Transformers as Effective Fields

A quantum-physics theory of AI

Feature *State* ***Architecture*** *Learning* *Evaluation*

Changqing Fu, presented at MLCollective, November 2025

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

Generative Models

Dataset

Model

Training

Text,
Shape,
Texture,
Style
...

Inference

What happened inside?

Take Home Message: Transformer is not a mere engineering coincidence, but is grounded in principles from (quantum) physics.
Today's Goal: I don't see this idea ANYwhere after asking many people. Please correct it or guide me to any existing papers.

Plan

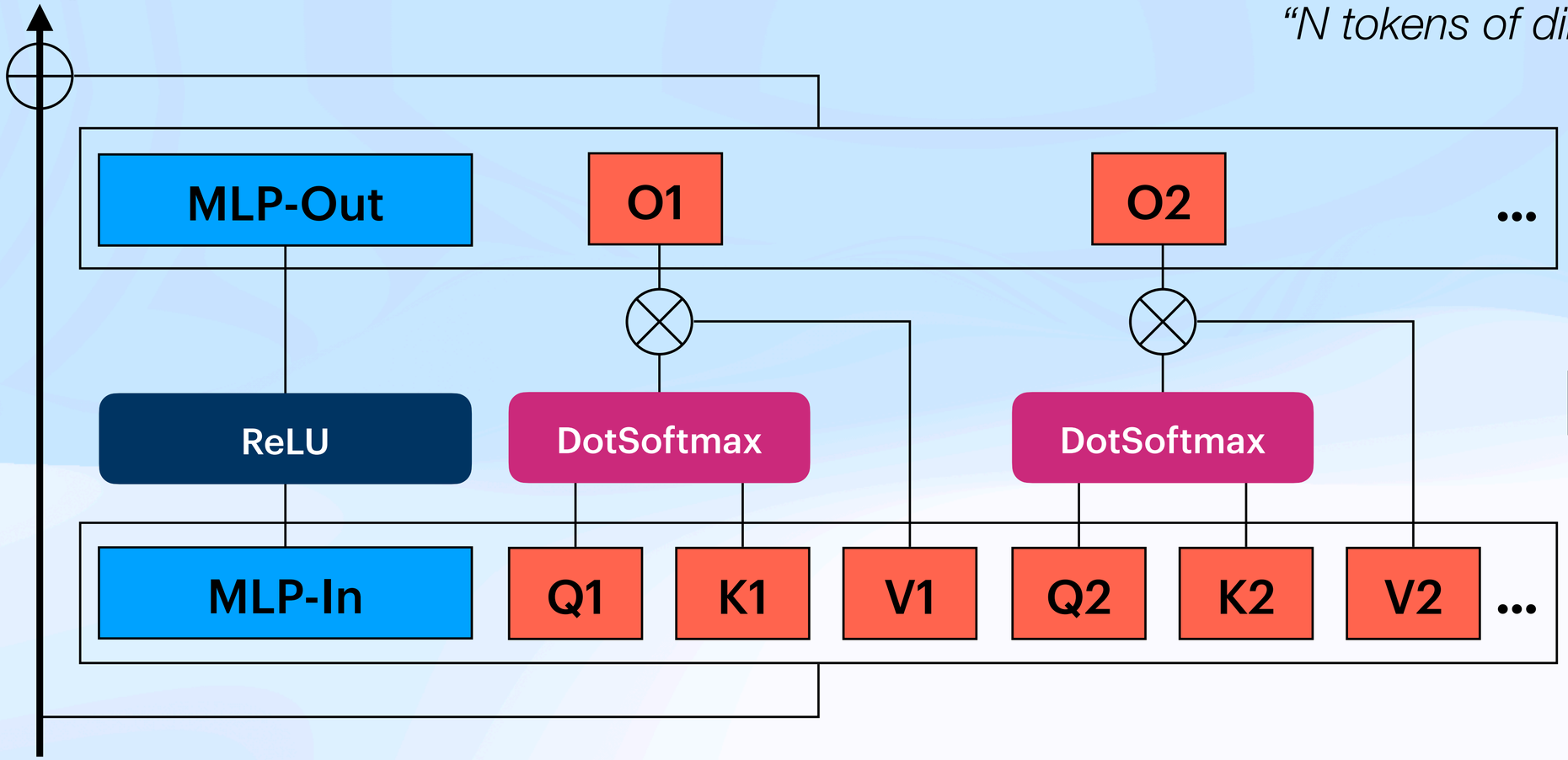
- What are Transformers
- Potential Field of Linear Transformer
- Why is the Field Effective?
- Axiomize Softmax / Activation Functions

Transformers

"Scaling Vision Transformers to 22 Billion Parameters", ICML 2023

$$\boldsymbol{X} : [0,1] \to \mathbb{R}^{N \times C}$$

"N tokens of dimension C"



Discrete $\boldsymbol{X}(t+1) = \boldsymbol{X}(t) + \boldsymbol{W}'(t) \circ \lambda \circ \boldsymbol{W}(t)(\boldsymbol{X}(t))$

Continuous $\dot{\boldsymbol{X}} = \boldsymbol{W}' \circ \lambda \circ \boldsymbol{W}(\boldsymbol{X})$

$$\boldsymbol{X} \in \mathbb{C}^{N \times C}, \boldsymbol{W}_{\text{in},i}, \boldsymbol{W}_{\text{out},i}, \boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i, \boldsymbol{O}_i \in \mathbb{C}^{C \times r}$$

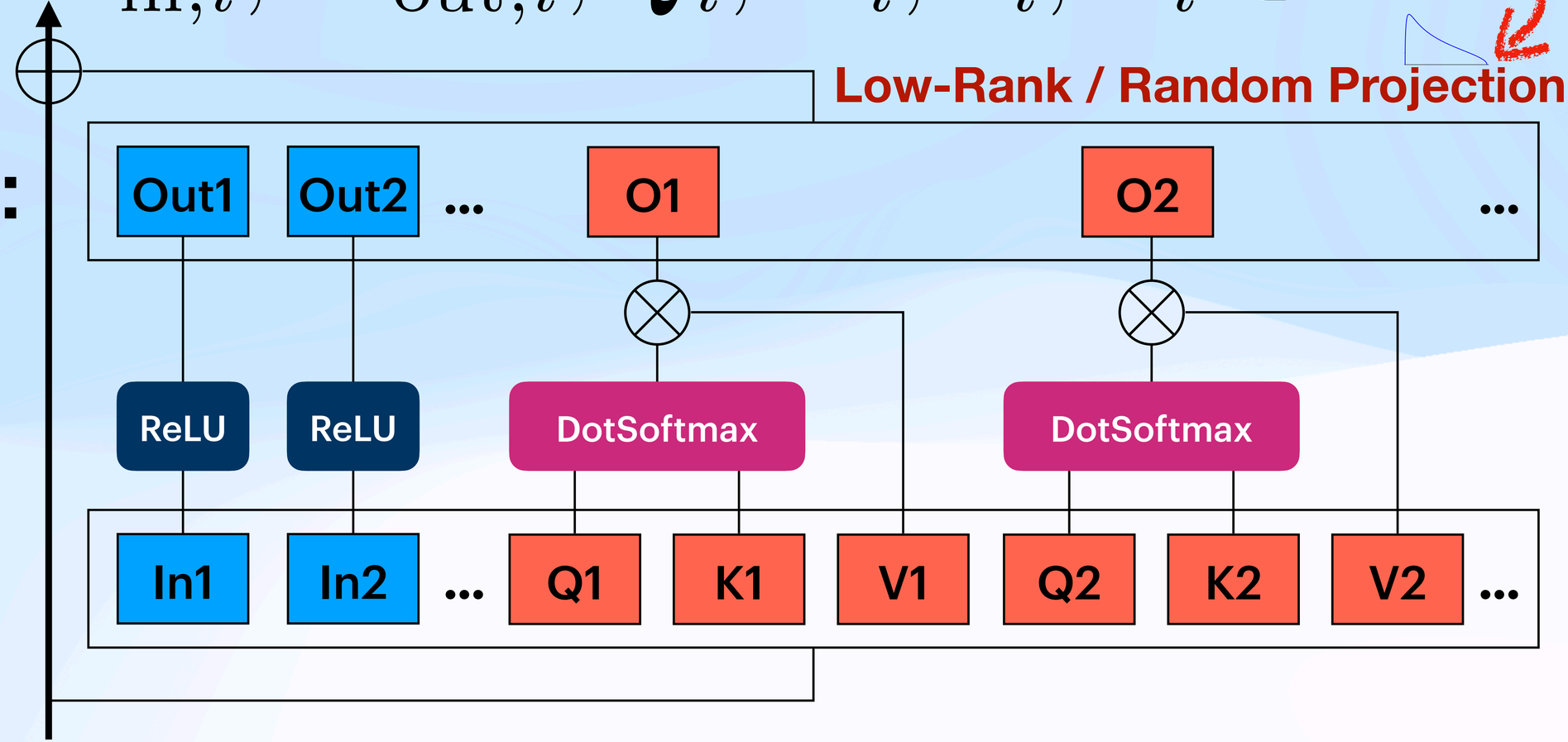
Low-Rank / Random Projection

***Simplification:**

***MultiHeadMLP**

=

Wide MLP



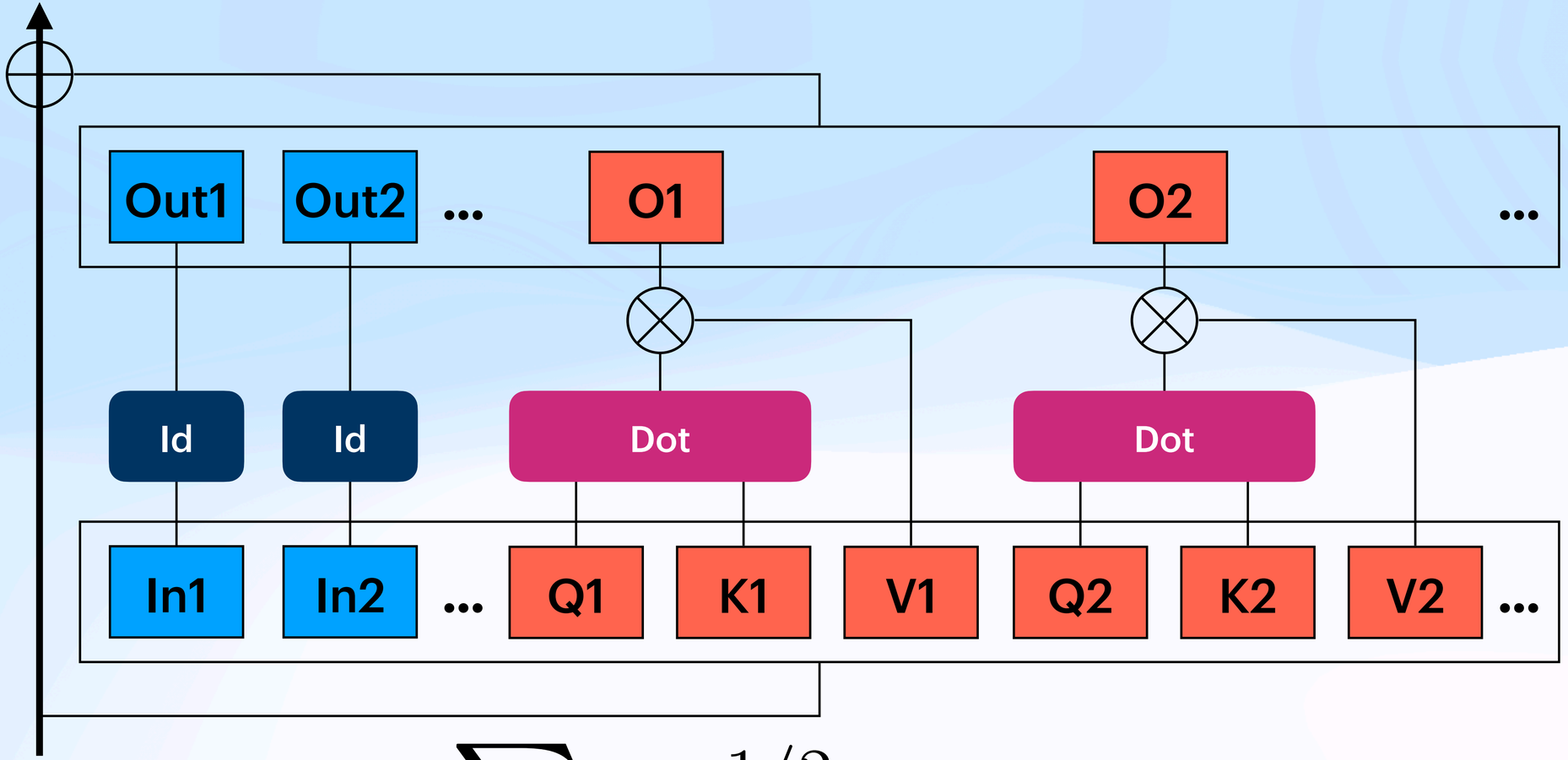
$$\dot{\boldsymbol{X}} = \sum_i (\boldsymbol{X}\boldsymbol{W}_{\text{in},i})_+ \boldsymbol{W}_{\text{out},i}^\top + \sum_i \text{Softmax}(C^{-1/2}\boldsymbol{X}\boldsymbol{Q}_i\boldsymbol{K}_i^\top\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{V}_i\boldsymbol{O}_i^\top$$

Linear Transformers*

**Without Loss of Generality*

* ReLU
=
Conic Proj-GD

* Softmax
=
LogSumExp Energy



$$\dot{X} = \sum_i X W_{\text{in},i} W^*_{\text{out},i} + \sum_i C^{-1/2} X Q_i K_i^* X^* X V_i O_i^*$$

Plan

- What are Transformers
- **Potential Field of Linear Transformer**
- Why is the Field Effective?
- Axiomize Softmax / Activation Functions

“In-Context Learning” Desires $\mathrm{i}\dot{\boldsymbol{X}} = \dfrac{\partial V}{\partial \boldsymbol{X}^*}$

* Complex Field = Dissipative + Conservative ***Previous MLC Talks*
* Otherwise we will get self-adjoint / symmetric weights

** Wirtinger Calculus*

$$V(\boldsymbol{X}) = \frac{\mathrm{i}}{2}\sum_{i=1}^{n}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*) + \frac{\mathrm{i}}{2\sqrt{C_{\mathrm{A}}}}\sum_{i=1}^{n/2}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{A}i}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{\mathrm{B}i}^*\boldsymbol{X}^*)$$

$$\frac{\partial V}{\partial \boldsymbol{X}^*} = \textbf{Transformers Architecture!}$$

Key Observation $V = V_2 + V_4 + O(\|\boldsymbol{X}\|^6), \boldsymbol{X}/\hbar \overset{\text{replace}}{\to} \boldsymbol{X}$

Why Is $\mathrm{i}\dot{\boldsymbol{X}} = \frac{\partial V}{\partial \boldsymbol{X}^*}$ Canonical?

Euler-Lagrange Eq
Schrödinger Eq
Feynman Path Integral

$$Z = \int \exp(\mathrm{i}S(\boldsymbol{X}))\mathcal{D}\boldsymbol{X}$$

$$S(\boldsymbol{X}) := \int_0^1 \mathrm{Tr}\left(\frac{\mathrm{i}}{2}(\dot{\boldsymbol{X}}\boldsymbol{X}^* - \boldsymbol{X}\dot{\boldsymbol{X}}^*) - V(\boldsymbol{X})\right)\mathrm{d}t$$

$$\boldsymbol{X} = \frac{q+ip}{\sqrt{2}}, \quad \boldsymbol{X}^* = \frac{q-ip}{\sqrt{2}}, \quad \frac{1}{2}(\dot{\boldsymbol{X}}\boldsymbol{X}^* - \boldsymbol{X}\dot{\boldsymbol{X}}^*) = \frac{\mathrm{i}}{2}(\boldsymbol{X}^*\dot{\boldsymbol{X}} - \dot{\boldsymbol{X}}^*\boldsymbol{X}) = p\dot{q}$$

Legendre Transform

$$L^*(p) = \max_{\dot{q}} p\dot{q} - L(\dot{q})$$

$$\omega = \frac{\mathrm{i}}{2}((\mathrm{d}\boldsymbol{X})\boldsymbol{X}^* - \boldsymbol{X}\,\mathrm{d}\boldsymbol{X}^*) \quad \mathrm{d}\omega = \mathrm{i}\,\mathrm{d}\boldsymbol{X} \wedge \mathrm{d}\boldsymbol{X}^*$$

Symplectic Form

Plan

- What are Transformers
- Potential Field of Linear Transformer
- **Why is the Field Effective?**
- Axiomize Softmax / Activation Functions

Why Is V "Effective"?

NO MORE TERMS!!!

Axiom 1: Real Analyticity $V = V_2 + V_4 + O(\|\boldsymbol{X}\|^6), \boldsymbol{X}/\hbar \overset{\text{replace}}{\to} \boldsymbol{X}$

(Multi-Scale, Free+Interactive terms, RG / Phase Transition)

Axiom 2: Left Permutation Symmetry $V(\boldsymbol{X}^*\boldsymbol{X}) : \mathbb{C}^{N\times N} \to \mathbb{C}$

(Word Exchangeability, *broken by RoPE)

Axiom 3: Right Learnable U(1) Symmetry $V(\{\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*\}_i) : \mathbb{C}^n \to \mathbb{C}$

(Hamiltonian Matrix/RMT)

$$V(\boldsymbol{X}) = \frac{\mathrm{i}}{2}\sum_{i=1}^{n} \mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*) + \frac{\mathrm{i}}{2\sqrt{C_\mathrm{A}}}\sum_{i=1}^{n/2} \mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{A}i}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{\mathrm{B}i}^*\boldsymbol{X}^*)$$

This is the Minimal Interactive Polynomial Field!

Why Is V “Effective”?

NO MORE TERMS!!!

Axiom 1: Real Analyticity $V = V_2 + V_4 + O(\|\boldsymbol{X}\|^6), \boldsymbol{X}/\hbar \overset{\text{replace}}{\to} \boldsymbol{X}$

(Multi-Scale, Free+Interactive terms, RG / Phase Transition)

Axiom 2: Left Permutation Symmetry $V(\boldsymbol{X}^*\boldsymbol{X}) : \mathbb{C}^{N\times N} \to \mathbb{C}$

(Word Exchangeability, *broken by RoPE)

Axiom 3: Right Learnable U(1) Symmetry $V(\{\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*\}_i) : \mathbb{C}^n \to \mathbb{C}$

(Hamiltonian Matrix/RMT)

$$V(\boldsymbol{P}\boldsymbol{X}) = V(\boldsymbol{X}), \forall \boldsymbol{P} \in S_N$$

$$\forall i, V(\boldsymbol{X}\boldsymbol{Q}^*) = V(\boldsymbol{X}), \forall \boldsymbol{Q} \text{ s.t. } [\boldsymbol{Q}, \boldsymbol{W}_i] = 0$$

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

Why Is V "Effective"?

NO MORE TERMS!!!

Axiom 1: Real Analyticity $V = V_2 + V_4 + O(\|\boldsymbol{X}\|^6), \boldsymbol{X}/\hbar \overset{\text{replace}}{\to} \boldsymbol{X}$

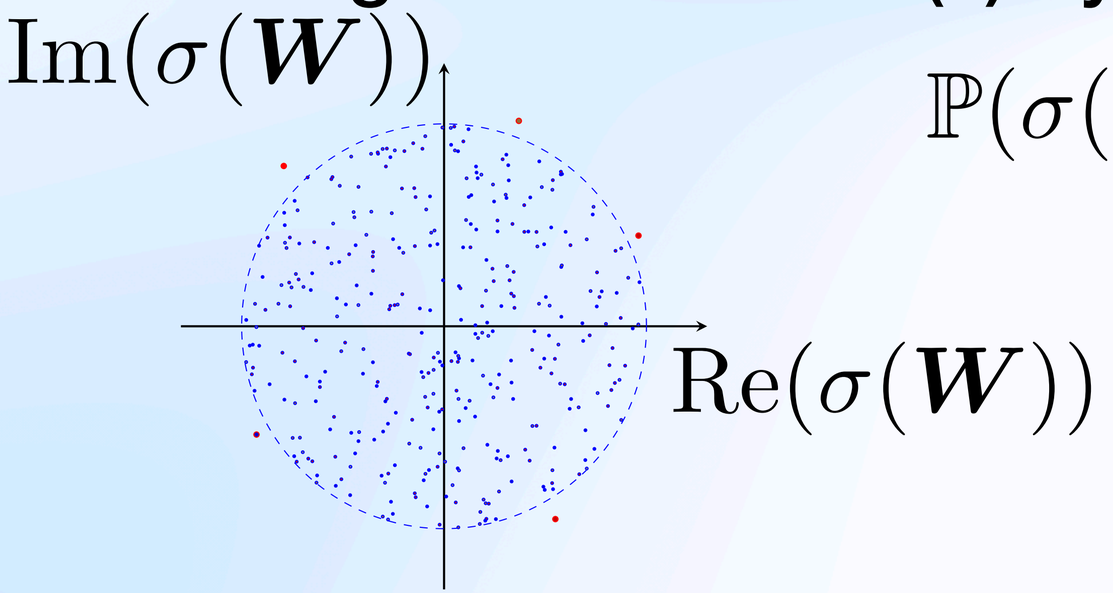
(Multi-Scale, Free+Interactive terms, RG / Phase Transition)

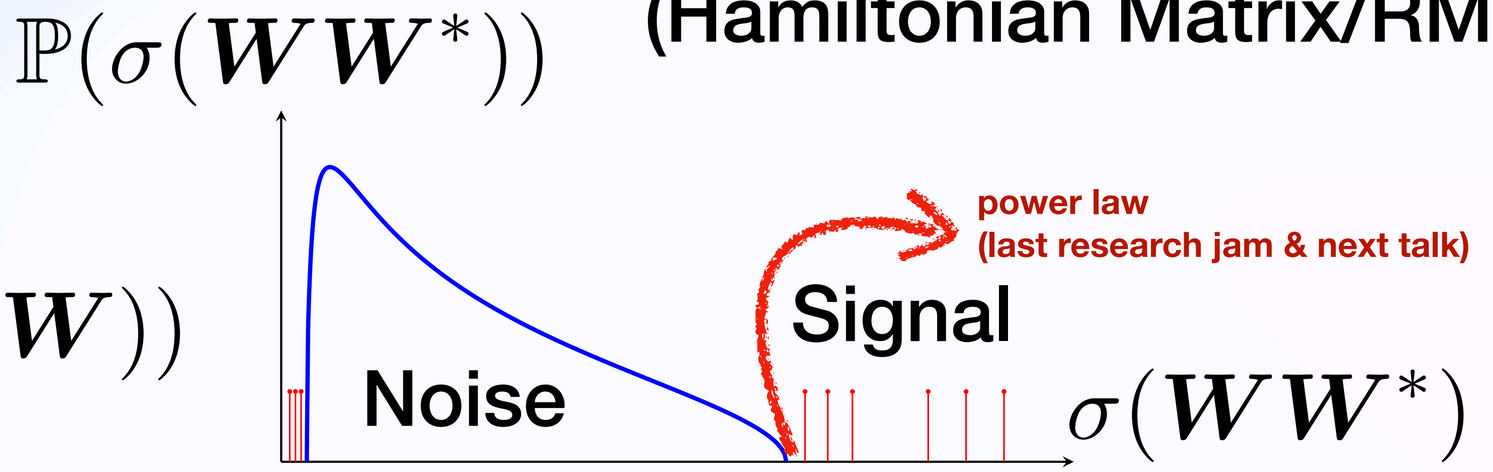
Axiom 2: Left Permutation Symmetry $V(\boldsymbol{X}^*\boldsymbol{X}) : \mathbb{C}^{N\times N} \to \mathbb{C}$

(Word Exchangeability, *broken by RoPE)

Axiom 3: Right Learnable U(1) Symmetry $V(\{\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*\}_i) : \mathbb{C}^n \to \mathbb{C}$

(Hamiltonian Matrix/RMT)





Plan

- What are Transformers
- Potential Field of Linear Transformer
- Why is the Field Effective?
- **Axiomize Softmax / Activation Functions**

Q1: What About Softmax?

Why Softmax = LogSumExp Potential?

$$V(\boldsymbol{X}) = \frac{\mathrm{i}}{2} \sum_{i=1}^{\mathrm{nHead}} \sum_{j=1}^{N} \log \sum_{k=1}^{N} \exp([\frac{1}{\sqrt{r}} \boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*]_{jk})$$

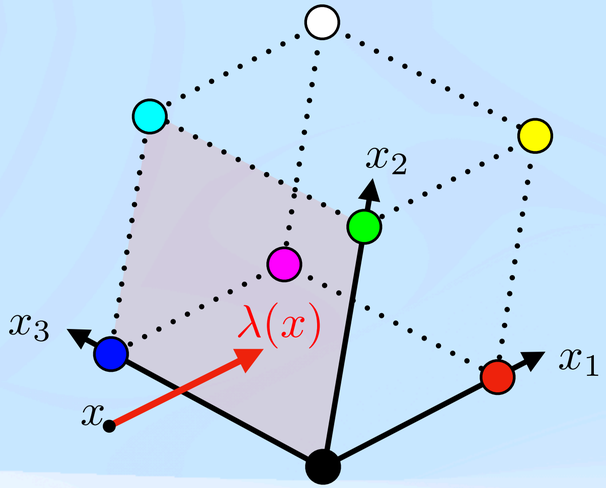
$$\frac{\partial V}{\partial \boldsymbol{X}^*} = \frac{\mathrm{i}}{2} \sum_{i=1}^{\mathrm{nHead}} \mathrm{Softmax}(\frac{1}{\sqrt{r}} \boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*)\boldsymbol{X}\boldsymbol{W}_i$$

$$\min_{p \in \Delta^{n-1}} \mathbb{E}_p[E] + \tau \mathbb{E}_p[\log p] \implies p = \mathrm{Softmax}(-E/\tau)$$

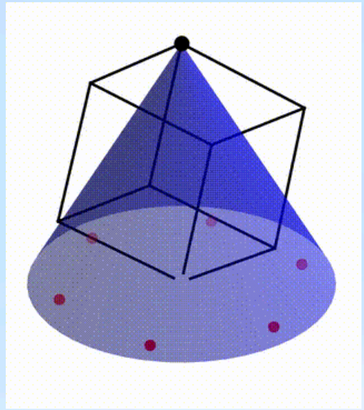
“Energy + Entropic Regularization”

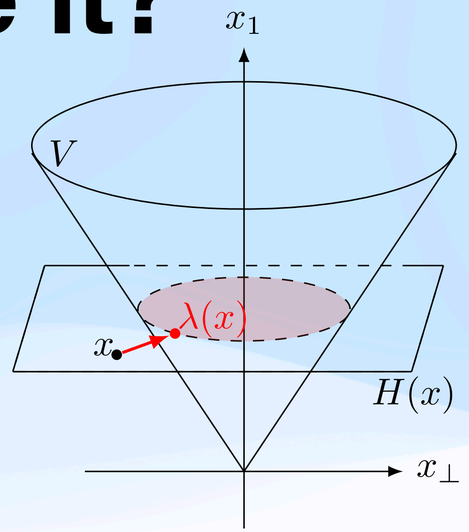
Q2: What About Activation Functions? (Previous Works)

Why ReLU-MLP = Projected Gradient Descent?
Or Even: How to improve it?



ReLU = Orthant Projection





CoLU

	Axiom	Property	Formulation
•	**Component-Wise**	$\forall i, \lambda\pi_i = \pi_i\lambda$	$\lambda(x_i) = \lambda(x)_i$
•	Positive 1-Homogeneity	$\forall t > 0, \lambda(tx) = t\lambda(x)$	$\lambda \mid_\Omega = \mathrm{id}$
•	Idempotence/Projection	$\lambda\lambda = \lambda$	$\lambda(\Omega^c) = \partial\Omega$

Permutation ❌

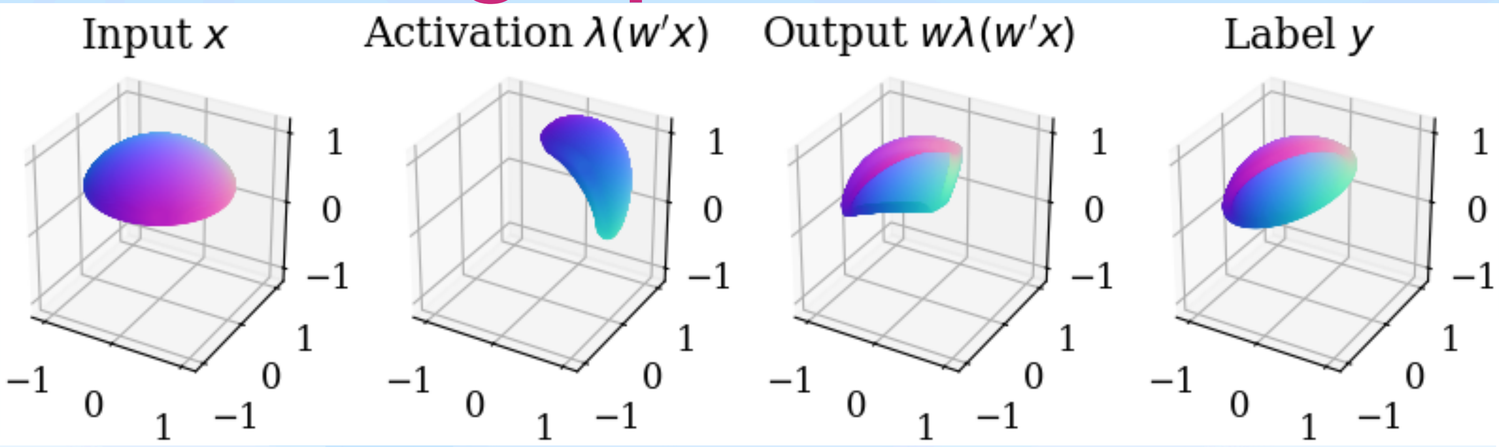
Permutation+Rotation ✅

PSD -> Conic Programming

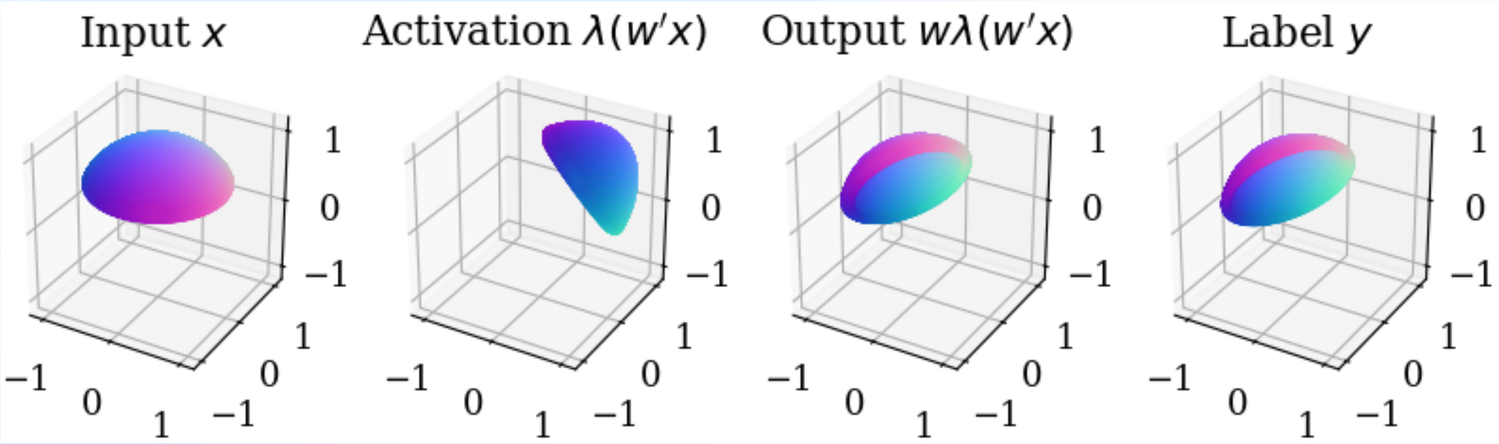
Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

Embedding Space Favors Orthogonal Symmetry

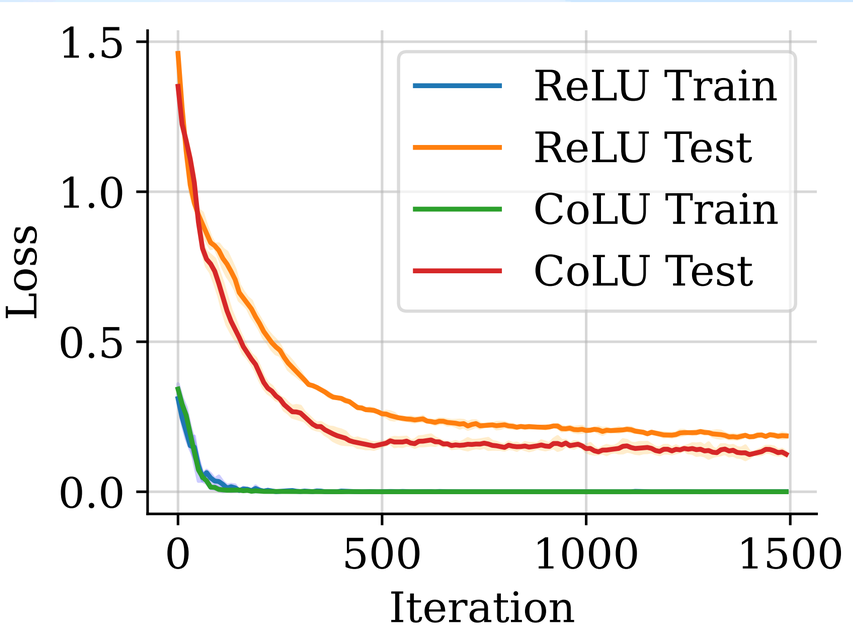


ReLU: permutation symmetry



CoLU: orthogonal/rotary symmetry

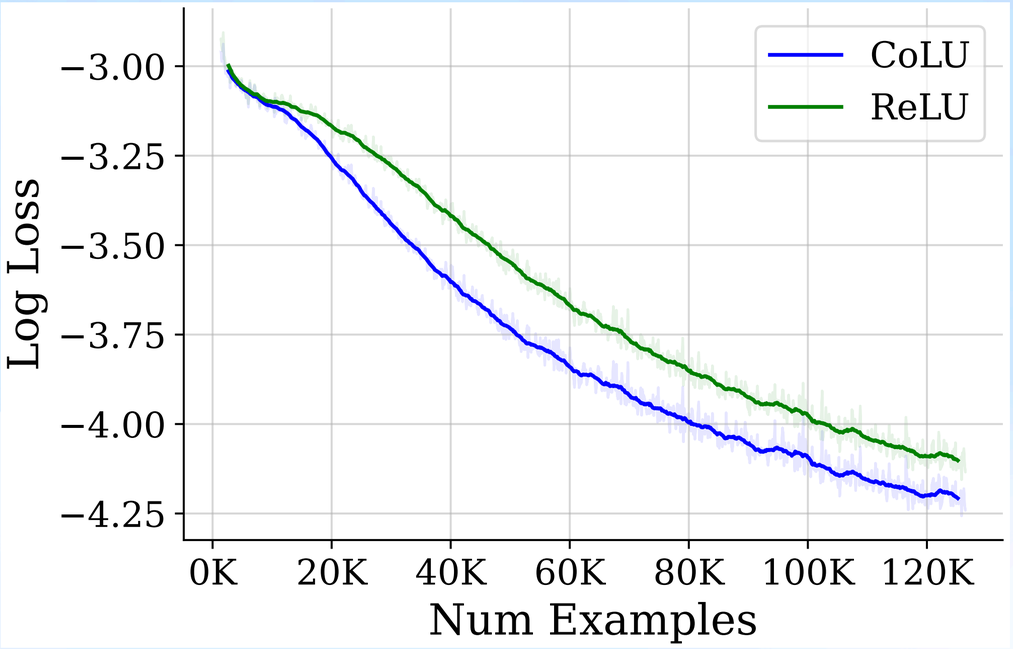
- Minimal Example (C=3)
- Improved Generalization



Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

Embedding Space Favors Orthogonal Symmetry



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. VISAPP 2024.
Conic activation functions. UniReps@NeurIPS 2024. PMLR 2025.

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

What happened inside?

Take Home Message: Transformer is not a mere engineering coincidence, but is grounded in principles from (quantum) physics.
Today's Goal: I don't see this idea ANYwhere after asking many people. Please correct it or guide me to any existing papers.

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

Email: evergreencqfu at gmail

Thank you!