenges:

- The reduced system is a (2+8)-dimensional object, which is still too large for a general classification of long-time behavior
- The function $F(R_2, R_2)$ is very complicated (cubic in R_2 with coefficients involving all combinations $A_{ij}V_{k\ell}$)

Case Studies: Prologue

Recall that on the OA manifold we have $\zeta(t) = R_2(t)$ and hence,

$$\frac{d}{dt}\zeta(t) = \frac{d}{dt}R_2(t) = F(R_2, R_2)$$

Thus, understanding the long-time limit of dynamics on the OA manifold reduces to understanding the behavior of the complex ODE $\dot{R}_2 = F(R_2, R_2)$.

Challenges:

- The reduced system is a (2+8)-dimensional object, which is still too large for a general classification of long-time behavior
- The function $F(R_2, R_2)$ is very complicated (cubic in R_2 with coefficients involving all combinations $A_{ij}V_{k\ell}$)

Solutions:

- Perform case studies on subsets of the parameter space to highlight the diversity of dynamics, even on the OA manifold
- Use computer algebra software to reduce F in the paramter regimes we care about.

Case Studies: $V = I$, A Arbitrary

$$A := \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad V := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Induces the dynamics:

$$\dot{R}_2 = \frac{1}{4}\left(1 - |R_2|^2\right)\Bigg(\left(a_+^d + ia_-^o\right) R_2 + \left(a_-^d + ia_+^o\right)\Bigg)$$

Case Studies: $V = I$, A Arbitrary

$$A := \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad V := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Induces the dynamics:

$$\dot{R}_2 = \frac{1}{4}\left(1 - |R_2|^2\right)\left(\left(a_+^d + ia_-^o\right) R_2 + \left(a_-^d + ia_+^o\right)\right)$$

- Full Clustering If $\mathrm{tr}(A) > 0$ or $\det(A) \leq 0$, then the OA dynamics satisfy $(\rho, \phi) \to (1, \phi_\infty)$
- Partial Clustering if $A + A^\top \prec 0$, then the OA dynamics satisfy $(\rho, \phi) \to (\rho_\infty, \phi_\infty)$ with $\rho_\infty < 1$.

Case Studies: $V = I$, A Arbitrary

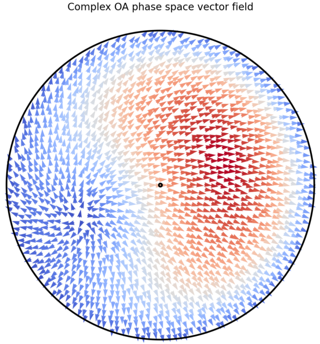


Figure: OA phase portrait for full clustering.

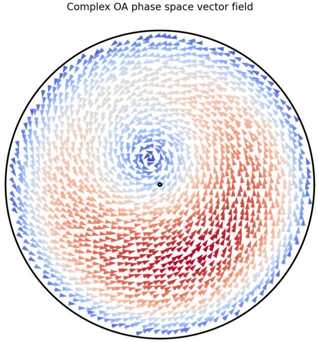


Figure: OA phase portrait for partial clustering.

Case Studies: $A = I$, V Symmetric

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad V := \begin{pmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{pmatrix}$$

Induces the dynamics:

$$\begin{aligned} \dot{R}_2 = \frac{1}{4}\Big(& v_+^d \left(1 - |R_2|^2\right) R_2 \\ & + 2\left(-v_+^o + i v_-^d\right) (R_2)^2 \\ & + \left(v_+^o + i v_-^d\right) \left(1 + |R_2|^2\right)\Big), \end{aligned}$$

Case Studies: $A = I$, V Symmetric

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad V := \begin{pmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{pmatrix}$$

Induces the dynamics:

$$\begin{aligned} \dot{R}_2 = \frac{1}{4}\Big(& v_+^d \left(1 - |R_2|^2\right) R_2 \\ & + 2\left(-v_+^o + i v_-^d\right) (R_2)^2 \\ & + \left(v_+^o + i v_-^d\right) \left(1 + |R_2|^2\right)\Big), \end{aligned}$$

- Full Clustering If $\lambda_{\max}(V) \geq 0$ then the OA dynamics satisfy $(\rho, \phi) \to (1, \phi_\infty)$
- Partial Clustering if $\lambda_{\max}(V) < 0$, then the OA dynamics satisfy $(\rho, \phi) \to (\rho_\infty, \phi_\infty)$ with $\rho_\infty < 1$.

Case Studies: $A = I$, V Symmetric

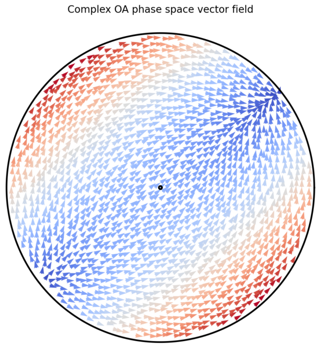


Figure: OA phase portrait for full clustering.

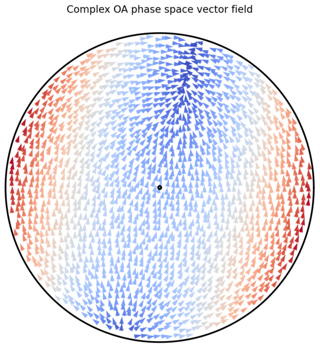


Figure: OA phase portrait for partial clustering.

Case Studies: Non-Commuting A and V

$$A := \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix},$$

$$V := \begin{pmatrix} v_{11} & v_{12} \\ -\frac{a_{22}}{a_{11}} v_{12} & \frac{a_{11}}{a_{22}} v_{11} \end{pmatrix}$$

Induces the dynamics:

$$\dot{\phi} = K(\phi) + (\rho - 1) Q(\rho, \phi)$$

The matrices do not commute if $a_{22} \neq a_{11}$ and $v_{12} \neq 0$.

Case Studies: Non-Commuting A and V

$$A := \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix},$$

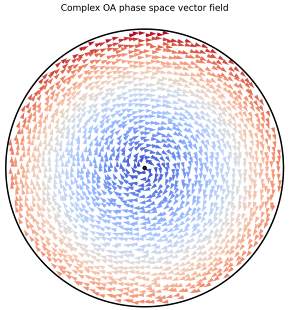
$$V := \begin{pmatrix} v_{11} & v_{12} \\ -\frac{a_{22}}{a_{11}} v_{12} & \frac{a_{11}}{a_{22}} v_{11} \end{pmatrix}$$

Induces the dynamics:

$$\dot{\phi} = K(\phi) + (\rho - 1) Q(\rho, \phi)$$

The matrices do not commute if $a_{22} \neq a_{11}$ and $v_{12} \neq 0$.

Cyclic behavior at clustering: If $|K(\phi)| > 0$ and $|Q(\rho, \phi)| < \infty$, then $\rho \to 1$ while ϕ never converges



Case Studies: Hamiltonian Bifurcation

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$V := \begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & -v_{11} \end{pmatrix}$$

Induces the dynamics:

$$\dot{R}_2 = \frac{v_{11}}{2}\left(1 + |R_2|^2 - 2(R_2)^2\right) - \frac{iv_{12}}{2}\left(|R_2|^2 + 3\right) R_2$$

Case Studies: Hamiltonian Bifurcation

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$V := \begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & -v_{11} \end{pmatrix}$$

Induces the dynamics:

$$\dot{R}_2 = \frac{v_{11}}{2}\left(1 + |R_2|^2 - 2(R_2)^2\right) - \frac{iv_{12}}{2}\left(|R_2|^2 + 3\right) R_2$$

Saddle-Node bifurcation in the parameters v_{11} and v_{12}.

- Periodic Trajectories: If $|v_{11}| < |v_{12}|$ then ρ is always bounded away from $\{0, 1\}$, and the trajectory $(\rho(t), \phi(t))$ forms a closed periodic orbit in $(0, 1) \times \mathbb{T}^1$
- Full Clustering: If $|v_{11}| \geq |v_{12}|$, then the OA dynamics satisfy $(\rho, \phi) \to (1, \phi_\infty)$.

Case Studies: Hamiltonian Bifurcation

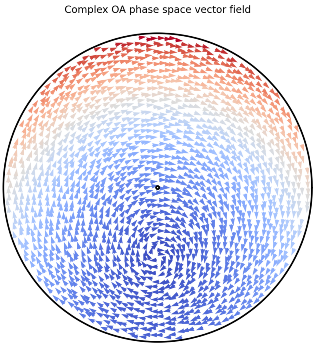


Figure: Cyclic Regime.

Case Studies: Hamiltonian Bifurcation

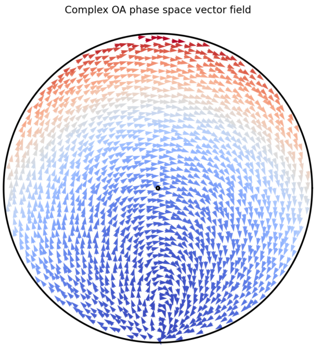


Figure: Critical Point

Case Studies: Hamiltonian Bifurcation

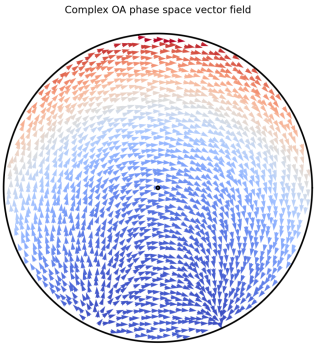


Figure: Clustering Regime.

Conclusion

- We have connected techniques from the Kuramoto model literature to self-attention dynamics via the WS transform and the OA ansatz.

Conclusion

- We have connected techniques from the Kuramoto model literature to self-attention dynamics via the WS transform and the OA ansatz.
- We have identified a low-dimensional invariant manifold for the linear self-attention dynamics, and demonstrated stability conditions off the manifold.

Conclusion

- We have connected techniques from the Kuramoto model literature to self-attention dynamics via the WS transform and the OA ansatz.
- We have identified a low-dimensional invariant manifold for the linear self-attention dynamics, and demonstrated stability conditions off the manifold.
- Our reduction allows for rapid identification of convergence, clustering, and cycling in the long-time limit via phase portraits and analysis of the resulting ODE.

Conclusion

- We have connected techniques from the Kuramoto model literature to self-attention dynamics via the WS transform and the OA ansatz.
- We have identified a low-dimensional invariant manifold for the linear self-attention dynamics, and demonstrated stability conditions off the manifold.
- Our reduction allows for rapid identification of convergence, clustering, and cycling in the long-time limit via phase portraits and analysis of the resulting ODE.
- In our case studies we have found Hamiltonian-like systems which exhibit bifurcations in the parameter matrices A and V.