Unifying Ising Models, Prime Distribution, and Transformer Optimization via β Scaling

William Chuang

1 Introduction

This paper explores the deep interconnections between three seemingly disparate topics: the Ising model and phase transitions in statistical mechanics, the distribution of prime numbers in number theory, and the optimization of self-attention in Transformer architectures via the dynamic scaling of β. I demonstrate that:

1. The Ising model provides a natural framework to study prime distributions, viewing them as phase transitions.
2. The statistical structure of primes can be used to analyze energy landscapes in both physical and neural network models.
3. A dynamical approach to determining the optimal β factor in self-attention—based on phase transition behavior—can significantly reduce the computational cost of training Transformers.

I further investigate how neural networks embedded in hyperbolic spaces relate to these connections, suggesting that hyperbolic embeddings provide a natural topology for unifying these domains.

2 From Ising Models and Phase Transitions to Prime Distribution and Transformer Optimization

2.1 The Ising Model and Phase Transitions

The classical Ising model consists of spin variables $S_i \in \{-1, 1\}$ interacting via the Hamiltonian:

$$H = -J \sum_{\langle i,j \rangle} S_i S_j - h \sum_i S_i, \tag{1}$$

where J is the interaction strength and h is an external magnetic field.

The system undergoes a phase transition at a critical temperature T_c where long-range correlations emerge, characterized by a singularity in the free energy:

$$F = -\frac{1}{\beta} \ln Z_N, \quad Z_N = \sum_S e^{-\beta H(S)}. \tag{2}$$

2.2 Phase Transitions and Prime Number Distribution

A remarkable connection exists between statistical physics and number theory: the nontrivial zeros of the Riemann zeta function,

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}, \tag{3}$$

can be viewed as the locations of phase transitions in an appropriately defined statistical mechanical system.

Through this lens, primes play the role of fundamental energy states, and their statistical fluctuations resemble those found in spin systems at criticality. The Lee-Yang theorem, which ensures phase transitions in Ising-like models occur at purely imaginary zeros of the partition function, has deep parallels with the Riemann Hypothesis, which posits that nontrivial zeros of $\zeta(s)$ lie on the critical line $\Re(s) = 1/2$.

2.3 Transformer Optimization via Dynamic β Scaling

In Transformers, self-attention computes attention scores via:

$$A_{ij} = \frac{(XW_Q)_i (XW_K)_j^T}{\sqrt{d_k}}, \tag{4}$$

where the denominator $1/\sqrt{d_k}$ is the traditional scaling factor introduced in "Attention Is All You Need". However, this choice is heuristic and does not consider the topology of the network.

A more rigorous approach involves determining β dynamically using the Ising model framework. Specifically, I hypothesize that the optimal β is analogous to the critical temperature T_c:

$$\beta_{\text{opt}} = \frac{1}{T_c}. \tag{5}$$

At T_c, correlations span the entire network, allowing global information propagation with minimal redundancy, reducing training costs.

Remark: While some may hesitate to view self-attention as an approximation of a stack of Ising models, given that after applying the Softmax function, the resulting attention matrix undergoes a subsequent multiplication with V, the key insight lies in treating the Softmax component as a system exhibiting Ising-like behavior. Once the critical temperature T_c is reached, a phase transition occurs, leading to an effectively infinite correlation length across the entire Softmax structure. This super-correlation state ensures that the attention mechanism globally propagates information, significantly decreasing training time and computational cost. Although fine-tuning might be necessary to optimize second-order effects introduced by the multiplication with V, the dominant component of the transformer's behavior is dictated by the Softmax layers, making the Ising model analogy a powerful tool for understanding and optimizing the system.

Remark 2: While the exponential in the Softmax function is primarily responsible for inducing Ising-like interactions among the spin variables, the subsequent multiplication with the value matrix V plays a dual role. In a single self-attention layer, V acts merely as a projection that maps the correlated state into a new representational space, without directly adding further interaction terms. However, in modern Transformer architectures—where it is common to have between 6 and 24 layers (and in some cases even deeper, such as 12 layers in BERT-base, 24 in BERT-large, or up to 96 in models like GPT-3)—the output of one self-attention layer, after being modulated by V, is fed into the next. This cascading of layers creates an effective stack of Ising-like transformations, with each application of V contributing to the evolving interaction landscape. For instance, in GPT-3, the 96 self-attention layers form a composition of functions, where each layer refines the output of the previous one, exemplifying the power of deep, layered processing. In such a multi-layer setup, the role of V is not merely a projection, but part of a compositional chain that refines and propagates the global correlations established by the Softmax, underscoring the hierarchical nature of information propagation in Transformer models.

Remark 3: It is important to note that phase transitions are not exclusive to quantum systems. Classical spin gases or spin glasses also exhibit phase transitions driven by thermal fluctuations. In our framework, the effective Hamiltonian derived from the composition of self-attention layers,

$$H_{\text{eff}} = \sum_{l=1}^{L} H^{(l)},$$

with L (e.g., 96 in GPT-3) representing the number of layers, is constructed from classical spin-like variables S_i (corresponding to discrete floating-point representations). The resulting Boltzmann distribution,

$$P(S) \propto e^{-\beta H_{\text{eff}}},$$

is defined over these classical degrees of freedom, implying that the emergent phase transition is a *classical* one rather than a quantum phase transition.

Mathematically, our derivation via the Trotter–Suzuki formula shows that the layered composition

$$\prod_{l=1}^{L} e^{-\beta H^{(l)}}$$

is well-approximated by a single exponential $e^{-\beta H_{\text{eff}}}$ under the assumptions of weak non-commutativity and bounded Hamiltonians. Since the variables S_i in each $H^{(l)}$ are classical, the effective large-scale Ising model governing GPT-3 is inherently classical. Thus, the phase transition observed in such architectures is best described as a classical phase transition.

Aspect	**Classical Phase Transition**	**Quantum Phase Transition**
Driving Parameter	Temperature, external fields	Quantum fluctuations (e.g., transverse field)
Nature of Fluctuations	Thermal fluctuations	Quantum fluctuations (entanglement, superposition)
Typical Models	Ising, Potts, spin glasses	Quantum Ising, Heisenberg, Bose-Hubbard
Order Parameter	Magnetization, density, etc.	Similar observables, modulated by quantum coherence
Mathematical Framework	Partition functions over classical states	Path integrals and ground state analyses

Table 1: Comparison of Classical and Quantum Phase Transitions

3 Using Prime Distribution to View Ising Models and Transformers

Since primes dictate the distribution of energy levels in statistical physics models, they can also provide a natural way to analyze the energy landscape of Ising models and Transformer networks.

3.1 Prime Gaps and Phase Transitions

The gaps between consecutive prime numbers behave analogously to domain walls in spin models. The distribution of these gaps follows a statistical pattern that resembles the critical behavior of the Ising model. This suggests that:

- Prime number statistics can be used to construct phase transition models.
- Transformers, when optimized, might exhibit patterns in attention weights that reflect the same statistical regularities.

3.2 Entropy, Free Energy, and Neural Networks

If primes dictate an energy landscape, the entropy of a Transformer network can be analyzed in terms of number-theoretic quantities. Given the entropy formulation:

$$S = -\sum_i P_i \ln P_i, \tag{6}$$

where P_i are normalized attention probabilities, a prime-based approach to entropy estimation could yield novel optimization techniques.

4 Using β and Neural Network Topology to Unify the Three Perspectives

4.1 Viewing Neural Networks as Ising Models

Since Transformers can be mapped to stacks of Ising models, one might consider the question: can they be viewed as a *single* large Ising model?

Applying mean-field theory, one can approximate interactions with an effective Hamiltonian:

$$H_{\text{eff}} = -J_{\text{eff}} \sum_i S_i S_{\text{eff}}, \tag{7}$$

where S_{eff} represents a global embedding. If Transformer weights are embedded in hyperbolic spaces, this induces a new type of connectivity that could lead to:

- Faster convergence due to hyperbolic distance minimization.
- More efficient attention mechanisms by exploiting the curved geometry.

4.2 Hyperbolic Neural Networks and β Optimization

Embedding Transformer weights in hyperbolic space suggests a natural way to optimize β. In a hyperbolic model, distances grow exponentially, meaning that long-range dependencies can be captured efficiently. If one assumes that:

$$\beta_{\text{opt}} \sim \frac{1}{\sqrt{\text{curvature}}}, \tag{8}$$

then optimal β scales dynamically with the curvature of the embedding space. This would allow Transformers to be trained in a fundamentally more efficient manner.

5 Concrete Example: Hyperbolic Attention Scaling

Consider a Transformer with a hyperbolic attention mechanism where the attention weights are computed using:

$$A_{ij} \propto e^{-\beta d_{\mathbb{H}}(X_i, X_j)}, \tag{9}$$

where $d_{\mathbb{H}}$ is the hyperbolic distance.

6 Conjecture: (Dynamic Scaling of β in Hyperbolic Neural Networks)

Let κ denote the curvature of the hyperbolic space in which a neural network is embedded, and let λ represent a topological factor incorporating connectivity properties such as fugacity (from statistical physics) and the average number of neighboring nodes influencing a given node in the network. Then, the optimal scaling factor β for self-attention mechanisms in Transformers should dynamically adjust according to:

$$\beta_{\text{opt}} \sim \frac{1}{\sqrt{|\kappa| \cdot \lambda}}. \tag{10}$$

This conjecture suggests that the scaling of attention scores is not merely a function of dimensionality (d_k), as traditionally used, but rather an emergent property of both the curvature of the space and the interaction structure of the network. Specifically:

- The curvature κ determines how distances scale within the space, affecting long-range dependencies.
- The topological factor λ encodes the effective connectivity and interaction strength between nodes, akin to fugacity in statistical physics.

If this relationship holds, then adjusting β dynamically in accordance with the underlying geometry and topology of the neural network should lead to more efficient convergence and reduced computational costs in Transformer training.

7 Conclusion

This note has provided a framework to unify Ising models, prime distributions, and Transformer optimization by exploring how:

- The Ising model naturally describes phase transitions in prime number distributions.
- Prime distributions offer insights into statistical physics models and neural network energy landscapes.
- Neural networks, when viewed through an Ising model lens, suggest a dynamically optimized β that minimizes training costs.
- Hyperbolic embeddings provide a compelling geometric structure for modeling these phenomena.

Future work should explore Monte Carlo simulations of these ideas, as well as empirical validation in large-scale neural network training.