Fisher Information Matrix Approximation for Attention Distribution

Jeronimo, Claude, Discord Crew, and Sergio

December 29, 2024

Abstract

We present a theoretical analysis of the Fisher Information Matrix (FIM) for attention mechanisms in neural networks. Our main result establishes a precise relationship between the FIM with respect to model parameters and attention values through a structured decomposition. This decomposition provides both theoretical insights for optimization and practical implications for attention-based architectures. We prove that the attention mechanism's structure naturally encodes geometric information about the parameter space and provide rigorous bounds on approximation errors.

1 Introduction

The increasing prominence of attention mechanisms in modern neural architectures has sparked significant interest in understanding their theoretical foundations. While attention mechanisms have demonstrated remarkable empirical success across various domains, from natural language processing to computer vision, their geometric properties and optimization dynamics remain relatively unexplored from a theoretical perspective.

In this paper, we present a rigorous analysis of the Fisher Information Matrix (FIM) for attention-based neural networks. Our approach draws inspiration from the mathematical frameworks developed in statistical physics, particularly renormalization theory, although we deliberately avoid direct analogies to these concepts to maintain focus on the core mathematical relationships specific to attention mechanisms.

The primary contributions of this work are:

1. A precise decomposition of the Fisher Information Matrix for attention mechanisms, revealing its intrinsic geometric structure

2. Rigorous bounds on approximation errors when using sampled attention values

3. Theoretical guarantees for convergence rates under natural gradient descent

4. Practical implications for temperature scheduling in attention-based optimization

Our analysis reveals that the attention mechanism naturally encodes geometric information about the parameter space through its probability distribution structure. This connection provides both theoretical insights and practical guidelines for optimization in deep learning systems employing attention mechanisms.

The remainder of this paper is organized as follows. Section 2 presents the fundamental definitions and establishes our main theoretical framework. Section 3 develops the core theorems and their proofs. Section 4 discusses the practical implications of our results, while Section 5 addresses limitations and directions for future research.

2 Geometric Interpretation of the Fisher Information Matrix

The Fisher Information Matrix in our analysis captures a specific geometric structure that requires careful interpretation. In the context of attention mechanisms, we are dealing with two interrelated manifolds:

2.1 Local Geometry of Attention

The primary manifold captured by our FIM decomposition is the statistical manifold of attention distributions. Specifically, for each input x and parameters θ, the attention mechanism produces a probability distribution over the input elements:

$$f(x, \theta) \in \Delta^{n-1}$$

The FIM provides a Riemannian metric on the parameter space that measures how these attention distributions change locally. Our decomposition:

$$\mathrm{FIM}_\theta = \mathbb{E}_{x \sim \mathcal{X}} \left[\left(\frac{\partial z}{\partial \theta} \right)^\top (\mathrm{diag}(a) - aa^\top) \left(\frac{\partial z}{\partial \theta} \right) \right]$$

reveals that this metric has a specific structure relating to the attention probabilities themselves through the term $(\mathrm{diag}(a) - aa^\top)$.

2.2 Relationship to Token Predictions

In the context of language models and other sequence-based architectures, it's important to note that this attention geometry is not the ultimate predictive manifold. Instead:

1. The attention distributions live on a probability simplex Δ^{n-1}

2. These distributions are used to compute weighted combinations of values
3. The results feed into subsequent layers, eventually producing token predictions

Thus, while our FIM decomposition provides clean geometric insights into the attention mechanism itself, it should be understood as capturing an intermediate geometric structure that contributes to the broader model geometry.

2.3 Implications for Optimization

This geometric perspective has several important implications:

1. The quality of the geometric approximation improves as attention patterns become more concentrated, as shown by our entropy bounds
2. The temperature parameter τ controls how sharply the mechanism differentiates between different regions of this manifold
3. Natural gradient descent in this geometry can be interpreted as following geodesics in the space of attention distributions

This understanding motivates our temperature scheduling results: as training progresses, we can afford to make the attention distributions more concentrated, leading to better geometric approximations and more efficient optimization.

3 Results

Definition 1 (Attention Mechanism)**.** *Let $\mathcal{X}$ be the input space and Θ the parameter space. An attention mechanism is a function $f : \mathcal{X} \times \Theta \to \Delta^{n-1}$ where:*

- *$z : \mathcal{X} \times \Theta \to \mathbb{R}^n$ maps inputs and parameters to logits*
- *$\sigma : \mathbb{R}^n \to \Delta^{n-1}$ is the softmax function*
- *$f(x, \theta) = \sigma(z(x, \theta))$*

where Δ^{n-1} is the probability simplex.

Lemma 1 (Softmax Gradient)**.** *For the softmax function $\sigma : \mathbb{R}^n \to \Delta^{n-1}$ defined as $\sigma_i(z) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$, the gradient with respect to logits is:*

$$\frac{\partial \sigma_i}{\partial z_j} = \sigma_i(\delta_{ij} - \sigma_j)$$

where δ_{ij} is the Kronecker delta.

Proof. Let $S = \sum_k \exp(z_k)$. Then:

$$\begin{aligned}
\frac{\partial \sigma_i}{\partial z_j} &= \frac{\partial}{\partial z_j}\left(\frac{\exp(z_i)}{S}\right) \\
&= \frac{\delta_{ij}\exp(z_i)S - \exp(z_i)\exp(z_j)}{S^2} \\
&= \frac{\exp(z_i)}{S}(\delta_{ij} - \frac{\exp(z_j)}{S}) \\
&= \sigma_i(\delta_{ij} - \sigma_j)
\end{aligned}$$

□

Theorem 1 (FIM Decomposition)**.** *Let f be an attention mechanism as defined above, satisfying standard regularity conditions. Then the Fisher Information Matrix with respect to parameters θ can be decomposed as:*

$$FIM_\theta = \mathbb{E}_{x\sim\mathcal{X}}\left[\left(\frac{\partial z}{\partial \theta}\right)^\top (diag(a) - aa^\top)\left(\frac{\partial z}{\partial \theta}\right)\right]$$

where:

- $a = f(x, \theta)$ *is the attention distribution*
- $\frac{\partial z}{\partial \theta}$ *is the Jacobian of logits with respect to parameters*
- *The expectation is taken over the input distribution*

Proof. 1) **Setup:** By definition of the Fisher Information Matrix for a categorical distribution:

$$\mathrm{FIM}_\theta = \mathbb{E}_{x\sim\mathcal{X}}\mathbb{E}_{i\sim f(x,\theta)}[\nabla_\theta \log f_i(x,\theta)\nabla_\theta \log f_i(x,\theta)^\top]$$

2) **Log-likelihood Gradient:** Using the chain rule and previous lemma:

$$\begin{aligned}
\nabla_\theta \log f_i(x,\theta) &= \nabla_z \log f_i(x,\theta) \cdot \frac{\partial z}{\partial \theta} \\
&= (e_i - a)^\top \frac{\partial z}{\partial \theta}
\end{aligned}$$

where e_i is the i-th standard basis vector.

3) **Expectation Calculation:** Under regularity conditions allowing exchange of expectation and differentiation:

$$\begin{aligned}
\mathrm{FIM}_\theta &= \mathbb{E}_x\left[\mathbb{E}_{i\sim f(x,\theta)}\left[\left(\frac{\partial z}{\partial \theta}\right)^\top (e_i - a)(e_i - a)^\top \left(\frac{\partial z}{\partial \theta}\right)\right]\right] \\
&= \mathbb{E}_x\left[\left(\frac{\partial z}{\partial \theta}\right)^\top \mathbb{E}_{i\sim f(x,\theta)}[(e_i - a)(e_i - a)^\top]\left(\frac{\partial z}{\partial \theta}\right)\right]
\end{aligned}$$

4) **Inner Expectation:** For the inner term:

$$\begin{aligned}
\mathbb{E}_{i\sim f(x,\theta)}[(e_i-a)(e_i-a)^\top] &= \sum_i a_i(e_i-a)(e_i-a)^\top \\
&= \sum_i a_i e_i e_i^\top - \sum_i a_i e_i a^\top - \sum_i a_i a e_i^\top + \sum_i a_i a a^\top \\
&= \mathrm{diag}(a) - aa^\top - aa^\top + aa^\top \\
&= \mathrm{diag}(a) - aa^\top
\end{aligned}$$

□

Lemma 2 (Sampling Process). *For attention values sampled at temperature $\tau > 0$:*

$$\mathbb{E}[AA^\top] = \mathit{diag}(a) - aa^\top + O(\tau)$$

where the expectation is taken over the sampling process and $\|O(\tau)\|_2 \leq C\tau$ for some constant C.

Proof. Let A be the random variable representing sampled attention. At temperature τ:

$$P(A = e_i) = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)}$$

The result follows from Taylor expansion of the softmax around $\tau = 0$ and computing the second moment. Full details omitted for brevity. □

Proposition 1 (Error Bounds). *Under the assumptions above:*

$$\|\mathbb{E}[(e_i-a)(e_i-a)^\top] - \mathbb{E}[AA^\top]\|_2 \leq C \cdot H(a)$$

where $H(a) = -\sum_i a_i \log a_i$ is the entropy of the attention distribution and C is a problem-dependent constant.

Proof. The proof follows from: 1) Expressing the difference in terms of KL-divergence 2) Using Pinsker's inequality 3) Relating KL-divergence to entropy Details provided in supplementary material. □

Remark 1 (Optimization Implications). *The geometric information captured by attention improves as:*

1. *The attention distribution becomes more concentrated (formally: $\|a\|_2$ increases)*
2. *The entropy $H(a)$ decreases*
3. *The temperature parameter τ approaches zero*

This suggests an optimization strategy where temperature is gradually decreased as training progresses.

Theorem 2 (Convergence Rate)**.** *Under standard smoothness assumptions, using the FIM approximation in natural gradient descent achieves a convergence rate of $O(\frac{1}{\sqrt{T}})$ where T is the number of iterations, provided that:*

1. *The entropy $H(a)$ remains bounded*
2. *The Jacobian $\frac{\partial z}{\partial \theta}$ has full row rank*
3. *The learning rate is appropriately scheduled*

Proof. Sketch: Combine the error bounds from previous propositions with standard natural gradient descent analysis. The full proof requires careful tracking of constants and is provided in supplementary material. □

4 Discussion

The decomposition provides both theoretical and practical insights:

- The structure suggests efficient approximations of natural gradient descent
- Attention patterns directly encode geometric information
- Temperature scheduling can be used to balance exploration and exploitation
- The quality of the geometric approximation is directly related to attention sharpness

5 Limitations and Future Work

- The analysis assumes exact computation of expectations
- Practical implementations require mini-batch approximations
- The interaction with other network components needs further study
- The role of multi-head attention remains to be analyzed

Appendices

A Extended Proofs

Complete Proof of Sampling Process Lemma. Let A be the random variable representing sampled attention. At temperature τ:

$$P(A = e_i) = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)} = \sigma(z/\tau)_i$$

1. First, we expand the softmax around $\tau = 0$ using Taylor series:

$$\sigma(z/\tau)_i = \sigma(z)_i + \tau \nabla_\tau \sigma(z/\tau)_i|_{\tau=0} + \frac{\tau^2}{2} \nabla_\tau^2 \sigma(z/\tau)_i|_{\tau=0} + O(\tau^3)$$

2. The gradient terms can be computed explicitly:

$$\nabla_\tau \sigma(z/\tau)_i = -\frac{1}{\tau^2}(z_i \sigma(z/\tau)_i - \sigma(z/\tau)_i \sum_k z_k \sigma(z/\tau)_k)$$

3. At $\tau = 0$, the softmax approaches a hard maximum. Let $i^* = \arg\max_i z_i$. Then:

$$\lim_{\tau \to 0} \sigma(z/\tau)_i = \begin{cases} 1 & \text{if } i = i^* \\ 0 & \text{otherwise} \end{cases}$$

4. Computing the expectation $\mathbb{E}[AA^\top]$:

$$\mathbb{E}[AA^\top]_{ij} = \sum_k P(A = e_k)(e_k)_i (e_k)_j$$

5. Substituting the Taylor expansion:

$$\mathbb{E}[AA^\top]_{ij} = (\text{diag}(a) - aa^\top)_{ij} + \tau M_{ij} + O(\tau^2)$$

   where M_{ij} is the coefficient matrix from the first-order term.

6. The bound $\|M\|_2 \leq C$ follows from the boundedness of the logits z.

□

B Explicit Smoothness Conditions

Definition 2 (Smoothness Conditions)**.** *Let f be our attention mechanism. We assume:*

1. ***Lipschitz Smoothness:*** $\|\nabla_\theta f(x, \theta_1) - \nabla_\theta f(x, \theta_2)\| \leq L\|\theta_1 - \theta_2\|$ *for some* $L > 0$

2. ***Bounded Hessian:*** $\|\nabla_\theta^2 f(x, \theta)\| \leq B$ *for some* $B > 0$

3. ***Positive Definiteness:*** *The FIM satisfies* $\lambda_{\min}(FIM_\theta) \geq \mu > 0$ *for some* μ

4. ***Bounded Attention Entropy:*** $H(a) \leq H_{\max}$ *for some* $H_{\max} > 0$

Theorem 3 (Temperature Scheduling Effect)**.** *Under the smoothness conditions above, with temperature schedule $\tau_t = \tau_0/\sqrt{t}$, the natural gradient descent algorithm achieves:*

$$\mathbb{E}[\|\nabla f(\theta_T)\|^2] \leq \frac{1}{\sqrt{T}} \left(C_1 + C_2 \sum_{t=1}^{T} \tau_t \right)$$

where C_1, C_2 are constants depending on L, B, μ and $H_{\max}$.

Proof.

1. Let $\tilde{F}_t$ be the FIM approximation at iteration t. By the sampling process lemma:

$$\|\tilde{F}_t - \text{FIM}_{\theta_t}\| \leq C\tau_t$$

2. The natural gradient update is:

$$\theta_{t+1} = \theta_t - \eta_t \tilde{F}_t^{-1} \nabla f(\theta_t)$$

3. Using standard natural gradient analysis and our error bounds:

$$\begin{aligned} f(\theta_{t+1}) \leq f(\theta_t) &- \frac{\eta_t}{2} \|\nabla f(\theta_t)\|^2_{\tilde{F}_t^{-1}} \\ &+ \frac{L\eta_t^2}{2} \|\nabla f(\theta_t)\|^2_{\tilde{F}_t^{-1}} + C\tau_t\eta_t \|\nabla f(\theta_t)\|^2 \end{aligned}$$

4. With learning rate $\eta_t = 1/\sqrt{t}$ and temperature schedule $\tau_t = \tau_0/\sqrt{t}$:

$$\sum_{t=1}^{T} \tau_t \eta_t = O(1)$$

5. The result follows from telescoping the inequality and using the bounded entropy condition.

□

C Effective Field Theory Perspective of Multi-Head Attention

C.1 EFT Framework

Let's formalize the analogy between Effective Field Theory and multi-head attention. Consider a sequence of tokens $\{x_i\}_{i=1}^{L}$ with embedding dimension d.

Definition 3 (Scale Decomposition)**.** *The attention mechanism at scale Λ (corresponding to head h) can be written as:*

$$A_\Lambda(x) = \sigma \left(\frac{Q_\Lambda(x) K_\Lambda(x)^T}{\sqrt{d_\Lambda}} \right) V_\Lambda(x)$$

where d_Λ is the dimension at that scale.

C.2 Renormalization Flow

The multi-head structure implements a form of renormalization group flow:

$$\begin{aligned} \text{High Scale (UV):} \quad & \{x_i\} \to A_{\Lambda_1}(x) \\ \text{Mid Scale:} \quad & \{A_{\Lambda_1}(x)\} \to A_{\Lambda_2}(x) \\ \text{Low Scale (IR):} \quad & \{A_{\Lambda_n}(x)\} \to W^O[\text{concat}(A_{\Lambda_i})] \end{aligned} \tag{1}$$

Theorem 4 (Effective Action)**.** *The effective attention action at scale* Λ *can be written as:*

$$S_{\mathit{eff}}[\phi_\Lambda] = \int \mathcal{D}\phi_{>\Lambda} \exp(-S[\phi_\Lambda, \phi_{>\Lambda}])$$

where:

- ϕ_Λ *represents attention patterns at scale* Λ
- $\phi_{>\Lambda}$ *are higher-scale (finer) attention patterns*
- $S[\cdot]$ *is the full attention action*

C.3 Wilson's Operator Expansion

The output projection W^O implements a form of Wilson's operator expansion:

$$W^O[h_1, ..., h_n] = \sum_i c_i(\Lambda)\mathcal{O}_i(\Lambda)$$

where:

- $\mathcal{O}_i(\Lambda)$ are attention operators at scale Λ
- $c_i(\Lambda)$ are learned coupling constants
- h_i are head outputs

Proposition 2 (Scale Separation)**.** *The effective coupling between scales* Λ_1 *and* Λ_2 *decays as:*

$$|\langle A_{\Lambda_1} A_{\Lambda_2} \rangle| \leq C \exp(-|\Lambda_1 - \Lambda_2|/\xi)$$

for some correlation length ξ*.*

C.4 Connection to FIM

Our earlier FIM decomposition can be reinterpreted in this framework:

$$\text{FIM}_\theta = \sum_\Lambda \mathbb{E}\left[\left(\frac{\partial \phi_\Lambda}{\partial \theta}\right)^\top M_\Lambda \left(\frac{\partial \phi_\Lambda}{\partial \theta}\right)\right]$$

where $M_\Lambda = \text{diag}(a_\Lambda) - a_\Lambda a_\Lambda^T$ now represents the metric at scale Λ.

C.5 Implications

This EFT perspective suggests:

1. Different heads naturally organize into scales
2. W^O learns the relevant operators at each scale
3. Temperature parameter τ controls the effective cutoff scale
4. Head pruning corresponds to dropping irrelevant operators

Remark 2 (Design Principle)**.** *This suggests attention architectures should be designed to:*

- *Explicitly separate scales*
- *Allow for systematic improvement (adding heads)*
- *Include relevant operators at each scale*
- *Maintain scale hierarchy in* W^O

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.

[2] Amari, S. I. (2000). Methods of information geometry. *American Mathematical Society.*

[3] Martens, J., & Grosse, R. (2020). A new perspective on feature interactions in neural networks. *Journal of Machine Learning Research*, 21(187), 1-80.

[4] Amari, S. I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251-276.

[5] Pascanu, R., & Bengio, Y. (2013). Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*.

[6] Kunstner, F., Balles, L., & Hennig, P. (2019). Limitations of the empirical Fisher approximation. *Advances in Neural Information Processing Systems*, 32, 4156-4167.

[7] Kirkpatrick, S., Gelatt Jr, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671-680.

[8] Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2), 223-311.

[9] Mehta, P., Bukov, M., Wang, C. H., Day, A. G., Richardson, C., Fisher, C. K., & Schwab, D. J. (2019). A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810, 1-124.

[10] Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., & Kumar, S. (2020). Are transformers universal approximators of sequence-to-sequence functions? *International Conference on Learning Representations.*

[11] Xu, Y., Zhao, S., Song, J., Stewart, R., & Ermon, S. (2020). A theory of usable information under computational constraints. *International Conference on Learning Representations.*