Self-attention Networks Localize When QK-eigenspectrum Concentrates

Han Bao [1]   **Ryuichiro Hataya** [2]   **Ryo Karakida** [3]

Abstract

The self-attention mechanism prevails in modern machine learning. It has an interesting functionality of adaptively selecting tokens from an input sequence by modulating the degree of attention localization, which many researchers speculate is the basis of the powerful model performance but complicates the underlying mechanism of the learning dynamics. In recent years, mainly two arguments have connected attention localization to the model performances. One is the rank collapse, where the embedded tokens by a self-attention block become very similar across different tokens, leading to a less expressive network. The other is the entropy collapse, where the attention probability approaches non-uniform and entails low entropy, making the learning dynamics more likely to be trapped in plateaus. These two failure modes may apparently contradict each other because the rank and entropy collapses are relevant to uniform and non-uniform attention, respectively. To this end, we characterize the notion of attention localization by the eigenspectrum of query-key parameter matrices and reveal that a small eigenspectrum variance leads attention to be localized. Interestingly, the small eigenspectrum variance can prevent both rank and entropy collapses, leading to better model expressivity and trainability.

1. Introduction

Transformers have been widely adopted in language modeling (Vaswani et al., 2017), vision tasks (Dosovitskiy et al., 2021; Touvron et al., 2021), and speech recognition (Likhomanenko et al., 2021). A crucial building block in transformers is the attention mechanism, dating back to Graves (2013), which was initially designed to capture long-range signals in sequential inputs by mixing individual tokens but has also been leveraged to capture general struc-

tures of input data. After the fully-attention-based language model has appeared (Brown et al., 2020), the research community gets interested in the functionality and benefits of the attention. To mention a few, transformers implicitly prefer hierarchical interpretations of input sequences (Kharitonov & Chaabouni, 2021); store relational knowledge in MLP layers as an associative memory (Meng et al., 2022); its computational graphs tend to be tree-structured (Murty et al., 2023b); suddenly capture tree structures of inputs after long training epochs (Murty et al., 2023a). Theoretically, training dynamics analysis explains how to learn spatially correlated patches by vision transformers (ViT) (Jelassi et al., 2022), select dominant tokens (Tian et al., 2023a), store information as an associative memory (Bietti et al., 2023), and select max-margin tokens (Tarzanagh et al., 2023b), whereas Xie et al. (2022) explains the in-context learning as a process of concept learning in Bayesian inference.

Among many aspects of attention, we specifically focus on *localization*—for a query token, self-attention can select a few relevant tokens only (which we call *localized* attention) or select many tokens *uniformly*. As attention can be regarded as a token mixer, it plays a pivotal role in studying how it selects tokens to reveal the characteristics of the token embeddings. To this end, we have the following research questions: (Q1) *When* is self-attention localized or uniform? (Q2) *How* does localization affect model performances?

Along this line, previous studies mainly investigated from the model expressivity and training stability perspectives. On the one hand, Dong et al. (2021) and Noci et al. (2022) initiated the discussion of attention localization and theoretically showed that a network with self-attention layers without skip connections exponentially loses the rank of hidden layers; the fact indicates that *the model expressivity shall be immediately lost with more self-attention layers being stacked*. On the other hand, Zhai et al. (2023) empirically found that attention entropy—averaged Shannon entropy of an attention probability matrix—correlates with the training stability. Specifically, a training loss curve tends to fall into a plateau when the attention entropy is low. Since higher entropy indicates near-uniform attention weights, their finding apparently suggests that *localized attention may lead the learning dynamics to a plateau*. Up until now, these two failure modes have been discussed independently with slightly different notions of attention

---

*Equal contribution [1] Kyoto University [2] RIKEN AIP [3] AIST. Correspondence to: Han Bao <bao@i.kyoto-u.ac.jp>.
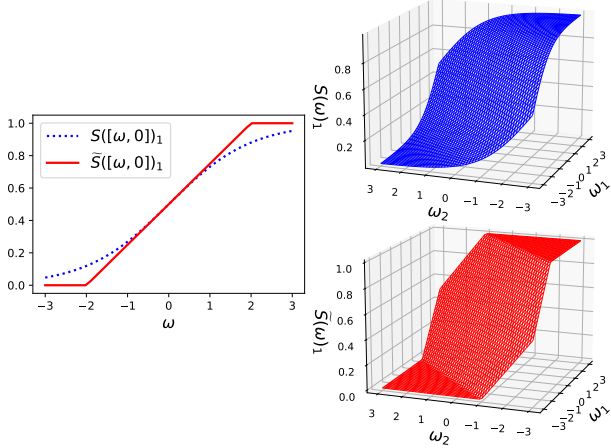
Figure 1: Comparison of softmax $\mathbf{S}$ and the piecewise approximation $\widetilde{\mathbf{S}}$ for two-dimensional inputs.

localization, and hence, our understanding of the blessing and curse of attention localization remains elusive.

To better comprehend, we characterize self-attention patterns by attention parameter matrices to reconcile the two collapse modes. We formulate the concept of localization by *signal propagation probability* (Section 3), which describes how likely the signal of a specific input token propagates to the gradient of a training objective. If the signal propagation probability is high for a few numbers of tokens only, attention is regarded to be localized, and the learning dynamics is dominated by them. We show that the localization mode can be characterized by the eigenspectrum of attention weight matrices (Section 4). Specifically, attention is localized in the above sense when the eigenspectrum of the query-key parameter matrix has a non-zero mean and a small variance. Furthermore, the small eigenspectrum variance is relevant to both the rank collapse and entropy collapse (Section 5), and thus, we give a unified perspective of the two notions of attention collapse. For this reason, we argue that attention collapse and its performance can be viewed more transparently based on the eigenspectrum variance. Lastly, we indirectly observed the correlation of the eigenspectrum and the model performance in the experiments with the WikiText dataset (Merity et al., 2016) by introducing a regularization scheme called LOCATER.

2. Setup

Notation. We write vectors with all zeros and ones by $\mathbf{0}$ and $\mathbf{1}$, respectively, whereas the i-th one-hot vector is written as $\mathbf{e}^i$. A vector is written in bold-face like $\mathbf{a}$, and its i-th scalar element is written by non-bold a_i. A matrix is written by capital bold-face like $\mathbf{A}$, and $\mathbf{A}_i$ denotes its i-th column vector unless otherwise noted. The identity matrix is denoted by $\mathbf{I}$. The Hadamard product of $\mathbf{A}$ and $\mathbf{A}$ is written by $\mathbf{A}^{\odot 2} := \mathbf{A} \odot \mathbf{A}$. We write the error function as erf and

use $\mathrm{erf}(-z) = -\mathrm{erf}(z)$ without explicitly mentioning it. Infinitesimal asymptotic orders $o(\cdot)$ are with respect to the sequence length T unless otherwise noted.

Transformer. Let $\mathbf{X} := [\mathbf{x}_1 \ \mathbf{x}_2 \ \ldots \ \mathbf{x}_T] \in \mathbb{R}^{d \times T}$ be an input with T tokens, defined later shortly. We suppose that all input sequences have the same length T, and T is occasionally taken sufficiently large. The ℓ-th (single-head) self-attention layer is defined as

$$\mathbf{A}^\ell := \mathbf{S}\left(\frac{(\mathbf{X}^{\ell-1})^\top \mathbf{W}_{\mathrm{QK}} \mathbf{X}^{\ell-1}}{\lambda} \right), \tag{1}$$

$$\mathbf{U}^\ell := \mathbf{W}_{\mathrm{V}} \mathbf{X}^{\ell-1} \mathbf{A}^\ell, \tag{2}$$

where $\mathbf{W}_{\mathrm{V}} \in \mathbb{R}^{d \times d}$ is the value parameters, $\mathbf{W}_{\mathrm{QK}}(:= \mathbf{W}_{\mathrm{Q}}^\top \mathbf{W}_{\mathrm{K}}) \in \mathbb{R}^{d \times d}$ is the query-key parameters (with joint parametrization), $\lambda > 0$ is temperature, commonly $\lambda = \sqrt{d}$, and $\mathbf{S} : \mathbb{R}^T \to \mathbb{R}^T$ is the softmax applied for each row. In this way, each input token in $\mathbf{X}^{\ell-1}$ (embedded by $\mathbf{W}_{\mathrm{V}}$) is mixed by $\mathbf{A}^\ell$. Then, the transformer block (without layer normalization (Ba et al., 2016)) is defined as

$$\mathbf{Z}^\ell := \mathbf{U}^\ell + \mathbf{X}^{\ell-1},$$
$$\mathbf{H}^\ell := \mathbf{W}_{\mathrm{F}_2} \sigma(\mathbf{W}_{\mathrm{F}_1} \mathbf{Z}^\ell),$$
$$\mathbf{X}^\ell := \mathbf{H}^\ell + \mathbf{Z}^\ell,$$

where $\mathbf{H}^\ell$ is a feed-forward network with parameters $\mathbf{W}_{\mathrm{F}_1}, \mathbf{W}_{\mathrm{F}_2} \in \mathbb{R}^{d \times d}$ and an (element-wise) activation $\sigma : \mathbb{R} \to \mathbb{R}$. We omit the token embedding layer and set $\mathbf{X}^0 := \mathbf{X}$. There are two common variants of layer normalization positions, Post-LN (Vaswani et al., 2017) and Pre-LN (Xiong et al., 2020), which are applied token-wise after the residual connections ($\mathbf{Z}^\ell$ and $\mathbf{X}^{\ell+1}$) and before the inputs ($\mathbf{X}^\ell$ and $\mathbf{Z}^\ell$), respectively. Then, the transformer block $\mathbf{X}^\ell$ is stacked L times and $\mathbf{F}(\mathbf{X}) := \mathbf{X}^L \in \mathbb{R}^{d \times T}$ is the output.

Learning task. We focus on causal language modeling, where a model predicts the next token given contextual tokens. Formally, given T contextual tokens $\mathbf{X} \in \mathbb{R}^{d \times T}$, the prediction target is the $(T+1)$-th token $\mathbf{y} := \mathbf{x}_{T+1} \in \mathbb{R}^d$. With the squared loss, the objective is written as follows:

$$J(\mathbf{\Theta}) := \frac{1}{2} \mathbb{E} \|\mathbf{y} - \mathbf{F}(\mathbf{X})_T\|^2,$$

where $\mathbf{\Theta} := (\mathbf{W}_{\mathrm{V}}, \mathbf{W}_{\mathrm{QK}}, \mathbf{W}_{\mathrm{F}_1}, \mathbf{W}_{\mathrm{F}_2})$ denotes the model parameter set, and the expectation is taken over input sequences $(\mathbf{X}, \mathbf{y})$. Here, our decoding procedure in consideration is to simply choose the embedded last token $\mathbf{F}(\mathbf{X})_T \in \mathbb{R}^T$. The parameters $\mathbf{\Theta}$ are learned by minimizing J. Note that our analysis considers optimizing the query-key parameters jointly. Although such joint parameterization is less common in practice, it is convenient for the theoretical derivation of the gradients and has been used

in several previous studies (Jelassi et al., 2022; Tian et al., 2023a). Interested readers may refer to a recent work revealing that the joint and separate QK-parametrization lead to different implicit regularizations (Tarzanagh et al., 2023a).

Picewise linear approximation of softmax.   In this article, we choose to approximate the softmax function $\mathbf{S}$ by linearization. This is for the convenience of computing Gaussian moments while keeping the attention structure as close to the original softmax as possible in Section 4. For an input $\boldsymbol{\omega} \in \mathbb{R}^T$, the softmax function is defined as

$$S(\boldsymbol{\omega})_i := \frac{\exp(\omega_i)}{\sum_{j \in [T]} \exp(\omega_j)} \quad \text{for all } i \in [T].$$

For linearization, the Taylor expansion of $S(\boldsymbol{\omega})_i$ around the origin $\langle \boldsymbol{\gamma}^i, \omega \rangle + \gamma_0^i$ is used, where

$$\boldsymbol{\gamma}^i := \nabla_i \mathbf{S}(\mathbf{0}) = \frac{1}{T}\mathbf{e}^i - \frac{1}{T^2}\mathbf{1}, \quad \gamma_0^i := S(\mathbf{0})_i = \frac{1}{T}.$$

Then, we approximate $\mathbf{S}$ by the piecewise linear function such that $S(\boldsymbol{\omega})_i \approx \max\{0, \min\{1, \langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i\}\} = \langle \widetilde{\boldsymbol{\gamma}}^i, \boldsymbol{\omega} \rangle + \widetilde{\gamma}_0^i$, where

$$(\widetilde{\boldsymbol{\gamma}}^i, \widetilde{\gamma}_0^i) = \begin{cases} (\mathbf{0}, 0) & \text{if } \langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i < 0, \\ (\boldsymbol{\gamma}^i, \gamma_0^i) & \text{if } \langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i \in [0,1], \\ (\mathbf{0}, 1) & \text{if } \langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i > 1. \end{cases}$$

In the vector form, the piecewise approximation $\mathbf{S}(\boldsymbol{\omega}) \approx \widetilde{\mathbf{S}}(\boldsymbol{\omega})$ is given by

$$\widetilde{\mathbf{S}}(\boldsymbol{\omega}) = \boldsymbol{\Gamma}^\top \boldsymbol{\omega} + \widetilde{\boldsymbol{\gamma}}_0, \quad \text{where } \begin{cases} \boldsymbol{\Gamma} := [\widetilde{\boldsymbol{\gamma}}^1 \ \widetilde{\boldsymbol{\gamma}}^2 \ \ldots \ \widetilde{\boldsymbol{\gamma}}^T], \\ \widetilde{\boldsymbol{\gamma}}_0 = [\widetilde{\gamma}_0^1, \widetilde{\gamma}_0^2, \ldots, \widetilde{\gamma}_0^T]^\top. \end{cases}$$

For notational simplicity, the column vectors of $\boldsymbol{\Gamma}$ are exceptionally denoted by $\widetilde{\boldsymbol{\gamma}}^i$ with superscripts, for which the α-th element is written by $\widetilde{\gamma}_\alpha^i$. The difference between $\mathbf{S}$ and $\widetilde{\mathbf{S}}$ is illustrated in Fig. 1. Note that a popular alternative to softmax, sparsemax (Martins & Astudillo, 2016), is also a piecewise linear function, although the functional form is slightly different from $\widetilde{\mathbf{S}}$.

Remark 1. *Each $(\widetilde{\boldsymbol{\gamma}}^i, \widetilde{\gamma}_0^i)$ depends on the softmax input $\boldsymbol{\omega}$, unlike the Taylor coefficient $(\boldsymbol{\gamma}^i, \gamma_0^i)$ being independent from $\boldsymbol{\omega}$. This point matters particularly when we take expectations of terms involving $(\widetilde{\boldsymbol{\gamma}}^i, \widetilde{\gamma}_0^i)$.*

Remark 2. *When T is sufficiently large, the coefficient vector $\boldsymbol{\gamma}^i = T^{-1}\mathbf{e}^i + o(1)$. In this regime, $\gamma_\alpha^i = o(1)$ for any $\alpha \in [T] \setminus \{i\}$, so $\boldsymbol{\gamma}^i$ behaves like a selector of the i-th input. Hence, $\gamma_i^i = \gamma_0^i = T^{-1} + o(1)$.*

3. Signal propagation probability

We analyze how much each token contributes to the learning dynamics. To this end, we formalize how much the signal

of a specific input token $\mathbf{x}_i$ propagates to the gradient ∇J. Remark that this notion is slightly different from the contribution of an input token $\mathbf{x}_i$ to the model output $\mathbf{F}(\mathbf{X})_j$ analyzed by Kobayashi et al. (2023) recently.

Uniform vs. localized softmax.   The piecewise linear approximation implies that the i-th input signal is propagated to the subsequent blocks when $\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i \in [0,1]$; otherwise, $\widetilde{S}(\boldsymbol{\omega})_i = \langle \widetilde{\boldsymbol{\gamma}}^i, \boldsymbol{\omega} \rangle + \widetilde{\gamma}_0^i = \widetilde{\gamma}_0^i$, which hinders the input token $\mathbf{x}_i$ from contributing to the self-attention layer (2). Thus, we will focus on the following quantity.

Definition 1 (Signal propagation probability). *Suppose that $\mathbf{W}_{\mathrm{QK}}$ is independent from $\mathbf{X}$. For $i \in [T]$, the* signal propagation probability *of the i-th token is defined as follows:*

$$\rho_i := \mathbb{P} \left\{ \langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i \in [0,1] \right\},$$

where $\boldsymbol{\omega} := \mathbf{X}^\top \mathbf{W}_{\mathrm{QK}} \mathbf{x}_T / \lambda$ and the randomness originates solely from the input tokens $\mathbf{X}$.

When only a few ρ_i are significantly larger than zero, we can interpret it as localized softmax; in this case, the self-attention (2) is dominated by a small number of tokens. By contrast, most of the tokens contribute to self-attention almost equally if ρ_i takes a similar value across different i; this situation is interpreted as uniform softmax.

Through the lens of gradient.   The signal propagation probability naturally arises in the gradient as a quantity characterizing how each token contributes to the gradient of the loss function J. Since the learning dynamics of causal language modeling is governed by the gradient flow of J, we can benefit from deriving the gradient of J to see how attention affects the learning dynamics. This is an important step to analyze the learning dynamics and implicit bias of self-attention layers, and we choose to measure the localization and uniformity through the signal propagation probability instead of other measures like the entropy.

To keep the derivation concise, we consider a 1-layer transformer (where we drop the superscripts ℓ) without layer normalization and simplify the feed-forward net $\mathbf{H}$ by supposing the identity activation. With the approximated softmax $\widetilde{\mathbf{S}}$, the transformer can be written as follows:

$$\begin{aligned} \mathbf{F}(\mathbf{X})_T &= \mathbf{W}_{\mathrm{F}} \{ \mathbf{W}_{\mathrm{V}} \mathbf{X} \widetilde{\mathbf{S}}(\boldsymbol{\omega}) + \mathbf{x}_T \} \\ &= \mathbf{W}_{\mathrm{F}} \{ \mathbf{W}_{\mathrm{V}} \mathbf{X} \boldsymbol{\Gamma}^\top \boldsymbol{\omega} + \mathbf{W}_{\mathrm{V}} \mathbf{X} \widetilde{\boldsymbol{\gamma}}_0 + \mathbf{x}_T \}, \end{aligned}$$

where $\mathbf{W}_{\mathrm{F}} := \mathbf{W}_{\mathrm{F}_2} \mathbf{W}_{\mathrm{F}_1} + \mathbf{I}$ and $\boldsymbol{\omega} := \mathbf{X}^\top \mathbf{W}_{\mathrm{QK}} \mathbf{x}_T / \lambda$. For this architecture, the QK-gradient is computed:

$$\begin{aligned} \nabla_{\mathbf{W}_{\mathrm{QK}}} J &= \lambda^{-2} \, \mathbb{E}[\mathbf{X} \boldsymbol{\Gamma} \mathbf{P} \boldsymbol{\Gamma}^\top \mathbf{X}^\top \mathbf{W}_{\mathrm{QK}} \mathbf{x}_T \mathbf{x}_T^\top] \\ &\quad + \lambda^{-1} \, \mathbb{E}[\mathbf{X} \boldsymbol{\Gamma} \mathbf{P} \widetilde{\boldsymbol{\gamma}}_0 \mathbf{x}_T^\top] + \lambda^{-1} \, \mathbb{E}[\mathbf{X} \boldsymbol{\Gamma} \mathbf{q} \mathbf{x}_T^\top], \end{aligned} \tag{3}$$
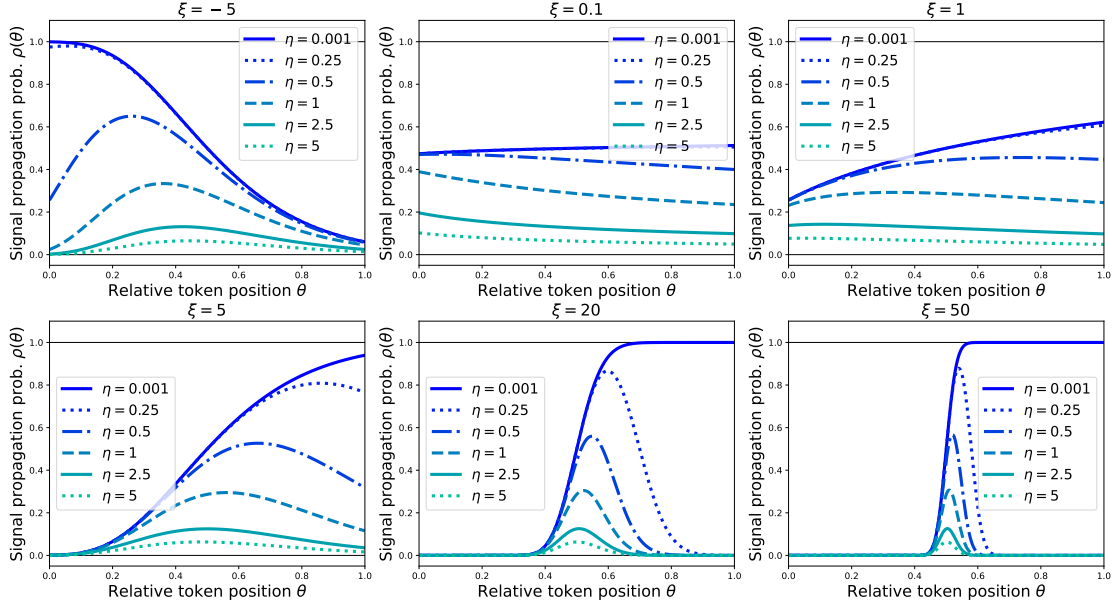
Figure 2: The theoretical plots of the signal propagation probability $\rho(\theta)$ with different $\xi = \text{tr}(\mathbf{W})/\sqrt{\text{tr}(\mathbf{W}^2)}$ and $\eta = \sqrt{\text{tr}(\mathbf{W}^2)}/\lambda^2$. The vertical axes indicate relative token position $\theta = i/T$ (i: token index, T: number of tokens). Smaller θ close to zero and larger θ close to one correspond to early-site and late-site tokens, respectively.

where

$$\mathbf{P} := \mathbf{X}^\top \mathbf{W}_V^\top \mathbf{W}_F^\top \mathbf{W}_F \mathbf{W}_V \mathbf{X},$$

$$\mathbf{q} := \mathbf{X}^\top \mathbf{W}_V^\top \mathbf{W}_F^\top (\mathbf{W}_F \mathbf{x}_T - \mathbf{y}).$$

When T is sufficiently large, we can drop asymptotically negligible terms with respect to T as detailed in Appendix B, and the QK-gradient (3) is simplified as follows:

$$\frac{1}{\lambda^2} \sum_{i,j,\alpha,\beta \in [T]} \mathbb{E}\left[\widetilde{\gamma}_\alpha^i \widetilde{\gamma}_\beta^j (\mathbf{x}_i^\top \check{\mathbf{P}} \mathbf{x}_j)(\mathbf{x}_\beta^\top \mathbf{W}_{QK} \mathbf{x}_T) \mathbf{x}_\alpha \mathbf{x}_T^\top \right], \quad (4)$$

where $\check{\mathbf{P}} := \mathbf{W}_V^\top \mathbf{W}_F^\top \mathbf{W}_F \mathbf{W}_V$.

Now, in the gradient term (4), the summands with $\widetilde{\gamma}_i^i$ (introduced in Section 2) are asymptotically dominant over those with $\widetilde{\gamma}_\alpha^i$ (with $\alpha \neq i$) because $\widetilde{\gamma}_i^i = T^{-1}$ and $\widetilde{\gamma}_\alpha^i = o(T^{-1})$. Additionally, the i-th signal propagates when $\widetilde{\gamma}_i^i > 0$, which holds iff $\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i \in [0,1]$ by definition. Therefore, we are motivated to check the condition $\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i \in [0,1]$ to see whether $\mathbf{x}_i$ contributes to the gradient (3). The signal propagation probability ρ_i characterizes its strength.

Summary. In this section, we introduced the signal propagation probability ρ_i, which characterizes how likely a given token $\mathbf{x}_i$ contributes to the learning dynamics. Specifically, $\widetilde{\gamma}^i \neq \mathbf{0}$ holds more likely with larger ρ_i, where $\mathbf{x}_i$ contributes to the QK-gradient (3). Subsequently, we will analyze the quantity ρ_i to see the behavior of the probability vector $\boldsymbol{\rho} \in [0,1]^T$. When does the mass of $\boldsymbol{\rho}$ concentrate to only a few tokens or scatter across most of the tokens?

4. When does attention localize?

We derive the signal propagation probability ρ_i based on the following synthetic data model for the sake of clarity.

Assumption 1 (Random walk). *The tokens* $(\mathbf{x}_t)_{t \geq 1}$ *are generated by the following Gaussian random walk:*

$$\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \mathbf{x}_{t+1} \sim \mathcal{N}(\mathbf{x}_t, \boldsymbol{\Sigma}).$$

We discuss the validity of this assumption at the end of this section. To derive ρ_i, we resort to the Gaussian approximation of $\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i$. Define

$$\mu^i := \mathbb{E}[\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i] \text{ and } v^i := \mathbb{V}[\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i].$$

We approximately suppose that $\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i \sim \mathcal{N}(\mu^i, v^i)$. Then, the signal propagation probability is approximated:

$$\rho_i \approx \frac{1}{2}\left\{ \text{erf}\left(\frac{1 - \mu^i}{\sqrt{2v^i}} \right) + \text{erf}\left(\frac{\mu^i}{\sqrt{2v^i}} \right) \right\}.$$

To leverage this formula, we derive μ^i and v^i.

Lemma 1. *Suppose that* $\mathbf{W}_{QK}$ *is symmetric and independent from* $\mathbf{X}$*, and let* $\mathbf{W} := \mathbf{W}_{QK}\boldsymbol{\Sigma}$*. Under Assumption 1, for* $i \in [T]$*, the mean* μ^i *and variance* v^i *of* $\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i$ *with the input* $\boldsymbol{\omega} := \mathbf{X}^\top \mathbf{W}_{QK} \mathbf{x}_T / \lambda$ *are given as follows:*

$$\mu^i = \left(\frac{i}{T} - \frac{1}{2} \right) \frac{\text{tr}(\mathbf{W})}{\lambda} + o(1),$$

$$v^i = \left(\frac{2i^2}{T^2} + \frac{7}{12} \right) \frac{\text{tr}(\mathbf{W}^2)}{\lambda^2} + o(1).$$

The proof is given in Appendix C. The symmetry of $\mathbf{W}_{QK}$ is assumed for convenience. In the case of asymmetric $\mathbf{W}_{QK}$, we can redefine the signal propagation probability with the symmetrized matrix $(\mathbf{W}_{QK} + \mathbf{W}_{QK}^{\top})/2$.

Recall that $\mathrm{tr}(\mathbf{W}) = \sum_{i \in [d]} w_i$ if we write the eigenvalues of $\mathbf{W}$ by $(w_1, w_2, \ldots, w_d)$. If $\mathbf{W}$ is real diagonalizable, $\mathrm{tr}(\mathbf{W}^2) = \sum_{i \in [d]} w_i^2$ holds, and $\mathrm{tr}(\mathbf{W})^2 \le d \, \mathrm{tr}(\mathbf{W}^2)$ follows due to Jensen's inequality. This implies that

$$-\sqrt{d \, \mathrm{tr}(\mathbf{W}^2)} \le \mathrm{tr}(\mathbf{W}) \le \sqrt{d \, \mathrm{tr}(\mathbf{W}^2)}. \qquad (5)$$

Moreover, μ^i and v^i are determined by the relative token location i/T. By continuously extending i/T to $\theta \in [0, 1]$, the signal propagation probability ρ_i can be extended to $\rho : [0, 1] \to [0, 1]$, defined over relative token locations:

$$\rho(\theta) := \Phi\left(\left(\theta - \frac{1}{2}\right)\xi; \theta\right) - \Phi\left(\left(\theta - \frac{1}{2}\right)\xi - \frac{1}{\eta}; \theta\right), \quad (6)$$

where

$$\xi := \frac{\mathrm{tr}(\mathbf{W})}{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}, \quad \eta := \frac{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}{\lambda},$$

$$\Phi(z; \theta) := \frac{1}{2}\mathrm{erf}\left(\frac{z}{\sqrt{2(2\theta^2 + \frac{7}{12})}}\right),$$

and the parameter ranges are $\xi \in [-\sqrt{d}, \sqrt{d}]$ (due to the bound (5)) and $\eta \in (0, \infty)$. Here, ξ and η can be regarded independent (when $\mathbf{W}$ is independent from $\mathbf{X}$) because the eigenspectrum scale $\mathrm{tr}(\mathbf{W}^2)$ can be modulated within the bound (5) once the eigenspectrum of $\mathbf{W}$ is given.

Figure 2 numerically illustrates $\rho(\theta)$ with different ξ and η. From these figures, we obtain a couple of observations.

- *Localization.* $\rho(\theta)$ concentrates on fewer tokens as η increases (see $|\xi| \ge 5$). By contrast, $\rho(\theta)$ behaves relatively uniformly regardless of η for small $|\xi| \le 1$.
- *Late-/middle-/early-site focus.* Focus on small η such as $\eta = 0.001$. As ξ increases to a large positive, $\rho(\theta)$ puts positive weights for only late-site tokens, i.e., $\theta > 0.5$. By contrast, as ξ decreases to a negative, $\rho(\theta)$ focuses on early-site tokens, i.e., $\theta < 0.5$. When η increases (see $\eta \ge 0.5$), $\rho(\theta)$ localizes around $\theta = 0.5$ with sufficiently large ξ (say, $|\xi| \ge 5$), which indicates middle-site focus.
- *Vanishing signal.* As η increases, $\rho(\theta)$ degenerates to zero for any $\theta \in [0, 1]$ regardless of ξ.

How ρ behaves at the limit. Subsequently, we claim the above observations formally, which is proven in Appendix C.

Lemma 2. $\rho(\theta)$ satisfies the following properties.

1. *(Late-/middle-site) As $(\xi, \eta) \to (\infty, 0)$ with $\xi\eta \to r$,*

$$\rho(\theta) \to \begin{cases} \mathbb{1}_{\{\theta \ge \frac{1}{2}\}} & \text{if } 0 \le r \le 2 \\ \mathbb{1}_{\{\frac{1}{2} \le \theta \le \frac{1}{2} + \frac{1}{r}\}} & \text{if } r > 2 \end{cases}.$$
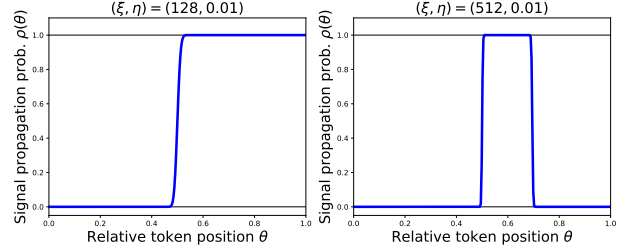

Figure 3: The theoretical plots of $\rho(\theta)$. For each $\xi = 128, 512$, the product value $\xi\eta = 1.28, 5.12$, respectively. The latter is sufficiently larger than the localization threshold $r = 2$ and localized.
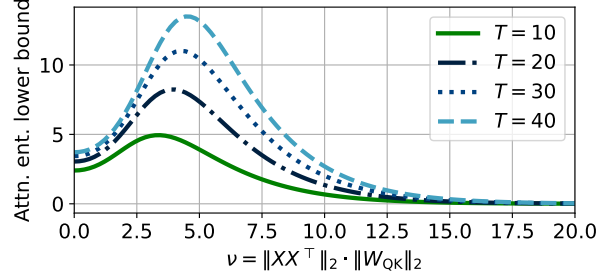

Figure 4: Entropy lower bound (9) by Zhai et al. (2023).

2. *(Early-/middle-site) As $(\xi, \eta) \to (-\infty, 0)$ with $\xi\eta \to r$,*

$$\rho(\theta) \to \begin{cases} \mathbb{1}_{\{\theta \le \frac{1}{2}\}} & \text{if } -2 \le r < 0 \\ \mathbb{1}_{\{\frac{1}{2} + \frac{1}{r} \le \theta \le \frac{1}{2}\}} & \text{if } r < -2 \end{cases}.$$

3. *(Uniformity) Fix η as a finite value. As $|\xi| \to 0$, $|\rho'(\theta)| \to 0$ for any $\theta \in [0, 1]$.*

4. *(Vanishing signal) Fix ξ as a finite value. As $\eta \to \infty$, $\rho(\theta) \to 0$ for any $\theta \in [0, 1]$.*

From late-/middle-/early-site focus in Lemma 2, we see interestingly that $\rho(\theta)$ localizes when $\xi\eta = \mathrm{tr}(\mathbf{W})/\lambda$ asymptotically deviates from zero significantly so that $|r| \gg 2$. At this limit, $\rho(\theta)$ concentrates on $\theta = 0.5$, inducing the middle-site focus. Conversely, attention becomes relatively uniform when $\xi\eta = \mathrm{tr}(\mathbf{W})/\lambda$ is kept close to zero. In Fig. 3, we numerically illustrate this regime: ρ localizes at the middle site when $\eta = 0.02$ (i.e., $\xi\eta = 5.12$).

Let us investigate the limiting condition of (ξ, η) for localization: When is (ξ, η) close to the limit $(\infty, 0)$ while $\xi\eta \to r \gg 2$? Here, we focus on the *eigenspectrum* of $\mathbf{W}$ by regarding its eigenvalues $(w_i)_{i \in [d]}$ as being sampled from a distribution with the mean $\mathrm{tr}(\mathbf{W}) = \sum_{i \in [d]} w_i$ and scale $\mathrm{tr}(\mathbf{W}^2) = \sum_{i \in [d]} w_i^2$ (supposing that $\mathbf{W}$ is real diagonalizable). First, $\eta \to 0$ indicates that the scale $\mathrm{tr}(\mathbf{W}^2)$ should be close to zero. Next, $\xi\eta (= \mathrm{tr}(\mathbf{W})/\lambda) \to r \gg 2$ means that $\mathrm{tr}(\mathbf{W}) \gg 2\lambda$ at the limit, i.e., $\mathrm{tr}(\mathbf{W})$ should be significantly away from zero. By combining them, we tell that ρ localizes when the eigenspectrum concentrates around a non-zero mean. This happens more likely when the embedding dimension d is excessively large to make

the eigenvalue sum $\operatorname{tr}(\mathbf{W})$ bounded away from zero while keeping the scale $\operatorname{tr}(\mathbf{W}^2)$ close to zero (i.e., keeping every eigenvalue close to zero). Thus, a larger embedding dimension d is beneficial to drive attention to localization.

Next, from the claim of uniformity in Lemma 2, we tell that ρ fluctuates less and less with ξ closer to zero. Hence, $\rho(\theta)$ takes a similar value across different token positions θ in this limit. When $\operatorname{tr}(\mathbf{W}) \to 0$, ρ attains this limit.

Summary. Wrapping up this section, we obtain answers to (Q1) posed in Section 1 under the random walk model.

A1: When does attention localize?

- ρ *localizes* when $\operatorname{tr}(\mathbf{W}^2)$ is close to zero while $|\operatorname{tr}(\mathbf{W})|$ is significantly bounded away from zero, i.e., $\mathbf{W}$-eigenspectrum *concentrates to a non-zero mean*.
- ρ is *uniform* when $\operatorname{tr}(\mathbf{W})$ is close to zero while $\operatorname{tr}(\mathbf{W}^2)$ remains finite, i.e., $\mathbf{W}$-eigenspectrum has *the zero mean and a finite variance*.
- ρ *degenerates* to zero uniformly when $\operatorname{tr}(\mathbf{W}^2)$ is sufficiently large, i.e., $\mathbf{W}$-eigenspectrum has *an infinitely large variance*.

Remark 3 (Validity of random walk). *The analysis results in this section largely owe to the random walk assumption in Assumption 1. This is reminiscent of the random walk model of word vectors proposed by Arora et al. (2016): they suppose that a discourse vector that governs the latent topic in a sentence evolves along a random walk, and observed word vectors are emitted depending on their affinity to the discourse vector. Under this model, they showed that the PMI (pointwise mutual information) between two words is close to their scaled inner product (Arora et al., 2016, Theorem 2.2). Hence, the random walk model may be semantically plausible in such a geometrical sense.*

Their proof essentially benefits from the random walk to claim that two discourse vectors are "independent" (Arora et al., 2016, Eq. (2.11)), for which a sufficient tiny drift order $O(d^{-1/2})$ matters. That said, we suppose the Gaussianity in addition to their random walk. We exploit this to approximately get the closed form of $\rho(\theta)$ (6), which is a limitation of our analysis.

5. Are different collapse regimes reconcilable?

We discuss the results of our analysis in Section 4 and the previous arguments related to attention uniformity.

Connection to rank collapse. Dong et al. (2021) showed that self-attention blocks $\mathbf{U}^\ell$ (see Eq. (2)) converges to a rank-1 matrix $\mathbf{z}\mathbf{1}^\top$ (for some $\mathbf{z}$) with $L \to \infty$ without skip connections or feed-forward blocks, which is called

rank collapse.[1] They argued the importance of avoiding rank collapse for better expressivity because each token embedding in a rank-1 self-attention block degenerates to the same. Hence, the rank collapse is related to the failure mode attributed to the uniformity *after* mixing key tokens by attention, which is slightly different from what we are concerned about—how each token contributes *during* mixing by attention (through the gradient, as discussed in Section 3).

Nonetheless, the uniformity of Dong et al. (2021) can be connected to the perspective of the $\mathbf{W}$-eigenspectrum. Dong et al. (2021, Theorem 2.2) proved that the convergence rate to a rank-1 matrix slows down when the matrix ℓ_1-norm $\|\mathbf{W}_{\mathrm{QK}}\|_1$, i.e., the ℓ_1-norm of the vectorized matrix, is large. Because we can draw the following connection between $\|\mathbf{W}_{\mathrm{QK}}\|_1$ and $|\operatorname{tr}(\mathbf{W})|$:

$$\frac{|\operatorname{tr}(\mathbf{W})|}{\sqrt{d}\|\mathbf{\Sigma}\|_2} \leq \|\mathbf{W}_{\mathrm{QK}}\|_2 \leq \|\mathbf{W}_{\mathrm{QK}}\|_{\mathrm{F}} \leq \|\mathbf{W}_{\mathrm{QK}}\|_1, \quad (7)$$

where the first inequality is due to the bound (5) and the Cauchy–Schwarz inequality, it is sufficient (but not necessary!) to increase $|\operatorname{tr}(\mathbf{W})|$ under fixed $\operatorname{tr}(\mathbf{W}^2)$ to mitigate the rank collapse. This is equivalent to reducing the eigenspectrum variance

$$d^2\mathbb{V}[w_i] = d^2(\mathbb{E}[w_i^2] - \mathbb{E}[w_i]^2) = d\operatorname{tr}(\mathbf{W}^2) - |\operatorname{tr}(\mathbf{W})|^2. \tag{8}$$

Hence, *minimizing the $\mathbf{W}$-eigenspectrum variance leads to better expressivity*.

Connection to entropy collapse. Zhai et al. (2023) introduced a concept called *entropy collapse*, in which the average Shannon entropy of the columns of the attention matrix $\mathbf{A}^\ell$ (see Eq. (1)) shrinks. Intuitively speaking, low attention entropy induces localized attention. This notion of localization is akin to ours because the attention entropy measures the uniformity the attention is applied to input tokens *during* mixing. They empirically observed that the training loss falls into a plateau with low attention entropy, which causes training instability of transformers, and hence advocate for keeping attention less peaked during training.

In Zhai et al. (2023, Theorem 3.1), the attention entropy is asymptotically lower-bounded for large T by

$$\ln(1 + T\exp(-\nu)) + \frac{\nu\exp(-\nu/2)}{T^{-1} + \exp(-\nu)}, \qquad (9)$$

where $\nu := \|\mathbf{X}\mathbf{X}^\top\|_2\|\mathbf{W}_{\mathrm{QK}}\|_2$. This lower bound is unimodal in ν and vanishes at $\|\mathbf{W}_{\mathrm{QK}}\|_2 \to \infty$ (see Fig. 4), so the attention entropy tends to be higher when $\|\mathbf{W}_{\mathrm{QK}}\|_2$ is small. If $|\operatorname{tr}(\mathbf{W})|$ is not too small, $\|\mathbf{W}_{\mathrm{QK}}\|_2$ is lower-bounded (see Eq. (7)) and the attention entropy may be kept

---

[1]Note that our matrix notation is different from the one used in Dong et al. (2021) so that we chose to let each column of $\mathbf{X}$ store a token, whereas they let each row of $\mathbf{X}$ store a token.
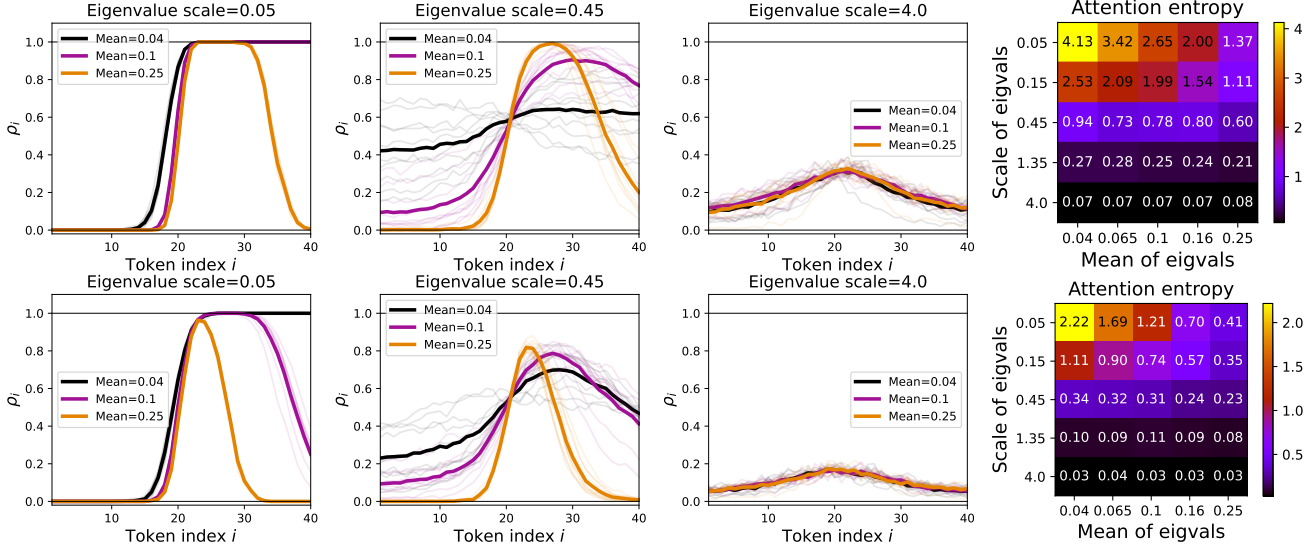
Figure 5: Simulated signal propagation probability. In the **top** and **bottom** rows, the results for the isotropic and anisotropic covariances (the details in the text) are shown, respectively. (**Left**) Signal propagation probability ρ_i computed over repeatedly sampled 300 random walks (Assumption 1) with 40 tokens. For each line, $\mathbf{W}_{\mathrm{QK}}$ ($d = 128$) is sampled 10 times with the corresponding mean and scale of the eigenvalue distribution, and the averaged ρ_i is denoted by the bold line. (**Right**) The attention entropy (Zhai et al., 2023) is computed for $\mathbf{W}_{\mathrm{QK}}$ with different eigenvalue mean-scale pairs.

close to the peak of the lower bound (9). To mitigate the entropy collapse, it is sufficient (but not necessary!) to decrease $\mathrm{tr}(\mathbf{W}^2)$ under fixed $\mathrm{tr}(\mathbf{W})$ (which is equivalent to minimizing the eigenspectrum variance by Eq. (8)) because of the bound

$$
\begin{aligned}
\|\mathbf{\Sigma}^{-1}\|_{\mathrm{F}} \sqrt{\mathrm{tr}(\mathbf{W}^2)} &= \|\mathbf{\Sigma}^{-1}\|_{\mathrm{F}} \|\mathbf{W}\|_{\mathrm{F}} \\
&\geq \|\mathbf{W}_{\mathrm{QK}}\|_{\mathrm{F}} \geq \|\mathbf{W}_{\mathrm{QK}}\|_2,
\end{aligned}
\tag{10}
$$

where the first inequality is due to the Cauchy–Schwarz inequality. Hence, *minimizing the $\mathbf{W}$-eigenspectrum variance helps the model to avoid the entropy collapse.*

Rank collapse vs. entropy collapse. At first sight, the two notions of collapse seem to contradict each other because avoiding the rank collapse leads to diverse token embeddings, whereas avoiding the entropy collapse leads to a uniform token mixer. Indeed, the matrix ℓ_1-norm (that decreases under the rank collapse) and the spectral norm (that increases under the entropy collapse) are equivalent norms, and the two modes appear to be incompatible.

However, as we discussed above, this trade-off is reconcilable from the viewpoint of the $\mathbf{W}$-eigenspectrum. Setting the eigenspectrum mean to be bounded away from zero, we can avoid the rank collapse owing to the bound (7). Under a fixed eigenspectrum mean, minimizing the eigenspectrum scale (equivalently, minimizing its variance (8)) leads to high attention entropy due to the bound (10) and the unimodal shape of the entropy lower bound (9). This variance minimization is nothing else but the condition of attention localization. Eventually, $\rho(\theta)$ localizes and attends to spe-

cific sites of tokens, as we showed in Lemma 2. Hence, the signal propagation probability offers us a better view of localization. Let us summarize our second take-home.

A2: How does attention localization impact?

- *Better expressivity*: If $|\mathrm{tr}(\mathbf{W})|$ is maximized for a fixed $\mathrm{tr}(\mathbf{W}^2)$, the convergence to the rank collapse becomes slow.

- *High attention entropy*: If $\mathrm{tr}(\mathbf{W}^2)$ is minimized for a fixed $|\mathrm{tr}(\mathbf{W})|$ bounded away from zero, the attention entropy is increased.

Both of the above are attributed to minimizing the $\mathbf{W}$-eigenspectrum variance (8).

Finally, let us reiterate that the minimization of the eigenspectrum variance is merely a sufficient but *not necessary* condition of the aforementioned better expressivity and high attention entropy. The apparent contradiction between the rank and entropy collapses could be reconciled by other mechanisms in practice. The aim of this work is to elucidate the connection between the collapse phenomena and the eigenspectrum.

Numerical simulation. To see the relationship between the eigenspectrum, ρ, and attention entropy, we simulated the signal propagation probability ρ_i using synthesized random walks following Assumption 1 with the isotropic $\mathbf{\Sigma} = \mathbf{I}$ and anisotropic $\mathbf{\Sigma}$. To obtain an anisotropic $\mathbf{\Sigma}$, we first sampled $\mathbf{R} \in \mathbb{R}^{d \times d}$ from element-wise $\mathrm{Unif}(-2.5, 2.5)$ and computed $\mathbf{\Sigma} = \mathbf{R}^{\top}\mathbf{R}/d$. We sampled 300 sequences
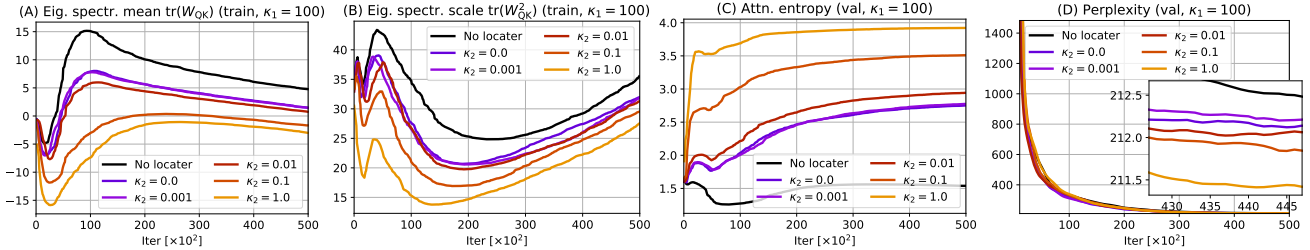
Figure 6: Experimental results of language modeling (WikiText-2) with $d = 128$ and 1-layer transformers, fixed $\kappa_1 = 100$, and varying regularization intensity κ_2. With stronger κ_2, the eigenspectrum scale shrinks **(B)**, the attention entropy increases **(C)**, and the perplexity improves **(D)**. The rightmost figure magnifies the x-range, where the perplexity attains the minimum.
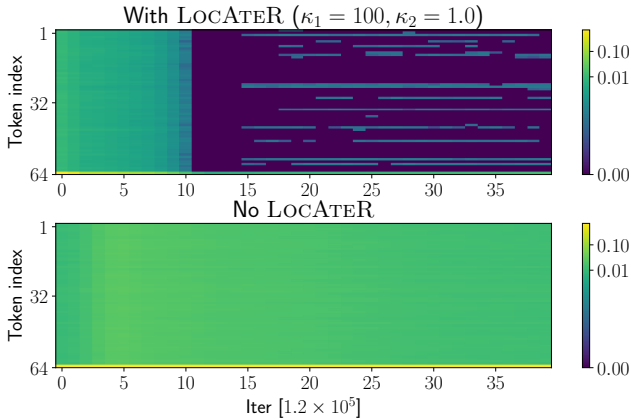

Figure 7: The signal propagation probabilities are shown at each iteration over 50000 iterations. **(Top)** LOCATER with $\kappa_1 = 100$ and $\kappa_2 = 1$. A couple of light and dark horizontal stripes correspond to the attention localization. **(Bottom)** No LOCATER. Overall, the signal propagation probability is uniform at each time.

with 40 tokens, and obtained $\mathbf{W}_{QK}$ by generating 128 eigenvalues $(w_i)_{i\in[d]}$ from $\mathcal{N}(\text{mean}, \text{scale}^2)$ and composed with a sampled orthogonal basis matrix $\mathbf{B}$, by the eigendecomposition formula $\mathbf{W}_{QK} = \mathbf{B}\operatorname{diag}((w_i)_i)\mathbf{B}^\top$. The signal propagation probability was averaged over 300 sequences.

Figure 5 shows the averaged ρ_i with different mean-scale pairs and the corresponding attention entropy in the rightmost figure. As seen, ρ_i localizes with smaller scales and larger means, which is consistent with the conclusion in Section 4. This trend supports the validity of $\mathbf{W}_{QK}$-eigenspectrum as a proxy to $\mathbf{W}$-eigenspectrum. Moreover, we observe that $\mathbf{W}_{QK}$-eigenspectrum with a fixed mean and scale leads to higher attention entropy.

6. Intervening attention localization

To empirically see the impact of localization on the model performance, we propose a method to control the degree of attention localization. We focus on the eigenspectrum of $\mathbf{W}_{QK}$ instead of $\mathbf{W} = \mathbf{W}_{QK}\boldsymbol{\Sigma}$ because $\boldsymbol{\Sigma}$ does not change during training, and the numerical simulation showed that $\mathbf{W}_{QK}$ behaves as a reasonable proxy to $\mathbf{W}$ (see Section 5).

We minimize the loss function J while minimizing the eigenspectrum scale and maintaining the mean to a fixed level:

$$\min_{\boldsymbol{\Theta}} \left\{ J(\boldsymbol{\Theta}) + \kappa_1 \operatorname{tr}(\mathbf{W}_{QK}^\top \mathbf{W}_{QK}) + \kappa_2 (\operatorname{tr}(\mathbf{W}_{QK}) - 1)^2 \right\},$$

where $\kappa_1, \kappa_2 > 0$ are the regularization strengths. Here, we allow $\mathbf{W}_{QK}$ to be asymmetric, and the eigenspectrum scale is represented by $\operatorname{tr}(\mathbf{W}_{QK}^\top \mathbf{W}_{QK})$. The regularization terms can be optimized fairly easily thanks to the following derivative formulae (for $\mathbf{W}_{QK} = \mathbf{W}_Q^\top \mathbf{W}_K$):

$$\nabla_{\mathbf{W}_Q} \operatorname{tr}(\mathbf{W}_{QK}^\top \mathbf{W}_{QK}) = 2\mathbf{W}_K \mathbf{W}_K^\top \mathbf{W}_Q,$$

$$\nabla_{\mathbf{W}_K} \operatorname{tr}(\mathbf{W}_{QK}^\top \mathbf{W}_{QK}) = 2\mathbf{W}_Q \mathbf{W}_Q^\top \mathbf{W}_K,$$

$$\nabla_{\mathbf{W}_Q}[(\operatorname{tr}(\mathbf{W}_{QK}) - 1)^2] = [\operatorname{tr}(\mathbf{W}_{QK}) - 1]\operatorname{tr}(\mathbf{W}_{QK})\mathbf{W}_K,$$

$$\nabla_{\mathbf{W}_K}[(\operatorname{tr}(\mathbf{W}_{QK}) - 1)^2] = [\operatorname{tr}(\mathbf{W}_{QK}) - 1]\operatorname{tr}(\mathbf{W}_{QK})\mathbf{W}_Q.$$

Since this whole objective drives the eigenspectrum scale to a small value, the signal vanishing can be avoided automatically. We call this regularization scheme LOCATER (LOCalized ATtEntion Regularization).

7. Experiments

We aim to observe the correlation between the eigenspectrum and localization. To this end, we train transformers with LOCATER and varying κ_1, κ_2, and see how the model performances and attention foci change over time.

Setup. We used fairseq v0.12.2 (Ott et al., 2019), which is a toolkit oriented for sequence modeling, to implement and train transformers. The basic training scheme was inherited from fairseq-cli/train.py. The model is a 1-layer transformer with a single-head self-attention and Post-LN (default), and the input embedding dimension, attention embedding dimension, and feed-forward net embedding dimension are set to 128 altogether (namely, $d = 128$).[2] Note that we use the standard softmax attention in the experiments, although our theoretical analysis was conducted with its piecewise approximation. Input data

---

[2]The experimental results with deeper transformers are shown in Appendix D. The overall trends remain alike.

were transformed into 64 tokens (namely, $T = 64$) with batch size 64. The optimizer is Adam (Kingma & Ba, 2015) with default parameters and no clip norm, and the weight decay with 0.01 is used. The learning rate is fixed to 2.5×10^{-5} without any scheduling. The FP16 quantizer was applied to reduce memory usage. All the other configs remain to be the same as the default in `fairseq-cli/train.py`. Under this config, we updated the model with 50000 iters.

Language modeling. We conducted the language modeling task. The dataset we used is WikiText-2 (Merity et al., 2016), which is a collection of high-quality Wikipedia articles. We conduced the experiments with fixed scale regularization strength $\kappa_1 = 100$ and varying mean regularization strength κ_2 from $0, 10^{-3}, 10^{-2}, \ldots, 10^0$.

The results are shown in Fig. 6, in which stronger regularizers tend to make the eigenspectrum scale smaller. This, in turn, maintains the attention entropy higher during the updates entirely, and eventually, the model achieves better perplexity. While the better model performance with higher attention entropy has already been observed by Zhai et al. (2023), we also showed that a smaller eigenspectrum scale contributes to higher attention entropy. This empirically corroborates that attention localization leads to better model performance, probably because the attention mechanism appropriately selects relevant tokens during training.

Figure 7 shows the signal propagation probability at each training iteration. We compute the signal propagation probability of token i by counting the frequency of $\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i \in [0, 1]$ in a given batch. LOCATER entails salient horizontal stripes, each corresponding to attended tokens. Yet, the stripes do not appear in "bulk" as we analyzed in Section 4 because our synthetic data model in Assumption 1 does not perfectly align with real datasets. Nevertheless, our experiments evidently contrast the localized and uniformed attention depending on the eigenspectrum scale because no salient stripes are observed without LOCATER.

In Figs. 6 and 7, we observe different learning phases for the first 10^4 and the rest iters. Indeed, Tian et al. (2023b) observed similar phenomena and explained that it is due to the different convergence speeds between attention weights corresponding to informative and non-informative tokens. The relationship between the $\mathbf{W}_{\mathrm{QK}}$-eigenspectrum and this dynamics is beyond our scope and left for future work.

8. Conclusion and limitation

We revealed that attention localizes when the eigenspectrum of $\mathbf{W}$ concentrates to a non-zero mean, or equivalently, with larger eigenspectrum mean $\mathrm{tr}(\mathbf{W})$ and smaller scale $\mathrm{tr}(\mathbf{W}^2)$. Based on it, LOCATER was proposed to shrink the scale $\mathrm{tr}(\mathbf{W}_{\mathrm{QK}}^2)$ while maintaining the mean $\mathrm{tr}(\mathbf{W}_{\mathrm{QK}})$. Interestingly, maximizing the scale is related to mitigating

both rank collapse and entropy collapse, and hence, the two apparently contradictory failure modes can be reconciled. The experiments on a real-world dataset corroborate it, though the random walk model is not perfectly satisfied.

We recognize three limitations of this work. First, we rely on the strong random walk model. Although the Gaussianity may be reasonable because of usual initialization schemes of transformer embedding layers, it is interesting to consider an alternative model to capture token correlations better. Second, the formal analysis is mainly restricted to 1-layer transformers. Recent studies often consider gradient explosion in the large-depth limit from the viewpoint of layer normalization (Xiong et al., 2020; Takase et al., 2023b) and initialization (Bachlechner et al., 2021; Takase et al., 2023a). It must be fruitful to integrate these perspectives to our gradient analysis through Eq. (3). Third, why attention localization leads to better model performance still remains elusive. Whereas localization is related to avoiding rank collapse (and hence higher model expressivity), we need additional effort to fully understand the mechanism.

Last but not least, let us mention concurrent work on understanding the attention mechanism. Geshkovski et al. (2023) argued that a trained multi-layer self-attention network exhibits a layer-wise dynamics similar to the Kuramoto model and token embeddings converge to a few "leader" tokens depending on the structure determined by the self-attention parameter matrices. Li et al. (2024) proved that the learning dynamics of a 1-layer self-attention network yields a query-key parameter matrix capturing the token-pair frequencies. Together with our work, it has become an interesting direction to study the implicit bias of attention through parameter eigenspectra.

Impact Statement

Transformer architectures have prevailed in recent years, yet the internal functionality has not been transparently understood. Our work reconciles the apparently contradictory two collapse modes, rank collapse and entropy collapse, with a new perspective of the parameter eigenspectrum. We believe this result pushes forward the understanding of transformers.

Acknowledgments

References

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Lin-*

guistics, 4:385–399, 2016.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bachlechner, T., Majumder, B. P., Mao, H., Cottrell, G., and McAuley, J. ReZero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence*, pp. 1352–1361. PMLR, 2021.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36, 2023.

Brookes, M. The matrix reference manual. http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html, 1998. [Online; accessed 01-September-2023].

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901, 2020.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2793–2803. PMLR, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2023.

Graves, A. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Jelassi, S., Sander, M., and Li, Y. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Kharitonov, E. and Chaabouni, R. What they do when in doubt: a study of inductive biases in seq2seq learners. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Kobayashi, G., Kuribayashi, T., Yokoi, S., and Inui, K. Analyzing feed-forward blocks in transformers through the lens of attention map. *arXiv preprint arXiv:2302.00456*, 2023.

Li, Y., Huang, Y., Ildiz, M. E., Rawat, A. S., and Oymak, S. Mechanics of next token prediction with self-attention. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 685–693. PMLR, 2024.

Likhomanenko, T., Xu, Q., Kahn, J., Synnaeve, G., and Collobert, R. slimIPL: Language-model-free iterative pseudo-labeling. In *Proc. Interspeech 2021*, pp. 741–745, 2021.

Martins, A. and Astudillo, R. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1614–1623. PMLR, 2016.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Murty, S., Sharma, P., Andreas, J., and Manning, C. Grokking of hierarchical structure in vanilla transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 439–448. Association for Computational Linguistics, 2023a.

Murty, S., Sharma, P., Andreas, J., and Manning, C. D. Characterizing intrinsic compositionality in transformers with tree projections. In *Proceedings of the 11th International Conference on Learning Representations*, 2023b.

Noci, L., Anagnostidis, S., Biggio, L., Orvieto, A., Singh, S. P., and Lucchi, A. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35: 27198–27211, 2022.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, 2019.

Takase, S., Kiyono, S., Kobayashi, S., and Suzuki, J. Spike no more: Stabilizing the pre-training of large language models. *arXiv preprint arXiv:2312.16903*, 2023a.

Takase, S., Kiyono, S., Kobayashi, S., and Suzuki, J. B2T connection: Serving stability and performance in deep transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3078–3095, 2023b.

Tarzanagh, D. A., Li, Y., Thrampoulidis, C., and Oymak, S. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023a.

Tarzanagh, D. A., Li, Y., Zhang, X., and Oymak, S. Max-margin token selection in attention mechanism. *Advances in Neural Information Processing Systems*, 36, 2023b.

Tian, Y., Wang, Y., Chen, B., and Du, S. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems*, 36, 2023a.

Tian, Y., Wang, Y., Zhang, Z., Chen, B., and Du, S. JoMA: Demystifying multilayer transformers via JOint Dynamics of MLP and Attention. *arXiv preprint arXiv:2310.00535*, 2023b.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:6000–6010, 2017.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit Bayesian inference. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10524–10533. PMLR, 2020.

Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., and Susskind, J. M. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 40770–40803. PMLR, 2023.

A. Helper lemmas

Lemma 3. *Let* $\mathbf{W} \in \mathbb{R}^{d \times d}$ *be a symmetric matrix. Fix* $\mathbf{a}, \boldsymbol{\mu} \in \mathbb{R}^d$ *and* $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ *be a covariance matrix. For* $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$, *the following moment formulae hold:*

$$\mathbb{E}[\mathbf{x}^\top \mathbf{W} \mathbf{x}] = \operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}) + \mathbf{m}^\top \mathbf{W}\mathbf{m} \tag{11}$$

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma} + \mathbf{m}\mathbf{m}^\top \tag{12}$$

$$\mathbb{E}[\mathbf{a}^\top \mathbf{W}\mathbf{x}\mathbf{x}^\top \mathbf{W}\mathbf{x}] = 2\mathbf{a}^\top \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\mathbf{m} + \mathbf{a}^\top \mathbf{W}\mathbf{m}\{\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}) + \mathbf{m}^\top \mathbf{W}\mathbf{m}\} \tag{13}$$

$$\mathbb{E}[\mathbf{x}^\top \mathbf{W}\mathbf{x}\mathbf{x}^\top \mathbf{W}\mathbf{x}] = 2\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}) + \operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})^2 + 4\mathbf{m}^\top \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\mathbf{m} + 2\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})\mathbf{m}^\top \mathbf{W}\mathbf{m} + \mathbf{m}^\top \mathbf{W}\mathbf{m}\mathbf{m}^\top \mathbf{W}\mathbf{m} \tag{14}$$

The formulae in Lemma 3 are standard and cropped from Brookes (1998).

Lemma 4. *Let* $\mathbf{W} \in \mathbb{R}^{d \times d}$ *be a symmetric matrix. For* $i \leq j$, *suppose that* $\mathbf{x}_i, \mathbf{x}_j$ *follow Assumption 1. Then, the following formulae hold:*

$$\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_i] = (i-1)\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}) \tag{15}$$

$$\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_i\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_i] = (i^2 - 2i + 2)\{2\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}) + \operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})^2\} \tag{16}$$

$$\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_j\mathbf{x}_j^\top \mathbf{W}\mathbf{x}_i] = (i^2 + ij - 3i - j + 4)\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}) + (i^2 - 2i + 2)\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})^2 \tag{17}$$

$$\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_j\mathbf{x}_j^\top \mathbf{W}\mathbf{x}_j] = (ij - i - j + 2)\{2\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}) + \operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})^2\} \tag{18}$$

The formulae in Lemma 4 can be shown by recursively applying Lemma 3. We omit the proofs since they are elementary.

B. Omitted derivations

B.1. QK-gradient

Here, we complement the derivations of the QK-gradient terms shown in Section 3. To get Eq. (3), we compute $\nabla_{\mathbf{W}_{\mathrm{QK}}} J$:

$$\begin{aligned}
\nabla_{\mathbf{W}_{\mathrm{QK}}} J &= \frac{1}{2}\mathbb{E}[\nabla_{\mathbf{W}_{\mathrm{QK}}}\|\mathbf{y} - \mathbf{F}(\mathbf{X})_T\|^2] \\
&= \frac{1}{2}\mathbb{E}[\nabla_{\mathbf{W}_{\mathrm{QK}}}\{\mathbf{F}(\mathbf{X})_T^\top \mathbf{F}(\mathbf{X})_T - 2\mathbf{y}^\top \mathbf{F}(\mathbf{X})_T\}] \\
&= \frac{1}{2}\mathbb{E}[\boldsymbol{\omega}^\top \boldsymbol{\Gamma}\mathbf{X}^\top \mathbf{W}_{\mathrm{V}}^\top \mathbf{W}_{\mathrm{F}}^\top \mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X}\boldsymbol{\Gamma}^\top \boldsymbol{\omega} + 2(\mathbf{W}_{\mathrm{V}}\mathbf{X}\widetilde{\boldsymbol{\gamma}}_0 + \mathbf{x}_T)^\top \mathbf{W}_{\mathrm{F}}^\top \mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X}\boldsymbol{\Gamma}^\top \boldsymbol{\omega} - 2\mathbf{y}^\top \mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X}\boldsymbol{\Gamma}^\top \boldsymbol{\omega}] \\
&= \frac{1}{2\lambda^2}\mathbb{E}[\mathbf{x}_T^\top \mathbf{W}_{\mathrm{QK}}^\top \mathbf{X}\boldsymbol{\Gamma}\mathbf{P}\boldsymbol{\Gamma}^\top \mathbf{X}^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T] + \frac{1}{\lambda}\mathbb{E}[(\widetilde{\boldsymbol{\gamma}}_0)^\top \mathbf{P}\boldsymbol{\Gamma}^\top \mathbf{X}^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T] + \frac{1}{\lambda}\mathbb{E}[\mathbf{q}^\top \boldsymbol{\Gamma}^\top \mathbf{X}^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T] \\
&= \frac{1}{\lambda^2}\mathbb{E}[\mathbf{X}\boldsymbol{\Gamma}\mathbf{P}\boldsymbol{\Gamma}^\top \mathbf{X}^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T\mathbf{x}_T^\top] + \frac{1}{\lambda}\mathbb{E}[\mathbf{X}\boldsymbol{\Gamma}\mathbf{P}\widetilde{\boldsymbol{\gamma}}_0\mathbf{x}_T^\top] + \frac{1}{\lambda}\mathbb{E}[\mathbf{X}\boldsymbol{\Gamma}\mathbf{q}\mathbf{x}_T^\top].
\end{aligned}$$

By expanding the first term of $\nabla_{\mathbf{W}_{\mathrm{QK}}} J$, we get the following:

$$\frac{1}{\lambda^2}\sum_{i,j,\alpha,\beta \in [T]} \mathbb{E}\left[\widetilde{\gamma}_\alpha^i \widetilde{\gamma}_\beta^j (\mathbf{x}_i^\top \check{\mathbf{P}}\mathbf{x}_j)(\mathbf{x}_\beta^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T)\mathbf{x}_\alpha \mathbf{x}_T^\top\right], \tag{19}$$

where $\check{\mathbf{P}} := \mathbf{W}_{\mathrm{V}}^\top \mathbf{W}_{\mathrm{F}}^\top \mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}$. Similarly, by expanding the second and third terms of $\nabla_{\mathbf{W}_{\mathrm{QK}}} J$, we get the following terms, respectively:

$$\frac{1}{\lambda}\sum_{i,\alpha,\beta \in [T]} \mathbb{E}\left[\widetilde{\gamma}_\alpha^i \widetilde{\gamma}_0^\beta (\mathbf{x}_i^\top \check{\mathbf{P}}\mathbf{x}_\beta)\mathbf{x}_\alpha \mathbf{x}_T^\top\right], \tag{20}$$

$$\frac{1}{\lambda}\sum_{i,\alpha \in [T]} \mathbb{E}\left[\widetilde{\gamma}_\alpha^i \{\mathbf{x}_i^\top \mathbf{W}_{\mathrm{V}}^\top \mathbf{W}_{\mathrm{F}}^\top (\mathbf{W}_{\mathrm{F}}\mathbf{x}_T - \mathbf{y})\}\mathbf{x}_\alpha \mathbf{x}_T^\top\right]. \tag{21}$$

To get Eq. (19), we expand the first term of $\nabla_{\mathbf{W}_{\mathrm{QK}}} J$:

$$
\begin{aligned}
\mathbb{E}[\mathbf{X}\boldsymbol{\Gamma}\mathbf{P}\boldsymbol{\Gamma}^\top\mathbf{X}^\top\mathbf{W}_{\mathrm{QK}}\mathbf{x}_T\mathbf{x}_T^\top] &= \mathbb{E}[\mathbf{X}\boldsymbol{\Gamma}\mathbf{X}^\top\mathbf{W}_{\mathrm{V}}^\top\mathbf{W}_{\mathrm{F}}^\top\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X}\boldsymbol{\Gamma}^\top\mathbf{X}^\top\mathbf{W}_{\mathrm{QK}}\mathbf{x}_T\mathbf{x}_T^\top] \\
&= \mathbb{E}\left[\{(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X})(\mathbf{X}\boldsymbol{\Gamma})^\top\}^\top\{(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X})(\mathbf{X}\boldsymbol{\Gamma})^\top\}\mathbf{W}_{\mathrm{QK}}\mathbf{x}_T\mathbf{x}_T^\top\right] \\
&= \mathbb{E}\left[\left\{\sum_{i\in[T]}(\mathbf{X}\widetilde{\boldsymbol{\gamma}}^i)(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{x}_i)^\top\right\}\left\{\sum_{j\in[T]}(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{x}_j)(\mathbf{X}\widetilde{\boldsymbol{\gamma}}^j)^\top\right\}\mathbf{W}_{\mathrm{QK}}\mathbf{x}_T\mathbf{x}_T^\top\right] \\
&= \mathbb{E}\left[\sum_{i,j\in[T]}(\mathbf{X}\widetilde{\boldsymbol{\gamma}}^i)\{(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{x}_i)^\top(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{x}_j)\}(\mathbf{X}\widetilde{\boldsymbol{\gamma}}^j)^\top\mathbf{W}_{\mathrm{QK}}\mathbf{x}_T\mathbf{x}_T^\top\right] \\
&= \mathbb{E}\left[\sum_{\alpha,\beta\in[T]}\left\{\sum_{i,j\in[T]}\mathbf{x}_i^\top\check{\mathbf{P}}\mathbf{x}_j\right\}\widetilde{\gamma}_\alpha^i\widetilde{\gamma}_\beta^j\mathbf{x}_\alpha\mathbf{x}_\beta^\top\mathbf{W}_{\mathrm{QK}}\mathbf{x}_T\mathbf{x}_T^\top\right] \\
&= \sum_{i,j,\alpha,\beta\in[T]}\mathbb{E}[\widetilde{\gamma}_\alpha^i\widetilde{\gamma}_\beta^j(\mathbf{x}_i^\top\check{\mathbf{P}}\mathbf{x}_j)(\mathbf{x}_\beta^\top\mathbf{W}_{\mathrm{QK}}\mathbf{x}_T)\mathbf{x}_\alpha\mathbf{x}_T^\top].
\end{aligned}
$$

To get Eq. (20), we expand the second term of $\nabla_{\mathbf{W}_{\mathrm{QK}}} J$:

$$
\begin{aligned}
\mathbb{E}[\mathbf{X}\boldsymbol{\Gamma}\mathbf{P}\widetilde{\boldsymbol{\gamma}}_0\mathbf{x}_T^\top] &= \mathbb{E}[\mathbf{X}\boldsymbol{\Gamma}\mathbf{X}^\top\mathbf{W}_{\mathrm{V}}^\top\mathbf{W}_{\mathrm{F}}^\top\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X}\widetilde{\boldsymbol{\gamma}}_0\mathbf{x}_T^\top] \\
&= \mathbb{E}[\{(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X})(\mathbf{X}\boldsymbol{\Gamma})^\top\}^\top\{(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X})(\widetilde{\boldsymbol{\gamma}}_0\mathbf{x}_T^\top)\}] \\
&= \mathbb{E}\left[\left\{\sum_{i\in[T]}(\mathbf{X}\widetilde{\boldsymbol{\gamma}}^i)(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{x}_i)^\top\right\}\left\{\sum_{\beta\in[T]}(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{x}_\beta)(\widetilde{\gamma}_0^\beta\mathbf{x}_T^\top)\right\}\right] \\
&= \mathbb{E}\left[\sum_{i,\beta\in[T]}(\mathbf{X}\widetilde{\boldsymbol{\gamma}}^i)\{(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{x}_i)^\top(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{x}_\beta)\}(\widetilde{\gamma}_0^\beta\mathbf{x}_T^\top)\right] \\
&= \mathbb{E}\left[\sum_{\alpha\in[T]}\left\{\sum_{i,\beta\in[T]}\mathbf{x}_i^\top\check{\mathbf{P}}\mathbf{x}_\beta\right\}\widetilde{\gamma}_\alpha^i\widetilde{\gamma}_0^\beta\mathbf{x}_\alpha\mathbf{x}_T^\top\right] \\
&= \sum_{i,\alpha,\beta\in[T]}\mathbb{E}[\widetilde{\gamma}_\alpha^i\widetilde{\gamma}_0^\beta(\mathbf{x}_i^\top\check{\mathbf{P}}\mathbf{x}_\beta)\mathbf{x}_\alpha\mathbf{x}_T^\top].
\end{aligned}
$$

To get Eq. (21), we expand the third term of $\nabla_{\mathbf{W}_{\mathrm{QK}}} J$:

$$
\begin{aligned}
\mathbb{E}[\mathbf{X}\boldsymbol{\Gamma}\mathbf{q}\mathbf{x}_T^\top] &= \mathbb{E}[\mathbf{X}\boldsymbol{\Gamma}\mathbf{X}^\top\mathbf{W}_{\mathrm{V}}^\top\mathbf{W}_{\mathrm{F}}^\top(\mathbf{W}_{\mathrm{F}}\mathbf{x}_T - \mathbf{y})\mathbf{x}_T^\top] \\
&= \mathbb{E}[\{(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{X})(\mathbf{X}\boldsymbol{\Gamma})^\top\}^\top(\mathbf{W}_{\mathrm{F}}\mathbf{x}_T - \mathbf{y})\mathbf{x}_T^\top] \\
&= \mathbb{E}\left[\left\{\sum_{i\in[T]}(\mathbf{X}\widetilde{\boldsymbol{\gamma}}^i)(\mathbf{W}_{\mathrm{F}}\mathbf{W}_{\mathrm{V}}\mathbf{x}_i)^\top\right\}(\mathbf{W}_{\mathrm{F}}\mathbf{x}_T - \mathbf{y})\mathbf{x}_T^\top\right] \\
&= \mathbb{E}\left[\left\{\sum_{i,\alpha\in[T]}\widetilde{\gamma}_\alpha^i\mathbf{x}_\alpha\mathbf{x}_i^\top\mathbf{W}_{\mathrm{V}}^\top\mathbf{W}_{\mathrm{F}}^\top\right\}(\mathbf{W}_{\mathrm{F}}\mathbf{x}_T - \mathbf{y})\mathbf{x}_T^\top\right] \\
&= \sum_{i,\alpha\in[T]}\mathbb{E}[\widetilde{\gamma}_\alpha^i\{\mathbf{x}_i^\top\mathbf{W}_{\mathrm{V}}^\top\mathbf{W}_{\mathrm{F}}^\top(\mathbf{W}_{\mathrm{F}}\mathbf{x}_T - \mathbf{y})\}\mathbf{x}_\alpha\mathbf{x}_T^\top].
\end{aligned}
$$

B.2. Order evaluation of QK-gradient

The orders of the QK-gradient terms (19), (20), and (21) are evaluated. In this subsection, we assume that the covariance matrix in Assumption 1 is $\boldsymbol{\Sigma} = \mathbf{I}$ for simplicity. The following evaluation still applies with minor modifications for a general

Σ. For Eq. (19), the Cauchy–Schwarz inequality implies that

$$
\text{Eq. (19)} = \left| \sum_{i,j,\alpha,\beta} \mathbb{E}[\widetilde{\gamma}_\alpha^i \widetilde{\gamma}_\beta^j (\mathbf{x}_i^\top \check{\mathbf{P}} \mathbf{x}_j)(\mathbf{x}_\beta^\top \mathbf{W}_{\mathrm{QK}} \mathbf{x}_T) \mathbf{x}_\alpha \mathbf{x}_T^\top] \right|^2
$$

$$
\leq \underbrace{\left\{ \sum_{i,j,\alpha,\beta} \mathbb{E}[(\widetilde{\gamma}_\alpha^i \widetilde{\gamma}_\beta^j)^2] \right\}}_{(A)} \underbrace{\left\{ \sum_{i,j,\alpha,\beta} \mathbb{E}[(\mathbf{x}_i^\top \check{\mathbf{P}} \mathbf{x}_j)^2] \right\}}_{(B)} \underbrace{\left\{ \sum_{i,j,\alpha,\beta} \mathbb{E}[(\mathbf{x}_\beta^\top \mathbf{W}_{\mathrm{QK}} \mathbf{x}_T)^2] \right\}}_{(C)} \underbrace{\left\{ \sum_{i,j,\alpha,\beta} \mathbb{E}[(\mathbf{x}_\alpha \mathbf{x}_T^\top)^{\odot 2}] \right\}}_{(D)}.
$$

For (A), we have $\widetilde{\gamma}_\alpha^i, \widetilde{\gamma}_\beta^j \leq T^{-1}$ by definition, and hence (A) $= O(1)$. For (B), by using Eq. (17),

$$
\text{(B)} = T^2 \sum_{i,j} \mathbb{E}[\mathbf{x}_i^\top \check{\mathbf{P}} \mathbf{x}_j \mathbf{x}_j^\top \check{\mathbf{P}} \mathbf{x}_i] = T^2 \sum_{i,j} \{(i^2 + ij - 3i - j + 4) \operatorname{tr}(\check{\mathbf{P}}^2) + (i^2 - 2i + 2) \operatorname{tr}(\check{\mathbf{P}})^2\} = O(T^6).
$$

For (C), by following the same computation as Eq. (22),

$$
\text{(C)} = T^3 \sum_{\beta} \mathbb{E}[\mathbf{x}_\beta^\top \mathbf{W}_{\mathrm{QK}} \mathbf{x}_T \mathbf{x}_T^\top \mathbf{W}_{\mathrm{QK}} \mathbf{x}_\beta]
$$

$$
= T^3 \sum_{\beta} \{(\beta^2 + (T-3)\beta - (T-4)) \operatorname{tr}(\mathbf{W}_{\mathrm{QK}}^2) + (\beta^2 - 2\beta + 2) \operatorname{tr}(\mathbf{W}_{\mathrm{QK}})^2\} = O(T^6).
$$

For (D), its (i,j)-element can be evaluated as follows (no matter whether $i = j$ or not):

$$
\text{(D)}_{ij} = T^3 \sum_{\alpha} \mathbb{E}[x_{\alpha,i}^2 x_{T,j}^2] = T^3 \sum_{\alpha} \mathbb{E}[x_{\alpha,i}^2 \{(T-\alpha) + x_{\alpha,j}^2\}] = O(T^5) \sum_{\alpha} \mathbb{E}[x_{\alpha,i}^2] + T^3 \sum_{\alpha} \mathbb{E}[x_{\alpha,i}^2 x_{\alpha,j}^2] = O(T^6).
$$

By plugging them back, we now confirmed that $|\text{Eq. (19)}| = O(T^8)$.

The orders of Eqs. (20) and (21) can be evaluated similarly and the detailed evaluations are omitted.

$$
|\text{Eq. (20)}|^2 = \left| \sum_{i,\alpha,\beta} \mathbb{E}[\widetilde{\gamma}_\alpha^i \widetilde{\gamma}_0^\beta (\mathbf{x}_i^\top \check{\mathbf{P}} \mathbf{x}_\beta) \mathbf{x}_\alpha \mathbf{x}_T^\top] \right|^2
$$

$$
\leq \left\{ \sum_{i,\alpha,\beta} \mathbb{E}[(\widetilde{\gamma}_\alpha^i \widetilde{\gamma}_0^\beta)^2] \right\} \left\{ \sum_{i,\alpha,\beta} \mathbb{E}[(\mathbf{x}_i^\top \check{\mathbf{P}} \mathbf{x}_\beta)^2] \right\} \left\{ \sum_{i,\alpha,\beta} \mathbb{E}[(\mathbf{x}_\alpha \mathbf{x}_T^\top)^{\odot 2}] \right\}
$$

$$
= O(T) \cdot O(T^4) \cdot O(T^5)
$$

$$
= O(T^{10})
$$

$$
\implies |\text{Eq. (20)}| = O(T^5).
$$

$$
|\text{Eq. (21)}|^2 = \left| \sum_{i,\alpha} \mathbb{E}[\widetilde{\gamma}_\alpha^i \{\mathbf{x}_i^\top \mathbf{W}_{\mathrm{V}}^\top \mathbf{W}_{\mathrm{F}}^\top (\mathbf{W}_{\mathrm{F}} \mathbf{x}_T - \mathbf{y})\} \mathbf{x}_\alpha \mathbf{x}_T^\top] \right|^2
$$

$$
\leq \left\{ \sum_{i,\alpha} \mathbb{E}[(\widetilde{\gamma}_\alpha^i)^2] \right\} \left\{ \sum_{i,\alpha} \mathbb{E}[(\mathbf{x}_i^\top \mathbf{W}_{\mathrm{V}}^\top \mathbf{W}_{\mathrm{F}}^\top (\mathbf{W}_{\mathrm{F}} \mathbf{x}_T - \mathbf{y}))^2] \right\} \left\{ \sum_{i,\alpha} \mathbb{E}[(\mathbf{x}_\alpha \mathbf{x}_T^\top)^{\odot 2}] \right\}
$$

$$
= O(1) \cdot O(T^4) \cdot O(T^4)
$$

$$
= O(T^8)
$$

$$
\implies |\text{Eq. (21)}| = O(T^4).
$$

Hence, we have $|\text{Eq. (19)}| = O(T^8)$, $|\text{Eq. (20)}| = O(T^5)$, and $|\text{Eq. (21)}| = O(T^4)$, which implies that the QK-gradient (3) is asymptotically dominated by Eq. (19).

C. Proofs

Lemma 1. *Suppose that* $\mathbf{W}_{\mathrm{QK}}$ *is symmetric and independent from* $\mathbf{X}$*, and let* $\mathbf{W} := \mathbf{W}_{\mathrm{QK}}\mathbf{\Sigma}$*. Under Assumption 1, for* $i \in [T]$*, the mean* μ^i *and variance* v^i *of* $\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega} \rangle + \gamma_0^i$ *with the input* $\boldsymbol{\omega} := \mathbf{X}^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T / \lambda$ *are given as follows:*

$$\mu^i = \left(\frac{i}{T} - \frac{1}{2} \right) \frac{\mathrm{tr}(\mathbf{W})}{\lambda} + o(1),$$

$$v^i = \left(\frac{2i^2}{T^2} + \frac{7}{12} \right) \frac{\mathrm{tr}(\mathbf{W}^2)}{\lambda^2} + o(1).$$

Proof. To derive the mean, we use Eq. (15).

$$\mu^i = \frac{1}{\lambda T} \mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T] - \frac{1}{\lambda T^2} \sum_{j \in [T]} \mathbb{E}[\mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T] + o(1)$$

$$= \frac{i-1}{\lambda T} \mathrm{tr}(\mathbf{W}) - \frac{\sum_{j \in [T]}(j-1)}{\lambda T^2} \mathrm{tr}(\mathbf{W}) + o(1)$$

$$= \left(i - \frac{T+1}{2} \right) \frac{\mathrm{tr}(\mathbf{W})}{\lambda T} + o(1).$$

Note that $\gamma_0^i = o(1)$.

To derive the variance, we first compute $\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T]$ (for $i \leq j \leq T$).

$$\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T] = \mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}(\mathbf{x}_T \mathbf{x}_T^\top)\mathbf{W}_{\mathrm{QK}}\mathbf{x}_j]$$

$$= \mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\{(T-j)\mathbf{\Sigma} + \mathbf{x}_j \mathbf{x}_j^\top\}\mathbf{W}_{\mathrm{QK}}\mathbf{x}_j]$$

$$= (T-j)\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{\Sigma}\mathbf{W}_{\mathrm{QK}}\mathbf{x}_j] + \mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_j \mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_j] \tag{22}$$

$$= (T-j)(i-1)\mathrm{tr}(\mathbf{W}^2) + (ij - i - j + 2)\{2\,\mathrm{tr}(\mathbf{W}^2) + \mathrm{tr}(\mathbf{W})^2\}$$

$$= (ij + (T-2)i - j - (T-4))\mathrm{tr}(\mathbf{W}^2) + (ij - i - j + 2)\mathrm{tr}(\mathbf{W})^2,$$

where Eq. (11) is used recursively at the second identity and Eqs. (15) and (18) are used at the fourth identity. Then, the expectation of the squared term is expanded:

$$\mathbb{E}[\langle \boldsymbol{\gamma}^i, \mathbf{X}^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \rangle^2]$$

$$= \mathbb{E}\left[\left(\frac{1}{T}\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T - \frac{1}{T^2} \sum_{j \in [T]} \mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \right)^2 \right]$$

$$= \mathbb{E}\left[\frac{1}{T^2}\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T - \frac{2}{T^3} \sum_{j \in [T]} \mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T + \frac{1}{T^4} \sum_{j,j' \in [T]} \mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_{j'}^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \right]$$

$$= \frac{1}{T^2} \underbrace{\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T]}_{(A)}$$

$$- \frac{1}{T^3} \underbrace{2\,\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T]}_{(B1)} - \frac{1}{T^3} \underbrace{2\sum_{j>i}\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T]}_{(B2)} - \frac{1}{T^3} \underbrace{2\sum_{j<i}\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T]}_{(B3)}$$

$$+ \frac{1}{T^4} \underbrace{\sum_{j \in [T]}\mathbb{E}[\mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T]}_{(C1)} + \frac{1}{T^4} \underbrace{2\sum_{j<j'}\mathbb{E}[\mathbf{x}_j^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T \mathbf{x}_{j'}^\top \mathbf{W}_{\mathrm{QK}}\mathbf{x}_T]}_{(C2)}.$$

Each term is computed by using Eq. (22) multiple times.

$$
\begin{aligned}
\text{(A)} &= (i^2 + (T-3)i - (T-4))\operatorname{tr}(\mathbf{W}^2) + (i^2 - 2i + 2)\operatorname{tr}(\mathbf{W})^2 \\
&= (i^2 + Ti)\operatorname{tr}(\mathbf{W}^2) + i^2 \operatorname{tr}(\mathbf{W})^2 + o(T^2), \\
\text{(B1)} &= o(T^3), \\
\text{(B2)} &= 2\sum_{j>i}\{(ij + (T-2)i - j - (T-4))\operatorname{tr}(\mathbf{W}^2) + (ij - i - j + 2)\operatorname{tr}(\mathbf{W})^2\} \\
&= (T^2 i - 2Ti^2 - i^3)\operatorname{tr}(\mathbf{W}^2) + (T^2 i - i^3)\operatorname{tr}(\mathbf{W})^2 + o(T^3), \\
\text{(B3)} &= 2\sum_{j<i}\{(ij + (T-2)j - i - (T-4))\operatorname{tr}(\mathbf{W}^2) + (ij - i - j + 2)\operatorname{tr}(\mathbf{W})^2\} \\
&= (Ti^2 + i^3)\operatorname{tr}(\mathbf{W}^2) + i^3 \operatorname{tr}(\mathbf{W})^2 + o(T^3), \\
\text{(C1)} &= \sum_{j\in[T]}\{(j^2 + (T-3)j - (T-4))\operatorname{tr}(\mathbf{W}^2) + (j^2 - 2j + 2)\operatorname{tr}(\mathbf{W})^2\} \\
&= o(T^4), \\
\text{(C2)} &= 2\sum_{j<j'}\{(jj' + (T-2)j - j' - (T-4))\operatorname{tr}(\mathbf{W}^2) + (jj' - j - j' + 2)\operatorname{tr}(\mathbf{W})^2\} \\
&= 2\sum_{j<j'} jj'\{\operatorname{tr}(\mathbf{W}^2) + \operatorname{tr}(\mathbf{W})^2\} + 2\sum_{j<j'} Tj\operatorname{tr}(\mathbf{W}^2) + o(T^4) \\
&= \sum_{j\in[T]} j(T-j)(T+j+1)\{\operatorname{tr}(\mathbf{W}^2) + \operatorname{tr}(\mathbf{W})^2\} + 2T\sum_{j\in[T]}(T-j)j\operatorname{tr}(\mathbf{W}^2) + o(T^4) \\
&= \sum_{j\in[T]}(T^2 j - j^3)\{\operatorname{tr}(\mathbf{W}^2) + \operatorname{tr}(\mathbf{W})^2\} + \frac{T^4}{3}\operatorname{tr}(\mathbf{W}^2) + o(T^4) \\
&= \frac{7T^4}{12}\operatorname{tr}(\mathbf{W}^2) + \frac{T^4}{4}\operatorname{tr}(\mathbf{W})^2 + o(T^4).
\end{aligned}
$$

By plugging them back,

$$
\mathbb{E}[\langle \boldsymbol{\gamma}^i, \mathbf{X}^\top \mathbf{W}_{\text{QK}}\mathbf{x}_T\rangle^2] = \left(\frac{7}{12} + \frac{2i^2}{T^2}\right)\operatorname{tr}(\mathbf{W}^2) + \left(\frac{1}{4} - \frac{i}{T} + \frac{i^2}{T^2}\right)\operatorname{tr}(\mathbf{W})^2 + o(1).
$$

Hence, the variance is derived:

$$
\begin{aligned}
v^i &= \mathbb{V}[\langle \boldsymbol{\gamma}^i, \boldsymbol{\omega}\rangle] \\
&= \frac{1}{\lambda^2}\mathbb{E}[\langle \boldsymbol{\gamma}^i, \mathbf{X}^\top \mathbf{W}_{\text{QK}}\mathbf{x}_T\rangle^2] - (\mu^i)^2 \\
&= \frac{1}{\lambda^2}\left(\frac{7}{12} + \frac{2i^2}{T^2}\right)\operatorname{tr}(\mathbf{W}^2) + o(1).
\end{aligned}
$$

$\square$

Lemma 2. $\rho(\theta)$ *satisfies the following properties.*

1. (Late-/middle-site) As $(\xi, \eta) \to (\infty, 0)$ with $\xi\eta \to r$,

$$
\rho(\theta) \to \begin{cases} \mathbb{1}_{\{\theta \geq \frac{1}{2}\}} & \text{if } 0 \leq r \leq 2 \\ \mathbb{1}_{\{\frac{1}{2} \leq \theta \leq \frac{1}{2} + \frac{1}{r}\}} & \text{if } r > 2 \end{cases}.
$$

2. (Early-/middle-site) As $(\xi, \eta) \to (-\infty, 0)$ with $\xi\eta \to r$,

$$
\rho(\theta) \to \begin{cases} \mathbb{1}_{\{\theta \leq \frac{1}{2}\}} & \text{if } -2 \leq r < 0 \\ \mathbb{1}_{\{\frac{1}{2} + \frac{1}{r} \leq \theta \leq \frac{1}{2}\}} & \text{if } r < -2 \end{cases}.
$$

3. *(Uniformity)* Fix η as a finite value. As $|\xi| \to 0$, $|\rho'(\theta)| \to 0$ for any $\theta \in [0, 1]$.

4. *(Vanishing signal)* Fix ξ as a finite value. As $\eta \to \infty$, $\rho(\theta) \to 0$ for any $\theta \in [0, 1]$.

Proof. To see 1: We first see that as $\xi \to \infty$,

$$\Phi\left(\left(\theta - \frac{1}{2}\right)\xi; \theta\right) \to \begin{cases} \frac{1}{2} & \text{if } \theta > \frac{1}{2} \\ 0 & \text{if } \theta = \frac{1}{2} \\ -\frac{1}{2} & \text{if } \theta < \frac{1}{2} \end{cases}.$$

In addition, as $\xi \to \infty$ and $\eta \to 0$ with $\xi\eta \to r \in [0, 2]$,

$$\Phi\left(\left(\theta - \frac{1}{2}\right)\xi - \frac{1}{\eta}; \theta\right) \to \frac{1}{2}\text{erf}\left(\frac{(\theta - \frac{1}{2})r - 1}{\eta\sqrt{2(2\theta^2 + \frac{7}{12})}}\right) \to -\frac{1}{2}.$$

By combining them, $\rho(\theta) \to \mathbb{1}_{\{\theta \geq \frac{1}{2}\}}$ at the limit. If $r > 2$,

$$\Phi\left(\left(\theta - \frac{1}{2}\right)\xi - \frac{1}{\eta}; \theta\right) \to \frac{1}{2}\text{erf}\left(\frac{(\theta - \frac{1}{2})r - 1}{\eta\sqrt{2(2\theta^2 + \frac{7}{12})}}\right) \to \begin{cases} -\frac{1}{2} & \text{if } \theta < \frac{1}{2} + \frac{1}{r} \\ 0 & \text{if } \theta = \frac{1}{2} + \frac{1}{r} \\ \frac{1}{2} & \text{if } \theta > \frac{1}{2} + \frac{1}{r} \end{cases},$$

and $\rho(\theta) \to \mathbb{1}_{\{\frac{1}{2} \leq \theta \leq \frac{1}{2} + \frac{1}{r}\}}$ at the limit.

We can see 2 in the same way as 1.

To see 3: First, compute $\rho'(\theta)$ by using $\frac{d}{dz}\text{erf}(z) = \frac{2}{\sqrt{\pi}}\exp(-z^2)$:

$$\rho'(\theta) = \frac{1}{\sqrt{\pi}}\exp\left(-\frac{((\theta - \frac{1}{2})\xi)^2}{2(2\theta^2 + \frac{7}{12})}\right)\frac{d}{d\theta}\left\{\frac{(\theta - \frac{1}{2})\xi}{\sqrt{2(2\theta^2 + \frac{7}{12})}}\right\} - \frac{1}{\sqrt{\pi}}\exp\left(-\frac{((\theta - \frac{1}{2})\xi - \frac{1}{\eta})^2}{2(2\theta^2 + \frac{7}{12})}\right)\frac{d}{d\theta}\left\{\frac{(\theta - \frac{1}{2})\xi - \frac{1}{\eta}}{\sqrt{2(2\theta^2 + \frac{7}{12})}}\right\}$$

$$= \left[\frac{1}{\sqrt{\pi}}\exp\left(-\frac{((\theta - \frac{1}{2})\xi)^2}{2(2\theta^2 + \frac{7}{12})}\right)\frac{4\theta^2 - \theta + \frac{5}{3}}{(2(2\theta^2 + \frac{7}{12}))^{3/2}} - \frac{1}{\sqrt{\pi}}\exp\left(-\frac{((\theta - \frac{1}{2})\xi - \frac{1}{\eta})^2}{2(2\theta^2 + \frac{7}{12})}\right)\frac{4\theta^2 - \theta + \frac{5}{3} - \frac{1}{\eta}}{(2(2\theta^2 + \frac{7}{12}))^{3/2}}\right]\xi.$$

By noting that $0 < \exp(-z^2) \leq 1$,

$$|\rho'(\theta)| \leq \frac{|\xi|}{\sqrt{\pi}}\left|\frac{4\theta^2 - \theta + \frac{5}{3}}{(2(2\theta^2 + \frac{7}{12}))^{3/2}} - \frac{4\theta^2 - \theta + \frac{5}{3} - \frac{1}{\eta}}{(2(2\theta^2 + \frac{7}{12}))^{3/2}}\right|$$

$$= \frac{|\xi|}{\sqrt{\pi}}\frac{1}{(2(2\theta^2 + \frac{7}{12}))^{3/2}\eta}$$

$$\to 0 \text{ as } |\xi| \to 0.$$

To see 4: For finite ξ,

$$\lim_{\eta \to \infty}\Phi\left(\left(\theta - \frac{1}{2}\right)\xi - \frac{1}{\eta}\right) = \Phi\left(\left(\theta - \frac{1}{2}\right)\xi\right),$$

which indicates that $\rho(\theta) \to 0$ at the limit $\eta \to \infty$. $\qquad \square$

D. Additional experiments

Here, we show additional results of the language modeling task with 1-/3-/6-layer transformers with different embedding dimensions $d = 32, 128$. For $d = 128$, the configurations remain the same except for the number of decoder layers as in Section 7. For $d = 32$, we used the learning rate 0.0001 (instead of 0.000025 used for $d = 128$), and the other configurations remain the same. The results are shown in Fig. 8 ($d = 32$, 1-layers), Fig. 9 ($d = 32$, 3-layers), Fig. 10 ($d = 32$, 6-layers),
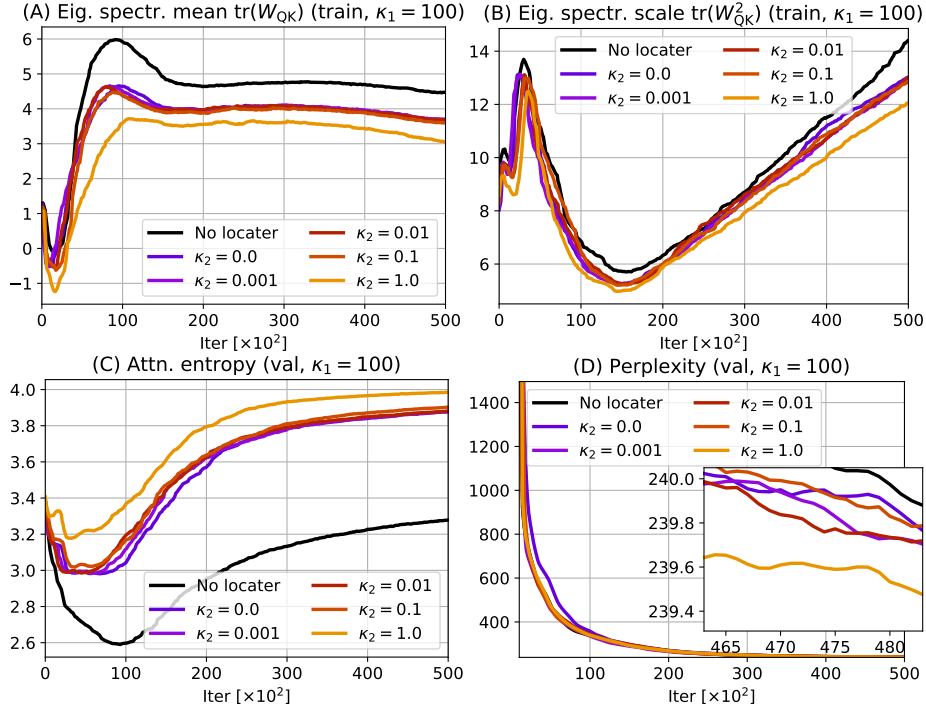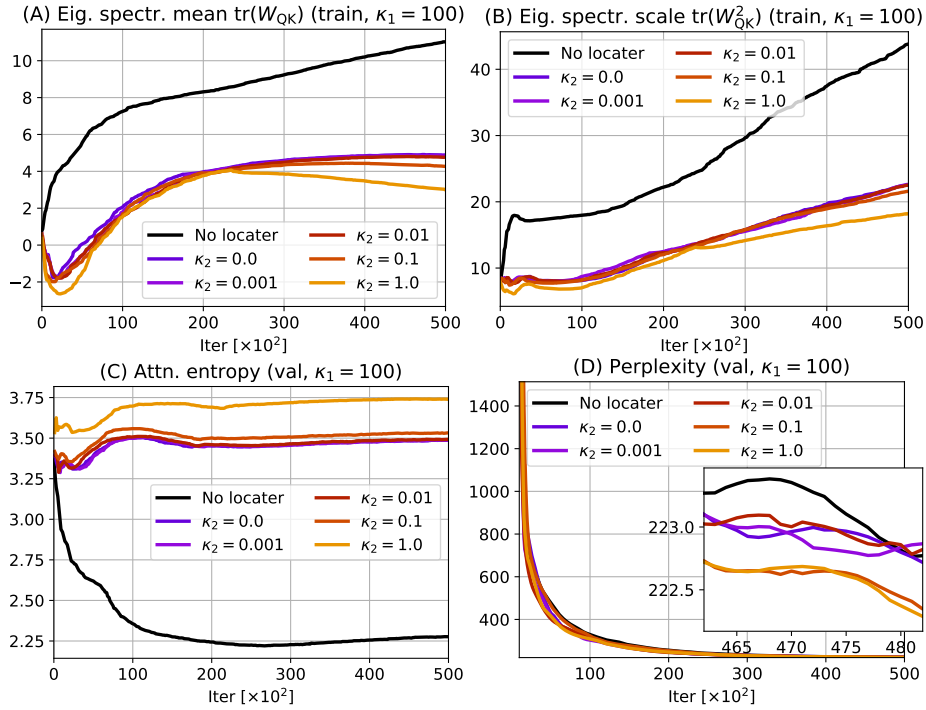
Figure 8: Experimental results of language modeling (WikiText-2) with $d = 32$ with 1-layers transformers, fixed $\kappa_1 = 100$, and varying regularization intensity κ_2. With stronger κ_2, the eigenspectrum scale shrinks **(B)**, the attention entropy increases **(C)**, and the perplexity improves **(D)**.
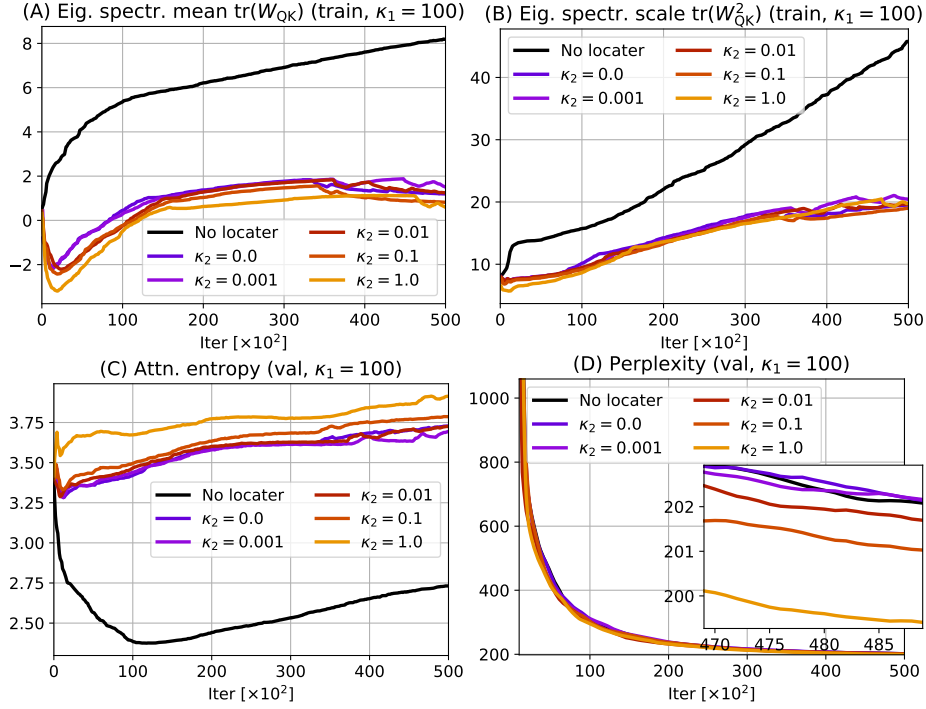

Figure 9: Experimental results of language modeling (WikiText-2) with $d = 32$ with 3-layers transformers, fixed $\kappa_1 = 100$, and varying regularization intensity κ_2. With stronger κ_2, the eigenspectrum scale shrinks **(B)**, the attention entropy increases **(C)**, and the perplexity improves **(D)**.
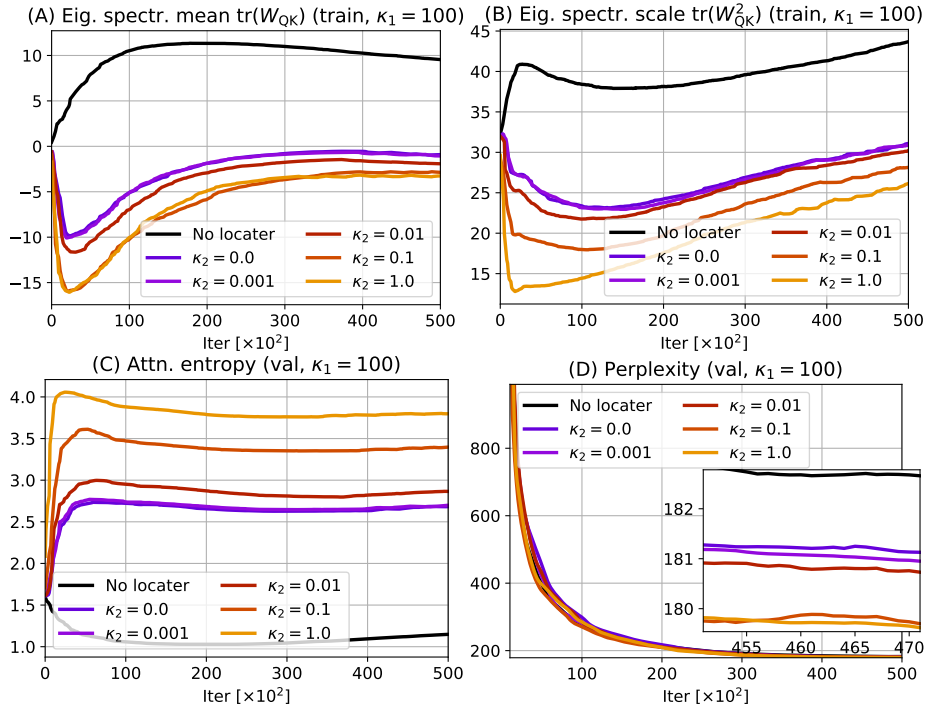
Figure 10: Experimental results of language modeling (WikiText-2) with $d = 32$ with 6-layers transformers, fixed $\kappa_1 = 100$, and varying regularization intensity κ_2. With stronger κ_2, the eigenspectrum scale shrinks (**B**), the attention entropy increases (**C**), and the perplexity improves (**D**).

Figure 11: Experimental results of language modeling (WikiText-2) with $d = 128$ with 3-layers transformers, fixed $\kappa_1 = 100$, and varying regularization intensity κ_2. With stronger κ_2, the eigenspectrum scale shrinks (**B**), the attention entropy increases (**C**), and the perplexity improves (**D**).
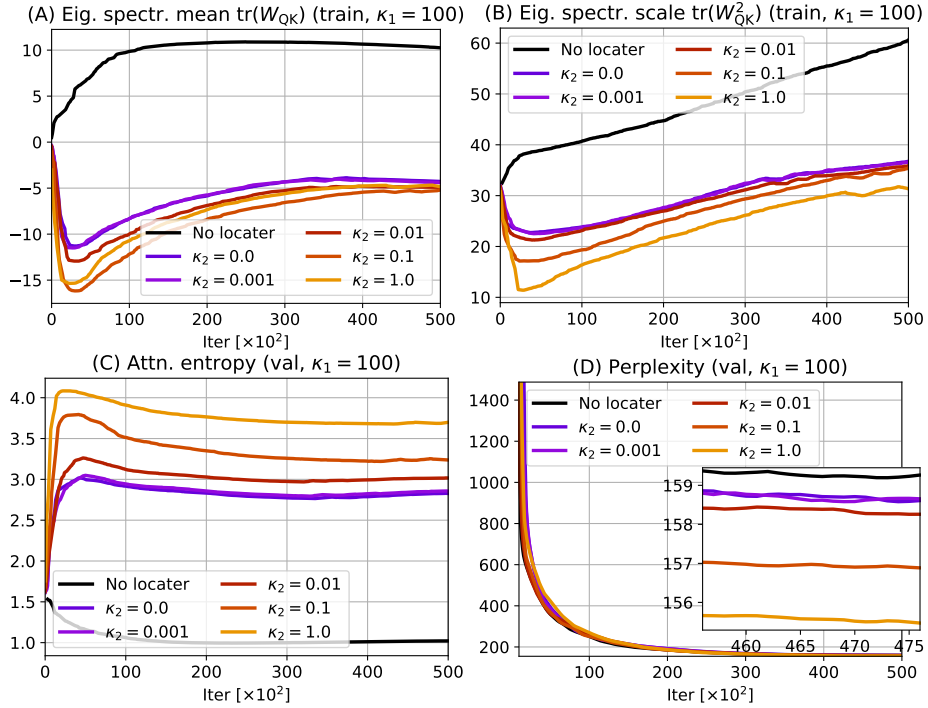
19

Figure 12: Experimental results of language modeling (WikiText-2) with $d = 128$ with 6-layers transformers, fixed $\kappa_1 = 100$, and varying regularization intensity κ_2. With stronger κ_2, the eigenspectrum scale shrinks **(B)**, the attention entropy increases **(C)**, and the perplexity improves **(D)**.

Fig. 11 ($d = 128$, 3-layers), and Fig. 12 ($d = 128$, 6-layers). Unlike the 1-layer case, we monitored the statistics of $\mathbf{W}_{\mathrm{QK}}$ in the first layer only. The overall trends are quite similar to the case of 1-layer transformers with $d = 128$ as seen in Fig. 6: As κ_2 increases, the eigenspectrum scale decreases, the attention entropy increases, and eventually, the perplexity improves (namely, decreases).