Holistic Theory of Attention: Bridging Fisher Information, Group Symmetries, Renormalization Group, and Eigenvector Learning

Jeronimo Aranda Barois

November 9, 2024

Abstract

This work presents a unified theory of attention mechanisms, linking them to fundamental concepts in statistical physics, differential geometry, and linear algebra. We show how attention mechanisms in transformers can be understood through the lens of **Fisher Information Geometry**, **Boltzmann statistics**, **Renormalization Group theory**, and **group symmetries** in block matrices. Specifically, we prove that attention mechanisms follow a process akin to **diffusion over time** with phase transitions in attention behavior. Furthermore, we derive the mathematical link between **attention weights** and **energy distributions** in the **Boltzmann-like framework**. Finally, we show that the **Fisher Information Matrix (FIM)** closely relates to attention patterns and that attention updates approximate a **natural gradient descent** due to the geometry of the Fisher Information. We also introduce Renormalization Group flows to explain multi-scale attention mechanisms and hierarchical structure learning.

1 Introduction

1.1 Motivation

Transformer architectures have achieved significant success in diverse machine learning tasks. However, the theoretical underpinnings of attention mechanisms remain underexplored. This work presents a unified framework that connects **attention mechanisms** to **statistical physics**, **geometric principles**, and **block matrix structures**, bridging the gap between machine learning, thermodynamics, and linear algebra.

1.2 Key Contributions

- Development of a **unified thermodynamic theory** connecting attention mechanisms to **Fisher Information**, **Renormalization Group (RG) theory**, and **Boltzmann statistics**.
- Proof that attention mechanisms exhibit **diffusion-like dynamics** and undergo **phase transitions** during training.
- Derivation of the **natural gradient descent** connection between **attention weights** and the **Fisher Information Matrix**.
- A formal proof showing that **attention heads** learn **eigenvectors** in **block matrices**, capturing **group symmetries**.
- Introduction of **Renormalization Group** flows to explain the scale-invariance and hierarchical structure formation during training.

2 Theoretical Framework

2.1 Fisher Information in Attention

Attention mechanisms in transformer models can be mathematically described through the **softmax function**, which resembles a **probabilistic distribution**. Given the attention weight between tokens i and j as:

$$\alpha_{ij} = \text{softmax}\left(\frac{q_i^T k_j}{\sqrt{d}}\right),$$

where q_i and k_j represent the query and key vectors, and d is the dimensionality of the vectors.

The **Fisher Information Matrix (FIM)**, which measures the **sensitivity** of model predictions to parameter changes, can be computed by taking the second derivative of the **log-likelihood** of the attention distribution. This connects the **attention mechanism** with **natural gradient descent**:

$$\theta_{\text{new}} = \theta - \eta F^{-1} \nabla L,$$

where F^{-1} is the inverse of the Fisher Information Matrix and ∇L is the gradient of the loss function.

2.2 Diffusion Process and Phase Transitions in Attention

We propose that attention behavior during training follows a **diffusion process**, with information propagating across layers and heads. Attention mechanisms exhibit **phase transitions** as training progresses, characterized by the following phases:

- **Exploration Phase**: High temperature leads to uniform, exploratory attention patterns.
- **Critical Phase**: At an optimal temperature T_c, attention patterns become **hierarchical** and **scale-invariant**.
- **Exploitation Phase**: Low temperature leads to **specialized**, sharp attention patterns as the model fine-tunes.

2.3 Renormalization Group (RG) and Attention Flows

The Renormalization Group (RG) approach, commonly used in statistical mechanics and quantum field theory, provides a powerful framework for understanding the scaling behaviors of physical systems. We hypothesize that the attention mechanism undergoes an RG-like flow, where the attention patterns at different layers or scales represent coarse-grained versions of the finer-scale attention patterns.

The RG transformation captures how the effective *scales* of attention evolve during training. As layers are stacked in transformers, attention heads aggregate information at coarser scales, and the **effective temperature** (T_{eff}) decreases as training progresses, driving the system toward sharper, more specialized attention patterns:

$$A_l \to A_{l+1}, \quad FIM_l \to FIM_{l+1},$$

where A_l and A_{l+1} represent attention patterns at successive layers, and the Fisher Information Matrix evolves as information is aggregated across layers.

2.4 Eigenvector Learning and Group Symmetries

We now focus on proving that attention mechanisms can learn **eigenvectors** and **group symmetries** in **block matrices**. Let us consider a block diagonal matrix B with two blocks A and C:

$$B = \begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix}.$$

The eigenvectors of this matrix B are simply the union of the eigenvectors of A and C. We show that attention mechanisms can approximate these eigenvectors by learning the **group symmetries** present in the block structure.

3 Proof: Attention Mechanisms Learning Eigenvectors in Block Matrices

3.1 Eigenvalue Problem for Block Matrices

Consider a block matrix B with known blocks A and C, each having eigenvectors v_A and v_C. The matrix B has eigenvalues and eigenvectors formed from those of A and C. Specifically, the **eigenvalue problem** for B is:

$$Bv = \lambda v,$$

where v is the eigenvector, and λ is the corresponding eigenvalue. For the block matrix B, the eigenvectors of B are:

$$v_B = \begin{bmatrix} v_A \\ v_C \end{bmatrix},$$

where v_A and v_C are the eigenvectors of blocks A and C, respectively.

3.2 Attention Mechanism as a Diffusion Process

We hypothesize that attention heads in a transformer learn to focus on different **blocks** of a matrix, corresponding to the eigenvectors of the blocks. The attention mechanism essentially diffuses information across layers, progressively focusing on the structure of the matrix. This diffusion process mirrors the learning of the eigenvectors associated with each block:

$$P(i \to j) = \frac{\exp\left(\frac{q_i^T k_j}{\sqrt{d}}\right)}{Z_A},$$

where the attention weight between tokens i and j is modeled probabilistically, much like a **Boltzmann distribution**. As the model trains, attention heads specialize in learning different components of the eigenvector structure of A and C.

3.3 Fisher Information Matrix (FIM) and Attention Updates

The **Fisher Information Matrix (FIM)** quantifies the **parameter sensitivity** and provides a measure for the **natural gradient**. Since the attention mechanism computes a softmax over similarity scores between queries and keys, the FIM captures how the attention updates reflect the geometry of the model's parameter

space. This shows that attention heads, when trained, approximate the natural gradient flow, helping the model converge to optimal solutions in terms of learning the group symmetries and eigenvectors of the underlying data structure.

4 Conclusion

We have presented a unified theoretical framework connecting attention mechanisms to Fisher Information, Boltzmann statistics, and Renormalization Group theory. Our results show that attention mechanisms:

- Exhibit diffusion-like dynamics and undergo phase transitions in behavior.
- Learn hierarchical structures and eigenvectors in block matrices, capturing group symmetries.
- Approximate natural gradient descent due to the geometric properties of the Fisher Information Matrix.
- Evolve according to Renormalization Group flows, helping to scale attention across multiple levels of abstraction.

These insights provide a deeper theoretical understanding of transformer architectures and open the door to more efficient training and novel architectures.