Recurrent Self-Attention Dynamics: An Energy-Agnostic Perspective from Jacobians

Akiyoshi Tomihari & Ryo Karakida

Artificial Intelligence Research Center, AIST, Japan

AIST

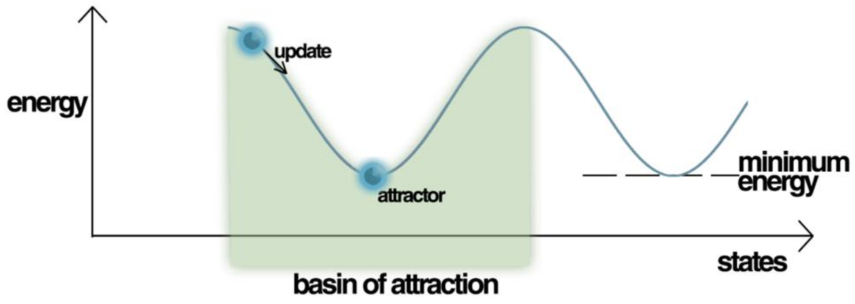
NEURAL INFORMATION PROCESSING SYSTEMS

Theoretical Analysis of Self-attention

- Self-attention (SA) is a core mechanism in Transformers
  - Transformers are the backbone of modern language models

$$\mathrm{SA}_h(\boldsymbol{X}) := \mathrm{softmax}(\beta \boldsymbol{X}\boldsymbol{W}_h^Q \boldsymbol{W}_h^{K\top} \boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{W}_h^V$$

- Theoretical analysis of SA is challenging
  - Mainly due to the Softmax function
- A major approach involves energy-based analysis
  - SA can be viewed as minimizing an implicit or explicit energy function



Copied from An Energy-Based Perspective on Attention Mechanisms in Transformers

Energy-based analysis of SA

- SA can be viewed as minimizing an energy function [Geshkovski et al., 2023, 2024, Bruno et al., 2025]
- Often formulated using continuous equations and a particle-based interpretation of tokens

$$\dot{\boldsymbol{X}} = \mathrm{Proj}_{\boldsymbol{X}} \left(\mathrm{softmax}(\beta \boldsymbol{X} \boldsymbol{W}^Q \boldsymbol{W}^{K\top} \boldsymbol{X}^\top) \boldsymbol{X} \boldsymbol{W}^V \right)$$

- To define an energy function, constraints are typically assumed
  - Weight constraints and single-head constraint

$$\boldsymbol{W}^Q \boldsymbol{W}^{K\top} = \boldsymbol{W}^V = \boldsymbol{W}^{V\top} \text{ or } \boldsymbol{W}^Q = \boldsymbol{W}^K = \boldsymbol{W}^V = \boldsymbol{I}_D$$

- However, these constraints deviate from practical scenarios

Our approach

1. Relax constraints in energy-based formulations
   - remove symmetry constraints on weights
   - extend beyond single-head SA
2. Apply Jacobian-based analysis to SA
   - Captures linear stability without requiring an energy function
   - Compute Lyapunov exponents from Jacobians

→ **Our study reveals Jacobians and Lyapunov exponents as fundamental tools for realistic SA.**

Revisit Energy-based analysis

What is a more realistic setting for energy guarantees?

- Remove symmetry constraints on weights (W^Q, W^K)

Proposition 4.1. *Consider the continuous-time dynamics for single-head SA equipped with projection* (3). *The energy function*

$$E_{single}(\boldsymbol{X}) = -\sum_{i,j} \exp\left(\beta \boldsymbol{X}_{[i,:]}^\top \boldsymbol{W}^Q \boldsymbol{W}^{K\top} \boldsymbol{X}_{[j,:]}\right) \tag{10}$$

is monotonically decreasing as $dE_{single}(\boldsymbol{X})/dt \leq 0$ *under the condition:*

$$\boldsymbol{W}^V = (\boldsymbol{W}^{Q\top} \boldsymbol{W}^K + \boldsymbol{W}^Q \boldsymbol{W}^{K\top})/2. \tag{11}$$

- Extend beyond single-head self-attention

Proposition 4.2. *Consider the continuous-time dynamics for multi-head SA without projection:* $d\boldsymbol{X}/dt = \sum_{h=1}^{H} SA_h(\boldsymbol{X})$. *An energy function*

$$E_{multi}(\boldsymbol{X}) = -\sum_{h} \sum_{i,j} \exp\left(\beta \boldsymbol{X}_{[i,:]}^\top \boldsymbol{W}_h^Q \boldsymbol{W}_h^{K\top} \boldsymbol{X}_{[j,:]}\right) \tag{12}$$

is monotonically decreasing as $dE_{multi}(\boldsymbol{X})/dt \leq 0$ *under the condition*

$$\boldsymbol{W}_h^V = (\boldsymbol{W}_h^{Q\top} \boldsymbol{W}_h^K + \boldsymbol{W}_h^Q \boldsymbol{W}_h^{K\top})/2, \;\; \boldsymbol{W}_h^Q \boldsymbol{W}_h^{K\top} = \boldsymbol{U}_{1,h} \boldsymbol{U}_{2,h}^\top, \tag{13}$$

where $\boldsymbol{U}_{1(2),h} \in \mathbb{R}^{D \times D/(2H)}$ ($h \in [1, H]$) *satisfies the orthogonality condition* $\boldsymbol{U}_{k,h}^\top \boldsymbol{U}_{k',h'} = \delta_{hh'} \delta_{kk'} \boldsymbol{I}_{D/(2H)}$.

Apply Jacobian-based analysis to SA

- More general than energy-based analysis:
  - captures **linear stability**
  - detects **non-stationary dynamics** more easily
- **Normalization** plays a crucial role in discrete systems
  - no counterpart in continuous-time dynamics

Proposition 5.1. *Suppose that, in the update of ItrSA* (9)*, the input to the normalization layer satisfies* $\|\boldsymbol{X}_{[i,:]} + \eta \Delta \boldsymbol{X}_{[i,:]}\| \geq R$ *for all* $i \in [1, S]$*. Then, the spectral norm of the Jacobian satisfies*

$$\left\| \frac{\partial \operatorname{RMSNorm}(\boldsymbol{X} + \eta \Delta \boldsymbol{X})}{\partial \boldsymbol{X}} \right\|_2 \leq \frac{\max_j(|\gamma_j|)}{R} \left(1 + |\eta| \, \|\boldsymbol{J}_{MSA}(\boldsymbol{X})\|_2\right), \tag{14}$$

where $\boldsymbol{J}_{MSA}(\boldsymbol{X}) := \partial \operatorname{MSA}(\boldsymbol{X}) / \partial \boldsymbol{X}$ *denotes the Jacobian of* MSA.

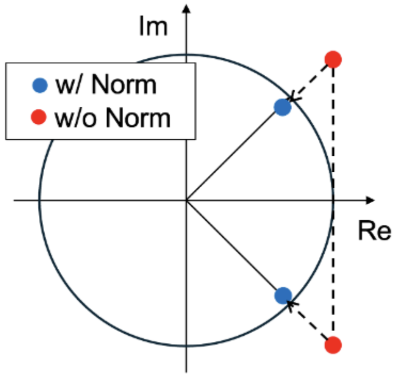
- Normalization operators suppress the Jacobian eigenvalues

Example: oscillatory components

$$
\begin{aligned}
\dot{\boldsymbol{x}} &= \boldsymbol{\Omega}\boldsymbol{x}, & \boldsymbol{J}(\boldsymbol{x}) &= \boldsymbol{\Omega} & &\text{(continuous)} \\
\boldsymbol{x}^{(t+1)} &= (\boldsymbol{I}_D + \eta\boldsymbol{\Omega})\boldsymbol{x}^{(t)}, & \boldsymbol{J}(\boldsymbol{x}^{(t)}) &= \boldsymbol{I}_D + \eta\boldsymbol{\Omega} & &\text{(discrete w/o Norm)} \\
\boldsymbol{x}^{(t+1)} &= \Pi\big((\boldsymbol{I}_D + \eta\boldsymbol{\Omega})\boldsymbol{x}^{(t)}\big), & \boldsymbol{J}(\boldsymbol{x}^{(t)}) &= \left(\boldsymbol{I}_D - \frac{\boldsymbol{y}\boldsymbol{y}^\top}{\|\boldsymbol{y}\|^2}\right)\frac{\boldsymbol{I}_D + \eta\boldsymbol{\Omega}}{\|\boldsymbol{y}\|} & &\text{(discrete w/ Norm)}
\end{aligned}
$$

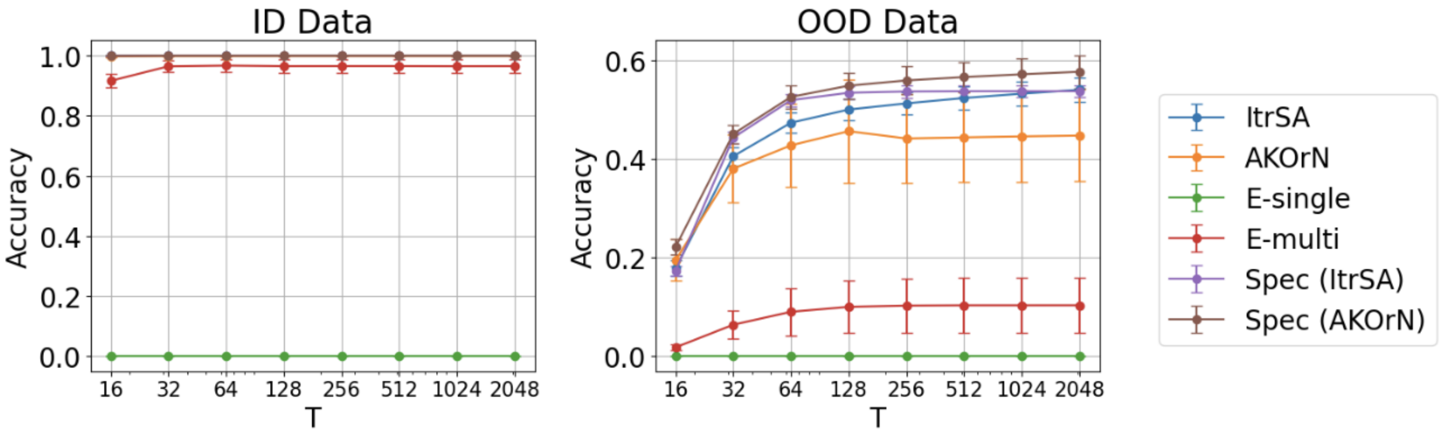
Ω : anti-symmetric matrix

- The continuous systems can exhibit stability
- Discrete counterparts require normalization for stability



(b) Effect of normalization on eigen-values in oscillatory case

Application: Regularization

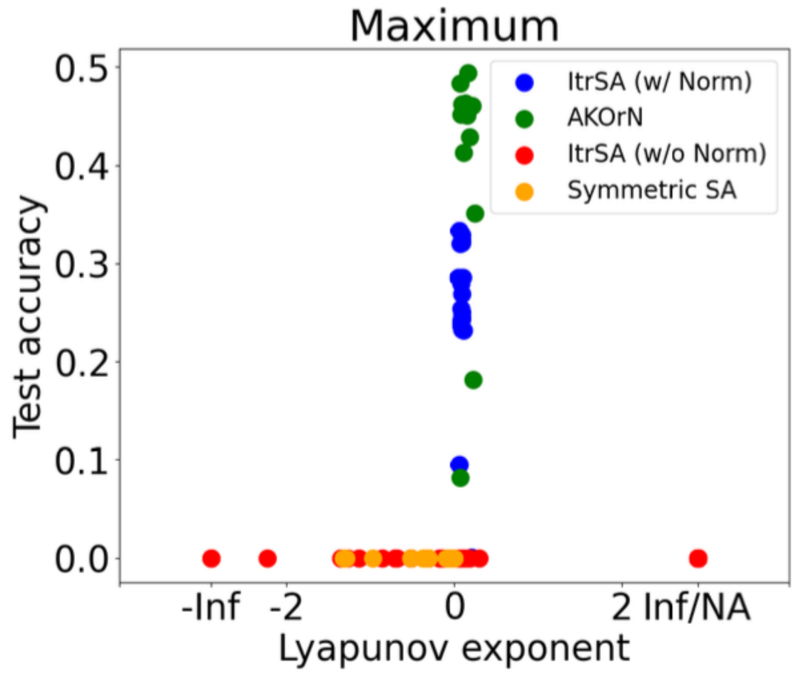


- Apply energy-based and Jacobian spectral regularization
  - Regularization based on each approach
- Energy-based regularization performs worse than the original
- Jacobian spectral regularization outperforms both

Lyapunov Exponent and Criticality

- Measures local divergence/convergence rate in dynamics
  - Positive → instability, Negative → convergence, Zero → criticality

Normalization
→ exponents toward zero & higher acc.



- Successful models:
  max exponent ≈ +0.1
  → near-critical, from the chaotic side

Conclusion

- Relax constraints in energy-based formulations
- Extend Jacobian-based analysis to SA
- Empirically, strong SA models have Lyapunov exponent ≈ 0
- Jacobians and Lyapunov exponent emerge as fundamental tools for realistic SA architectures

Paper link: