Transformers Are Optimal Effective Fields

Changqing Fu
changqingfu@changqingfu.com

Abstract

Are representations in the Transformer architecture provably optimal? We present an axiomatic theory of the Transformer architecture. First, we show that a complex-valued Transformer with linear attention and linear feed-forward residual blocks is uniquely defined by a potential field governed by leading linear and interactive terms. Then, we characterize a Softmax Transformer via a canonical axiomatic construction. The practical implications include a non-exhaustive unification of existing Transformer variants within a single formalism, and a principled foundation for future architecture search.

1 Introduction

Deriving the pivotal Transformer architecture and its many variants [Vaswani et al., 2017, Dehghani et al., 2019, Narang et al., 2021] in a constructive and systematic way naturally motivates the study of its variational form. From CNNs to Transformers, the interactions within the graph of tokens shift from local and deterministic to long-range and dynamic. Applying symmetries to CNNs has inspired geometric deep learning [Bronstein et al., 2021]. Likewise, it is necessary to relax geometric constraints on more principled domains whenever they are too restrictive, leaving enough flexibility for the model to learn. This echoes the principle: *Good models are those with the least geometric structures.* In the language of physics: *asymmetrical effects must have asymmetrical causes* [Curie, 1894]. In machine learning, this idea aligns with *the bitter lesson* [Sutton, 2019], which states that models tend to become less hand-designed over time. Another motivation arises from the continual increase in computing power (Moore's Law), which makes it feasible to implement generalist models that do not overfit across a wide variety of test data. Growing computational capacity, and its associated power consumption, encourages algorithm designers to allocate computation within a model as efficiently as possible. Developing a foundational theory of Transformers thus remains an open question of great theoretical and practical values.

2 Background

Notation Let $[n] = \{0, 1, \dots, n-1\}$ be the set of n indices. The discrete state is written as $x(t, \omega, \sigma) \in \mathbb{F}$, where $t \in \mathcal{T} := [T]$ is the discrete layer index, $\omega \in \Omega := [N]$ indexes particles (words/pixels), and $\sigma \in \Sigma := [C]$ indexes neurons or vector field components. The number field $\mathbb{F}$ can either be real $\mathbb{R}$ or complex $\mathbb{C}$. The symbols T, N, and C denote the depth, the sequence length (number of particles), and the feature space dimension (network width), respectively. In a continuous-layer setting we inherit the notation $\mathcal{T} := [0, 1]$, and write $\dot{x} := \frac{\mathrm{d}x}{\mathrm{d}t}$ for the layer-wise derivative and omit the layer as an implicit variable. For brevity, we assume Σ, Ω are finite dimensional and denote the discrete state as a layer-dependent matrix $\boldsymbol{X} \in \mathbb{F}^{N \times C}$ whose row and column entries are $\boldsymbol{X}_{\omega\sigma} := x(\omega, \sigma)$, and denote each particle at $\omega \in \Omega$ as a vector $x_\omega \in \mathbb{F}^C$ in the feature space whose components are $(x_\omega)_\sigma := x(\omega, \sigma)$. The parameters are the layer-dependent matrices $\boldsymbol{W} = \boldsymbol{W}(t) \in \mathbb{F}^{C \times C}$. We use $\odot$ for element-wise (Hadamard) operations. The Lie bracket $[f, g] := fg - gf$ denotes the commutator. $\boldsymbol{W}^\top$ and $\boldsymbol{W}^*$ are the real and conjugate (Hermitian) transposes of $\boldsymbol{W}$ respectively. $\|\cdot\|_\mathrm{F}$ denotes the Frobenius norm. $U(C)$ is the C-dimensional unitary

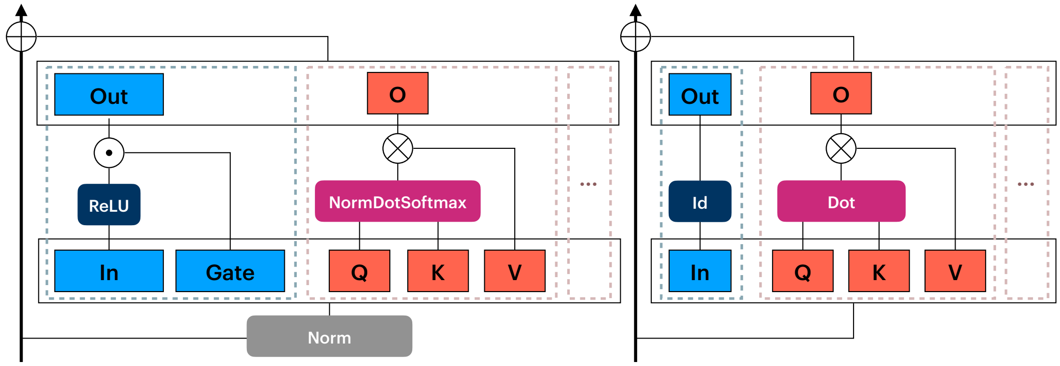

Figure 1: **Left:** a practical parallelized Transformer block with GLU activation. The wide MLP can be equivalently separated into multiple narrow heads. **Right:** a simplified Transformer block eliminating activation functions, normalizations and softmax.

group, S_N is the N-dimensional permutation group, and $C_{U(C)}(\boldsymbol{W}) := \{\boldsymbol{V} \in U(C) : [\boldsymbol{V}, \boldsymbol{W}] = 0\}$ is the centralizer of $\boldsymbol{W}$ within $U(C)$.

2.1 Transformer as ODE

Figure 1 (left) visualizes the architecture of a parallelized Transformer [Dehghani et al., 2023] as a two-layer ResNet with mixed nonlinearities. Suppose the parameters of the i-th attention head are $\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i, \boldsymbol{O}_i \in \mathbb{F}^{C \times C_{\mathrm{A}}}$, the parameters of the MLP are $\boldsymbol{W}_{\mathrm{gate}}, \boldsymbol{W}_{\mathrm{in}}, \boldsymbol{W}_{\mathrm{out}} \in \mathbb{F}^{C \times C_{\mathrm{MLP}}}$, and the number field is real $\mathbb{F} = \mathbb{R}$, we define the model with normalization omitted:

Definition 1 (Transformers)**.** The ODE form of a Transformer is defined as

$$\dot{\boldsymbol{X}} = ((\boldsymbol{X}\boldsymbol{W}_{\mathrm{gate}}) \odot \mathrm{ReLU}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{in}}))\boldsymbol{W}_{\mathrm{out}} + \sum_{i=1}^{n} \mathrm{Softmax}(C_{\mathrm{A}}^{-1/2}\boldsymbol{X}\boldsymbol{Q}_i\boldsymbol{K}_i^\top\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{V}_i\boldsymbol{O}_i^\top. \tag{1}$$

It can be more concisely written as $\dot{\boldsymbol{X}} = \lambda(\boldsymbol{X}\boldsymbol{W}_{\mathrm{all}})\boldsymbol{W}'_{\mathrm{all}}$ where $\lambda := \lambda_{\mathrm{MLP}} \oplus \lambda_{\mathrm{A}}$, or in expanded form, $\dot{\boldsymbol{X}} = \lambda_{\mathrm{MLP}}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{MLP}})\boldsymbol{W}'_{\mathrm{MLP}} + \lambda_{\mathrm{A}}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{A}})\boldsymbol{W}'_{\mathrm{A}}$, where $\boldsymbol{W}_{\mathrm{all}} := [\boldsymbol{W}_{\mathrm{MLP}} \mid \boldsymbol{W}_{\mathrm{A}}] = [\boldsymbol{W}_{\mathrm{gate}} \mid \boldsymbol{W}_{\mathrm{in}} \mid \boldsymbol{Q}_1 \mid \boldsymbol{K}_1 \mid \boldsymbol{V}_1 \mid \dots \mid \boldsymbol{Q}_n \mid \boldsymbol{K}_n \mid \boldsymbol{V}_n]$ and $\boldsymbol{W}'_{\mathrm{all}} := [\boldsymbol{W}'_{\mathrm{MLP}} \mid \boldsymbol{W}'_{\mathrm{A}}] = [\boldsymbol{W}_{\mathrm{out}} \mid \boldsymbol{O}_1 \mid \dots \mid \boldsymbol{O}_n]$.

The more common Transformer architecture is a discrete-layer operator-splitting scheme of this ODE. The split scheme is approximately equal to the parallelized scheme when the nonlinearities commute. Otherwise they are approximately equal when the model is deep.

2.2 Comparison Between Parallelized and Split Transformers

The discrete layer is a forward (Euler) step $\boldsymbol{X}(t+1) = \boldsymbol{X}(t) + \lambda(\boldsymbol{X}(t)\boldsymbol{W}(t))\boldsymbol{W}'(t)$ and the following approximation links between parallelized and common Transformers.

$$(1 + h\lambda_{\mathrm{MLP}} + h\lambda_{\mathrm{A}}) \approx e^{h\lambda_{\mathrm{MLP}} + h\lambda_{\mathrm{A}}} \approx e^{h\lambda_{\mathrm{MLP}}}e^{h\lambda_{\mathrm{A}}} \approx (1 + h\lambda_{\mathrm{MLP}})(1 + h\lambda_{\mathrm{A}}) \tag{2}$$

In equation 2, the left/right hand side is a layer of the parallelized/split Transformer respectively. The first and last approximations becomes equal as $h \to 0$ whereas the second equality holds if and only if λ_{MLP} and λ_{A} commute ($[\lambda_{\mathrm{MLP}}, \lambda_{\mathrm{A}}] = 0$). In the general case when the ODE generators don't commute, deeper models prove to be beneficial in mixing the nonlinear terms, summarized as Lemma 2, as a consequence of Lemma 1.

Lemma 1 (Lie-Trotter)**.** *The equality holds:* $e^{\lambda_{MLP} + \lambda_A} = \lim_{T\to\infty}(e^{\frac{\lambda_{MLP}}{T}}e^{\frac{\lambda_A}{T}})^T$.

Proof. Applying the Baker-Campbell-Hausdorff expansion [Varadarajan, 2013] $e^{\lambda_{\mathrm{MLP}}t}e^{\lambda_{\mathrm{A}}t} = e^{(\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{A}})t + \frac{1}{2}[\lambda_{\mathrm{MLP}}, \lambda_{\mathrm{A}}]t^2 + O(t^3)}$ gives the leading non-commutative correction. □

Lemma 2 (The Benefit of Depth)**.** *Suppose* $\lambda_{MLP}^{\boldsymbol{W}(t)}(\boldsymbol{X}) \coloneqq \lambda_{MLP}(\boldsymbol{X}\boldsymbol{W}(t))\boldsymbol{W}'(t)$ *and* $\lambda_{A}^{\boldsymbol{W}(t)}(\boldsymbol{X}) \coloneqq \lambda_{A}(\boldsymbol{X}\boldsymbol{W}(t))\boldsymbol{W}'(t)$*, then the parallel and the split schemes asymptotically coincide:*

$$\lim_{T\to\infty}\prod_{t=0}^{T-1}(1+\lambda_{MLP}^{\boldsymbol{W}(t/T)})(1+\lambda_{A}^{\boldsymbol{W}(t/T)}) = \lim_{T\to\infty}\prod_{t=0}^{T-1}(1+\lambda_{MLP}^{\boldsymbol{W}(t/T)}+\lambda_{A}^{\boldsymbol{W}(t/T)}).$$

2.3 Variational Formulation

To formulate the optimality of the ODE, we recall its potential/action form.

Definition 2 (Potential/Action)**.** The potential, or potential field, of an ODE (whenever exists) is a function $V : \mathbb{C}^{N\times C} \to \mathbb{C}$ such that the ODE coincides with the Euler-Lagrange equation $\mathrm{i}\dot{\boldsymbol{X}} = \frac{\partial V}{\partial \boldsymbol{X}^*}$ (or equivalently $-\mathrm{i}\dot{\boldsymbol{X}}^* = \frac{\partial V}{\partial \boldsymbol{X}}$). It extremizes the action of the path $\boldsymbol{X} : [0,1] \to \mathbb{C}^{N\times C}$ defined as

$$S(\boldsymbol{X}) \coloneqq \int_0^1 \mathrm{Tr}(\frac{\mathrm{i}}{2}(\dot{\boldsymbol{X}}\boldsymbol{X}^* - \boldsymbol{X}\dot{\boldsymbol{X}}^*) - V(\boldsymbol{X}))\,\mathrm{d}t. \tag{3}$$

The real and imaginary parts of V are the conservative and dissipative parts respectively.

3 Optimality of Transformers

We first study a simplified complex ODE whose potential exists, visualized in Figure 1 (right), and restrict to the real case in equation 1 for computational convenience.

3.1 Linearized Transformers

Consider the complex-valued case $\mathbb{F} = \mathbb{C}$ (use the Hermitian transpose instead of transpose), and consider a linearized Transformer without nonlinear functions (activation functions, normalization, softmax, etc.) and without gating layers.

Definition 3 (Linearized Transformers)**.** A linearized Transformer ODE is defined as the equation

$$\dot{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{W}_{\text{in}}\boldsymbol{W}_{\text{out}}^* + \sum_{i=1}^{n} C_{\text{A}}^{-1/2}\boldsymbol{X}\boldsymbol{Q}_i\boldsymbol{K}_i^*\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{V}_i\boldsymbol{O}_i^*. \tag{4}$$

From this definition, we can convert the ODE into a potential field form by the following lemma.

Lemma 3 (Linearized Transformers As Fields)**.** *Suppose* n *is even, and then there exists parameters* $\boldsymbol{W}_{in}, \boldsymbol{W}_{out}, \boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i$ *such that equation 4 is associated with the field*

$$V(\boldsymbol{X}) = \frac{\mathrm{i}}{2}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*) + \frac{\mathrm{i}}{2\sqrt{C_A}}\sum_{i=1}^{n/2}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}_{Ai}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{Bi}^*\boldsymbol{X}^*).$$

Proof. Calculate $\partial V/\partial \boldsymbol{X}^* = \mathrm{i}\boldsymbol{X}\boldsymbol{W} + \frac{i}{\sqrt{C_\text{A}}}(\sum_{i=1}^{n/2}\boldsymbol{X}\boldsymbol{W}_{\text{A}i}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{\text{B}i}^*\boldsymbol{X}^* + \boldsymbol{X}\boldsymbol{W}_{\text{B}i}^*\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{\text{A}i}^*\boldsymbol{X}^*)$, then take $\boldsymbol{W} = \boldsymbol{W}_{\text{in}}\boldsymbol{W}_{\text{out}}^*$, $\boldsymbol{W}_{\text{A}i} = \boldsymbol{Q}_i\boldsymbol{K}_i^*$, $\boldsymbol{W}_{\text{B}i} = \boldsymbol{V}_i\boldsymbol{O}_i^*$. □

Note that $\boldsymbol{Q}_i\boldsymbol{K}_i^* = \boldsymbol{V}_{i+n/2}\boldsymbol{O}_{i+n/2}^*$ and $\boldsymbol{Q}_{i+n/2}\boldsymbol{K}_{i+n/2}^* = \boldsymbol{V}_i\boldsymbol{O}_i^*$, otherwise the potential does not exist, and we leave the general case as a future work. Likewise, we may also recover the softmax function out of the log-sum-exp potential.

3.2 Softmax, ReLU, and GLU Transformers

Lemma 4 (Softmax Transformers)**.** *Suppose* n *is even. There exists parameters such that the ODE* $\dot{\boldsymbol{X}} = \sum_{i=1}^{n}\mathrm{Softmax}(C_A^{-1/2}\boldsymbol{X}\boldsymbol{Q}_i\boldsymbol{K}_i^*\boldsymbol{X}^*)\boldsymbol{X}\boldsymbol{V}_i\boldsymbol{O}_i^*$ *is associated with the field* $V(\boldsymbol{X}) = \frac{\mathrm{i}}{2}\sum_{i=1}^{n/2}\sum_{j=1}^{N}\log\sum_{k=1}^{N}\exp([\frac{1}{\sqrt{C_A}}\boldsymbol{X}\boldsymbol{W}_{Ai}\boldsymbol{X}^*]_{jk})$.

Proof. $\partial V/\partial \boldsymbol{X}^* = \frac{\mathrm{i}}{2}\sum_{i=1}^{n/2}\mathrm{Softmax}(C_\text{A}^{-1/2}\boldsymbol{X}\boldsymbol{W}_{\text{A}i}\boldsymbol{X}^*)\boldsymbol{X}\boldsymbol{W}_{\text{A}i}$, and take $\boldsymbol{W}_{\text{A}i} = \boldsymbol{Q}_i\boldsymbol{K}_i^* = \boldsymbol{V}_i\boldsymbol{O}_i^*$. □

Note that the potential exists only when $\boldsymbol{Q}_i\boldsymbol{K}_i^* = \boldsymbol{V}_i\boldsymbol{O}_i^* = (\boldsymbol{Q}_{i+n/2}\boldsymbol{K}_{i+n/2}^*)^* = (\boldsymbol{V}_{i+n/2}\boldsymbol{O}_{i+n/2}^*)^*$, and we leave the general case as a future work. As a remark, the gated MLP [Shazeer, 2020] without ReLU $\dot{\boldsymbol{X}} = (\boldsymbol{X}\boldsymbol{W}_{\text{gate}}) \odot (\boldsymbol{X}\boldsymbol{W}_{\text{in}})\boldsymbol{W}_{\text{out}}$ can be recovered from $V(\boldsymbol{X}) = -\frac{1}{3}\sum_{jk}((\boldsymbol{X}\boldsymbol{W}_{\text{gate}}) \odot (\boldsymbol{X}\boldsymbol{W}_{\text{in}}) \odot (\boldsymbol{X}\boldsymbol{W}_{\text{out}}))_{jk}$ and $\dot{\boldsymbol{X}} = -\partial V/\partial \boldsymbol{X}$. We may also recover the ReLU-activated ODE as a projected gradient flow, regarding $\mathrm{ReLU}(\boldsymbol{X}) = \Pi_{\mathbb{R}_+}(\boldsymbol{X}) \coloneqq \min_{\boldsymbol{Y}\in\mathbb{R}_+}\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}^2$ as the projection operator and $\Pi_{\boldsymbol{W}_{\text{MLP}}}(\boldsymbol{X}) \coloneqq (\boldsymbol{X}\boldsymbol{W}_{\text{in}})_+\boldsymbol{W}_{\text{out}}^*$ as the projection onto an affine orthant parameterized by $\boldsymbol{W}_{\text{in}}, \boldsymbol{W}_{\text{out}}$. The ReLU^2 activation can then be recovered by tying the input and gating parameters.

3.3 Optimality of Matrix Potentials

To formalize in what sense the potentials are optimal, we assume necessary axioms on the possible forms of a potential field V of a matrix $\boldsymbol{X}$:

Axiom 1 (Analyticity)**.** $V(\boldsymbol{X})$ is analytic in $(\mathrm{Re}\boldsymbol{X}, \mathrm{Im}\boldsymbol{X})$.

Axiom 2 (Space Homogeneity)**.** At each t, $\forall \boldsymbol{P} \in S_N$ (permutation), there holds $V(\boldsymbol{P}\boldsymbol{X}) = V(\boldsymbol{X})$.

Axiom 3 (Feature Symmetries)**.** At each t, $\exists \boldsymbol{W}_1 \dots \boldsymbol{W}_n \in \mathbb{C}^{C\times C}$ such that $\forall \boldsymbol{R} \in \bigcap_{i=1}^n C_{U(C)}(\boldsymbol{W}_i)$, there holds $V(\boldsymbol{X}\boldsymbol{R}^*) = V(\boldsymbol{X})$.

Axiom 4 (Optional, Low-Rankedness)**.** $C_{\text{A}} \coloneqq \max_{1\le i\le n} \mathrm{Rank}(\boldsymbol{W}_i) \ll C$.

Intuitions Axiom 1 guarantees a smooth model which separates into multiple scales. Axiom 2 ensures exchangeability among N words and Axiom 3 ensures symmetries among a set of C_{A} learnable $U(1)$ symmetries, represented by the centralizer of parameter matrices $\boldsymbol{W}$. Note the gated MLP does not satisfy Axiom 3. Axiom 4 is optional but practical: from the symmetry-breaking principle [Curie, 1894], we assume that the number of such $U(1)$ symmetries is small.

Lemma 5 (Canonical Form of Matrix Fields)**.** *Under Axioms 1, 2, and 3, the matrix potential takes the form of* $V(\boldsymbol{X}) = \mathrm{Tr} f(\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{X}^*, \dots, \boldsymbol{X}\boldsymbol{W}_n\boldsymbol{X}^*)$ *for some analytic spectral function* $f : \mathbb{C}^n \to \mathbb{C}$.

Proof. First we check that the axioms hold for the form of V: for all $\boldsymbol{P}, \boldsymbol{R}$ in the condition, by definitions of unitary group and centralizers, we have $V(\boldsymbol{P}\boldsymbol{X}\boldsymbol{R}^*) = \mathrm{Tr} f(\{\boldsymbol{P}\boldsymbol{X}\boldsymbol{R}^*\boldsymbol{W}_i\boldsymbol{R}\boldsymbol{X}^*\boldsymbol{P}^*\}_{i=1}^n) = \mathrm{Tr} f(\{\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*\}_{i=1}^n) = V(\boldsymbol{X})$. Next we show the uniqueness of the form of V. By Axiom 1, the matrix potential has an expansion form $V(\boldsymbol{X}) = f_1(\boldsymbol{X})$ for some $f_1 : \mathbb{C}^{N\times C} \to \mathbb{C}$. By Axiom 2, the potential $V(\boldsymbol{X}) = f_2(\boldsymbol{X}^*\boldsymbol{X})$ for some $f_2 : \mathbb{C}^{C\times C} \to \mathbb{C}$. By Axiom 3, the potential $V(\boldsymbol{X}) = \mathrm{Tr} f(\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{X}^*, \boldsymbol{X}\boldsymbol{W}_2\boldsymbol{X}^*, \dots, \boldsymbol{X}\boldsymbol{W}_n\boldsymbol{X}^*)$ for $f : \mathbb{C} \to \mathbb{C}$. □

Theorem 6 (Optimality of Linearized Transformers)**.** *Under Axioms 1, 2, 3, Definition 3 is the first nontrivial interaction ODE.*

Proof. By Lemma 5, the leading terms of the analytic expansion of V are written as $V(\boldsymbol{X}) = \mathrm{Tr}(c_0\boldsymbol{I} + \sum_{i=1}^n c_{1,i}\boldsymbol{X}\boldsymbol{W}_{1,i}\boldsymbol{X}^* + \sum_{i=1}^n c_{2,i}\boldsymbol{X}\boldsymbol{W}_{2,i}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{3,i}\boldsymbol{X}^* + O(\|\boldsymbol{X}\|^6))$ where $c_0, c_{1,i}, c_{2,i}$ are coefficients of f, $\boldsymbol{W}_{\text{in},i}, \boldsymbol{W}_{\text{out},i}, \boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i, \boldsymbol{O}_i \in \mathbb{C}^{C\times C_{\text{A}}}$ and $C_{\text{MLP}} = nC_{\text{A}}$. The minimality vanishes the remainders $O(\|\boldsymbol{X}\|^6)) = 0$. The MLP term can be fused into $\sum_{i=1}^n \boldsymbol{X}\boldsymbol{W}_{\text{in},i}\boldsymbol{W}_{\text{out},i}^*\boldsymbol{X}^* \eqqcolon \boldsymbol{X}\boldsymbol{W}_{\text{in}}\boldsymbol{W}_{\text{out}}\boldsymbol{X}^*$ where $\boldsymbol{W}_{\text{in}}, \boldsymbol{W}_{\text{out}} \in \mathbb{C}^{C\times(nC_{\text{A}})}$. We conclude by applying Lemma 3. □

The implication of Theorem 6 is that the linearized Transformer in equation 4 is a minimal model, i.e. it is the *effective* interactive field. Softmax attention equation 1 from the log-sum-exp potential and many variants (powers, sigmoid, etc.) also satisfies the canonical form in Lemma 5. Finally, we comment that the width of the MLP and attention matrices are a result of the low-rank assumption.

Lemma 7 (Low-Rankness)**.** *Under Axiom 4, if* n *is sufficiently large, then* $C_A \ll C < C_{MLP}$.

Proof. $C_{\text{A}} \ll C$ comes directly from Axiom 4. Fix $C_{\text{A}}, C < nC_{\text{A}} = C_{\text{MLP}}$ when n is large. □

4 Conclusion and Perspectives

We have shown a construction of the Transformer ODE from variational principles. A limitation of this paper is axiom dependency: non-attentional models (MLP-Mixers, State Space Models, etc) that violate the symmetries require different axioms. Masked/Sparse attention (especially attentional masks [DeepSeek-AI, 2025]) and mixture-of-expert MLPs are left for future works.

References

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Pierre Curie. Sur la symétrie dans les phénomènes physiques, symétrie d'un champ électrique et d'un champ magnétique. *Journal de physique théorique et appliquée*, 3(1):393–415, 1894.

DeepSeek-AI. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention, 2025. URL https://github.com/deepseek-ai/DeepSeek-V3.2-Exp/blob/main/DeepSeek_V3_2.pdf.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.

Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Févry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. Do transformer modifications transfer across implementations and applications? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5758–5773, 2021.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Richard Sutton. The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html, 2019. Accessed: 2025-10-18.

Veeravalli S. Varadarajan. *Lie Groups, Lie Algebras, and Their Representations*, volume 102. Springer, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A Expanded Proofs

Lemma 1 (Lie-Trotter)**.** *The equality holds:* $e^{\lambda_{MLP}+\lambda_A} = \lim_{T\to\infty}(e^{\frac{\lambda_{MLP}}{T}} e^{\frac{\lambda_A}{T}})^T$.

Proof. This follows from the Baker-Campbell-Hausdorff (BCH) formula [Varadarajan, 2013]. The BCH expansion for the product of exponentials is given by $e^{\lambda_{\text{MLP}} t} e^{\lambda_{\text{A}} t} = e^{(\lambda_{\text{MLP}}+\lambda_{\text{A}})t+\frac{1}{2}[\lambda_{\text{MLP}},\lambda_{\text{A}}]t^2+O(t^3)}$, where $[\lambda_{\text{MLP}}, \lambda_{\text{A}}] = \lambda_{\text{MLP}}\lambda_{\text{A}} - \lambda_{\text{A}}\lambda_{\text{MLP}}$ is the commutator. Substituting $t = 1/T$, the product becomes $e^{\frac{\lambda_{\text{MLP}}}{T}} e^{\frac{\lambda_{\text{A}}}{T}} = e^{\frac{\lambda_{\text{MLP}}+\lambda_{\text{A}}}{T}+\frac{1}{2}[\lambda_{\text{MLP}},\lambda_{\text{A}}]\frac{1}{T^2}+O(\frac{1}{T^3})}$. Raising this to the T-th power, $(e^{\frac{\lambda_{\text{MLP}}+\lambda_{\text{A}}}{T}+\frac{1}{2}[\lambda_{\text{MLP}},\lambda_{\text{A}}]\frac{1}{T^2}+O(\frac{1}{T^3})})^T = e^{(\lambda_{\text{MLP}}+\lambda_{\text{A}})+\frac{1}{2}[\lambda_{\text{MLP}},\lambda_{\text{A}}]\frac{1}{T}+O(\frac{1}{T^2})}$. As $T \to \infty$, the higher-order terms vanish, and the limit converges to $e^{\lambda_{\text{MLP}}+\lambda_{\text{A}}}$, proving the lemma. $\square$

Lemma 2 (The Benefit of Depth)**.** *Suppose* $\lambda_{MLP}^{\boldsymbol{W}(t)}(\boldsymbol{X}) := \lambda_{MLP}(\boldsymbol{X}\boldsymbol{W}(t))\boldsymbol{W}'(t)$ *and* $\lambda_{A}^{\boldsymbol{W}(t)}(\boldsymbol{X}) := \lambda_A(\boldsymbol{X}\boldsymbol{W}(t))\boldsymbol{W}'(t)$*, then the parallel and the split schemes asymptotically coincide:*

$$\lim_{T\to\infty}\prod_{t=0}^{T-1}(1+\lambda_{MLP}^{\boldsymbol{W}(t/T)})(1+\lambda_{A}^{\boldsymbol{W}(t/T)}) = \lim_{T\to\infty}\prod_{t=0}^{T-1}(1+\lambda_{MLP}^{\boldsymbol{W}(t/T)}+\lambda_{A}^{\boldsymbol{W}(t/T)}).$$

Proof. This follows from Lemma 1 and the convergence of discretization schemes for the ODE $\dot{\boldsymbol{X}} = \lambda_{\text{MLP}}^{\boldsymbol{W}(t)}(\boldsymbol{X}) + \lambda_{\text{A}}^{\boldsymbol{W}(t)}(\boldsymbol{X})$, where the parameters $\boldsymbol{W}(t)$ make the operators time-dependent. The parallel scheme is the standard Euler discretization with step size $1/T$: each term is $1 + (\lambda_{\text{MLP}}^{\boldsymbol{W}(t/T)} + \lambda_{\text{A}}^{\boldsymbol{W}(t/T)})/T + O(1/T^2)$, and the product over T steps approximates the time-ordered exponential $\exp\int_0^1(\lambda_{\text{MLP}}^{\boldsymbol{W}(t)}(\boldsymbol{X}) + \lambda_{\text{A}}^{\boldsymbol{W}(t)}(\boldsymbol{X}))dt$ as $T\to\infty$, with error $O(1/T)$. The split scheme is a split-step Euler discretization: each term is $(1+\lambda_{\text{MLP}}^{\boldsymbol{W}(t/T)}/T)(1+\lambda_{\text{A}}^{\boldsymbol{W}(t/T)}/T) = 1 + (\lambda_{\text{MLP}}^{\boldsymbol{W}(t/T)} + \lambda_{\text{A}}^{\boldsymbol{W}(t/T)})/T + (\lambda_{\text{MLP}}^{\boldsymbol{W}(t/T)}\lambda_{\text{A}}^{\boldsymbol{W}(t/T)})/T^2 + O(1/T^2)$, and the product approximates the same time-ordered exponential by the Lie-Trotter formula (Lemma 1), with the extra commutative error term $(\lambda_{\text{MLP}}^{\boldsymbol{W}(t/T)}\lambda_{\text{A}}^{\boldsymbol{W}(t/T)})/T$ vanishing as $T\to\infty$ (total error $O(1/T)$). Thus, both schemes converge to the same limit. $\square$

Lemma 3 (Linearized Transformers As Fields)**.** *Suppose n is even, and then there exists parameters* $\boldsymbol{W}_{in}, \boldsymbol{W}_{out}, \boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i$ *such that equation 4 is associated with the field*

$$V(\boldsymbol{X}) = \frac{\mathrm{i}}{2}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*) + \frac{\mathrm{i}}{2\sqrt{C_A}}\sum_{i=1}^{n/2}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}_{Ai}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{Bi}^*\boldsymbol{X}^*).$$

Proof. We separate the potential into linear and interaction terms $V = V_{\text{lin}} + V_{\text{int}}$, and compute the Wirtinger derivative $\partial V/\partial\boldsymbol{X}^*$ to show it matches $\mathrm{i}\dot{\boldsymbol{X}}$. Treat $\boldsymbol{X}$ and $\boldsymbol{X}^*$ as independent variables.

For the linear term, the differential $\mathrm{d}V_{\text{lin}} = \frac{\mathrm{i}}{2}[\mathrm{Tr}(\mathrm{d}\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*) + \mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\,\mathrm{d}\boldsymbol{X}^*)]$. The term with $\mathrm{d}\boldsymbol{X}^*$ is $\frac{\mathrm{i}}{2}\mathrm{Tr}(\boldsymbol{X}\boldsymbol{W}\,\mathrm{d}\boldsymbol{X}^*) = \frac{\mathrm{i}}{2}\mathrm{Tr}(\mathrm{d}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W})$, so $\partial V_{\text{lin}}/\partial\boldsymbol{X}^* = \frac{\mathrm{i}}{2}\boldsymbol{X}\boldsymbol{W}$.

For the interaction term, consider one term $\mathrm{Tr}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^*)$. The differential is $\mathrm{Tr}(\mathrm{d}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^*) + \mathrm{Tr}(\boldsymbol{X}\boldsymbol{A}\,\mathrm{d}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^*) + \mathrm{Tr}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^*\,\mathrm{d}\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^*) + \mathrm{Tr}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}\,\mathrm{d}\boldsymbol{X}^*)$. The terms with $\mathrm{d}\boldsymbol{X}^*$ are $\mathrm{Tr}(\boldsymbol{X}\boldsymbol{A}\,\mathrm{d}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^*) = \mathrm{Tr}(\mathrm{d}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{A})$ and $\mathrm{Tr}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}\,\mathrm{d}\boldsymbol{X}^*) = \mathrm{Tr}(\mathrm{d}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B})$. Thus, the contribution to $\partial V_{\text{int}}/\partial\boldsymbol{X}^*$ is $\frac{\mathrm{i}}{2\sqrt{C_{\text{A}}}}\sum_{i=1}^{n/2}(\boldsymbol{X}\boldsymbol{W}_{\text{B}i}^*\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{\text{A}i} + \boldsymbol{X}\boldsymbol{W}_{\text{A}i}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_{\text{B}i}^*)$.

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

Now take $\boldsymbol{W} = \boldsymbol{W}_{\text{in}}\boldsymbol{W}_{\text{out}}^*$, $\boldsymbol{W}_{\text{A}i} = \boldsymbol{Q}_i\boldsymbol{K}_i^*$, $\boldsymbol{W}_{\text{B}i} = \boldsymbol{V}_i\boldsymbol{O}_i^*$. With the pairing condition $\boldsymbol{Q}_i\boldsymbol{K}_i^* = \boldsymbol{V}_{i+n/2}\boldsymbol{O}_{i+n/2}^*$ and $\boldsymbol{Q}_{i+n/2}\boldsymbol{K}_{i+n/2}^* = \boldsymbol{V}_i\boldsymbol{O}_i^*$ (ensuring symmetry), the derivative matches $\mathrm{i}\dot{\boldsymbol{X}}$ for the ODE after adjusting coefficients to account for the pairing. $\square$

Lemma 4 (Softmax Transformers)**.** *Suppose n is even. There exists parameters such that the ODE* $\dot{\boldsymbol{X}} = \sum_{i=1}^{n}\mathrm{Softmax}(C_A^{-1/2}\boldsymbol{X}\boldsymbol{Q}_i\boldsymbol{K}_i^*\boldsymbol{X}^*)\boldsymbol{X}\boldsymbol{V}_i\boldsymbol{O}_i^*$ *is associated with the field* $V(\boldsymbol{X}) = \frac{\mathrm{i}}{2}\sum_{i=1}^{n/2}\sum_{j=1}^{N}\log\sum_{k=1}^{N}\exp_{jk}^{\odot}(\frac{1}{\sqrt{C_A}}\boldsymbol{X}\boldsymbol{W}_{Ai}\boldsymbol{X}^*)$.

Proof. Let $y_{jk} := (\frac{1}{2\sqrt{C_\text{A}}}\boldsymbol{X}\boldsymbol{W}_{\text{A}i}\boldsymbol{X}^*)_{jk}$. The log-sum-exp function has gradient softmax: $\partial(\log\sum_k \exp y_{jk})/\partial y_{jk} = \exp y_{jk}/\sum_k \exp y_{jk} = \text{Softmax}(y_j)$ for $j = 1, \dots, N$. The differential of the term $\log\sum_k \exp y_{jk}$ involves $\partial/\partial \boldsymbol{X}^* y_{jk}$, which is similar to the interaction term in Lemma 3 but element-wise. The full derivative is $\mathrm{i}\sum_{i=1}^{n/2} \text{Softmax}(C_\text{A}^{-1/2}\boldsymbol{X}\boldsymbol{W}_{\text{A}i}\boldsymbol{X}^*)\boldsymbol{X}\boldsymbol{W}_{\text{A}i}$. Take $\boldsymbol{W}_{\text{A}i} = \boldsymbol{Q}_i\boldsymbol{K}_i^* = \boldsymbol{V}_i\boldsymbol{O}_i^*$, then the pairing condition $\boldsymbol{Q}_i\boldsymbol{K}_i^* = \boldsymbol{V}_i\boldsymbol{O}_i^* = (\boldsymbol{Q}_{i+n/2}\boldsymbol{K}_{i+n/2}^*)^* = (\boldsymbol{V}_{i+n/2}\boldsymbol{O}_{i+n/2}^*)^*$ ensures symmetry, and the architectures matches the ODE after coefficient adjustment. □

Lemma 5 (Canonical Form of Matrix Fields)**.** *Under Axioms 1, 2, and 3, the matrix potential takes the form of* $V(\boldsymbol{X}) = \text{Tr} f(\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{X}^*, \dots, \boldsymbol{X}\boldsymbol{W}_n\boldsymbol{X}^*)$ *for some analytic spectral function* $f : \mathbb{C}^n \to \mathbb{C}$.

Proof. First we check that the axioms hold for the form of V: for all $\boldsymbol{P}, \boldsymbol{R}$ in the condition, by definitions of unitary group and centralizers, we have $V(\boldsymbol{P}\boldsymbol{X}\boldsymbol{R}^*) = \text{Tr} f(\{\boldsymbol{P}\boldsymbol{X}\boldsymbol{R}^*\boldsymbol{W}_i\boldsymbol{R}\boldsymbol{X}^*\boldsymbol{P}^*\}_{i=1}^n) = \text{Tr} f(\{\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*\}_{i=1}^n) = V(\boldsymbol{X})$, assuming f is invariant under simultaneous conjugation of its arguments (i.e., f is spectral).

Next we show the uniqueness of the form of V. By Axiom 1, the matrix potential has an expansion form $V(\boldsymbol{X}) = f_1(\boldsymbol{X})$ for some $f_1 : \mathbb{C}^{N\times C} \to \mathbb{C}$. By Axiom 2, the potential is invariant under row permutations. One form that satisfies this invariance is $V(\boldsymbol{X}) = f_2(\boldsymbol{X}^*\boldsymbol{X})$ for some $f_2 : \mathbb{C}^{C\times C} \to \mathbb{C}$ (equivalently, $V(\boldsymbol{X}) = g(\boldsymbol{X}^*\boldsymbol{X})$ for some scalar-valued analytic $g : \mathbb{C}^{C\times C} \to \mathbb{C}$). By Axiom 3, the potential is additionally invariant under right multiplication by elements of the intersection of the centralizers. The general analytic function satisfying this additional invariance is a spectral function of the invariants $\{\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*\}_{i=1}^n$, leading to the form $V(\boldsymbol{X}) = \text{Tr} f(\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{X}^*, \boldsymbol{X}\boldsymbol{W}_2\boldsymbol{X}^*, \dots, \boldsymbol{X}\boldsymbol{W}_n\boldsymbol{X}^*)$ for some analytic spectral $f : (\mathbb{C}^{N\times N})^n \to \mathbb{C}$. □

Theorem 6 (Optimality of Linearized Transformers)**.** *Under Axioms 1, 2, 3, Definition 3 is the first nontrivial interaction ODE.*

Proof. By Lemma 5, the matrix potential is $V(\boldsymbol{X}) = \text{Tr} f(\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{X}^*, \dots, \boldsymbol{X}\boldsymbol{W}_n\boldsymbol{X}^*)$ for some analytic spectral function $f : \mathbb{C}^n \to \mathbb{C}$.

In the spirit of effective field theory, we consider the low-order expansion of f around the origin, truncating at the lowest nontrivial interactive terms (i.e., up to quadratic in the arguments $\{z_k\}$ of f, corresponding to quartic terms in V). Higher-order terms are irrelevant at low "energy" scales and can be neglected for the effective description. Thus,

$$V(\boldsymbol{X}) = \text{Tr}(c_0\boldsymbol{I} + \sum_{i=1}^n c_{1,i}\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^* + \sum_{i,j=1}^n c_{2,ij}\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_j\boldsymbol{X}^*),$$

where we have relabeled the $\boldsymbol{W}_{1,i}$ as $\boldsymbol{W}_i$ for simplicity, and dropped higher-order contributions.

The constant term provides no dynamics. The quadratic terms in V (linear in f) yield the feed-forward (MLP-like) component in the ODE, which can be fused: $\sum_{i=1}^n c_{1,i}\text{Tr}(\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*) = \text{Tr}(\boldsymbol{X}\boldsymbol{W}_\text{in}\boldsymbol{W}_\text{out}^*\boldsymbol{X}^*)$ by redefining $\boldsymbol{W}_\text{in}, \boldsymbol{W}_\text{out}$ appropriately (with $C_\text{MLP} = nC_\text{A}$ under Axiom 4).

The quartic terms in V (quadratic in f) represent the first nontrivial interactions. The general form is $\sum_{i,j} c_{2,ij}\text{Tr}(\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{W}_j\boldsymbol{X}^*)$. By relabeling indices and applying Lemma 3 (which requires pairing terms with n even and parameter conditions like $\boldsymbol{Q}_i\boldsymbol{K}_i^* = \boldsymbol{V}_{i+n/2}\boldsymbol{O}_{i+n/2}^*$ to ensure the potential exists), this matches the attention component of the linearized Transformer ODE under those constraints. Thus, Definition 3 (with the noted parameter tying) is the leading interactive ODE consistent with the axioms. □