Unraveling the Gradient Descent Dynamics of Transformers

Bingqing Song [1] Boran Han [2] Shuai Zhang [2] Jie Ding[1] Mingyi Hong[1]

[1]University of Minnesota [2]Amazon AWS AI

Introduction and Motivation

- Transformer models have been popular in solving NLP, CV tasks (Vaswani, 2017; Radford, 2019; Brown 2020).
- However, there is still a generic lack of theoretical understanding of the optimization of Transformer models.

Main Question

1. Which types of Transformer **architectures** allow Gradient Descent (GD) to achieve guaranteed convergence?
2. What is the **key factor** in the Transformer that enables the fast convergence?
3. Is there any **empirical evidence** to support the finding in the answers to the above two questions?

Contributions

- Theoretically prove that: For regression problem modeled by Transformer, with appropriate network size, structure and initialization, **global optimal solution** can be found.
- We demonstrates that the **activation function** and the **variables** to be optimized as key factors in the optimization of Transformer model.

Some observations: Different performance with Gaussian/Softmax Transformer

- **Task**: IMDb review classification and Pathfinder
- **Model**: Two-layer transformer with Gaussian/Softmax kernel.
- **Conclusion**: Training a transformer with Gaussian kernel is easier than Softmax kernel in some cases.
- **Question from the observations**: Why training Softmax transformers is not efficient in some cases? How to guarantee training a transformer with faster convergence?

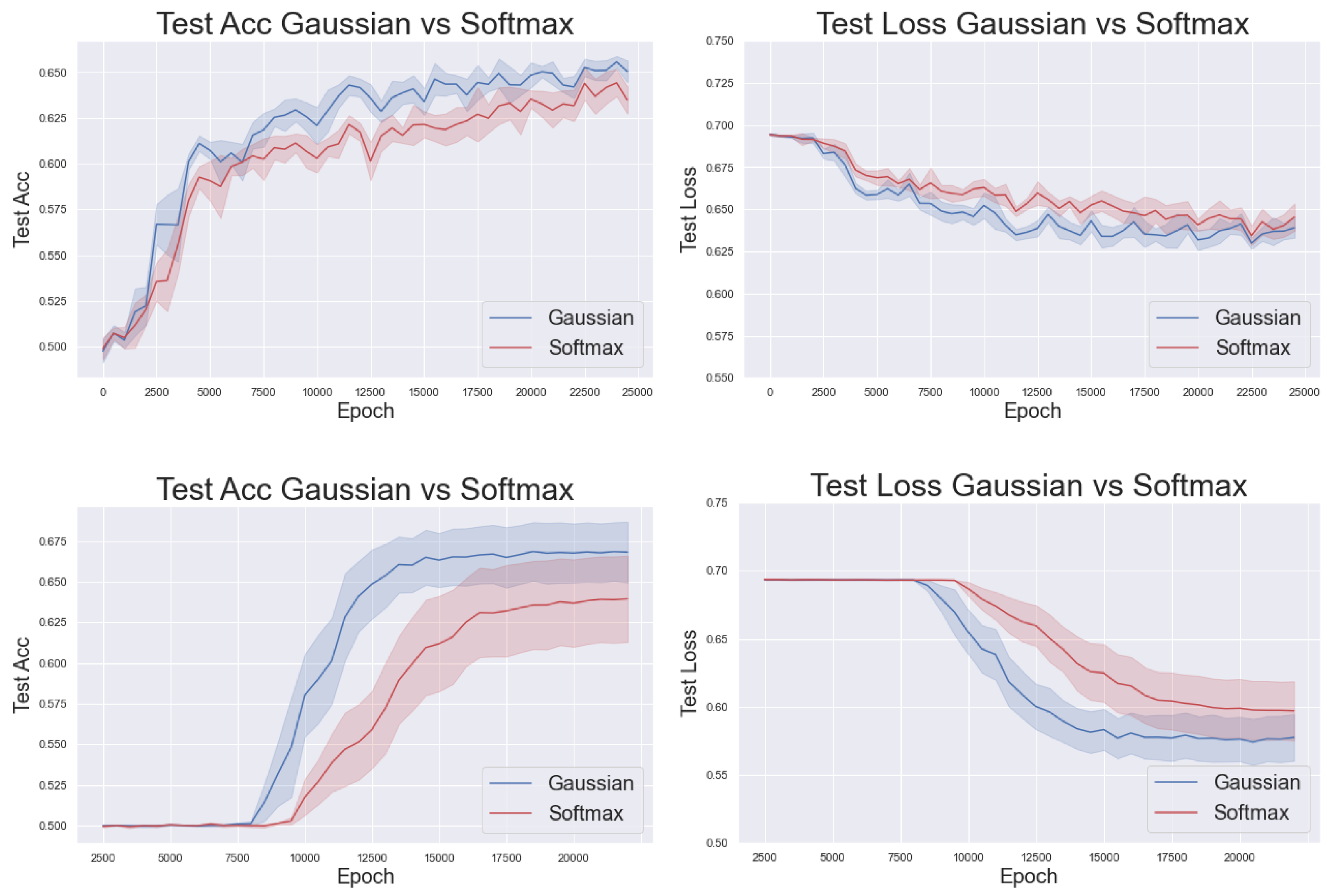

Figure 1. Test performance on text classfication and pathfinder task with different attention kernels. Optimization problem modeled by Softmax attention transformers can converge slower than Gaussian attention transformers

Problem Description

- **Dataset**: $\{(X_i, y_i)\}_{i=1}^N$, $X_i \in \mathbb{R}^{n\times D}$ is input sequence, and $y_i \in \mathbb{R}^n$ is label. n is sequence length, D is embedding dimension.
- **Model Structure**: One-layer Transformer with self-attention with $W_h^Q, W_h^K, W_h^Q \in \mathbb{R}^{D\times d}$, denoted as $\mathsf{MH}(M; X_i)$.
- **Two attention kernels:**
  - Softmax attention:

$$C_{ih} := \frac{X_i W_h^Q \left(X_i W_h^K\right)^\top}{\sqrt{d}} \in \mathbb{R}^{n\times n}, \tag{1}$$

$$S(W_h^Q, W_h^K; X_i) = \mathrm{Softmax}\left(C_{ih}\right). \tag{2}$$

  - Gaussian attention:

$$C_{ih} \in \mathbb{R}^{n\times n};\ (C_{ih})_{kj} = -\frac{\left\|X_{ik}.W_h^Q - X_{ij}.W_h^K\right\|^2}{2\sqrt{d}}, \tag{3}$$

$$\left(S(W_h^Q, W_h^K; X_i)\right)_{kj} = \exp\left((C_{ih})_{kj}\right). \tag{4}$$

- **Objective function:**

$$\min_M \frac{1}{2}\sum_{i=1}^N \|\mathsf{MH}(M; X_i) - y_i\|^2, \tag{5}$$

where $M := (W^Q, W^K, W^V)$ is the set of variables that can be optimized.

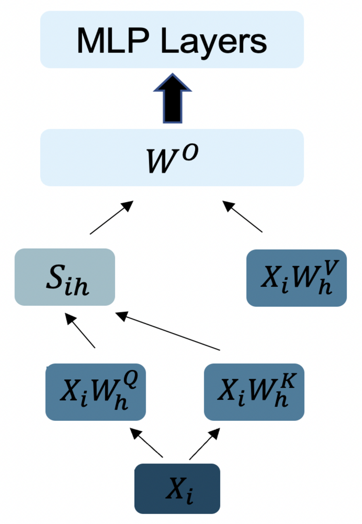

Figure 2. $\mathrm{Attention}(W_h^Q, W_h^K, W_h^V; X_i) := S(W_h^Q, W_h^K; X_i)X_iW_h^V$, $\mathsf{MH}(W^Q, W^K, W^V; X_i) := (\mathrm{head}_1, \ldots, \mathrm{head}_H)\cdot W^O$, where $\mathrm{head}_h := \mathrm{Attention}(W_h^Q, W_h^K, W_h^V; X_i), h = 1, \cdots, H$.

Convergence and Training Dynamic Analysis

Theorem 1: Solve Problem (5) with Gradient Decent update and $M = (W^Q, W^K, W^V)$. Suppose $HD \geqslant Nn$, then there exists initialization and stepsize η, such that at iteration t,

$$f\left(M_t; X\right) \leq (1-\eta\beta)^t f\left(M_0; X\right), \tag{6}$$

where $\beta = \sigma^2_{\min}(W^O)\sigma^2_{\min}(B_0^T) > 0$, $\sigma_{\min}(\cdot)$ is the smallest singular value. and matrix B_0 is only related to M_0 and input X.

Remark:

- Only when $HD \geqslant Nn$, $\beta > 0$ can hold, which means total number of features is **at least** sample size times sequence length.
- The initialization ensures variables start from a **near convex region**.
- Weights need to move 'a bit' in the **near convex region** to converge to global solution.

Theorem 2: Solve Problem (5) Gradient Decent update and $M = \{W^Q\}$. Suppose $HD \geqslant Nn$, then there exists initialization and stepsize η, such that

$$f\left(M_t; X\right) \leq (1-\eta\gamma)^t f\left(M_0; X\right) \qquad \text{Gaussian attention}$$

$$f\left(M_t; X\right) \leq f\left(M_0; X\right) - \eta\sum_{r=0}^{t-1}\|\nabla_{W^Q} f\left(M_r; X\right)\|_F \qquad \text{Softmax attention}$$

where γ is constant related to initialized weights and input.

Remark:

- Optimization problem modeled by classical Softmax attention transformer can be trained to global optimal solution.
- Softmax attention transformer can have more local solutions, which lead to worse performance than Gaussian kernel in some cases.
- The difference between attention kernels implies there is **vanishing gradient** issue in Softmax attention.

Empirical Results on Transformers with Different Attention Kernels

Setting:

- **Dataset:** Text Classification using the IMDb review dataset and Pathfinder.
- **Model:** 2-layer Transformer model with the following specifications: embedding dimension $D = 64$, hidden dimension $d = 128$, and number of attention heads $H = 2$.
- **Test Performance:**Compute test accuracy and test loss within the training steps with both Softmax and Gaussian kernel attention on both tasks.
- **Landscape Visualization:**
  - Obtain the model after $20,000$ training steps with Softmax kernel.
  - Proceed to train with additional 500 steps with Softmax/Gaussian kernel.
  - Choose two varying directions and plot the loss function when parameters change along the two directions.

Landscape Visualization

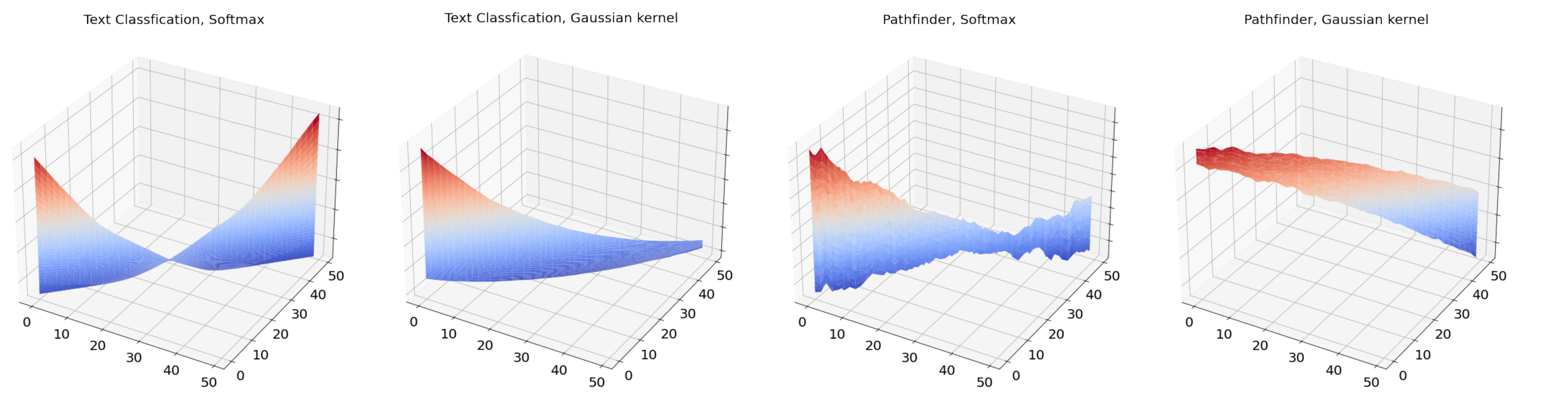

Figure 3. The loss landscapes on text classification task and Pathfinder task. For both tasks, we use the two-stage training in with the same training hyperparameters, while the only difference is the attention structure in the second training stage. The two axes represent the two directions (W^Q and W^K). With Softmax attention, the landscape appears more complicated compared with Gaussian kernel attention.

Conclusion:

- In both tasks, transformers with Gaussian attention exhibit a **more smooth** landscape compared to transformers with Softmax attention.
- The complicated landscape of transformers with Softmax attention indicates **more potential bad local solutions**.
- The landscape visualization provides evidence for Theorem 2, which explains the difference between different attention kernels.

References

[1] Xie, Sang Michael, et al. "An explanation of in-context learning as implicit bayesian inference." arXiv preprint arXiv:2111.02080 (2021).
[2] Nguyen, Quynh N., and Marco Mondelli. "Global convergence of deep networks with one wide layer followed by pyramidal topology." Advances in Neural Information Processing Systems 33 (2020): 11961-11972.
[3] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural inform
[4] Du, Simon, et al. "Gradient descent finds global minima of deep neural networks." International conference on machine learning. PMLR, 2019.
[5] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901