Transformers Are Optimal Effective Fields

A Variational Theory of Architecture Design

Changqing Fu, PSL University, France

Problem: Deep Learning architectures are often heuristic. "Why Attention?" is unanswered.

Hypothesis: Good models have the least geometric structure necessary (Symmetry Breaking).

The Approach: Treat the layer depth t as time. The Transformer is the optimal path (geodesic) of a particle system minimizing an action $S(\boldsymbol{X})$.

Axioms

Universality: Taylor expansion of interactions.

Isometry Invariance: Energy depends on intrinsic geometry.

Learnable Hamiltonians: Symmetry is broken only by parameters.

Result

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

Output: The Euler-Lagrange Equation = ***The Transformer ODE*** $\dot{\boldsymbol{X}} = \mathrm{MLP}(\boldsymbol{X}) + \mathrm{Attention}(\boldsymbol{X})$

	Dynamics Order	Interactions
MLP	(Piecewise) Linear	No
LinearAttn	Cubic	Yes

Renormalization → NO HIGHER TERMS!

From Theory to Practice

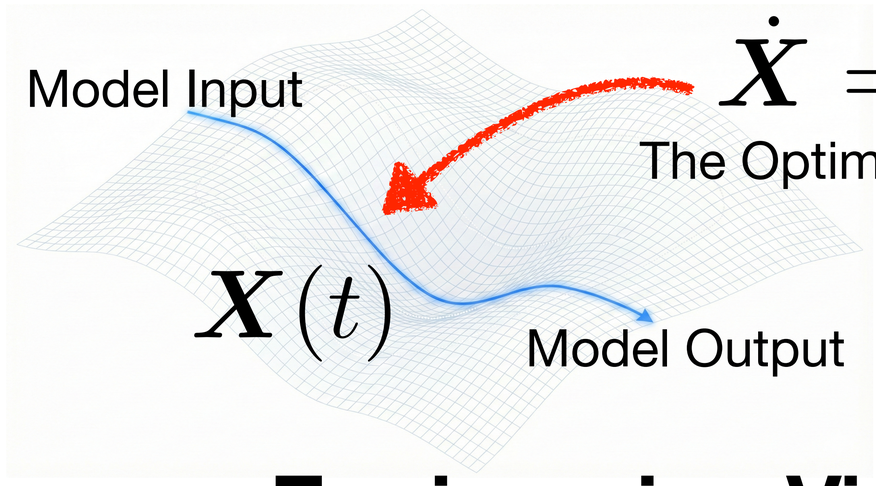
ReLU → Conic Projection [2]

GLU → Triple-Trace Potential

Softmax Attention → Free Energy Potential

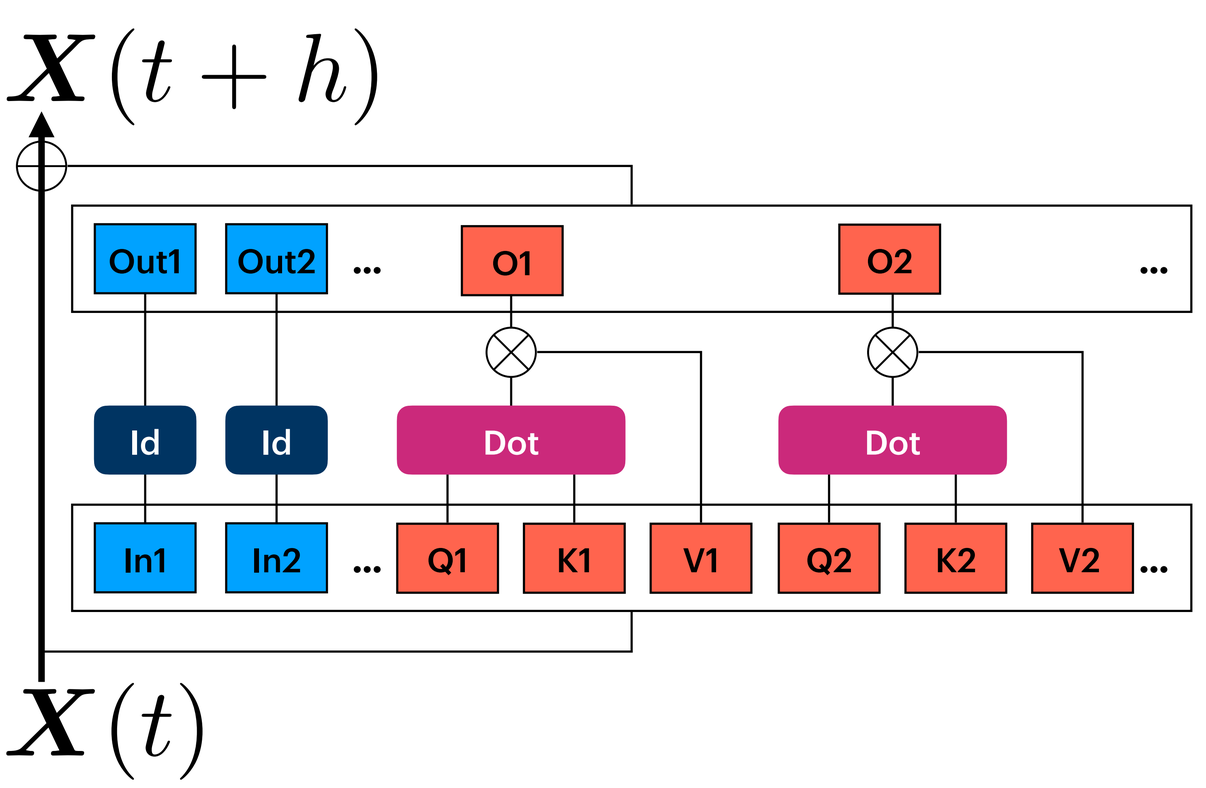
Sparse Attn → Non-commutative corrections

Physics View



Engineering View

$$\dot{\boldsymbol{X}} = \sum_i \boldsymbol{X}\boldsymbol{W}_{\mathrm{in},i}\boldsymbol{W}^*_{\mathrm{out},i} + \sum_i C^{-1/2}\boldsymbol{X}\boldsymbol{Q}_i\boldsymbol{K}_i^*\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{V}_i\boldsymbol{O}_i^*$$



The Geometry of Interaction

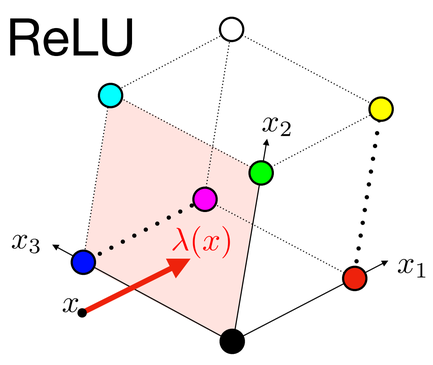
$$V(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{A}}\boldsymbol{X}^*)$$

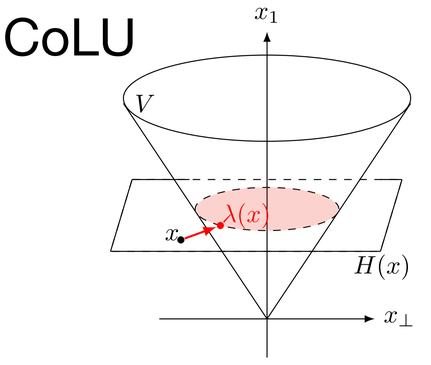
$$\boldsymbol{X}[1,:] = (\text{“The”})$$

$$V(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{A}}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{\mathrm{B}}\boldsymbol{X}^*)$$

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

The Geometry of Activation





Takeaway: This theory unifies GPT, ResNets, and Diffusion model architectures under a single Hamiltonian framework.

References:

[1]Transformers Are Optimal Effective Fields. PriGM@EurIPS, 2025.
[2]Conic Activation Functions. Fu and Cohen. UniReps@NeurIPS, PMLR 2024.

Paper