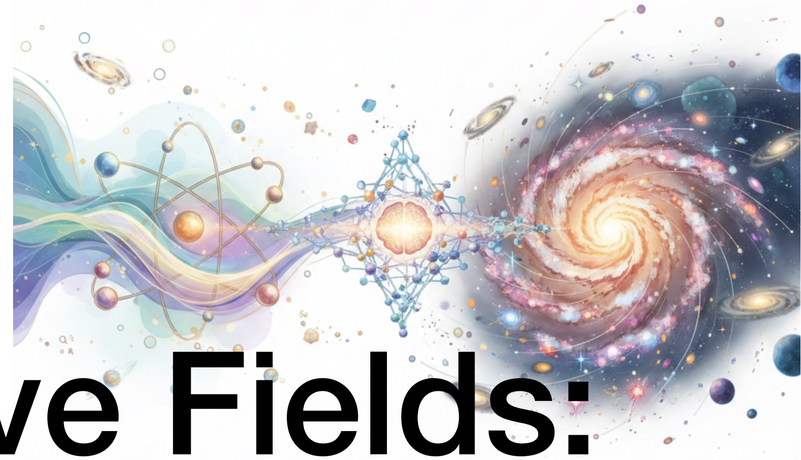


Transformers As Effective Fields: From Quantum Physics to AI

TropicalMathML Seminar, Yonsei University
February 23, 2026

Changqing Fu
PSL University

Generative AI through Transformers

what is the core argument?

PDF TransformerPriGM.pdf

Thought for 32s

The core argument of this work (the poster for the paper *"Transformers Are Optimal Effective Fields"* by Changqing Fu, PriGM@EurIPS 2025) is:

Transformers are not a heuristic design—they are the mathematically optimal ("geodesic") architecture that emerges automatically when you treat neural network layers as continuous-time evolution of a physical particle/field system and demand the least geometric structure necessary.

Image

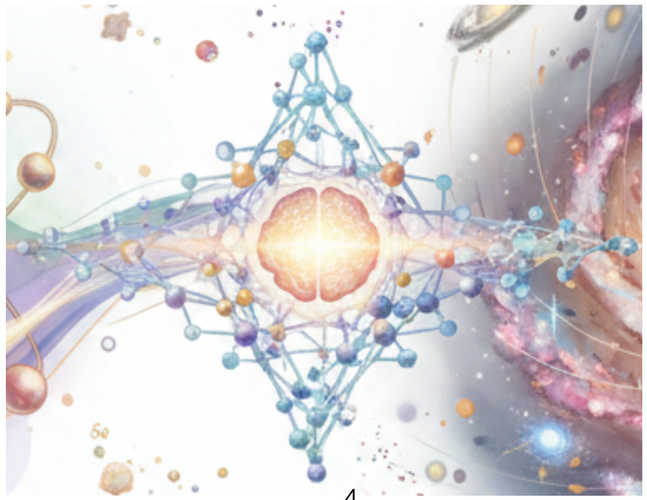
Text

Dauphine UNIVERSITÉ PARIS | PSL CEREMADE UMR CNRS 7534 PR[AI]RIE PaRis Artificial Intelligence Research InstituteE cnrs

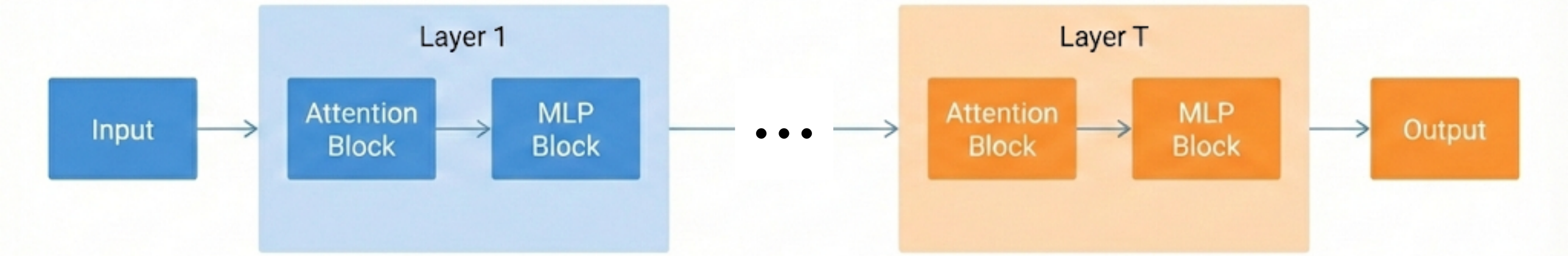
Plan

- Transformer and Least Action
  - Matrix Algebra, HighDim Geometry, MLP / SelfAttention
- Quantum Chaos / Statistical Physics
  - RMT, Double Descent / Scaling Law, Interactive Fields
- From QC to Transformers
  - Left/Right Symmetries, Real/Complex Multiplication, Predictions

What is a Transformer



“Attention Is All You Need”



$$\boldsymbol{X}(t+1) = \boldsymbol{X}(t) + \mathrm{Attention}(\boldsymbol{X}(t))$$

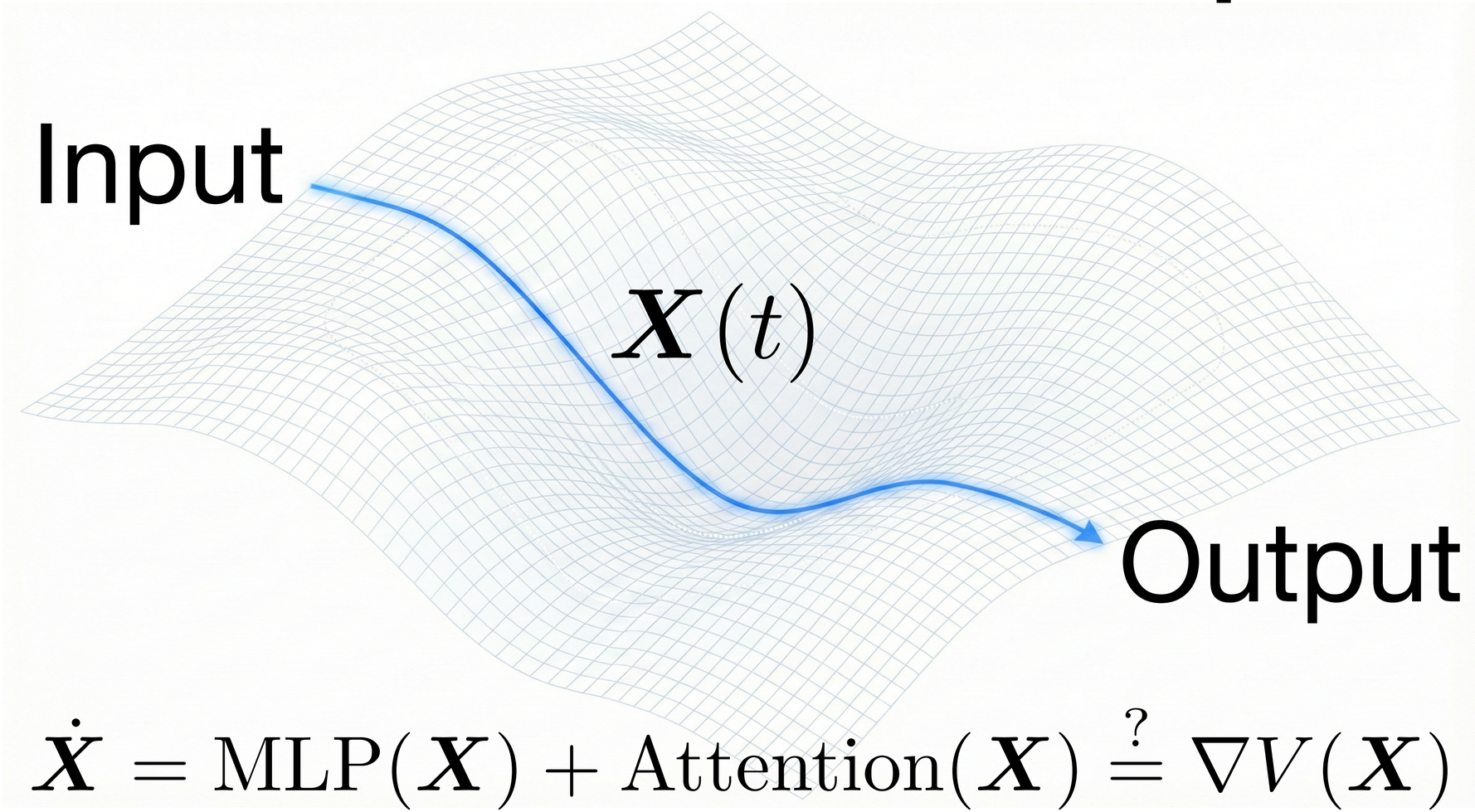
$$\boldsymbol{X}(t+2) = \boldsymbol{X}(t+1) + \mathrm{MLP}(\boldsymbol{X}(t+1))$$

Physics = Symmetry Breaking Principle

"Asymmetrical effects must have asymmetrical causes."

—Pierre Curie, 1894

Minimal Path Principle

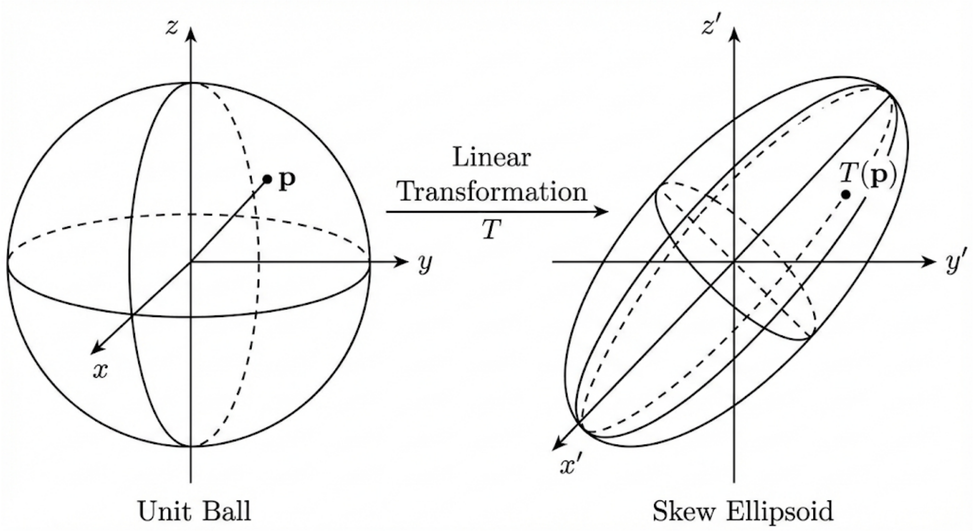


$$\dot{\boldsymbol{X}} = \mathrm{MLP}(\boldsymbol{X}) + \mathrm{Attention}(\boldsymbol{X}) \stackrel{?}{=} \nabla V(\boldsymbol{X})$$

This talk

From Word Vector To Sentence Matrix

$$\boldsymbol{X} = \begin{pmatrix} \text{“The”} \\ \text{“quick”} \\ \text{“brown”} \\ \dots \\ \text{word } N \end{pmatrix}$$



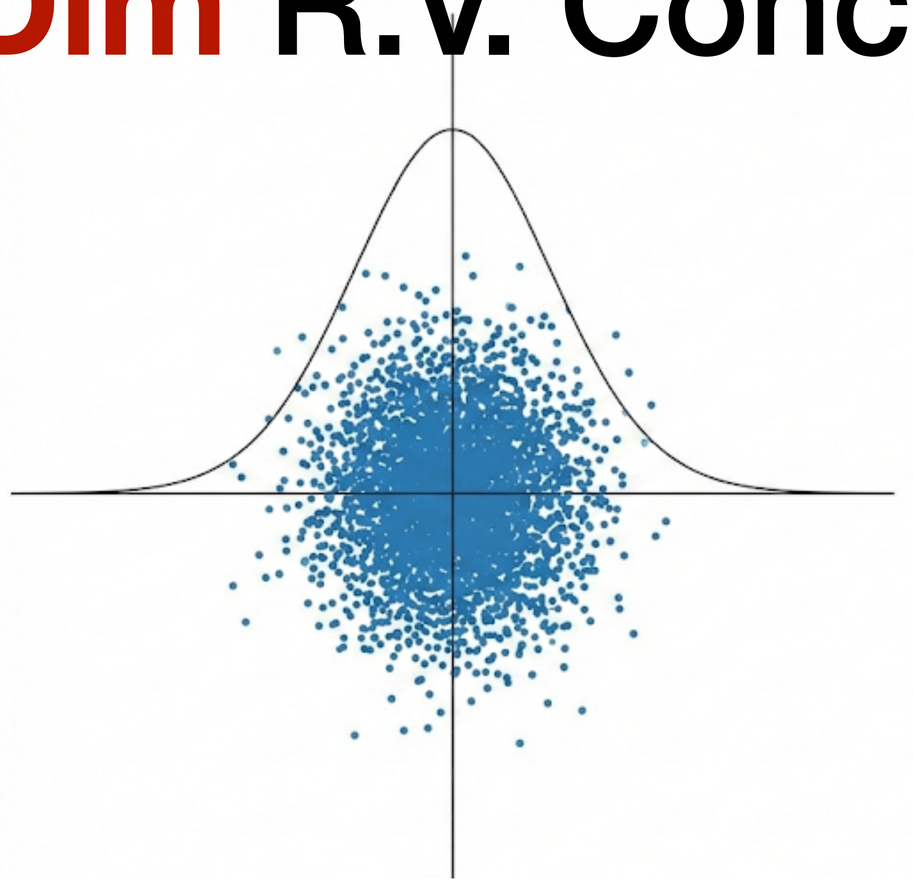
$$\boldsymbol{X} = \begin{pmatrix} \text{“Le”} \\ \text{“rapide”} \\ \text{“brune”} \\ \dots \\ \text{mot } N \end{pmatrix}$$

X = The quick brown fox jumps over the lazy dog.

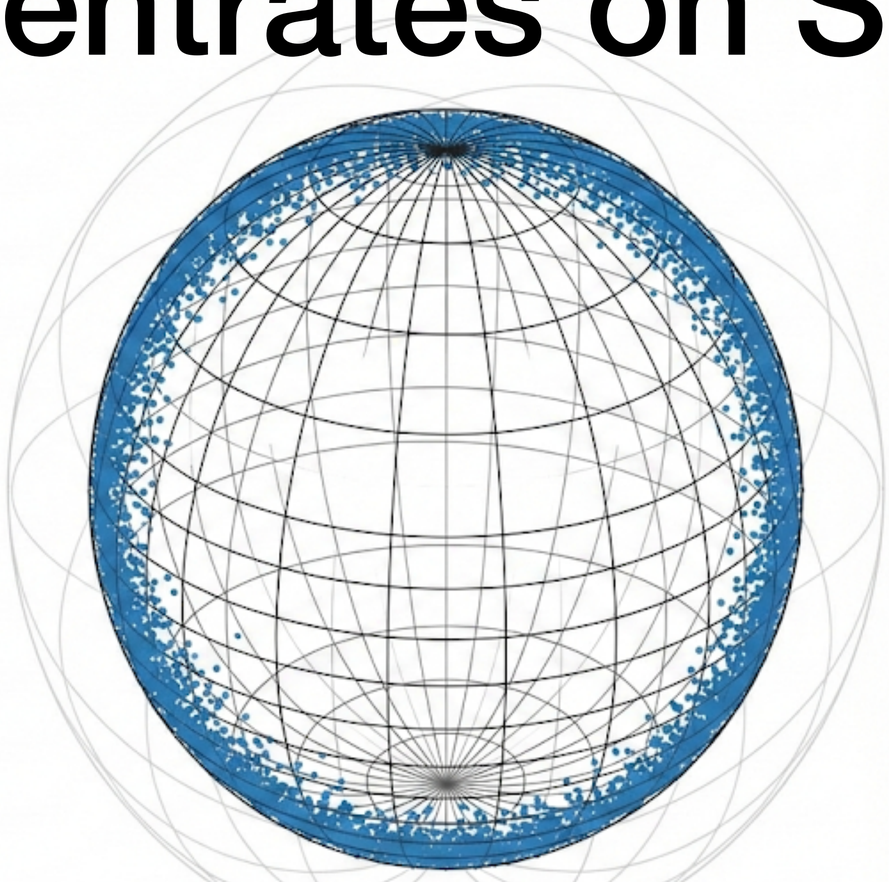
$$\sigma(\boldsymbol{W}) = \{\lambda_1, \lambda_2, \lambda_3\}$$

$$\boldsymbol{X}[1:] = [0.3\ 0.4\ 0.5] \xrightarrow{\boldsymbol{W} \in \mathbb{R}^{3\times 3}} \boldsymbol{X}[1:]\boldsymbol{W} = [0.7\ 0.8\ 0.9]$$

Normalization Layer Is Almost Linear:
HighDim R.V. Concentrates on Sphere



Dim = 2, 3, …



HighDim

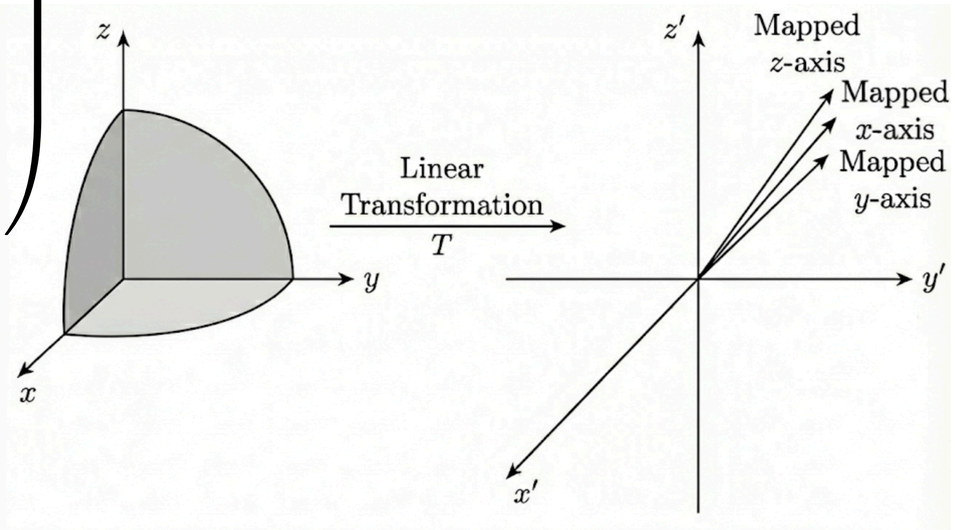
MLP = Learnable Conic Projection

X = The quick brown fox jumps over the lazy dog.

MLP(X) = The ultrasonic caramel unicorn bounces over the snoozing python.

$$\boldsymbol{X} = \begin{pmatrix} \text{“The”} \\ \text{“quick”} \\ \text{“brown”} \\ \dots \\ \text{word } N \end{pmatrix} \qquad (\boldsymbol{X}\boldsymbol{W}_{\text{in}})_+\boldsymbol{W}_{\text{out}} = \begin{pmatrix} \text{“The”} \\ \text{“ultrasonic”} \\ \text{“caramel”} \\ \dots \\ \text{word } N \end{pmatrix}$$

$$\boldsymbol{X}\boldsymbol{W}_{\text{in}} = \begin{pmatrix} \text{“Le”} \\ \text{“rapide”} \\ \text{“brune”} \\ \dots \\ \text{mot } N \end{pmatrix} \longrightarrow (\boldsymbol{X}\boldsymbol{W}_{\text{in}})_+ = \begin{pmatrix} \text{“Le”}_+ \\ \text{“rapide”}_+ \\ \text{“brune”}_+ \\ \dots \\ \text{mot } N \end{pmatrix}$$



ReLU MLP = Tropical Rational Map

X + MLP(X) = “Projected Gradient Flow”

$$\boldsymbol{X}[1:]\boldsymbol{W}_{\text{in}} = [0.7 \;\; -0.8 \;\; 0.9] \xrightarrow{\text{ReLU}} (\boldsymbol{X}[1:]\boldsymbol{W}_{\text{in}})_+ = [0.7 \;\; 0. \;\; 0.9]$$

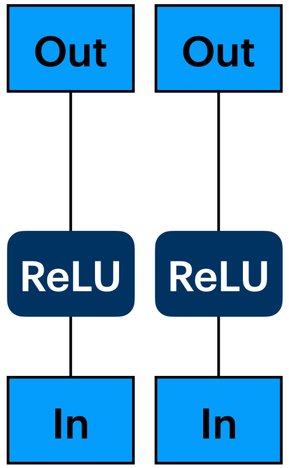
SelfAttention: Cubic/Nonlinear

$$\mathrm{LinearAttention}(\boldsymbol{X}) = \underbrace{(\boldsymbol{XQ})}_{\text{“query”}}\underbrace{(\boldsymbol{XK})^*}_{\text{“key”}}\underbrace{(\boldsymbol{XV})}_{\text{“value”}}\boldsymbol{O}^*$$

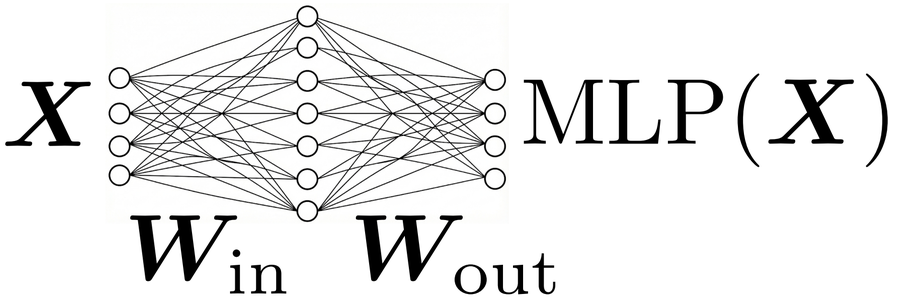
$$\mathrm{SoftmaxAttention}(\boldsymbol{X}) = \mathrm{RowSoftmax}(\boldsymbol{XQK}^*\boldsymbol{X}^*)\boldsymbol{XVO}^*$$

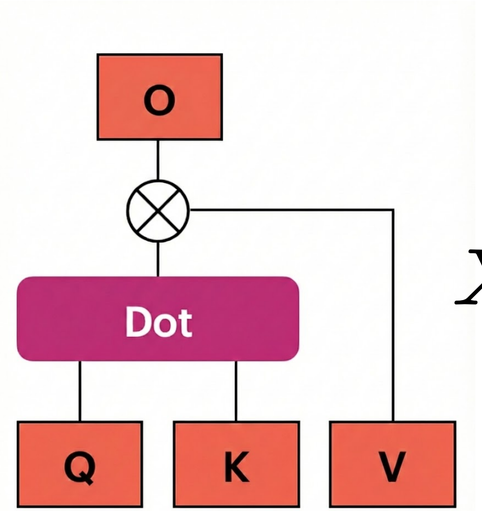
From MLP to Attention

$$\mathrm{MLP}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{W}_{\mathrm{in}}\boldsymbol{W}_{\mathrm{out}}^{*} \qquad \mathrm{Attention}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{Q}\boldsymbol{K}^{*}\boldsymbol{X}^{*}\boldsymbol{X}\boldsymbol{V}\boldsymbol{O}^{*}$$



MultiHead MLP = Wide MLP

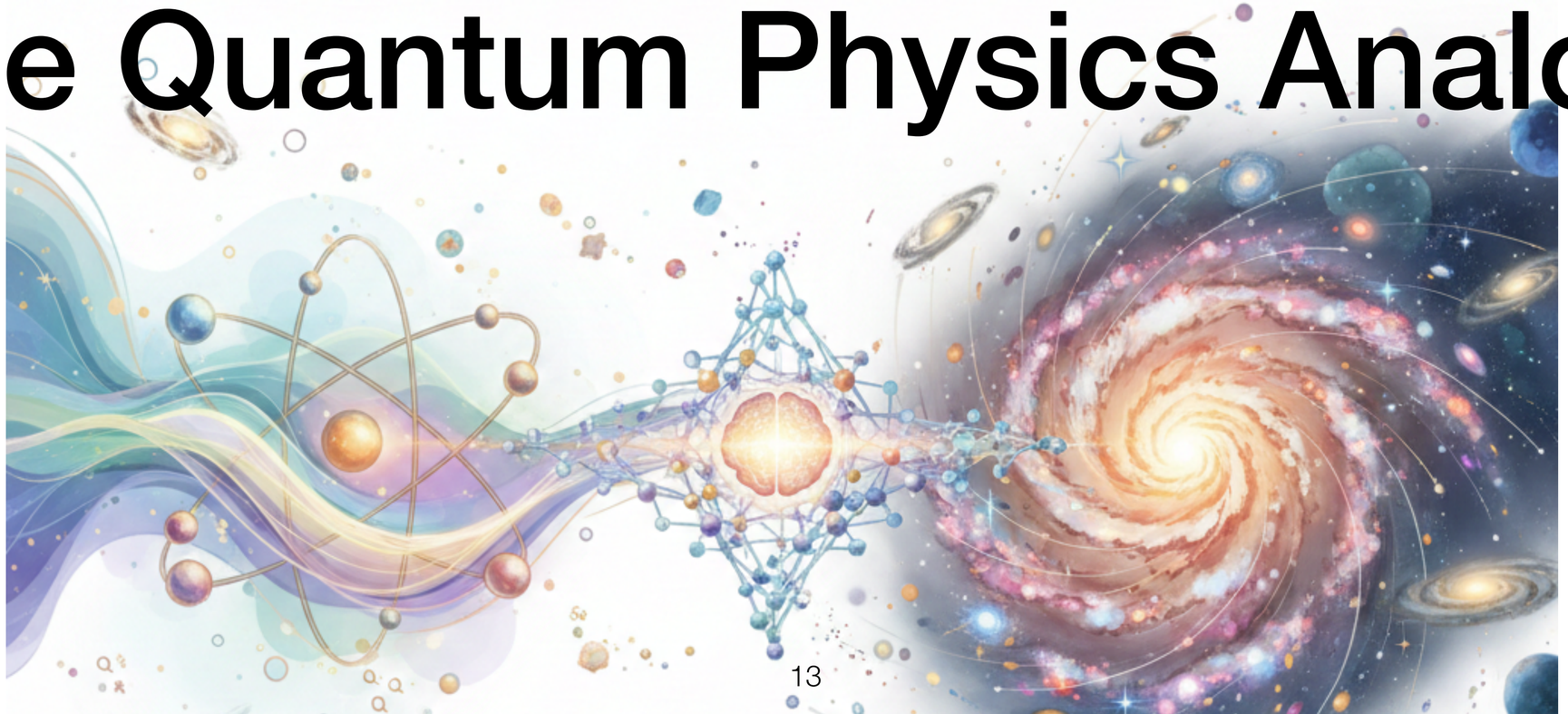




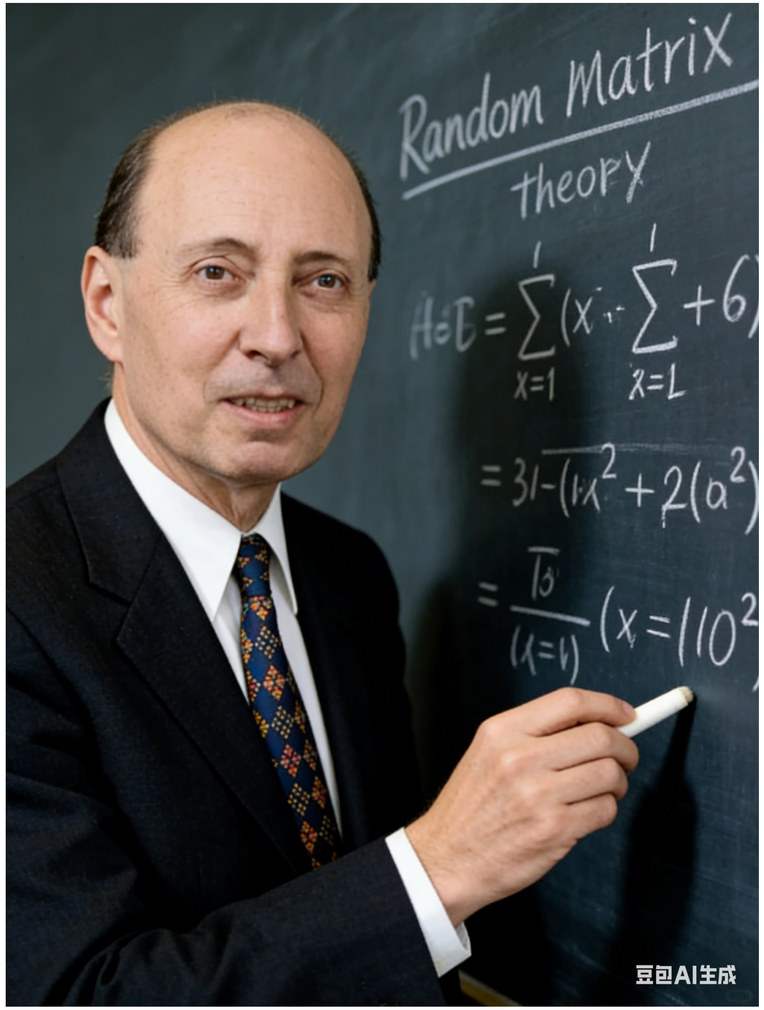
$$\boldsymbol{X} = \begin{pmatrix} \text{“The”} \\ \text{“quick”} \\ \text{“brown”} \\ \dots \\ \text{word } N \end{pmatrix}$$

Right (Token-Wise) vs Left (Neuron-Wise) Multiplication

The Quantum Physics Analogy



From Quantum Chaos to ML

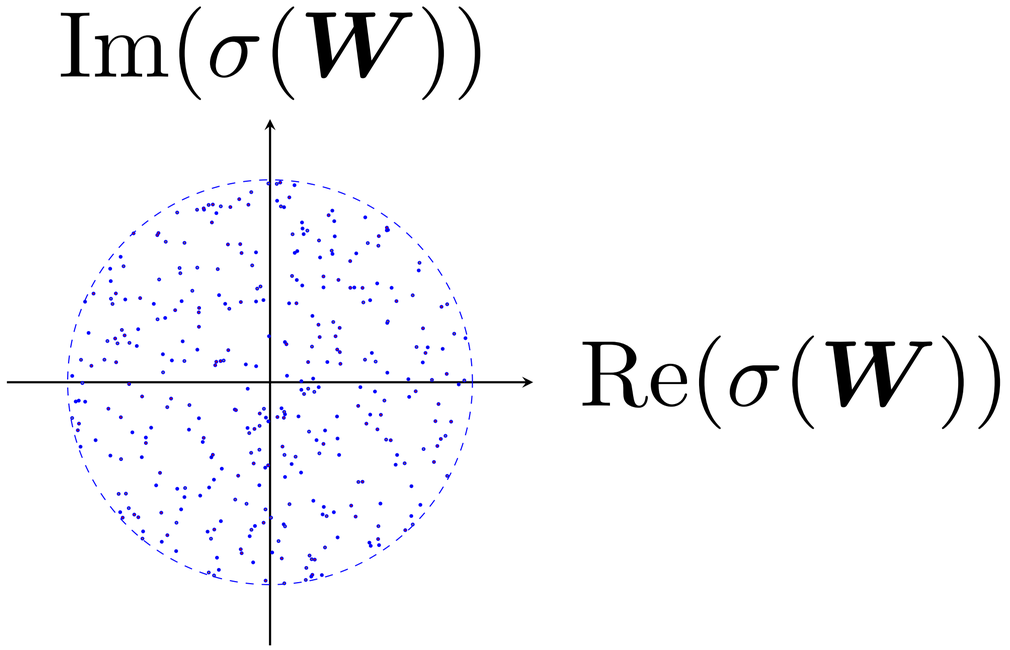


"Spectral statistics of complex systems are universal and determined solely by symmetry class — they are described by random matrix ensembles."

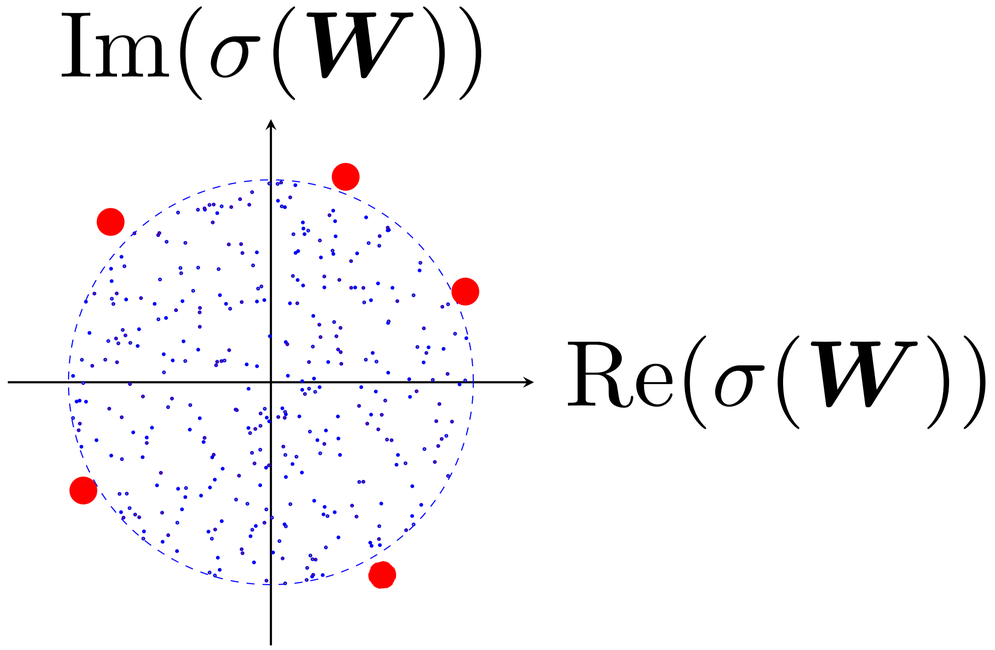
—Eugene Wigner, 1959

From Disordered Systems to ML

$$\sigma(\boldsymbol{W}) = \{\lambda_1, \lambda_2, \lambda_3, \ldots\} \subset \mathbb{C}$$



Before Learning (MaxEntropy)



After Learning (Correlated Weights)

Generalization and Double Descent

Model Size

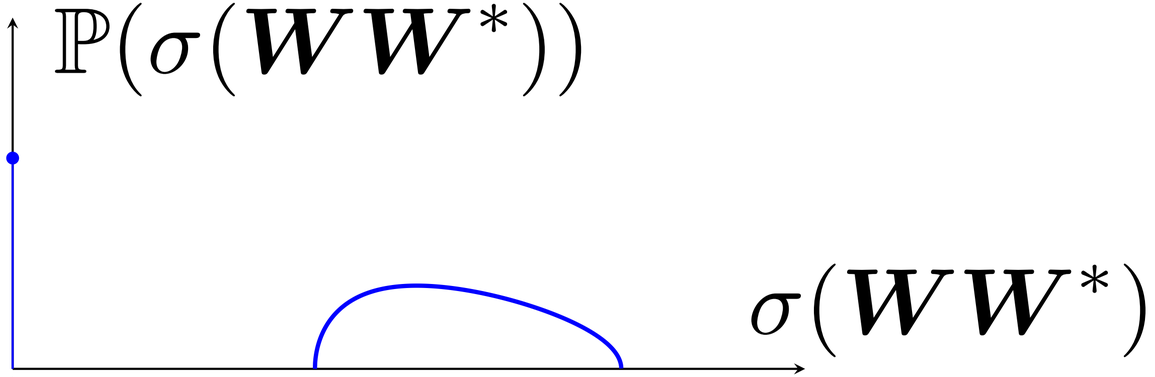
$\mathbb{P}(\sigma(\boldsymbol{W}\boldsymbol{W}^*))$

$\boldsymbol{W} \in \mathbb{R}^{C \times d}$

Tiny

c.f. Low-Rank QK / VO

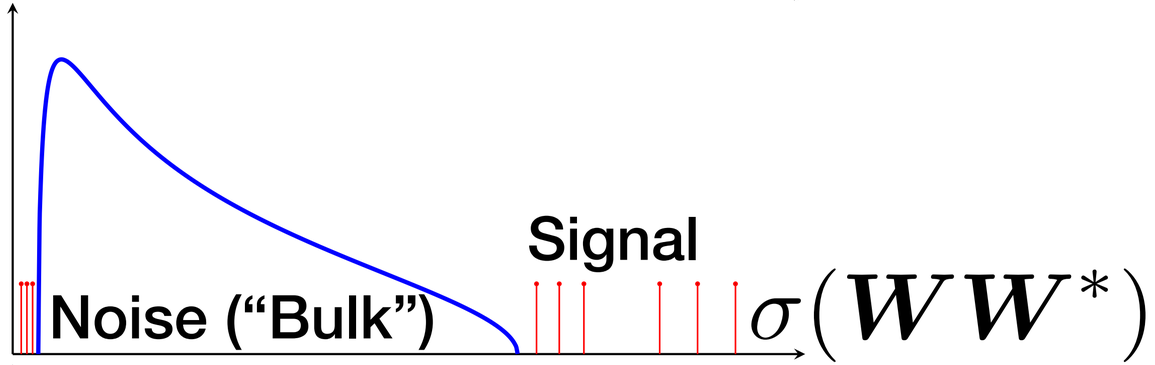
$d \ll C$



Small

(Bad / Singular Case)

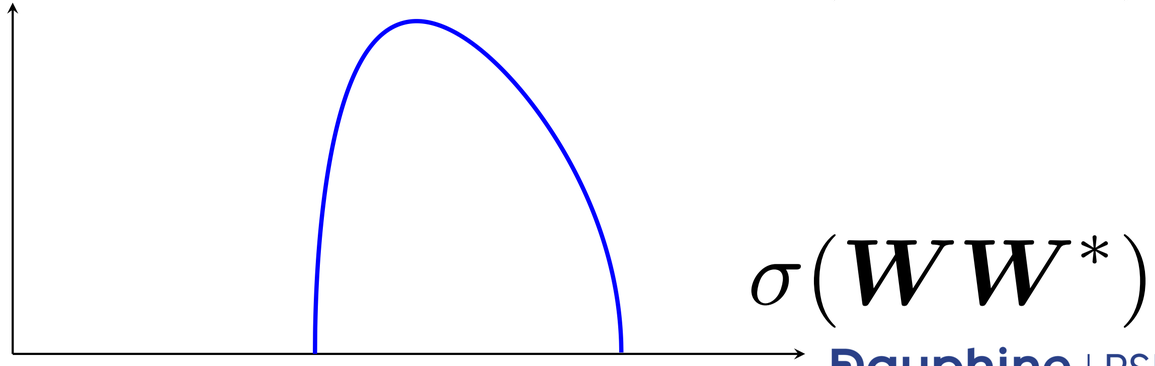
$d \sim C$



Large

c.f. Wide MLP: WinWout

$d \gg C$



MLP <-> Attention
Free <-> Interactive Systems

Quadratic Energy

$$V_2(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*)$$

$$\mathrm{MLP}(\boldsymbol{X}) = \frac{\partial V_2}{\partial \boldsymbol{X}^*} = \boldsymbol{X}\boldsymbol{W}$$

“Free”

Quartic Energy

$$V_4(\boldsymbol{X}) = \frac{1}{2}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*)$$

$$\mathrm{Attention}(\boldsymbol{X}) = \frac{\partial V_4}{\partial \boldsymbol{X}^*} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*\boldsymbol{W}$$

Interactive

Are these potential fields canonical?

We need:

$$V_2(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*)$$

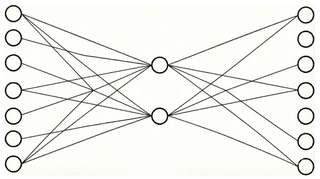
We also need:

$$V_4(\boldsymbol{X}) = \frac{1}{2}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*)$$

…but, what forms of $V(\boldsymbol{X}) = ?$ are **ALL** we need???

The Bridge: Symmetry

Symmetry / Invariance of Energy

Axiom	Invariance	Closed Form
Universality	$V = V_1 + V_2 + \dots$	$V_k(\boldsymbol{X}) = \mathrm{Tr}(\cdots \boldsymbol{W}_{N\times N} \boldsymbol{X} \boldsymbol{W}_{C\times C} \boldsymbol{X}^* \cdots)$ Double-Trace / Hilbert Space Contraction
Perm/Rot Symm	$V(U(N)\boldsymbol{X}) = V(\boldsymbol{X})$	$V(\boldsymbol{X}) = V(\boldsymbol{X}^*\boldsymbol{X})$
Learnable Symm	$V(\boldsymbol{X} C_{U(C)}(\boldsymbol{W})) = V(\boldsymbol{X})$	$V(\boldsymbol{X}) = \mathrm{Tr}(\prod_i \boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*)$
Low-Rank Geoffrey Hinton: “Hand Shaking Analogy”	$\boldsymbol{X}$  $\boldsymbol{X}\boldsymbol{V}\boldsymbol{O}^*$	$\boldsymbol{W} \leftarrow \{\boldsymbol{Q}\boldsymbol{K}^*, \boldsymbol{V}\boldsymbol{O}^*, \boldsymbol{W}_{\mathrm{in}}\boldsymbol{W}_{\mathrm{out}}^*\}$

Real and Complex-Valued Systems

Least Action: $S(\boldsymbol{X}) = \int (\text{Kinetic} - \text{Potential})\, \mathrm{d}t$

Kinetic Term $K(\boldsymbol{X}) = \dot{\boldsymbol{X}}\boldsymbol{X}^* - \boldsymbol{X}\dot{\boldsymbol{X}}^*$

Potential Term $V_2(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*)$

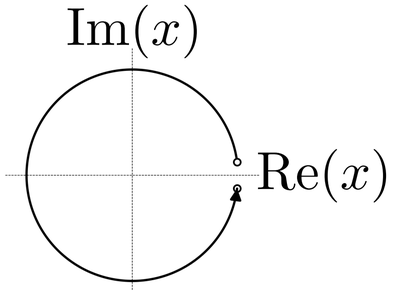
$V_4(\boldsymbol{X}) = \frac{1}{2}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*)$

Anti-Symplectic Terms $V_2(\boldsymbol{X}) = \mathrm{Tr}(\overline{\boldsymbol{X}}\boldsymbol{W}\boldsymbol{X}^*)$

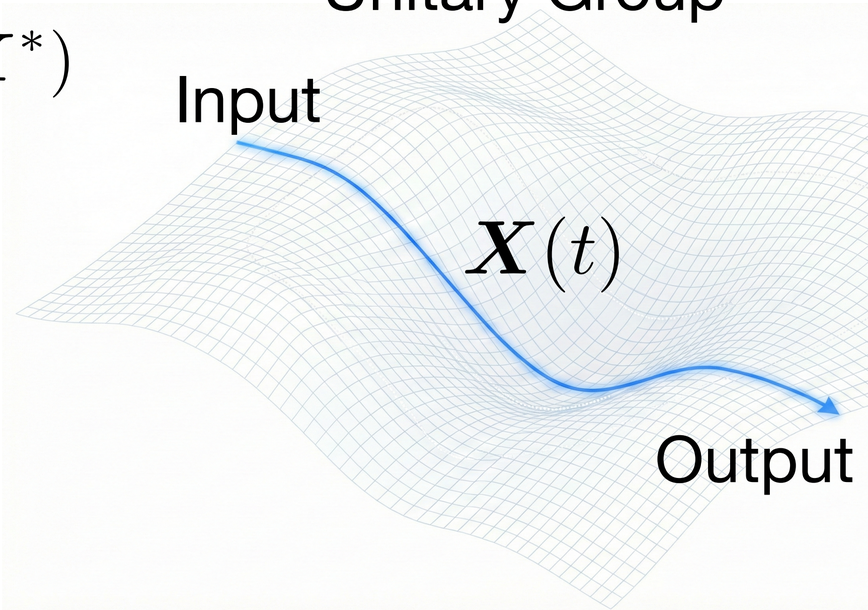
Euler-Lagrange Equation $\partial S / \partial \boldsymbol{X}^* = 0$

The Transformer ODE $\dot{\boldsymbol{X}} = \mathrm{MLP}(\boldsymbol{X}) + \mathrm{Attention}(\boldsymbol{X})$

Complex Multiplication



Unitary Group



Linear versus Softmax Attention: Effective and Entropic Interactions

Recall: LinearAttention $V_4(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*)$

$$V_{\mathrm{ent}}(\boldsymbol{X}) = \sum_{i=1}^{\mathrm{nHead}} \sum_{j=1}^{N} \log \sum_{k=1}^{N} \exp([\frac{1}{\sqrt{C}} \boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*]_{jk})$$

$$\frac{\partial V_{\mathrm{ent}}}{\partial \boldsymbol{X}^*} = \sum_{i=1}^{\mathrm{nHead}} \mathrm{RowSoftmax}(\frac{1}{\sqrt{C}} \boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*)\boldsymbol{X}\boldsymbol{W}_i$$

Renormalization

LinearAttention

“Effective Field” **Scale-Separated Interaction** **Hot Limit:** $\text{as } h \to 0$

$$V_{2+4}(h\boldsymbol{X}) = h^2 V_2(\boldsymbol{X}) + h^4 V_4(\boldsymbol{X}) + O(h^6)$$

NO MORE TERMS!!!

SoftmaxAttention

Entropic Regularization / “Free Energy” **Multi-Scale Interaction**

$$V_{2+\text{ent}}(h\boldsymbol{X}) = h^2 V_2(\boldsymbol{X}) + \sum \text{LogSum}_k \text{Exp}(h^2 [\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*]_{jk})$$

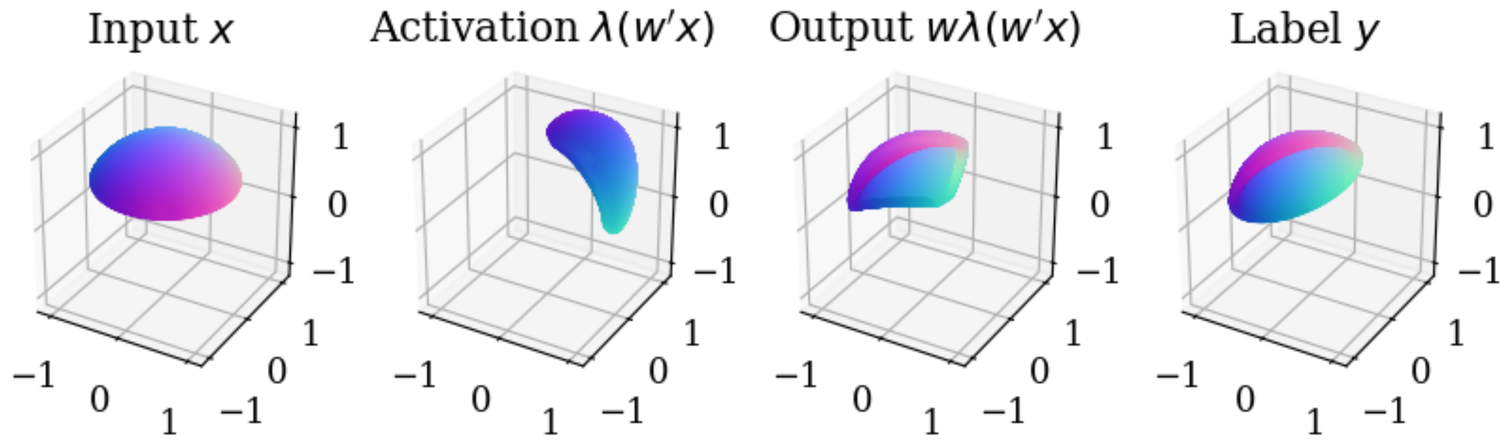
Head, Token i, j

Hot Limit: Vanishes Too

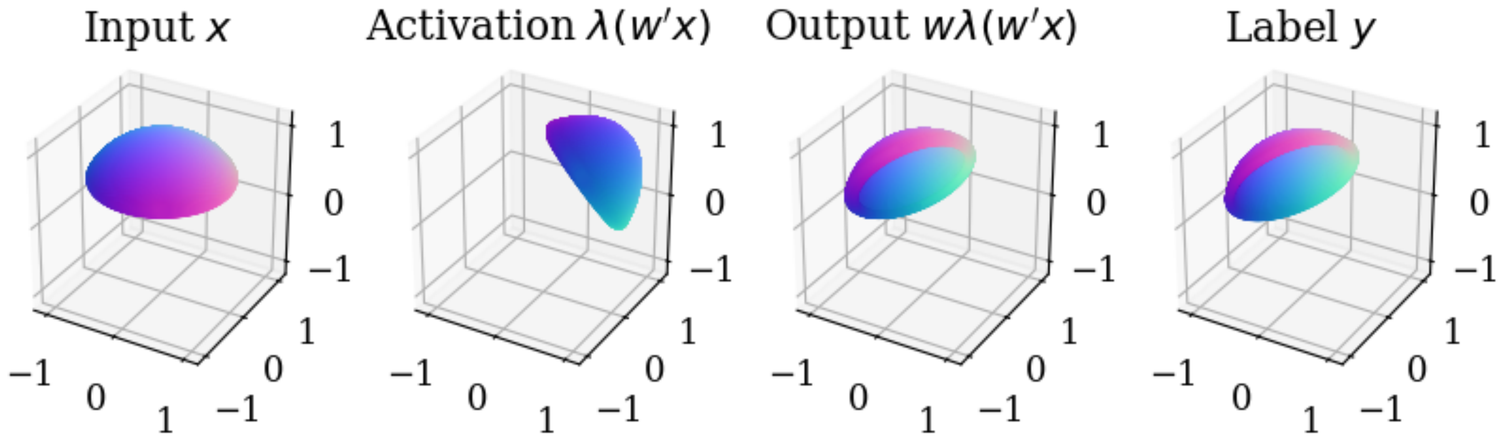
Cold Limit: $\text{as } h \to \infty$ **Concentrates on Max Token**

Applications: Predicting Structures

Embedding Space Favors Orthogonal Symmetry

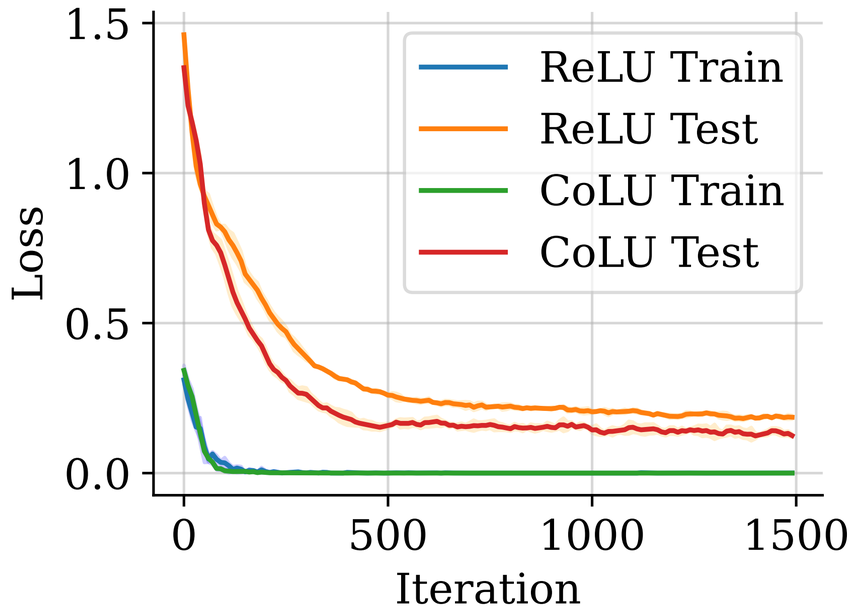


ReLU: permutation symmetry

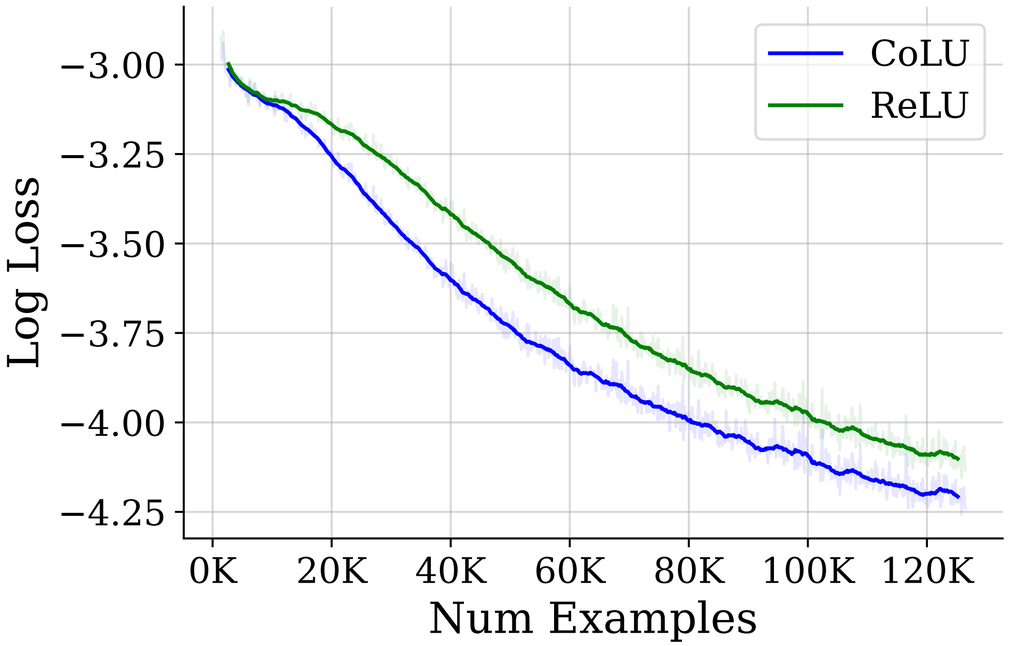


CoLU: orthogonal/rotary symmetry

- Minimal Example (C=3)
- Improved Generalization
- Accelerated Convergence
- Applies to ReLU-Attention



Scaling Up Conic Activations



Diffusion Transformer 0.8B (Oxford102)

GPT2 MLP (FineWeb10M)	ReLU	CoLU
Forward FLOPs	39.064M	39.101M
Test Loss	3.4569 ± 0.1182	**3.3804** ± 0.1159
ResNet-56 (CIFAR10)	ReLU	CoLU
Forward FLOPs	0.252M	0.257M
Test Accuracy	**92.7282** ± 0.357	93.5851 ± 0.442
Diffusion Model (CIFAR10)	ReLU	CoLU (Faster)
Train Loss	0.1653	**0.1458**
Early Samples		

Conic linear units: improved model fusion and rotational-symmetric generative model. VISIGRAPP 2024.
Conic activation functions. UniReps@NeurIPS 2024. PMLR 2025.

Gated Activation: Multi-Trace Model

$$V_3(\boldsymbol{X}) = (\boldsymbol{X}\boldsymbol{W}) \odot (\boldsymbol{X}\boldsymbol{W})(\boldsymbol{X}\boldsymbol{W})^*$$

$$\implies \partial V_3 / \partial \boldsymbol{X}^* = (\boldsymbol{X}\boldsymbol{W}) \odot (\boldsymbol{X}\boldsymbol{W})\boldsymbol{W}^*$$

$$\xrightarrow{\text{ReLU}} (\boldsymbol{X}\boldsymbol{W}) \odot (\boldsymbol{X}\boldsymbol{W})_+ \boldsymbol{W}^*$$

Sparse Attention: Non-Commutative Systems

Lemma 8 (Sparse Attention). *Under the pairwise non-commutative condition* $[\lambda_{MLP}, \lambda_A] = \beta\lambda_A$*, the sequential composition of the layers is equivalent to a single effective flow generated by:*

$$\exp(\lambda_{MLP})\exp(\lambda_A) = \exp\left(\lambda_{MLP} + G(\beta)\cdot\lambda_A\right)$$

where the effective interaction gate $G(\beta)$ *is given by the generating function of the Bernoulli numbers:* $G(\beta) = \frac{\beta}{1-e^{-\beta}}$*. As* $\beta \to -\infty$, $G(\beta) \sim |\beta| e^{-|\beta|} \to 0$.

“Fail to Commute between Tokens $\Longrightarrow$ Turn off the Gate”

Reference: https://changqingfu.com/pdf/transformer.pdf