Journal of Machine Learning Research 27 (2026) 1-32 Submitted 1/24; Revised 8/25; Published 4/26

Why "Classic" Transformers Are Shallow and A Depth-Enabling Technique

Yueyao Yu YUEYAOYU@LINK.CUHK.EDU.CN
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
Guangdong, China

Yin Zhang YINZHANG@CUHK.EDU.CN
School of Data Science
The Chinese University of Hong Kong, Shenzhen
Guangdong, China

Editor: Qiaozhu Mei

Abstract

Since its introduction in 2017, the Transformer has emerged as the leading neural network architecture, catalyzing revolutionary advancements in many AI disciplines. The key innovation in Transformer is a Self-Attention (SA) mechanism designed to capture contextual information. However, stacking up more layers of the same design has failed to produce trainable deeper Transformers. Thus far, various architectural modifications to the original design have been proposed to enable deeper depths for Transformer models, but a thorough understanding of this depth issue remains lacking. In this paper, we conduct a comprehensive investigation to substantiate the claim that the depth problem is caused by a phenomenon called *token similarity escalation*; that is, tokens grow increasingly alike after repeated applications of the SA mechanism. Our analysis reveals that, driven by the invariant leading eigenspace and large spectral gaps of attention matrices, token similarity provably escalates at a linear rate as the depth increases. This insight suggests a simple technique that surgically removes excessive token similarity without reducing the overall role of the SA mechanism, as is done by existing approaches. We perform a set of proof-of-concept, small-scale experiments to show the viability of the proposed depth-enabling technique.

Keywords: Transformer, Self-Attention mechanism, Token similarity, Escalation and De-escalation, Deep neural networks

1. Introduction

The Transformer architecture (Vaswani et al., 2017) for neural networks, incorporating self-attention (SA) mechanisms (Vaswani et al., 2017), residual connections (He et al., 2016),

layer normalizations (Ba et al., 2016) and conventional feedforward networks (Haykin, 1994), has revolutionized various areas of AI, including natural language processing, computer vision and beyond (Bommasani et al., 2021; Brown et al., 2020; Dosovitskiy et al., 2020; Devlin et al., 2019; Liu et al., 2021; Vaswani et al., 2017). One of the key strengths of Transformers lies in their scalability, enabling significant performance improvements through the use of larger models, more data, and increased computational resources (Brown et al., 2020; Radford et al., 2019; Touvron et al., 2023). However, further increasing the depth of transformers is by no means a task without obstacles. Some authors, including (Ethayarajh, 2019; Gao et al., 2019), have observed that the representation power of Transformer-based deep models is rather limited to the extent that the learned embeddings only occupy a small portion of the representation space.

1.1 Token Similarity

Recent investigations on deep Transformer models, including but not limited to (Dong et al., 2021; Ethayarajh, 2019; Gao et al., 2019; Li et al., 2020; Mu and Viswanath, 2018; Noci et al., 2022; Yan et al., 2022), have shed light on the occurrence of a phenomenon that has been called by different names, including *token uniformity*, *rank collapse*, *representation degeneration* and *representation anisotropy*. In this paper, we will use the term *token similarity* with a precise and distinctive definition. In plain language, token similarity means that as a representation matrix $X \in \mathbb{R}^{n\times d}$ traverses through layers of a Transformer model, the n rows, or tokens as they are called, in X grow increasingly similar to each other, thus substantially reducing the model's expressive capacity and hindering the training of the model (Noci et al., 2022).

To quantitatively measure token similarity, researchers have proposed a number of measurement methods. For example, the following cosine similarity (Ethayarajh, 2019) calculates the average cosine similarity between all pairs of tokens,

$$\mathbf{t}_{cos}(X) := \frac{2}{n^2-n}\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{x_i^T x_j}{\|x_i\|\|x_j\|}, \tag{1}$$

where $x_i \in \mathbb{R}^d$ is the i-th row of $X \in \mathbb{R}^{n\times d}$ and $\|\cdot\|$ is the Euclidean norm.

As a similarity measure, the cosine similarity can adequately fulfill its purpose. However, to facilitate our theoretical analysis we need to define a technically more manageable measure called *token similarity*.

Definition 1 *Given any non-zero matrix $X \in \mathbb{R}^{n\times d}$, the token similarity of X is*

$$\mathbf{t}_{sim}(X) := \|\Pi_{\mathbb{1}}X\|_F^2/\|X\|_F^2 \ \in \ [0,1]$$

where $\Pi_{\mathbb{1}} = \mathbb{1}\mathbb{1}^T/n \in \mathbb{R}^{n\times n}$ *and* $\mathbb{1}$ *is the vector of all ones in* $\mathbb{R}^n$*. The token diversity of* X *is*

$$\mathbf{t}_{div}(X) := \|\Pi_{\mathbb{1}}^{\perp} X\|_F^2/\|X\|_F^2 \ \in \ [0,1],$$

where $\Pi_{\mathbb{1}}^{\perp} = I - \Pi_{\mathbb{1}} \in \mathbb{R}^{n\times n}$*. Clearly, the similarity and diversity of* X *sum up to unity. For convenience, we will on occasions also use the notation:* $\Pi_1 = \Pi_{\mathbb{1}}$ *and* $\Pi_2 = \Pi_{\mathbb{1}}^{\perp}$.

We observe that $\mathbf{t}_{sim}(X) = 1$, or $\mathbf{t}_{div}(X) = 0$, if and only if $X = \mathbb{1}v^T$ for some $v \in \mathbb{R}^d$; that is, all rows of X are the same.

We start with examining how token similarity evolves in two well-known Transformer models: BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2019), both following the original (or classic) encoder architecture as proposed by (Vaswani et al., 2017). We utilized the package Hugging Face (Wolf et al., 2019) to conduct this experiment where we set the model depth to 100 for both BERT and ALBERT and initialize model weights by the package default. In Figure 1, we plot token similarity $\mathbf{t}_{sim}(X)$, cosine similarity $\mathbf{t}_{cos}(X)$ and, for BERT model, gradient norm (with respect to model weights) at each Transformer layer (or block, as is often called).[1] From Figure 1, we observe that, as the depth increases, in either model both token similarity and cosine similarity escalate to unity, where the escalation patterns for the two measures appear almost identical. We also observe that the gradient norm value in BERT remains roughly at a constant level, which indicates that in this case gradient vanishing or exploding is not occurring.

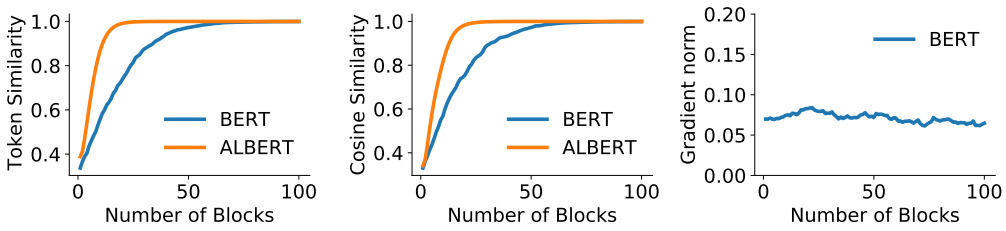

Figure 1: Values of token similarity (left), cosine similarity (middle), and gradient norm (right) at each block in 2 Transformer models at default initialization.

As mentioned earlier, it remains to be fully understood why classic Transformer architecture yields outstanding results with shallow models but becomes "useless" with deeper ones (Takase et al., 2023). Our experiments in Figure 1 strongly suggest that the root cause be token similarity escalation. In this paper, we conduct a systematic analysis to

1. We excluded ALBERT from the gradient norm experiment since it uses the same set of weights throughout all blocks.

substantiate this assertion and to address two fundamental questions: why token similarity escalates and how fast it escalates. Our theoretical and empirical results will provide definitive answers to these questions.

In the sequel, we will use capitalized Transformer(s) to refer to models based on the "classic" architecture as originally proposed by (Vaswani et al., 2017), while subsequently proposed modifications will be referred to as transformers in small letters.

For ease of exposition, we will assume the encoder architecture as the default architecture. Later in Section 2.6, we will discuss how to apply obtained results to the decoder architecture.

1.2 Related Works

A number of recent works have noticed depth-related problems with deep Transformers; for example, the learned word embeddings may degenerate into lying in a narrow cone (Ethayarajh, 2019). Dong et al. (2021) prove that in pure self-attention models without skip connections, the representation matrix would converge to a rank-one matrix with identical rows, a situation that they call token uniformity, while they also claim that skip-connections should be able to resolve this problem. However, more recent works (Yan et al., 2022; Noci et al., 2022) for example, have observed that residual models still encounter the same token uniformity problem. In this paper, we will use the term token similarity in order to utilize a distinct definition and avoid possible confusion with previously defined measures.

A variety of approaches have been explored to tackle difficulties caused by or related to token similarity. Some works have incorporated regularization terms into the training objectives (Gao et al., 2019; Wang et al., 2019a; Zhang et al., 2020) or introduced contrastive learning (Gao et al., 2021; Qiu et al., 2022) to alleviate the so-called anisotropy problem. Additionally, researchers have adopted post-processing strategies to normalize word or sentence embeddings (Huang et al., 2021; Mu and Viswanath, 2018) or transform learned representations into alternative distributions (Li et al., 2020; Yan et al., 2022) to obtain more isotropic representations. Moreover, Noci et al. (2022) use so-called residual branch scaling (Bachlechner et al., 2021) to slow down rank collapse. He et al. (2023) propose an approach termed "next-token prediction" to slow down rank collapse in skip-less Transformers (at a cost of prolonged training time). In essence, these techniques reduce, explicitly or implicitly, the role of self-attention relative to other operations.

Some researchers try to explain difficulties in deeper Transformers from the perspective of gradient instability. For example, authors such as (Wang et al., 2019b; Liu et al., 2020; Xiong et al., 2020) conclude that training deep Transformers of the original design is unstable, resulting in bad performance. Recently, Noci et al. (2022) prove that when X is rank-one, then certain gradient components would vanish at initialization. Zhai et al.

(2023) observe that the gradient is unstable when the so-called attention entropy collapses. On the other hand, Takase et al. (2023) give empirical evidence that at least in an encoder, the gradient norms in different blocks do not rise or fall significantly.

In the literature, the original Transformer architecture is now often called the post-norm (or post-LN) architecture where layer normalizations are applied after residual connections are added (Vaswani et al., 2017). To address difficulties in deep post-norm models, a variant called pre-norm (or pre-LN) architecture has been proposed (Wang et al., 2019b; Xiong et al., 2020), where layer normalizations are applied to the input of each sub-layer. It appears that contemporary large language models, for example (Touvron et al., 2023; Brown et al., 2020), have widely adopted the pre-norm architecture. Nevertheless, with shallow models a noticeable reduction in generalization performance has been reported for pre-norm models in comparison to their post-norm counterparts (Wang et al., 2019b; Xiong et al., 2020). In this paper, our analysis and experiments concentrate primarily on the classic, post-norm architecture.

1.3 Contributions

Despite some theoretical works (Dong et al., 2021; Noci et al., 2022) on factors influencing token similarity, our literature review suggests that there has been no analysis that accurately quantifies the token similarity dynamics in the original Transformer architecture: why escalation starts and how fast it develops. This work provides such a quantitative analysis for the first time.

- By analyzing a well-defined measure of token similarity dynamics, we prove that the SA-plus-residual sub-layer in the Transformer architecture increases token similarity by default upon standard initialization and in expectation. Our theory, together with strong empirical evidence, reveals why token similarity escalates and how fast it does so. That is, (i) the driving force behind the escalation is none else but the invariant leading eigenspace of and large spectral gaps in self-attention matrices (while other operations in the Transformer block do not interfere with the escalation process); and (ii) the similarity measure converges to 1 at a global linear rate which eventually accelerates to the local rate of $1/2$ under the standard setting.

- Based on the insights gained from our analysis, we propose a simple de-escalation strategy to remove excessive similarity and restore expressivity in deep Transformer models. Our preliminary experiments have confirmed the efficacy of the proposed strategy, substantially improving the training quality for very deep post-norm Transformers. In contrast to existing techniques, our proposed method does not discount,

explicitly or implicitly, the role of self-attention mechanism relative to other components.

1.4 Notation

In general, matrices are denoted by upper-case letters and vectors by lower-case letters. For any square matrix M, let $\lambda_i(M)$ be the i-th eigenvalue of M arranged in a descending order in magnitude unless otherwise specified. We use $[W]_{ij}$ to denote the ij-th element of a matrix W, and will do so similarly for vector elements as well. We denote $\mathbb{E}[\cdot] := \mathbb{E}_{[W]_{ij} \sim w}[\cdot]$. By default, the vector norm $\|\cdot\|$ is the Euclidean norm. The symbol $\mathbb{1}$ denotes the vector of all ones in $\mathbb{R}^n$, and $e = \mathbb{1}/\sqrt{n}$. We reiterate that capitalized Transformers are reserved for models based on the classic architecture, while later proposed modifications are referred to as transformers in small letters. We will occasionally use the notation $\Pi_1 \equiv \Pi_{\mathbb{1}}$ and $\Pi_2 \equiv \Pi_{\mathbb{1}}^{\perp}$. For brevity, we will use the acronym TSE for Token Similarity Escalation.

2. Analysis of TSE in Transformer

In general, a Transformer model can comprise both an encoder and a decoder. In this paper, our focus will be solely on the encoder side of Transformer models. Precisely speaking, our analysis is for Transformer encoders with random weights (which is the case at initialization).

2.1 Transformer Architecture

We will start with single-head Transformer layers and then show that our results extend to multi-head layers as well. An L-layer (encoder only) Transformer is the repeated composition of a layer function, say, $Y = \mathbf{postLN}(X)$, by L times:

$$Y = \mathbf{Transformer}(X) := \underbrace{\mathbf{postLN} \circ \cdots \circ \mathbf{postLN}}_{L \text{ times}}(X).$$

We formalize the Transformer layer function $Y = \mathbf{postLN}(X)$ into Algorithm 1 below, which follows exactly the original architecture as proposed in (Vaswani et al., 2017), and is organized into four steps for our treatment convenience. A few comments about the Transformer block in Algorithm 1 are in order.

- In Step 0, we allow the generality of utilizing different formulas for computing an attention matrix $P = P(X)$. Besides X, such formulas also involve their own learnable weights that are not explicitly shown.
- The matrix product term in Step 1 is called self-attention (SA) where the attention matrix $P(X)$ is applied to X itself to form a nonlinear operation that also includes

Algorithm 1: $Y = \mathbf{postLN}(X)$

Input: $X \in \mathbb{R}^{n\times d}$ (with weight matrices W, W_1, W_2 and scalar $\alpha > 0$).

0 Compute a row-stochastic, attention matrix $P = P(X) \in \mathbb{R}^{n\times n}$.

1 SA plus Residual: $Y_1 = X + \alpha PXW$.

2 Layer Normalization: $Y_2 = \mathbf{LN}(Y_1)$.

3 FFN plus Residual: $Y_3 = Y_2 + \phi(Y_2W_1)W_2$.

4 Layer Normalization: $Y_4 = \mathbf{LN}(Y_3)$.

Output: $Y := Y_4 \in \mathbb{R}^{n\times d}$

the multiplication by a weight matrix $W \in \mathbb{R}^{d\times d}$ from the right. Next to W, a positive parameter α is introduced to balance the contributions from the SA mechanism relative to the residual (or skip connection) term X. For weight matrices of a given norm, the smaller α is, the lesser role SA would play relative to residual.

- The term in Step 3 is a feedforward network (FFN) with activation function $\phi(\cdot)$ and two weight matrices $W_1 \in \mathbb{R}^{d\times q}$ and $W_2 \in \mathbb{R}^{q\times d}$ where the column size q is usually greater than d. In our experiments, we will use the popular ReLU function for activation by default unless otherwise specified.

- In Steps 2 and 4, layer-normalizations are applied row-wise to the input matrix. More specifically, for any vector x, $\mathbf{LN}(x) = \gamma(x - \mu)/\sigma + \lambda$ where μ is the mean and σ is the standard deviation of x, and γ and λ are learnable parameters.

The impact of weight matrix initializations on model training has been widely studied since the use of an initialization scheme may determine the success or failure of training, see a recent survey (Narkhede et al., 2022) on this subject. In our analysis and experiments, we will investigate the TSE phenomenon mostly under the widely used Xavier initialization (Glorot and Bengio, 2010).

In our experiments, we use the standard softmax formula (Bridle, 1990) to construct attention matrices P (though we have tried other formulas without notable differences), which is defined as follows. For $X \in \mathbb{R}^{n\times d}$, compute $M = XW_q(XW_k)^T/\sqrt{d} \in \mathbb{R}^{n\times n}$ and let

$$[P(X)]_{ij} := \exp\left([M]_{ij}\right) / \sum_{\ell=1}^{n} \exp\left([M]_{i\ell}\right). \tag{2}$$

where W_q and W_k are weight matrices with elements drawn uniformly from $(-1, 1)/\sqrt{d}$.

2.2 How Self-Attention Drives TSE

In this subsection, we conduct a comprehensive analysis on how token similarity escalates after each time the self-attention step $Y = X + \alpha PXW$ is carried out. We first motivate our analysis using a power-method-based intuition.

2.2.1 An intuitive interpretation

The reason behind TSE can be intuitively explained by extending the idea of the power method for computing the largest eigenvector of a matrix. For any stochastic matrix P, the largest eigenvalue is always 1 corresponding to the leading eigenspace $\text{span}\{\mathbb{1}\}$. From the convergence theory of power method, we know that if the second largest eigenvalue $|\lambda_2(P)| < 1$ and $\Pi_{\mathbb{1}} x \neq 0$, then $\{P^k x\}$ converges to $\text{span}\{\mathbb{1}\}$ as k goes to infinity. In a similar spirit, for a sequence of stochastic matrices $\{P_j\}$ for which $\{|\lambda_2(P_j)|\}$ is uniformly bounded away from 1, the sequence $\{P_k \cdots P_2 P_1 x\}$ will also converge to $\text{span}\{\mathbb{1}\}$ as k goes to infinity, under some additional technical condition (e.g., see Wolfowitz (1963); Coppersmith and Wu (2008)). Moreover, for any $\alpha > 0$ and a normalization sequence $\{c_k := 1/(1+\alpha)^k\} \subset \mathbb{R}$, one could also expect that as $k \to \infty$,

$$c_k(I + \alpha P_k) \cdots (I + \alpha P_2)(I + \alpha P_1)x \;\to\; \text{span}\{\mathbb{1}\},$$

since all matrices in the sequence $\{(I + \alpha P_j)/(1+\alpha)\}$ are still row-stochastic and having second-largest eigenvalues uniformly bounded away from 1, in view of

$$\frac{|1 + \alpha\lambda_i(P_j)|}{1+\alpha} \leq \frac{1 + \alpha|\lambda_i(P_j)|}{1+\alpha} < 1, \;\; \forall\, i > 1.$$

Now consider a sequence of SA operations (including the residual connection):

$$\text{SA}_i(X) = X + \alpha P_i X W_i, \;\; i = 1, 2, 3, \cdots$$

each of which is associated with a weight matrix $W_i \in \mathbb{R}^{d\times d}$ and a stochastic matrix $P_i \equiv P_i(X) \in \mathbb{R}^{n\times n}$. In addition to the SA operation, each Transformer block also utilizes other operations (feedforward network and layer normalizations). However, these other operations have no impact on TSE which is solely driven by the SA operation, as will be demonstrated later. Specifically, the driving force is the spectral properties of P_i, while the fact that P_i depends on X is inconsequential. Hence, roughly speaking, the escalation process can be characterized as

$$c_k(\text{SA}_k \circ \cdots \circ \text{SA}_2 \circ \text{SA}_1)(X) \to \text{span}\{\mathbb{1}e_j^T\}_{j=1}^d, \tag{3}$$

where $e_j \in \mathbb{R}^d$ is the unit vector with the j-th entry being unity, and $\{c_k\} \subset \mathbb{R}$ is a proper normalization sequence. In terms of the TSE behavior, each of the above SA_i operators can

be viewed as a linear operator with a fixed leading eigenspace given in the left-hand side of (3).

The above intuitive interpretation motivates us to conduct a comprehensive analysis on TSE. To do so, we need to develop a rigorous approach to analyzing a precisely defined quantity that is critical for our TSE analysis.

2.2.2 A theoretical analysis

We start by stating some basic assumptions and facts.

Assumption 1 *In the self-attention formula $Y = X + \alpha PXW$,*

1. *the matrix $W \in \mathbb{R}^{d\times d}$ is randomly initialized so that elements of W are all independent with mean-zero and variance σ^2;*
2. *the matrix P is row-stochastic (or right-stochastic or Markov) so that $P\mathbb{1} = \mathbb{1}$ with the spectral radius equal to one (for example, see Horn and Johnson (2012)).*

Under Assumption 1(1), there hold

$$\mathbb{E}[W] = 0 \in \mathbb{R}^{d\times d} \qquad \text{and} \qquad \mathbb{E}[WW^T] = d\sigma^2 I \in \mathbb{R}^{d\times d}. \tag{4}$$

We recall the definitions of token similarity, diversity and the relationship between them:

$$\mathbf{t}_{sim}(X) = \|\Pi_{\mathbb{1}} X\|_F^2/\|X\|_F^2, \quad \mathbf{t}_{div}(X) = \|\Pi_{\mathbb{1}}^{\perp} X\|_F^2/\|X\|_F^2, \quad \mathbf{t}_{sim}(X) + \mathbf{t}_{div}(X) = 1.$$

Token similarity of X equals 1 implies the rank of X being 1. However, it is entirely possible that when token similarity is fairly close to 1, say $\mathbf{t}_{sim}(X) = 0.99$, X is still numerically full-rank. Therefore, using the continuous quantity $\mathbf{t}_{sim}(X)$ to measure similarity is much more reasonable than using the discrete quantity rank as a measure of quality for representation (or embedding) matrices.

We now introduce a critical quantity called TSE rate, or simply escalation rate.

Definition 2 *For a pair of matrices $X, Y \in \mathbb{R}^{n\times d}$ so that $\mathbf{t}_{sim}(X), \mathbf{t}_{sim}(Y) \in (0, 1)$, the escalation rate from X to Y is*

$$r(X, Y) := \frac{1 - \mathbf{t}_{sim}(X)}{1 - \mathbf{t}_{sim}(Y)} = \frac{\mathbf{t}_{div}(X)}{\mathbf{t}_{div}(Y)}. \tag{5}$$

Clearly, $r(X, Y) > 1$ implies that Y has a higher similarity (or a lower diversity) than X does. For $Y = X + \alpha PXW$, we aim at analyzing the expected value of $r(X, Y)$ with respect to the random matrix W under Assumption 1. The following proposition gives a key identity for the TSE rate $r(X, Y)$ that facilitates our analysis.

Proposition 3 *Given $X, Y \in \mathbb{R}^{n\times d}$ with $\mathbf{t}_{sim}(X), \mathbf{t}_{sim}(Y) \in (0,1)$, let $r(X,Y)$ be the escalation rate defined in* (5)*. Then the following identity holds*

$$r(X,Y) = 1 + (\xi_1/\xi_2 - 1)\,\mathbf{t}_{sim}(X), \tag{6}$$

where

$$\xi_i \equiv \xi_i(X,Y) := \frac{\|\Pi_i Y\|_F^2}{\|\Pi_i X\|_F^2}, \quad i = 1, 2. \tag{7}$$

Therefore, $r(X,Y) > 1$ if and only if $\xi_1 > \xi_2$.

Proof We first note that $\|Y\|_F^2/\|X\|_F^2 = \xi_1 \mathbf{t}_{sim}(X) + \xi_2 \mathbf{t}_{div}(X)$. By direct calculations,

$$\frac{\mathbf{t}_{div}(X)}{\mathbf{t}_{div}(Y)} = \frac{\|Y\|_F^2/\|X\|_F^2}{\|\Pi_{\mathbb{1}}^{\perp} Y\|_F^2/\|\Pi_{\mathbb{1}}^{\perp} X\|_F^2} = \frac{\xi_2(1 - \mathbf{t}_{sim}(X)) + \xi_1 \mathbf{t}_{sim}(X)}{\xi_2} = 1 + (\xi_1/\xi_2 - 1)\mathbf{t}_{sim}(X).$$

The second statement is obvious. ■

Proposition 3 indicates that as long as $\xi_1 > \xi_2$, token similarity of Y will be larger than that of X; in other words, there happens an escalation in token similarity from X to Y. For the SA-plus-residual step $Y = X + \alpha PXW$, we aim to analyze the expected escalation rate with respect to W, that is,

$$\mathbb{E}[r(X,Y)] = 1 + (\mathbb{E}\,[\xi_1/\xi_2] - 1)\,\mathbf{t}_{sim}(X). \tag{8}$$

The roadmap of our analysis is as follows. In order to estimate $\mathbb{E}[\xi_1/\xi_2]$, which in general does not allow a closed-form formula, we show instead that under mild conditions ξ_1/ξ_2 is highly concentrated at $\mathbb{E}[\xi_1]/\mathbb{E}[\xi_2]$. Then it will suffice to estimate $\mathbb{E}[\xi_1/\xi_2]$ through the two individual expected values, $\mathbb{E}[\xi_1]$ and $\mathbb{E}[\xi_2]$, which are calculated in the lemma below.

Lemma 4 *Given $X \in \mathbb{R}^{n\times d}, P \in \mathbb{R}^{n\times n}, W \in \mathbb{R}^{d\times d}$ and $\alpha > 0$, let $Y = X + \alpha PXW$. Under Assumption 1, the expected values of ξ_i, $i = 1, 2$, with respect to W are, respectively,*

$$\mathbb{E}[\xi_i] = 1 + \alpha^2 d\sigma^2 \mu_i^2, \quad i = 1, 2, \tag{9}$$

where ξ_i are defined in (7)*, and*

$$\mu_i \equiv \mu_i(X,P) := \frac{\|\Pi_i PX\|_F}{\|\Pi_i X\|_F}, \quad i = 1, 2. \tag{10}$$

Furthermore, there hold the bounds for μ_1 and μ_2,

$$\mu_1^2 \geq (1-\omega)^2, \quad \mu_2^2 \leq \delta^2, \tag{11}$$

where

$$\omega = \|e^T P \Pi_{\mathbb{1}}^{\perp} X\|/\|e^T X\| \quad \text{and} \quad \delta = \|\Pi_{\mathbb{1}}^{\perp} P\|_2. \tag{12}$$

A proof for this lemma will be presented in the Appendix A.1.

Proposition 5 *Given $X \in \mathbb{R}^{n\times d}, P \in \mathbb{R}^{n\times n}, W \in \mathbb{R}^{d\times d}$ and $\alpha > 0$, let $Y = X + \alpha PXW$. Under Assumption 1, there holds*

$$\mathbb{E}[r(X,Y)] = 1 + \left(\frac{\alpha^2 d\sigma^2}{1+\alpha^2 d\sigma^2\mu_2^2}(\mu_1^2 - \mu_2^2) - \mathbb{E}[\eta]\right)\mathbf{t}_{sim}(X), \tag{13}$$

where

$$\eta := \mathbb{E}[\xi_1]/\mathbb{E}[\xi_2] - \xi_1/\xi_2, \tag{14}$$

The equality (13) follows from replacing $\mathbb{E}[\xi_1/\xi_2]$ by $\mathbb{E}[\xi_1]/\mathbb{E}[\xi_2] - \eta$ in (8), and then substituting in the expressions for $\mathbb{E}[\xi_1]$ and $\mathbb{E}[\xi_2]$ given in (9).

We will use some concentration inequalities to show that under suitable conditions the quantity $\mathbb{E}[\eta]$ will be amply small so that TSE occurs with overwhelming probability. Our main theoretical result is given as the following theorem. To state the result, we first note that when one of the two inequality in (11) is strict, the following condition must hold

$$\max\left\{\mu_1^2 - (1-\omega)^2, \delta^2 - \mu_2^2\right\} > 0. \tag{15}$$

Theorem 6 *Given $X \in \mathbb{R}^{n\times d}, P \in \mathbb{R}^{n\times n}, W \in \mathbb{R}^{d\times d}$ and $\alpha > 0$, let $Y = X + \alpha PXW$. Assume that the elements of $W \in \mathbb{R}^{d\times d}$ are independent, mean-zero, sub-gaussian random variables with sub-gaussian norm equal to $\sigma = 1/\sqrt{d}$, and (15) holds. Then there exists $d_* > 0$ such that for all $d \geq d_*$, the expected escalation rate satisfies*

$$\mathbb{E}[r(X,Y)] \geq 1 + \frac{\alpha^2}{1+\alpha^2\delta^2}\left((1-\omega)^2 - \delta^2\right)\mathbf{t}_{sim}(X) > 1 \tag{16}$$

whenever $\omega + \delta < 1$.

The proof for Theorem 6 will be given in Appendix A.4, built on multiple technical lemmas including an explicit, but rather involved, formula for d_* (see Lemma 15).

The most critical quantities in the bound (16) are ω and δ. Whenever these two numbers are small, TSE happens; and the smaller they are, the faster is the escalation rate. The following corollary considers a few special situations for the attention matrix P, indicating that spectral properties, especially spectral gaps, of P play a key role in the TSE process. This corollary can be directly verified from the definitions of the involved quantities.

Corollary 7 *Let the conditions in Theorem 6 hold. When P is doubly stochastic, then $\mu_1 = 1$ and $\omega = 0$. In this case, for $\alpha = 1$ the expected escalation rate estimate in (16) reduces to*

$$\mathbb{E}[r(X,Y)] \geq 1 + \frac{1-\delta^2}{1+\delta^2}\,\mathbf{t}_{sim}(X).$$

Moreover, when P is symmetric (thus doubly stochastic), then $\delta^2 = |\lambda_2(P)|^2$ and

$$\mathbb{E}[r(X,Y)] \geq 1 + \frac{1-|\lambda_2(P)|^2}{1+|\lambda_2(P)|^2}\,\mathbf{t}_{sim}(X), \tag{17}$$

where $\lambda_2(P)$ is the second largest eigenvalue of P in modulus, and $1-|\lambda_2(P)|^2$ serves as a measure of the spectral gap of P.

As we will empirically demonstrate later, in practice formula (17) turns out to be a better estimate for $\mathbb{E}[r(X,Y)]$ than the guaranteed lower bound in (16) which can be overly conservative. It is worth noting that as $\mathbf{t}_{sim}(X)$ becomes close to 1, the corresponding attention matrices $P(X)$ computed from the softmax formula (2) (and those from most other formulas as well) will be close to $\mathbb{1}\mathbb{1}^T/n$, which is symmetric with $\lambda_2(P) = 0$.

Aside from $\mathbf{t}_{sim}(X)$, the estimation formula in (17) is entirely determined by the spectral gap of P. If $|\lambda_2(P)| = 1$ (say, for $P = I$), then no escalation would happen. But if $|\lambda_2(P)| \ll 1$, the escalation rate can be large, which in fact is what happens in reality for random-like row-stochastic matrices P even though P is asymmetric in general. This phenomenon not only happens for P computed by the softmax formula (2), but also for P computed by other formulas as well.

Remark 8 *Some comments are due for Theorem 6 concerning the acceleration of TSE.*

- *The term $\mathbf{t}_{sim}(X)$ in* (16) *or* (17) *plays a role of accelerator for TSE. That is, the larger it is, the faster is the escalation since the factor in front of $\mathbf{t}_{sim}(X)$ does not vary significantly. As $\mathbf{t}_{sim}(X)$ goes to 1, both ω and δ converge to 0 so that the estimated escalation rate on the right-hand side of* (16) *approaches its maximum at $1+\alpha^2$.*

- *If we consider the convergence of similarity to 1 (or diversity to 0), then the rate is linear and asymptotically $1/(1+\alpha^2) = 1/2$ for $\alpha = 1$. Experiments show that the rate 1/2 occurs quite early in practice, resulting in a fast linear convergence.*

Next, we show that the above TSE result can be straightforwardly extended to the case of multi-head self-attention. For this purpose, it suffices to show that in the multi-head case, the expected escalation rate $\mathbb{E}[r(X,Y)]$ has an identical expression as in (13) for the single-head case except that in the former case, relevant quantities are the average over multiple heads. Now assume that the number of heads is h and d is divisible by h.

Proposition 9 *Given $X \in \mathbb{R}^{n\times d}$, $P_k \in \mathbb{R}^{n\times n}$ and $W_k \in \mathbb{R}^{d\times d/h}$, for $k = 1, \cdots, h$, and $\alpha > 0$, let*

$$Y = X + \alpha[P_1XW_1 \;\; P_2XW_2 \;\; \cdots \;\; P_hXW_h].$$

Under Assumption 1 (with column number d for W changed to d/h for all W_k),

$$\mathbb{E}[r(X,Y)] = 1 + \left(\frac{\alpha^2 d\sigma^2}{1+\alpha^2 d\sigma^2 \bar{\mu}_2^2}(\bar{\mu}_1^2 - \bar{\mu}_2^2) - \mathbb{E}[\eta] \right) \mathbf{t}_{sim}(X), \tag{18}$$

where η is defined as in (14) *and*

$$\bar{\mu}_i^2 := \frac{1}{h}\sum_{k=1}^{h} \frac{\|\Pi_i P_k X\|_F^2}{\|\Pi_i X\|_F^2}, \quad i = 1, 2.$$

The proof for this proposition is given in Appendix A.5.

Compared to (13) in Proposition 5, we see that the only difference in (18) is that μ_i^2 are replaced by $\bar{\mu}_i^2$ which are the average across multiple heads. Additionally, the bounds in (11) for μ_i^2 also hold for their average counterparts $\bar{\mu}_i^2$ once we replace $(1-\omega)^2$ and δ^2 by their corresponding average counterparts, that is, $\bar{\mu_1}^2 \geq \frac{1}{h}\sum_{k=1}^{h}(1-\omega_k)^2$ and $\bar{\mu_2}^2 \leq \frac{1}{h}\sum_{k=1}^{h}\delta_k^2$ where quantities with k-indices are still defined as in (12) except now associated with different P_k's.

Equipped with Proposition 9 and following the same line of arguments, we can readily derive the counterpart of Theorem 6 and thus extend our TSE analysis from the single-head case to multiple-head cases.

2.3 Other Steps Do Not Impact TSE

We first examine the FFN Step in Algorithm 1. We are not aware of any previous report that the classic FFN architecture has any involvement with TSE under any circumstances. This is easy to explain from the following column-space viewpoint. Essentially, TSE implies that the column space of X is moving towards the subspace $\text{span}\{\mathbb{1}\}$. However, in FFN, the right-multiplications of X by weight matrices do not change, in one way or another, the column space of X at all. For example, if $\Pi_{\mathbb{1}}^{\perp} X$ is small relative to $\Pi_{\mathbb{1}} X$, then so should be $(\Pi_{\mathbb{1}}^{\perp} X)W$ relative to $(\Pi_{\mathbb{1}} X)W$ for any generic W. Furthermore, no discernible reason is seen to expect any impact on TSE from the usual element-wise activation functions such as ReLU. On the contrary, it is evident that the subspace $\text{span}\{\mathbb{1}\}$ is invariant under any element-wise activation function $\phi : \mathbb{R} \to \mathbb{R}$. Therefore, such functions actually preserve similarity of X and do not interfere with the TSE process.

Next we examine the impact of layer normalizations on TSE. The layer normalization function $Y = \mathbf{LN}(X) \in \mathbb{R}^{n\times d}$ can be expressed (without scaling and shifting) as

$$Y = DX\left(I - \mathbb{1}_d \mathbb{1}_d^T/d\right),$$

where D is a diagonal matrix with $[D]_{ii} = 1/\sigma_i$, and σ_i^2 is the variance of the i-th row of X. Clearly, if $X = \mathbb{1}v^T$, then its column space $\text{span}\{\mathbb{1}\}$ is invariant under layer normalizations

since the D-matrix is a multiple of identity. This indicates that layer normalizations do not de-escalate high token similarity. Moreover, if the elements of X are independently random with a fixed variance, then the corresponding D-matrix will also be close to a multiple of identity, thus approximately preserving the column space of X. In either case, the TSE process is not observably interfered with by layer normalizations.

2.4 Experimental Verification

In this subsection, we provide some strong empirical evidence to corroborate our theoretical results. Our experiments are carried out on multi-head, Vision Transformers[2] that follows the block architecture given in Algorithm 1. We observe some key quantities at each block at the initial state where the input matrix $X \in \mathbb{R}^{n\times d}$ is randomly drawn from the standard normal distribution $\mathcal{N}(0,1)$ and the weight matrices W_k are randomly initialized from $\mathcal{N}\left(0, \frac{h}{d}I\right)$ (while other weight matrices are initialized using the default method in the code). The model parameters are set to $n = 64$, $d = 512$, $h = 8$, $\alpha = 1$ and the depth is set to 20. We always calculate attention matrices using the softmax formula (2)(though preliminary trials suggest that other formulas would essentially give the same results). We mention that the above experimentation setting, with different depth values, will be again used in the next section.

In the first experiment, we run 50 independent random trials. The average values of several important quantities are presented in Figure 2.

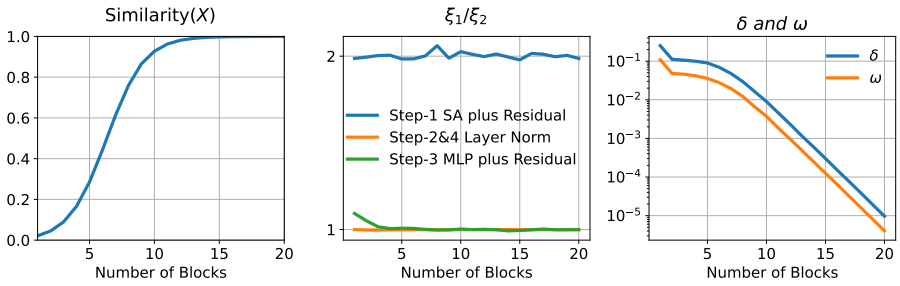

Figure 2: Average values of token similarity, ξ_1/ξ_2, δ and ω over 50 trials. At each multi-head block, δ and ω are also averaged across the eight heads.

From the left plot in Figure 2, we observe that similarity monotonically increases throughout all blocks, approaching the maximum before block 15, due to the fact that in Step 1 the (sampled) mean of ξ_1/ξ_2 stays around 2 from the very beginning, while the other three steps have a negligible impact on token similarity (since $\xi_1/\xi_2 \approx 1$), as can be

2. https://github.com/lucidrains/vit-pytorch

seen in the middle plot. Moreover, we see from the right plot that the quantities δ and ω are far less than 1 from the start and quickly approach 0 as the depth increases.

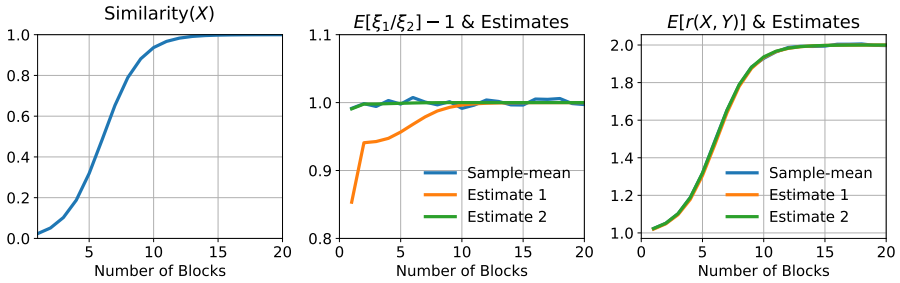

Figure 3: Average values over 1000 trials of token similarity, $\xi_1/\xi_2 - 1$ and $r(X, Y)$ and their estimates from (16) and (17) for the latter two quantities

In the second experiment, we only randomize the weight matrix W in Step 1 and run 1000 trials, while fixing all other quantities including X. This is exactly the same setting under which we derived expected values for relevant quantities in our analysis. In Figure 3, the left plot gives the average token similarity of X which looks visibly identical to the one given in the left plot of Figure 2.

In the middle plot of Figure 3, we present the sample mean of $\xi_1/\xi_2 - 1$ and the two estimates given in (16) (with $\alpha = 1$) and (17). Namely, Estimate 1 and 2 are, respectively,

$$\frac{(1-\omega)^2 - \delta^2}{1+\delta^2} \quad \text{and} \quad \frac{1 - |\lambda_2(P)|^2}{1 + |\lambda_2(P)|^2}.$$

It is interesting to observe that although Estimate 2 is theoretically valid only for symmetric attention matrix P, in the experiment it actually provides a closer approximation to $\mathbb{E}[\xi_1/\xi_2] - 1$ than the guaranteed lower bound in Estimate 1. This experiment exemplifies the argument that large spectral gaps of attention matrices are one of the real driving forces behind TSE, considering that Estimate 2 only depends on $\lambda_2(P)$.

In the right plot of Figure 3, we present the sample mean of $r(X, Y)$ and the corresponding two estimates. As we can see, the three curves are nearly identical. This is due to the fact that, in either (16) or (17), the token similarity term dominates the factor in front of it which varies relatively mildly.

2.5 Discussion

Our analysis reveals that the driving force behind TSE is two-fold: 1) the existence of the invariant leading eigenspace, span$\{\mathbb{1}\}$, for all attention matrices which are stochastic (or Markov), and 2) large spectral gaps commonly present in computed attention matrices.

Indeed, it has been established in (Bordenave et al., 2012) that, under mild conditions, n by n stochastic (or Markov) matrices generated from normalizations of i.i.d. nonnegative random variables almost surely have large spectral gaps of the order $1 - O(1/\sqrt{n})$ (see also the earlier work Chafaï (2010)). With randomly initialized weights in their construction, attention matrices are random (and stochastic) to also possess large spectral gaps with high probability, even though they may not necessarily or strictly satisfy all the required theoretical assumptions such as i.i.d. randomness.

Our similarity measure $\mathbf{t}_{sim}(\cdot)$ converges to 1 at a global linear rate while it itself also helps accelerate the rate. The asymptotical rate of convergence reaches $1/2$ when $\alpha = 1$. In view of the fact that $0.5^{10} \leq 10^{-3}$, this fast linear convergence explains why, under standard initializations, classic Transformers start to show some instability once the number of layers exceeds ten or so.

2.6 TSE in Decoder Models

Unlike BERT-based models, which adopt the Transformer encoder with bidirectional attention, GPT-based models are built on the Transformer decoder with unidirectional (causal) attention, see (Naveed et al., 2023). The difference is in the structure of attention matrices used. Decoder models use masked attention matrices that are lower-triangular, ensuring that each token can only attend to itself and preceding tokens.

It should be clear that our basic analysis applies equally to decoder models. Specifically, the basic result, Theorem 6, holds regardless of the structure of the attention matrix in use. On the other hand, whether TSE occurs, and how it behaves when it does occur, depends on the spectral gaps of the attention matrices involve. For decoder models, we need to examine the spectral gap of triangular attention matrices at random initialization.

To gain intuition about the spectral properties of triangular attention matrices, we consider the following realistic but simplified scenario. Let $P \in \mathbb{R}^{n\times n}$ be a random lower-triangular row-stochastic matrix constructed as follows. For each row $i = 1, \ldots, n$, draw i.i.d. random numbers $r_{i1}, \ldots, r_{ii}$ and let $s_i = \sum_{k=1}^{i} e^{r_{ik}}$. Then set

$$P_{ij} := \begin{cases} \dfrac{e^{r_{ij}}}{s_i}, & 1 \leq j \leq i, \\ 0, & i < j \leq n. \end{cases}$$

Since P is triangular, the eigenvalues of P are obviously its diagonal entries. It is straightforward to see that the expected value of the i-th eigenvalue is

$$\mathbb{E}[\lambda_i(P)] \equiv \mathbb{E}[P_{ii}] = \frac{1}{i}, \quad i = 1, \ldots, n.$$

In fact, because of the i.i.d. construction, all the i nonzero elements in the i-th row attain the same expectation $1/i$.

The simple analysis above offers a useful insight into decoder models: the spectral gaps of masked attention matrices consistently center around 0.5 when the model is randomly initialized, regardless of the specific initialization distribution (as long as it is i.i.d.). In contrast, the spectral gaps of full attention matrices in encoder models can vary depending on the type of random initialization used. One might therefore expect that, all else being equal, the TSE phenomenon is as likely to occur in decoder models as it is in encoder models. Indeed, we have observed similar TSE behaviors in a vision transformer (see Section 2.4 for model details), whether or not the triangular mask is applied.

In Section 3.1, we will present experimental results on two large-language models (LLMs): Qwen2 (Team, 2024) and Llama2 (Touvron et al., 2023), both adopting the decoder-only architecture. In addition, they also use a pre-norm architecture to suppress TSE by implicitly phasing out the self-attention (SA) mechanism; see Section 3.1 for more details.

3. Mitigation of TSE in Transformers

The analysis in the previous section underscores the issue of growing token similarity in post-norm Transformers. Many methods have been proposed to mitigate this problem, as discussed in our Related Work section. Implicitly or explicitly, these methods invariably lead to reducing the role of self-attention relative to residual. To take a simplistic view, in the SA-plus-residual step $X + \alpha P(X)XW$ one could explicitly diminish the size of α to slow down the progress of TSE; or one could instead modify the step into $X + P(\hat{X})\hat{X}W$ where $\hat{X}$ is related to X but has a smaller norm than X, thus implicitly reducing the role of self-attention.

3.1 Implicit Mitigation in Pre-norm

More recently, large-scale transformer models, such as those of (Touvron et al., 2023; Brown et al., 2020), opt for the pre-norm architecture. A pre-norm transformer block can be written as $Z = \mathbf{PreLN}(X)$ where Z is computed as follows:

$$\hat{X} = \mathbf{LN}(X), \quad Y = X + P(\hat{X})\hat{X}W, \quad Z = Y + \phi(\mathbf{LN}(Y)W_1)W_2, \tag{19}$$

From the above formulas with the usual random initializations, it is straightforward to verify that the expected value of $\|Y\|_F^2$ is larger than $\|X\|_F^2$, by following the same proof technique in the proof of Lemma 4. Specifically, under the standard weight initializations (where $\sigma = 1/\sqrt{d}$), for the above pre-norm transformer block there hold

$$\mathbb{E}[\|Y\|_F^2] = \|X\|_F^2 + \|P(\hat{X})\hat{X}\|_F^2 > \|X\|_F^2, \quad \mathbb{E}[\|Z\|_F^2] > \|Y\|_F^2. \tag{20}$$

According to (20), the output norm of the pre-norm block, $\|\mathbf{PreLN}(\cdot)\|_F$, is monotonically increasing in expectation, as is illustrated empirically in Figure 4 (second plot from the left).

Meanwhile, the norm of the self-attention input is always fixed at $\|\mathbf{LN}(\cdot)\|_F$ due to layer normalizations. Consequently, in the pre-norm architecture the role of the self-attention mechanism is progressively diminishing as the depth grows.

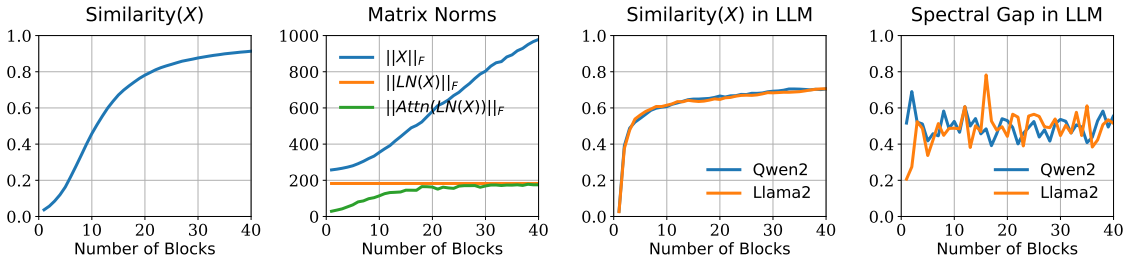

Figure 4: Token similarity and other quantities in three pre-norm transformer models. From left to right: (1) token similarity in a pre-norm vision transformer; (2) the corresponding Frobenius norms of X, $\mathbf{LN}(X)$ and $P\mathbf{LN}(X)W$ at each layer (see (19) for details); (3) token similarity in two LLMs that are based on pre-norm, decoder architecture; (4) spectral gaps of the corresponding triangular attention matrices in the LLMs.

In Figure 4, we illustrate the progression of token similarity and other related quantities across three pre-norm transformer models. The leftmost plot displays token similarity in a pre-norm vision transformer, which clearly increases through the layers but at a diminishing rate. This trend is attributed to a decreasing contribution from the attention operator relative to the residual term, as evidenced by the second plot from the left. The two plots on the right showcase results from two large language models (LLMs): Qwen2 Team (2024) and Llama2 Touvron et al. (2023), both employing pre-norm, decoder-only architectures. In these LLMs, token similarity rises rapidly at first, but the rate of increase quickly diminishes, resulting in token similarity remaining below 0.8 after 40 layers. This occurs despite the spectral gap of the triangular attention matrices in the decoders being approximately 0.5. These examples demonstrate that the pre-norm architecture can effectively curtail the escalation of token similarity at the cost of progressively suppressing the relative contribution of the self-attention (SA) mechanism. We also observe that the token similarity stabilizes at a lower level in the LLMs than in the vision transformer. This difference could be due to other differing architectural features such as different types of layer normalizations and activation functions.

3.2 A Simple De-escalation Strategy

We propose a simple strategy to counter the escalation of token similarity in deep Transformers, that is, to de-escalate the progressive growth in the subspace $\text{span}\{\mathbb{1}e_j^T\}_{j=1}^d$. Specifically, we will insert into Algorithm 1 a de-escalation step of the form

$$Y = (I - \tau\Pi_{\mathbb{1}})\, X, \quad \tau \in (0, 1]. \tag{21}$$

Alternatively speaking, we first project X onto $\text{span}\{\mathbb{1}e_j^T\}_{j=1}^d$ by applying the projection to all columns of X, and then subtract a portion of the projection from X. The parameter $\tau \in (0, 1]$ determines the removed portion. Particularly, $\tau = 0$ and $\tau = 1$ correspond to, respectively, no de-escalation and a complete de-escalation. Additionally, when $\tau = 1$ the operation becomes $Y = \Pi_{\mathbb{1}}^{\perp} X$ which is equivalent to the centralization of the columns of X, i.e., subtracting the column mean from each column. In principle, τ can be made a learnable parameter of the model, but the experimental results reported in Subsections 3.3 and 3.4 are all obtained using $\tau = 1$.

In Figure 5, we examine how the de-escalation operator (21) affects the dynamics of token similarity in the same Vision Transformer used in Subsection 2.4. We plot the token diversity values for layers 1 to 40, corresponding to the 6 different τ-values in $\{0, 0.1, \cdots, 0.5\}$. We observe that in this model the phenomenon of TSE appears to have been eliminated once τ reaches 0.4. In this particular test, operator (21) is applied to the layer output after the last step of Algorithm 1. We also experimented on inserting operator (21) into other possible locations in Algorithm 1, and observed almost identical plots.

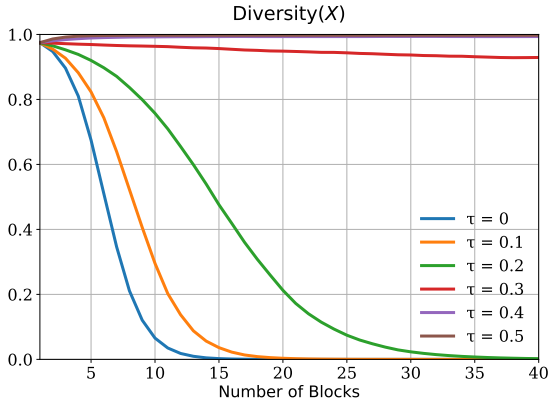

Figure 5: Token Diversity dynamics in a de-escalated Vision Transformer with softmax attentions corresponding to six de-escalation values τ. Each curve is the average values of 20 random runs.

In the following, we report proof-of-concept experiments to test whether adding the de-escalation step (21) will improve the training of post-norm transformers with deep depths.

On each test instance, we compare three model variants: post-norm model, pre-norm model, and ours where the de-escalation (21) is added within each Transformer block with $\tau = 1$ fixed without any tuning.

Two relatively small datasets, CIFAR10 (Krizhevsky et al., 2009) for image classification and WikiText-103 (Zhai et al., 2023) for natural language processing, will be used for our experiments due to constraints on available computing resources. All the experiments are carried out using PyTorch (Paszke et al., 2019) running on one Nvidia-V100 GPU. We emphasize that for each test instance we always run multiple trials, starting from different random initializations for model weights. All the reported values are the average of at least 3 trials.

3.3 Vision Transformer on CIFAR10

We apply the Vision Transformer model ViT (Dosovitskiy et al., 2020) to the CIFAR10 dataset with the prescribed three model variants. In this test, the de-escalation step (21) is added to the end of each post-norm block. We utilize the widely used optimizer AdamW (Loshchilov and Hutter, 2018) in which the hyper-parameters β_1 and β_2 are set to 0.9 and 0.999, respectively, along with a weight decay value of 0.1. A multi-step scheduler is employed with reduction factor of 1/5 at the 70% and 90% junctures of the training duration. Image patches are configured to be 4 by 4, and the gradient-sampling batch size is set to 128. Auto-augmentation (Cubuk et al., 2020) is enabled for the dataset. Further details about this ViT model's configurations are given in Table 1. With the depth 80, this tested model is qualified to be a deep transformer.

Table 1: Model size parameters for ViT

Depth	Hidden size	FFN size	Heads	Head size
80	192	384	8	24

The above settings are applied to all three model variants. There does exist a difference in the choice of learning rate lr. We used $lr = 10^{-4}$ for training the pre-norm and our models, while a smaller $lr = 0.5 * 10^{-4}$ was used for training the post-norm model in order to obtain a meaningful reduction in the loss value.

Training histories, over the course of 150 epochs, of the three ViT model variants on CIFAR10 are presented in Figure 6. From this figure, we can see a significant performance gap between the pure post-norm model and our de-escalated post-norm model.

In this particular case, our de-escalated post-norm model also obtained slightly lower training loss than its pre-norm counterpart, albeit such a test is too limited to make a meaningful comparison between the two. On the other hand, for the post-norm model the benefit of doing de-escalation appears unmistakable.

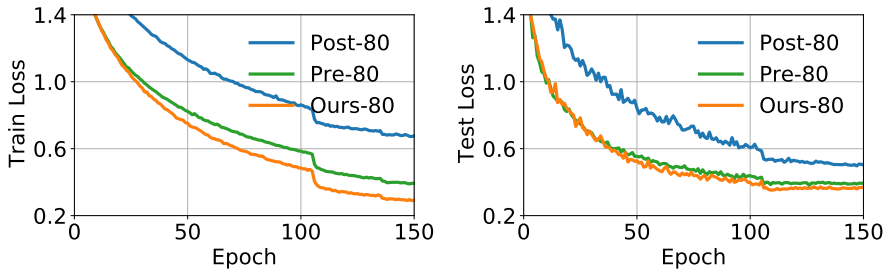

Figure 6: Histories of training and testing losses of three ViT model variants on the CIFAR10 dataset with auto-augmentation.

3.4 Transformer-XL on WikiText-103

We next employ the Transformer-XL model (Dai et al., 2019) to perform experiments on WikiText-103 dataset. In order to create a deep transformer model of a manageable size, we made the following modifications: increasing the model depth from 16 to 60 and decreasing the FFN hidden width (i.e., the column number of W_1 and W_2^T in Step 3 of Algorithm 1) from 2100 to 820. The resulting deep model has about 201M model parameters which is moderately larger than the original size of 151M. To train this larger model, we decrease the batch size from 60 to 40. In addition, we disable the dropout and gradient clipping options to have a more generic optimization process. Besides the afore-mentioned modifications, we keep all the model hyper-parameters and optimization settings intact exactly as specified in (Dai et al., 2019), which we refer to for further details. Figure 7 gives the training

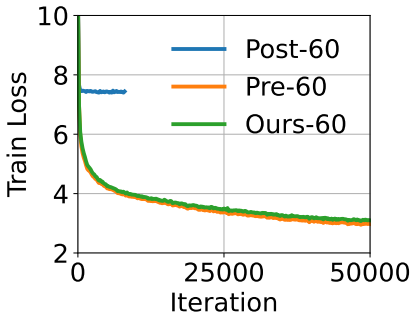

Figure 7: Training loss histories of 3 Transformer-XL variants on the WiKiText103 dataset over 50000 iterations.

performances of three Transformer-XL model variants of depth 60: post-norm, pre-norm, and our de-escalated post-norm model in which we add de-escalation operation (21) to the

input of each FFN block in the post-norm model. As we can see, the results are consistent with the previous experiment. There is a huge performance gap between the post-norm and our de-escalated post-norm models (the former essentially failed, thus was cut short); while the pre-norm model and our de-escalated post-norm model performed similarly.

3.5 Discussion

Our proposed framework offers flexibility in placing the de-escalation operation at different locations within the Transformer block, determining the value of de-escalation strength τ, and deciding whether de-escalation should be learnable or not. During our experimentation, we did try a number of combinations for these settings and found that our proposed de-escalation strategy is generally effective in addressing TSE-related issues. That said, real-world tasks are bound to be more intricate than our limited, small-scale experiments, for which adjustments and calibrations of all available settings are most likely needed to deal with different problem scenarios and characteristics.

4. Conclusion

We conduct a comprehensive analysis on the phenomenon of token similarity escalation (TSE) in classic Transformers which leads to loss of expressive power in deep models. Our theory reveals why and how TSE phenomenon occurs and what is the speed of escalation. Based on insights gained from our analysis, we propose a simple and linear de-escalation operation to surgically remove excessive similarity from token representations without necessarily suppressing the role of self-attention.

Proof-of-concept experiments show that, on small-scale transformers, the proposed de-escalation technique enables deep post-norm models to be trained as effectively as their pre-norm counterparts. The potential of the proposed strategy in large language models remains to be assessed.

Acknowledgments

This research was supported in part by Shenzhen Science and Technology Program (Grant Number GXWD20201231105722002-20200901175001001). The second author would like to thank his colleagues Professors Tiefeng Jiang and Jeff Yao for stimulating and useful discussions.

Appendix A. Proofs of Results in Section 2

A.1 Proof of Lemma 4

Proof Let "$\langle\cdot,\cdot\rangle$" denote the usual matrix inner product. For any given $Q \in \mathbb{R}^{n\times n}$,

$$\|Q(X+\alpha PXW)\|_F^2 = \|QX\|_F^2 + 2\alpha\langle QX, QPXW\rangle + \alpha^2\|QPXW\|_F^2.$$

Under Assumption 1 the expected value of the term linear in W vanishes. Hence,

$$\mathbb{E}\left[\|Q(X+\alpha PXW)\|_F^2\right] = \|QX\|_F^2 + \alpha^2\mathbb{E}[\|QPXW\|_F^2] = \|QX\|_F^2 + \alpha^2 d\sigma^2\|QPX\|_F^2,$$

in view of $\mathbb{E}[WW^T] = d\sigma^2 I$. Hence, the expressions for $\mathbb{E}[\xi_1]$ and $\mathbb{E}[\xi_2]$ follow from substituting the matrix Q by $\Pi_{\mathbb{1}}/\|\Pi_{\mathbb{1}}X\|_F$ and $\Pi_{\mathbb{1}}^\perp/\|\Pi_{\mathbb{1}}^\perp X\|_F$, respectively. We note that $ee^TPX = ee^TX$ if either $e^TP = e^T$ or $X = eu^T$.

Next, to derive the lower bound for μ_1. Rewriting $\Pi_{\mathbb{1}}PX = \Pi_{\mathbb{1}}X + \Pi_{\mathbb{1}}(P-I)X$, we calculate

$$\mu_1^2 = \frac{\|\Pi_{\mathbb{1}}X + \Pi_{\mathbb{1}}(P-I)X\|_F^2}{\|\Pi_{\mathbb{1}}X\|_F^2} = 1 + \frac{2\langle\Pi_{\mathbb{1}}X, \Pi_{\mathbb{1}}(P-I)X\rangle}{\|\Pi_{\mathbb{1}}X\|_F^2} + \frac{\|\Pi_{\mathbb{1}}(P-I)X\|_F^2}{\|\Pi_{\mathbb{1}}X\|_F^2}.$$

Since $\|\Pi_{\mathbb{1}}X\|_F = \|e^TX\|$ and $\|\Pi_{\mathbb{1}}(P-I)X\|_F = \|e^TP\Pi_{\mathbb{1}}^\perp X\|$, we recognize that the third term is ω^2. By applying Cauchy-Schwartz inequality to the middle term, we arrive at

$$\mu_1^2 \geq 1 - 2\omega + \omega^2 = (1-\omega)^2.$$

To obtain the upper bound of μ_2 by δ, we observe the equality

$$\Pi_{\mathbb{1}}^\perp P = \Pi_{\mathbb{1}}^\perp(P - ee^T) = \Pi_{\mathbb{1}}^\perp P\Pi_{\mathbb{1}}^\perp$$

which leads to

$$\mu_2 := \frac{\|\Pi_{\mathbb{1}}^\perp P\Pi_{\mathbb{1}}^\perp X\|_F}{\|\Pi_{\mathbb{1}}^\perp X\|_F} \leq \|\Pi_{\mathbb{1}}^\perp P\|_2 = \delta \tag{22}$$

after we invoke the inequality $\|AB\|_F \leq \|A\|_2\|B\|_F$. ■

A.2 Technical Lemmas

Lemma 10 *Let η be defined as in (14) and ξ_i, $i = 1, 2$, be defined as in (7) for $Y = X + \alpha PXW$. For all $t \in [0,1]$, there holds*

$$\mathbb{P}\left(|\eta| \geq t\right) \leq \mathbb{P}\left(\max_{i=1,2}|\xi_i - \mathbb{E}[\xi_i]| \geq \gamma t\right), \tag{23}$$

where

$$\gamma := \frac{\mathbb{E}[\xi_2]^2}{\mathbb{E}[\xi_1] + 2\mathbb{E}[\xi_2]} = \frac{(1+\alpha^2 d\sigma^2\mu_2^2)^2}{3 + \alpha^2 d\sigma^2(\mu_1^2 + 2\mu_2^2)}. \tag{24}$$

In particular, $\gamma \geq 1/4$ when P is doubly stochastic or $X = eu^T$ for some $u \in \mathbb{R}^d$, and $\alpha^2 = d\sigma^2 = 1$.

Proof By definition,

$$\eta = \frac{\mathbb{E}[\xi_1]}{\mathbb{E}[\xi_2]} - \frac{\xi_1}{\xi_2} = \frac{\mathbb{E}[\xi_1]\xi_2 - \xi_1\mathbb{E}[\xi_2]}{\mathbb{E}[\xi_2]\xi_2} = \frac{\mathbb{E}[\xi_1](\xi_2 - \mathbb{E}[\xi_2]) - \mathbb{E}[\xi_2](\xi_1 - \mathbb{E}[\xi_1])}{((\xi_2 - \mathbb{E}[\xi_2]) + \mathbb{E}[\xi_2])\mathbb{E}[\xi_2]}.$$

For any t, $|\eta| \geq t$ implies

$$t \leq \frac{(\mathbb{E}[\xi_1] + t\mathbb{E}[\xi_2])|\xi_2 - \mathbb{E}[\xi_2]| + \mathbb{E}[\xi_2]|\xi_1 - \mathbb{E}[\xi_1]|}{\mathbb{E}[\xi_2]^2},$$

which in turn implies that for $t \in [0, 1]$,

$$t \leq \frac{\mathbb{E}[\xi_1] + 2\mathbb{E}[\xi_2]}{\mathbb{E}[\xi_2]^2} \max_{i=1,2} |\xi_i - \mathbb{E}[\xi_i]| = \frac{1}{\gamma} \max_{i=1,2} |\xi_i - \mathbb{E}[\xi_i]|,$$

which proves (23) with γ defined in (24), while the second expression in (24) follows from (9).

Finally, we know that when P is doubly stochastic or $X = eu^T$, then $\mu_1 = 1$. Since γ increases monotonically with μ_2^2, it attains its minimum at $\mu_2 = 0$ which gives the minimum value $1/4$. ■

Now we need to estimate the concentration of ξ_1 and ξ_2, both being of the form $\|A + BW\|_F^2$, that is, sum of squares of a linear transformation of a random matrix W (which can also be viewed as a vector w). We will invoke the following two concentration results.

Lemma 11 *(General Hoeffding's inequality) Let $w \in \mathbb{R}^q$ be a random vector whose elements are independent, mean-zero, sub-gaussian random variables with sub-gaussian norm K. Then for any $a \in \mathbb{R}^q$ and $\epsilon \geq 0$ there holds*

$$\mathbb{P}\left(2\langle a, w\rangle \geq \epsilon\|a\|\right) \leq \exp\left(-c\frac{\epsilon^2}{K^2}\right), \tag{25}$$

where c is an absolute constant.

Lemma 12 *(Concentration of random vectors) Let $w \in \mathbb{R}^q$ be a random vector whose elements are independent, mean-zero, sub-gaussian random variables with sub-gaussian norm $K \leq 1$. Then for any given $B \in \mathbb{R}^{p\times q}$ and any $\epsilon \in (0, 1)$ there holds*

$$\mathbb{P}\left(\|Bw\|^2 - \|B\|_F^2 \geq \epsilon\|B\|_F^2\right) \leq \exp\left(-c\frac{\epsilon^2}{K^2}\right), \tag{26}$$

where c is an absolute constant.

Combining the above two results, we readily deduce that for some absolute constant c,

$$\mathbb{P}\left(\|Bw + a\|^2 - \mathbb{E}\|Bw + a\|^2 \geq \epsilon(\|B\|_F^2 + \|B^T a\|)\right) \leq 2\exp\left(-c\frac{\epsilon^2}{K^2}\right). \tag{27}$$

Remark 13 *A few remarks are in order.*

- *It should be noted that in these results the absolute constant c is generic in the sense that it can possibly have different values in different contexts. In doing so we avoid using multiple symbols for the sake of simplicity.*
- *Lemmas 11 and 12 are adopted from (Vershynin, 2018), see Theorem 2.6.3 and the proof of Theorem 6.3.2, respectively. These concentration inequalities are written in a form so that the right-hand sides are dimension-free.*
- *With regard to the values of sub-gaussian norm K (also called sub-gaussian parameter) in the above lemmas, it is known (see Example 2.5.8 in Vershynin (2018)) that one can take $K = \sigma$ for $w \sim N(0, \sigma^2 I)$ and $K = \|w\|_\infty$ if w is bounded while additional absolute constants, if any, can be absorbed into c.*

Lemma 14 *Let ξ_i, $i = 1, 2$, be defined as in Proposition 3 for $Y = X + \alpha PXW$. Assume that the elements of W are independent, mean-zero, sub-gaussian random variables with sub-gaussian norm $K = 1/\sqrt{d}$. Then for $i = 1, 2$, there holds*

$$\mathbb{P}\left(|\xi_i - \mathbb{E}[\xi_i]| \geq \epsilon \kappa_i\right) \leq 4 \exp\left(-c\epsilon^2 d\right), \tag{28}$$

where c is an absolute constant and $\kappa_i, i = 1, 2$, are constants dependent on the matrices $\Pi_i X$ and $\alpha \Pi_i PX$ but independent of W.

Proof This result follows directly from applying (27). ■

A.3 A Lower Bound for Dimension d

We derive a lower bound, d_*, for dimension d to ensure $\mathbb{E}[\eta] \leq \Delta$.

Lemma 15 *Assume the setting of Lemmas 10 and 14. Let γ be given in (24), and κ_1, κ_2 and c be the constants in (28). Let $\kappa_* := \max\{\kappa_1, \kappa_2\}$. For any $\Delta > 0$, define*

$$d_* := \frac{\pi}{c} \left(\frac{4\,\kappa_*}{\gamma\,\Delta}\right)^2. \tag{29}$$

Then for all $d \geq d_*$, there holds $\mathbb{E}[\eta] \leq \Delta$.

Proof By Lemma 10, for any $t \geq 0$,

$$\mathbb{P}(|\eta| \geq t) \;\leq\; \mathbb{P}\left(\max_{i=1,2} |\xi_i - \mathbb{E}[\xi_i]| \geq \gamma t\right).$$

Hence, using $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \geq t)\, dt$ for $Z \geq 0$,

$$\mathbb{E}[|\eta|] \;\leq\; \frac{1}{\gamma}\,\mathbb{E}\left[\max_{i=1,2} |\xi_i - \mathbb{E}[\xi_i]|\right].$$

By Lemma 14 and the union bound, for $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{i=1,2} |\xi_i - \mathbb{E}[\xi_i]| \geq \varepsilon\, \kappa_*\right) \;\leq\; 8e^{-c\,\varepsilon^2 d}.$$

Integrating with $t = \varepsilon\kappa_*$ gives

$$\mathbb{E}\left[\max_{i=1,2} |\xi_i - \mathbb{E}[\xi_i]|\right] \leq 8\int_0^\infty e^{-c(t/\kappa_*)^2 d}\, dt = 4\sqrt{\frac{\kappa_*^2\,\pi}{c\,d}}.$$

Therefore,

$$\mathbb{E}[|\eta|] \leq\; \frac{4\,\kappa_*}{\gamma}\sqrt{\frac{\pi}{c\,d}}.$$

If $d \geq d_*$ as is defined in (29), then $\mathbb{E}[|\eta|] \leq \Delta$, which implies $\mathbb{E}[\eta] \leq \Delta$. ■

A.4 Proof of Theorem 6

Proof Let $Y = X + \alpha PXW$ with $\sigma = 1/\sqrt{d}$. Then equation (13) can be written as

$$\mathbb{E}[r(X,Y)] = 1 + \left(\frac{\alpha^2}{1+\alpha^2\delta^2}((1-\omega)^2 - \delta^2) + \Delta - \mathbb{E}[\eta]\right)\mathbf{t}_{sim}(X),$$

with

$$\Delta = \frac{\alpha^2}{1+\alpha^2\mu_2^2}(\mu_1^2 - \mu_2^2) - \frac{\alpha^2}{1+\alpha^2\delta^2}((1-\omega)^2 - \delta^2) > 0, \tag{30}$$

where the positivity of Δ is ensured by (15). By Lemma 15, for $d \geq d_*$, there holds $\mathbb{E}[\eta] \leq \Delta$.

Dropping the nonnegative term $\Delta - \mathbb{E}[\eta]$ from the equality for $\mathbb{E}[r(X,Y)]$, we obtain the lower bound for $\mathbb{E}[r(X,Y)]$ in (16), and complete the proof. ■

To empirically observe the concentration of $\eta = \mathbb{E}[\xi_1]/\mathbb{E}[\xi_2] - \xi_1/\xi_2$ at zero, we conduct a set of experiments and present the results in Figure 8. Recall that η is a function of two matrices, $X, Y \in \mathbb{R}^{n\times d}$. We set $n = 100$ and run $d = 10, 20, 40$. For each d value, we generate a sequence of matrices of the form $X = ev^T + tQ \in \mathbb{R}^{100\times d}$ corresponding to a sequence of t values in $(0, 1]$, where the vector v and matrix Q are randomly chosen but otherwise fixed. We note that X is a perturbation to the rank-one matrix ev^T with t scaling the size of the perturbation. For each X, we generate $Y = X + PXW$ with 50 random samples of W drawn from $\mathcal{N}(0, I/d)$, and then compute the average value of η over the 50 samples. The attention matrix P is computed using the softmax formula (2).

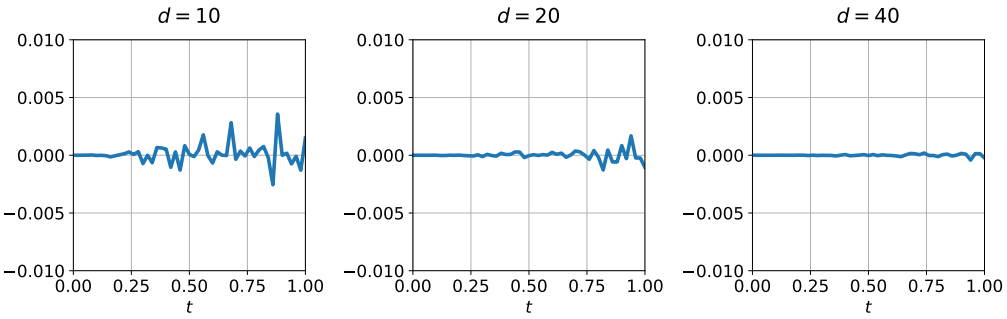

Figure 8: Average values of $\eta = \mathbb{E}[\xi_1]/\mathbb{E}[\xi_2] - \xi_1/\xi_2$ over 50 samples of $W \in \mathbb{R}^{d\times d}$ for $Y = X + PXW$ where $X = ev^T + tQ$ and P is from the softmax formula.

As we see from Figure 8, (a) when t is small (i.e., X close to rank-one), (sample mean) $\mathbb{E}[\eta]$ is close to 0 even for $d = 10$; (b) on the other hand, when t is close to 1, $\mathbb{E}[\eta]$ varies more significantly; and (c) as d increases, $\mathbb{E}[\eta]$ becomes closer to 0 even for larger t values. From these experiments, we observe that as d becomes larger, indeed η concentrates more at zero so that $\mathbb{E}[\eta]$ approaches 0.

A.5 Proof of Proposition 9

Lemma 16 *Given $X \in \mathbb{R}^{n\times d}$, $P_k \in \mathbb{R}^{n\times n}$ and $W_k \in \mathbb{R}^{d\times d/h}$, for $k = 1, \cdots, h$, and $\alpha > 0$, let*

$$Y = X + \alpha[P_1XW_1 \;\; P_2XW_2 \;\; \ldots \;\; P_hXW_h].$$

Then under Assumption 1(1),

$$\mathbb{E}[\xi_i] \equiv \mathbb{E}\left[\frac{\|\Pi_i Y\|_F^2}{\|\Pi_i X\|_F^2}\right] = 1 + \alpha^2 d\sigma^2 \bar{\mu}_i^2, \;\; i = 1, 2, \tag{31}$$

where

$$\bar{\mu}_i^2 := \frac{1}{h}\sum_{k=1}^{h} \frac{\|\Pi_i P_k X\|_F^2}{\|\Pi_i X\|_F^2}, \;\; i = 1, 2.$$

Proof The proof follows a similar line as in that of Lemma 4. First, we note that under Assumption 1(1), $\mathbb{E}[W_k] = 0$ and $\mathbb{E}[W_k W_k^T] = (d/h)\sigma^2 I_d$ for $k = 1, \cdots, h$.

Given $Q \in \mathbb{R}^{n\times n}$, noting that expected values of terms linear in W_k all vanish, we have

$$\mathbb{E}\left[\|Q(X + \alpha[P_1XW_1, \cdots, P_hXW_h])\|_F^2\right] = \|QX\|_F^2 + \alpha^2 \sum_{k=1}^{h} \mathbb{E}\left[\|QP_kXW_k\|_F^2\right],$$

In view of $\mathbb{E}[W_k W_k^T] = (d/h)\sigma^2 I$, we obtain

$$\mathbb{E}\left[\|Q(X + \alpha[P_1 X W_1, \cdots, P_h X W_h])\|_F^2\right] = \|QX\|_F^2 + \alpha^2 d\sigma^2 \frac{1}{h}\sum_{k=1}^{h} \|QP_k X\|_F^2.$$

Finally, the expressions for $\mathbb{E}[\xi_i]$ follow from substituting the matrix Q by $\Pi_i/\|\Pi_i X\|_F$, for $i = 1, 2$, which completes the proof. ■

Now Proposition 9 follows directly from Proposition 5 and Lemma 16.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence*, pages 1352–1361. PMLR, 2021.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Charles Bordenave, Pietro Caputo, and Djalil Chafaï. Circular law theorem for random markov matrices. *Probability Theory and Related Fields*, 152(3-4):751–779, 2012.

John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing: Algorithms, architectures and applications*, pages 227–236. Springer, 1990.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Djalil Chafaï. The dirichlet markov ensemble. *Journal of Multivariate Analysis*, 101(3): 555–567, 2010.

Don Coppersmith and Chai Wah Wu. Conditions for weak ergodicity of inhomogeneous markov chains. *Statistics & Probability Letters*, 78(17):3082–3085, 2008. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2008.05.012. URL `https://www.sciencedirect.com/science/article/pii/S016771520800271X`.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, 2019.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=SkEYojRqtm`.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

Bobby He, James Martens, Guodong Zhang, Aleksandar Botev, Andrew Brock, Samuel L Smith, and Yee Whye Teh. Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation. In *The Eleventh International Conference on Learning Representations*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. Whiteningbert: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2019.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, 2020.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5747–5763, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective post-processing for word representations. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Meenal V Narkhede, Prashant P Bartakke, and Mukul S Sutaone. A review on weight initialization strategies for neural networks. *Artificial intelligence review*, 55(1):291–322, 2022.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *NeurIPS 2022*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 813–823, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. B2t connection: Serving stability and performance in deep transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3078–3095, 2023.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018. Volume 47 of Cambridge Series in Statistical and Probabilistic Mathematics.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*, 2019a.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, 2019b.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

J. Wolfowitz. Products of indecomposable, aperiodic, stochastic matrices. *Proceedings of the American Mathematical Society*, 14(5):733–733, 1963.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020.

Hanqi Yan, Lin Gui, Wenjie Li, and Yulan He. Addressing token uniformity in transformers via singular value transformation. In *Uncertainty in Artificial Intelligence*, pages 2181–2191. PMLR, 2022.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse. *arXiv preprint arXiv:2303.06296*, 2023.

Zhong Zhang, Chongming Gao, Cong Xu, Rui Miao, Qinli Yang, and Junming Shao. Revisiting representation degeneration problem in language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 518–527, 2020.