Signal Propagation in Transformers: Theoretical Perspectives and the Role of Rank Collapse

Lorenzo Noci[*1]
lorenzo.noci@inf.ethz.ch

Sotiris Anagnostidis[*1]
sotirios.anagnostidis@inf.ethz.ch

Luca Biggio[*1,2]
luca.biggio@inf.ethz.ch

Antionio Orvieto[*1]
antonio.orvieto@inf.ethz.ch

Sidak Pal Singh[*1,3]
sidak.singh@inf.ethz.ch

Aurelien Lucchi[4]
aurelien.lucchi@unibas.ch

Abstract

Transformers have achieved remarkable success in several domains, ranging from natural language processing to computer vision. Nevertheless, it has been recently shown that stacking self-attention layers — the distinctive architectural component of Transformers — can result in rank collapse of the tokens' representations at initialization. The question of if and how rank collapse affects training is still largely unanswered, and its investigation is necessary for a more comprehensive understanding of this architecture. In this work, we shed new light on the causes and the effects of this phenomenon. First, we show that rank collapse of the tokens' representations hinders training by causing the gradients of the queries and keys to vanish at initialization. Furthermore, we provide a thorough description of the origin of rank collapse and discuss how to prevent it via an appropriate depth-dependent scaling of the residual branches. Finally, our analysis unveils that specific architectural hyperparameters affect the gradients of queries and values differently, leading to disproportionate gradient norms. This suggests an explanation for the widespread use of adaptive methods for Transformers' optimization.

1 Introduction

Since its first appearance in Vaswani et al. [2017], the Transformer architecture has revolutionized the field of Natural Language Processing (NLP), achieving remarkable success in tasks such as text classification [Yang et al., 2019], machine translation [Conneau and Lample, 2019], reading comprehension [Brown et al., 2020] and question answering [Raffel et al., 2019] among others. Recent efforts have effectively extended its applicability to computer vision [Dosovitskiy et al., 2020] and other domains [Baevski et al., 2020, Huang et al., 2018, Biggio et al., 2021, Polu et al., 2022], further popularizing it outside NLP.

The Transformer operates on inputs comprising a sequence of tokens. At its core, it relies on stacked attention layers, which compute a measure of relevance for the whole sequence by assigning token-wise importance weights — obtained by matrix multiplication of the *queries* and *keys*, and finally normalized with the softmax function. The output of an attention layer is then a linear combination

[1]Dept of Computer Science, ETH Zürich, [2]Robotics & ML, CSEM SA, Alpnach, Switzerland, [3]MPI for Intelligent Systems, Tübingen, [4]Department of Mathematics and Computer Science, University of Basel

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

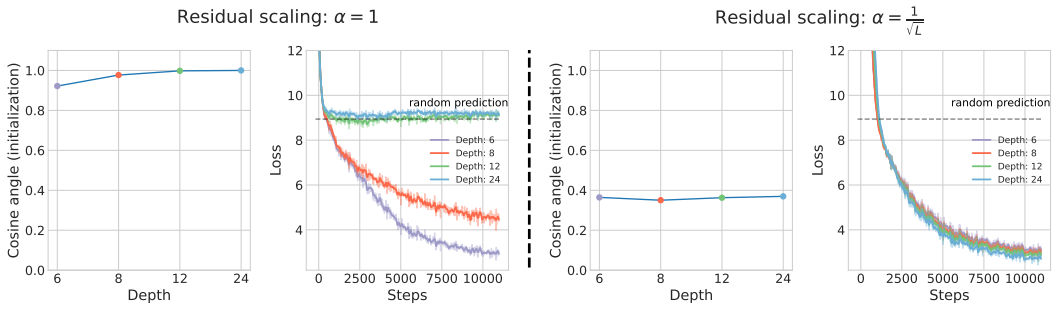

Figure 1: Evolution of the cosine of the angle between tokens for training POST-LN Transformers of increasing depth, with the Adam optimizer, for the IWSLT'14 De-En translation task. Unless adequate residual scaling is used at initialization, increasing depth leads to an increase in the tokens' alignment at initialization, which can inhibit training.

of the importance weights and the so-called *values*. Then, the architecture includes fully-connected sub-layers, residual connections [He et al., 2016], and layer normalization (LN), as illustrated in Fig. 2.

In the absence of residual connections, Dong et al. [2021] proved that at initialization the rank of the sequence representation collapses doubly exponentially with depth, and both layer normalization and fully connected layers can only partially alleviate the speed of degeneracy. Under *rank collapse*, the model does not distinguish between representations of different tokens, which are perfectly aligned in feature space at initialization. However, the precise implications of rank collapse in Transformers are not fully understood.

In this paper, we show that a high alignment of the tokens' representations at initialization — corresponding to rank collapse in the extreme case of perfect alignment — affects training by causing vanishingly small gradients of the queries and keys' parameter matrices. This problem severely diminishes the capabilities of the model to learn meaningful attention weights and is further exacerbated in very deep networks, where the rank deficiency — and hence the vanishing gradient problem of the queries and keys — affects several layers (see Fig. 1). In order to shed light on this problem, we take inspiration from the flourishing literature on signal propagation in random networks and start our analysis by computing the expected gradients of an attention layer with respect to the queries, keys, and values, which leads to Theorem 3.2 on the vanishing gradients for the queries and keys. From here, we pursue two different directions.

Firstly, we investigate under which conditions rank collapse can be avoided by studying the evolution of the input sequence in a Transformer at initialization. Our theory reveals that a depth-dependent scaling of the residual branches, beyond stabilizing the norm of the activations at initialization, also approximately preserves the cosine of the angle between tokens, hence stabilizing the rank of the propagating sequence. We show that this holds even in the infinite-depth limit.

Secondly, we illustrate that there are factors, other than the average tokens' correlation, that affect differently the gradient norm of the queries and keys compared to the values. In particular, the propagating sequence's squared norm has a linear dependence in the values, while a cubic one in the queries and keys, justifying the use of layer normalization. We also highlight a different dependence on the embedding dimension and the length of the input sequence, implying that the gradient norm of a subset of parameters can potentially be of different orders of magnitude, as empirically hinted by previous works [Liu et al., 2020]. Our analysis brings to light fundamental issues in the signal propagation in Transformers, opening the way for new, well-founded and motivated approaches to improve optimization in these models.

2 Background

Transformers. A Transformer architecture consists of L stacked attention blocks, as show in Fig. 2. Layer normalization is usually applied token-wise either after the residual connections or to the inputs

of the self-attention and position-wise feed-forward sub-layers, leading to the POST-LN [Vaswani et al., 2017] and PRE-LN [Wang et al., 2019, Xiong et al., 2020] variants respectively.

Formally, given an input sequence $\mathbf{X} \in \mathbb{R}^{n \times d_v}$, with n tokens of dimension d_v, the single-head unmasked scaled dot-product self-attention[1] is defined as:

$$\mathbf{S}^\ell := \mathbf{A}^\ell \mathbf{X}^\ell \mathbf{W}^V, \text{ where } \mathbf{A}^\ell = \text{softmax}\left(\frac{1}{\sqrt{d_k}} \mathbf{X}^\ell \mathbf{W}^Q \left(\mathbf{X}^\ell \mathbf{W}^K\right)^\top\right), \tag{1}$$

where the softmax function is applied independently across each row, and the superscript ℓ indexes the ℓ-th layer. The matrices $\mathbf{W}^Q, \mathbf{W}^K \in \mathbb{R}^{d_v \times d_k}$ and $\mathbf{W}^V \in \mathbb{R}^{d_v \times d_v}$ are learnable parameters, and each layer is initialized with an independent set of weights. In the literature, the matrices $\mathbf{X}^\ell \mathbf{W}^Q, \mathbf{X}^\ell \mathbf{W}^K, \mathbf{X}^\ell \mathbf{W}^V$ are referred to as queries, keys and values, respectively. The complete Transformer block, in the absence of layer normalization, can be written recursively as:

$$\mathbf{Z}^\ell = \alpha_1 \mathbf{S}^\ell + \mathbf{X}^\ell \tag{2}$$

$$\mathbf{Y}^\ell = \sigma(\mathbf{Z}^\ell \mathbf{W}^{F_1}) \mathbf{W}^{F_2} \tag{3}$$

$$\mathbf{X}^{\ell+1} = \alpha_2 \mathbf{Y}^\ell + \mathbf{Z}^\ell, \tag{4}$$

where the introduced α_1, α_2 parameters indicate the strength of the residual block, $\mathbf{W}^{F_1}, \mathbf{W}^{F_2} \in \mathbb{R}^{d_v \times d_v}$ [2] are matrices of learnable parameters; we set $\mathbf{X}^0 := \mathbf{X}$, and $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function. In our case, σ is the ReLU function, but we relax this assumption to the linear activation from Section 3.2 on.

At initialization, each weight is sampled independently from a distribution with zero-mean and variance $\sigma_v^2 = \frac{1}{d_v}$ for the values and feedforward weights[3], and $\sigma_k^2 = \frac{1}{d_k}$ for the queries and keys. This is the standard "Xavier" [Glorot and Bengio, 2010] or "He" [He et al., 2015] initialization, commonly used in deep learning.

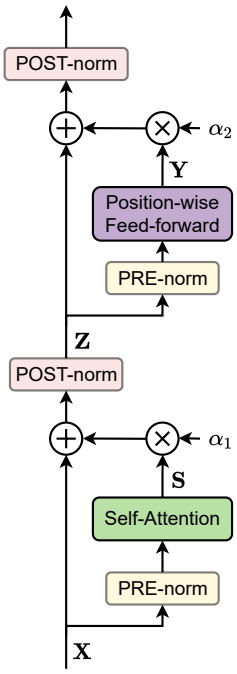

Figure 2: A single Transformer block.

Rank Collapse in Transformers. Interestingly, Dong et al. [2021] proved that when the residual branches are omitted, the matrix of the tokens' representations $\mathbf{X}^\ell$ converges to a rank-1 matrix in which all the representations are the same and equal to a vector $\boldsymbol{x} \in \mathbb{R}^{d_v}$, i.e. $\mathbf{X}^\ell \to \mathbf{1}_n \boldsymbol{x}^\top$, where $\mathbf{1}_{d_v}$ is the vector with all ones in $\mathbb{R}^{d_v}$. Note that this is a slightly stronger notion of a rank-1 matrix, as it implies that all the tokens' representations are both perfectly aligned and have the same norm. Indicating the inner product with the usual bracket notations $\langle \cdot, \cdot \rangle$, and the cosine of the angle between two tokens as $\theta_{k,k'}$, perfect alignment happens when $\langle \mathbf{X}_k^\ell, \mathbf{X}_{k'}^\ell \rangle = \left\|\mathbf{X}_k^\ell\right\| \left\|\mathbf{X}_k^\ell\right\| \cos\theta_{k,k'}$ with $\cos\theta_{k,k'} = 1$ for all $k, k' \in [n]$. Note that perfect alignment together with equal norm between all the tokens implies that all the representations are the same. One of our main contributions is to provide an explanation of how rank collapse affects the gradients of a Transformer at initialization.

Vanishing Gradient Problem. Traditionally considered one of the core issues that prevents successful training, the vanishing gradient problem has a long and rich history that dates back to before the popularization of deep learning [Hochreiter, 1991, Bengio et al., 1994]. In its essence, given a loss function $\mathcal{L} : \mathbb{R}^{n \times d_v} \to \mathbb{R}$, vanishing gradients occur when the norm of the gradient of the loss $\mathcal{L}$ with respect to the parameters of the network $\mathbf{W}$ — which we indicate as $\left\|\frac{\partial \mathcal{L}}{\partial \mathbf{W}}\right\|$ — is too small to provide enough backpropagating signal, thus hindering gradient-based optimization methods.

[1] Our analysis also easily generalizes to the case of cross-attention.

[2] In practice, one commonly uses $\mathbf{W}^{F_1} \in \mathbb{R}^{d_v \times d_F}, \mathbf{W}^{F_2} \in \mathbb{R}^{d_F \times d_v}$ where $d_F = \gamma d_v$, with $\gamma \in \{2, 4, 8\}$. Our results then hold up to a constant factor that depends on γ.

[3] One should explicitly write the layer dependence $\mathbf{W}^{Q,\ell}, \mathbf{W}^{K,\ell}, \mathbf{W}^{V,\ell}, \mathbf{W}^{F_1,\ell}, \mathbf{W}^{F_2,\ell}$. We at times suppress the ℓ index to improve readability. In case σ is the ReLU function, we set $\mathbf{W}^{F,1}$ to have variance $\frac{2}{d_v}$.

Despite extensive research toward understanding and overcoming the problem in disparate contexts [Glorot and Bengio, 2010, He et al., 2015, Hanin, 2018, Zhang et al., 2019], a formal explanation of its role in relatively new architectures such as Transformers is largely missing in the literature, with a few exceptions [Xiong et al., 2020, Wang et al., 2022, Huang et al., 2020]. In our paper (Section 3.1), we show how vanishing gradient occurs in conjunction with the rank collapse issue identified by Dong et al. [2021].

Signal Propagation in Random Networks at Initialization. After addressing the question on the effects of rank collapse, we take a step back and rigorously analyze its causes by looking at how the properties of the input sequence $\mathbf{X}$ are lost/preserved as it propagates through a randomly initialized Transformer. More specifically, we focus on two aspects of the propagating sequence: the expected Frobenius norm $\mathbb{E}\left\|\mathbf{X}^\ell\right\|^2$ and the expected inner product between different tokens $\mathbb{E}\langle \mathbf{X}_k, \mathbf{X}'_k\rangle$, with $k \neq k'$. The former is linked to a number of studies on the initialization of neural networks at the *edge of chaos* [Poole et al., 2016, Schoenholz et al., 2017], and vanishing/exploding gradients [Hanin, 2018]. The latter quantity describes how the geometry of the feature space changes after applying a Transformer block, and is related to the concept of *dynamical isometry* [Saxe et al., 2013]. To understand the evolution of the inner product, we analyze the following measure of correlation [Nachum et al., 2021, Cho and Saul, 2009]:

$$\rho^\ell_{kk'} := \frac{\mathbb{E}\langle \mathbf{X}^\ell_k, \mathbf{X}^\ell_{k'}\rangle}{\sqrt{\mathbb{E}\left\|\mathbf{X}^\ell_k\right\|^2 \mathbb{E}\left\|\mathbf{X}^\ell_{k'}\right\|^2}}. \tag{5}$$

Note that $\rho^\ell_{kk'} = 1$ if and only if the k-th and k'-th tokens are perfectly aligned ($\cos\theta_{kk'} = 1$). We stress that in our case — differently from the aforementioned works — instead of analyzing the relationship between two different data points, we study the relationship between tokens of the same sequence.

3 Theoretical Results

The goal of this section is twofold. In Lemma 3.1 and Theorem 3.2, we provide an explanation of the possible cause of vanishingly small gradients for queries and keys at initialization, namely the high correlations between the tokens representations in $\mathbf{X}^\ell$ as the depth increases. In Section 3.2, we show that the problem can be mitigated with an appropriate choice of the residual branch parameters α_1 and α_2 that inversely scales with the depth of the network L. Under the proposed scaling, the correlations are well behaved even in the infinite depth limit. Finally, in Section 3.3 we analyze the scaling of the gradients with respect to other network's parameters, and in 3.4 we draw some connections between our findings and optimization of Transformers.

3.1 Vanishing Gradients for Queries and Keys under Rank Collapse

To investigate the problem of vanishing gradients in the attention layers, we make use of the framework of matrix calculus [Magnus and Neudecker, 2019, Singh et al., 2021]. In particular, we compare the expected Frobenius norm of the gradient of a self-attention layer with respect to its parameters $\mathbb{E}\left\|\frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}}\right\|_F^2$, where here $\mathbf{W}$ indicates one of the keys, queries or values weight matrices. Due to the well-known difficulty of computing expectations of the softmax [Daunizeau, 2017, Shekhovtsov and Flach, 2018], throughout this manuscript, we make the simplifying assumption that the softmax output is the uniform distribution at initialization, i.e. the $n \times n$ matrix containing $\frac{1}{n}$ in each entry.

Assumption 3.1 (Uniform attention)**.** *We assume that* $\mathbf{A}^\ell = \frac{1}{n}\mathbf{1}_{n\times n}$,

where $\mathbf{1}_{n\times n}$ is the matrix with all entries equal to 1. Crucially, in Appendix A.5, we formally show that *this assumption holds almost surely* in the limit $d_k \to \infty$. There, we also experimentally show that even in the more realistic case where $d_k = d_v \approx 512$, the empirical simulations provide a surprisingly faithful approximation of the theoretical insights presented in this paper.

We define the mean token $\bar{\boldsymbol{x}}^\ell$ through its components $\bar{\boldsymbol{x}}^\ell_i = \frac{1}{n}\sum_{k=1}^n \mathbf{X}^\ell_{ki}$, $i \in [d_v]$. In the following theorem, we compute the expected gradients of an attention layer at initialization, and set the basis for our following analysis. We provide the results only for the queries, as the case for the keys is analogous.

Lemma 3.1. *Let $\mathbf{X}^\ell$ be the representations of the input sequence at the ℓ-th layer. Under the uniform-attention assumption, we have*

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^{V,\ell}}\right\|_F^2 = d_v n \mathbb{E}\|\bar{\boldsymbol{x}}^\ell\|^2 \; ; \tag{6}$$

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^{Q,\ell}}\right\|_F^2 = \frac{\sigma_v^2\sigma_k^2 d_v}{n^2} \cdot \mathbb{E}\left[\|\mathbf{X}^\ell\|_F^2 \cdot \|(\mathbf{X}^\ell)^\top \mathbf{X}^\ell - n\bar{\boldsymbol{x}}^\ell(\bar{\boldsymbol{x}}^\ell)^\top\|_F^2\right] \; ; \tag{7}$$

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}^\ell}{\partial \mathbf{X}^\ell}\right\|_F^2 \leq \frac{8\sigma_q^2\sigma_k^2\sigma_v^2 d_k d_v}{n} \cdot \mathbb{E}\left\|(\mathbf{X}^\ell)^\top \mathbf{X}^\ell - n\bar{\boldsymbol{x}}^\ell(\bar{\boldsymbol{x}}^\ell)^\top\right\|_F^2 + 2d_v^2\sigma_v^2 \; . \tag{8}$$

We defer the precise study of the scaling of these quantities as a function of n and d_v, d_k, to Section 3.3. At this stage, it is crucial to note that $\frac{1}{n}(\mathbf{X}^\ell)^\top \mathbf{X}^\ell - \bar{\boldsymbol{x}}^\ell(\bar{\boldsymbol{x}}^\ell)^\top$ is the centered empirical covariance matrix of the tokens' representations. It is easy to see that if $\mathbf{X}^\ell$ is a rank-1 matrix, then all the rows of $\mathbf{X}^\ell$ are proportional to a fixed d_v-dimensional vector, and the empirical covariance matrix has all zero entries. Introducing a differentiable loss function $\mathcal{L} : \mathbb{R}^{n\times d_v} \to \mathbb{R}$, we make the statement on vanishing gradients more formal in the following theorem:

Theorem 3.2 (Vanishing gradients under rank collapse)**.** *Suppose that the uniform-attention assumption holds. If additionally $\mathbf{X}^\ell$ for any $l \in [L]$ has rank-1, and there exists a vector $\boldsymbol{x} \in \mathbb{R}^d$ such that $\mathbf{X}^\ell = \mathbf{1}_n \boldsymbol{x}^T$, then:*

$$\mathbb{E}\left\|\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{Q,\ell}}\right\|_F^2 = 0, \quad \mathbb{E}\left\|\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{K,\ell}}\right\|_F^2 = 0, \tag{9}$$

where the expectation is taken over the weight matrices. This implies that these quantities are vanishing almost surely, due to the non-negativeness of the norm.

The proof simply relies on expanding the norm of the gradient of the loss with the aid of the chain rule and then bounding it by the product of the norms of each term of the chain. The final result holds with an application of Lemma 3.1, in which the rank-1 assumption makes $\mathbb{E}\left\|\frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^{Q,\ell}}\right\|$ vanish. The proof of the results presented in this section is deferred to Appendix A.1.

In light of Theorem 3.2, we can conclude that the issue of rank collapse originally identified in Dong et al. [2021] corresponds to an initialization in a region of vanishing gradient signal in the subspace of parameters identified by the queries and keys. How can this affect training? One may argue that if rank collapse does not happen in the very first layer, then the corresponding gradients are non-zero and the rank of the subsequent layers — affected by rank collapse — can be increased with the first few steps of gradient descent. In practice, we show empirically in Fig. 1 that escaping this pathological landscape is harder in deeper nets where rank collapse persists across several layers.

3.2 Forward Signal Propagation and the Importance of Scaling the Residual Branches

We now turn our attention to the study of the influence of skip connections in Transformers. Dong et al. [2021] showed that simply adding this architectural trick prevents rank collapse. Somewhat surprisingly, we show that while the claim holds for any finite depth, the average angle between different tokens quickly increases with just a few layers, and as $L \to \infty$ a Transformer can still lose rank unless the residual branches are adequately initialized. As Dong et al. [2021] showed that layer normalization does not avoid rank collapse, we omit it in our analysis. Firstly, we introduce two lemmas on the propagation of inner products (Lemma 3.2) and the norm (Lemma 3.3) of the tokens' representations.

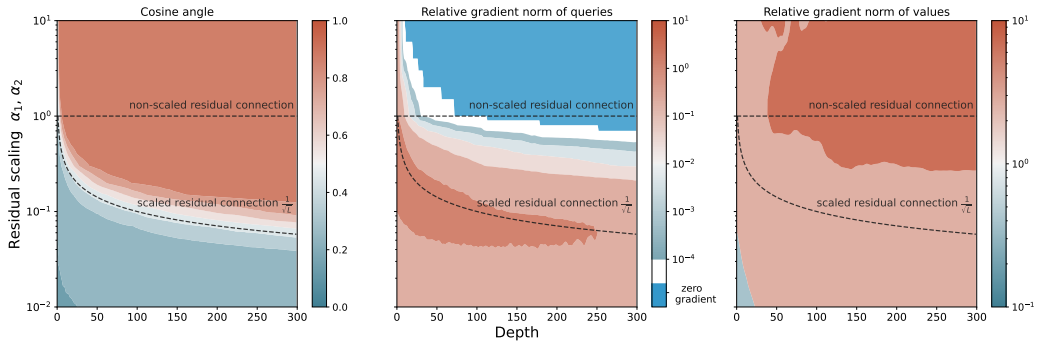

Figure 3: Effect of the residual scaling to the norm of the gradients of the network at initialization with respect to some loss. From left to right: (a) the cosine of the angle between tokens increases with depth. Note how larger values of α_1, α_2 imply a faster token alignment with depth (Theorem 3.3). Subplots (b) and (c) show the gradients of the queries-keys and values parameters respectively by increasing depth, compared to the corresponding norms of the first layer. Gradients for the queries-keys diminish with depth, while the opposite happens for the values. We use POST-LN to disentangle the effect of the variance of the input.

Lemma 3.2 (Propagation of inner products)**.** *Let $C(\mathbf{X}^\ell) = \sum_{k,k'} \langle \mathbf{X}_k^\ell, \mathbf{X}_{k'}^\ell \rangle$ and $\mathbf{X}$ the input sequence. Under the Assumption 3.1 and if σ is the linear activation function, we have that:*

$$\mathbb{E}\left[C(\mathbf{X}^L)\right] = (\alpha_2^2+1)^L(\alpha_1^2+1)^L C(\mathbf{X}). \tag{10}$$

hence, under the depth scaling for the residual block parameters $\alpha_1^2 = \frac{\tilde{\alpha}_1}{L}, \alpha_2^2 = \frac{\tilde{\alpha}_2}{L}$ with $\tilde{\alpha}_1, \tilde{\alpha}_2 \in \mathbb{R}$ independent of L, we have that:

$$\lim_{L\to\infty} \mathbb{E}[C(\mathbf{X}^L)] = e^{\tilde{\alpha}_1+\tilde{\alpha}_2} C(\mathbf{X}). \tag{11}$$

Note that $C(\mathbf{X}^\ell) = n^2 \left\|\bar{\boldsymbol{x}}^\ell\right\|^2$. The lemma on the propagation of the norm is slightly more involved:

Lemma 3.3 (Propagation of the norm)**.** *Let $\mathbf{X}^L$ be the representations of the input sequence at the final layer. Under the assumptions of Lemma 3.2, we have that:*

$$\mathbb{E}\left\|\mathbf{X}^L\right\|_F^2 = n(\alpha_2^2+1)^L\alpha_1^2 \sum_{k=0}^{L-1}(\alpha_1^2+1)^k \left\|\bar{\boldsymbol{x}}\right\|^2 + (\alpha_2^2+1)^L ||\mathbf{X}||_F^2, \tag{12}$$

hence, under the depth scaling for the residual block parameters $\alpha_1^2 = \frac{\tilde{\alpha}_1}{L}, \alpha_2^2 = \frac{\tilde{\alpha}_2}{L}$ with $\tilde{\alpha_1}, \tilde{\alpha_2} \in \mathbb{R}$ independent of L, we have that:

$$\lim_{L\to\infty} \mathbb{E}\left\|\mathbf{X}^L\right\|_F^2 = ne^{\tilde{\alpha}_2}(e^{\tilde{\alpha}_1}-1)\left\|\bar{\boldsymbol{x}}\right\|^2 + e^{\tilde{\alpha}_2}||\mathbf{X}||_F^2. \tag{13}$$

The proof of Lemma 3.3 consists in expanding $\mathbb{E}\left\|\mathbf{X}^L\right\|_F^2$ according to the defining equations for the Transformer, and simplifying the expression by using iterated expectations $\mathbb{E}\left\|\mathbf{X}^L\right\|_F^2 = \mathbb{E}[\mathbb{E}[\left\|\mathbf{X}^L\right\|_F^2 |\mathbf{X}^\ell]]$ to exploit the conditional independence between different layers, and then computing the expectations using the independence assumption on the weights. The expression on the right-hand side will then depend on $\mathbf{X}^\ell$ only through its norm $\left\|\mathbf{X}^\ell\right\|$ and the norm of the mean token $\left\|\bar{\boldsymbol{x}}^\ell\right\|^2$. Using Lemma 3.2 then allows us to unroll the recursion and get the final result. The complete proof, together with the proof of Lemma 3.2, can be found in Appendix A.3.

The previous Lemma provides theoretical justification that scaling the residual branches by setting the alpha parameters to be $\mathcal{O}(1/\sqrt{L})$ allows both the norm of the propagating input and the inner products between different tokens to be approximately preserved. Hence, the information contained in the input is not lost, even in the infinite depth limit.

Residual Scaling Preserves Correlations. We now prove that without the depth-dependent residual scaling (i.e. with $\alpha_1 = \alpha_2 = 1$) the correlation between the tokens quickly increases, and reaches perfect alignment in the infinite depth limit. More specifically, our argument shows that in this limit, the correlation between different tokens $\rho^{\ell}_{k,k'}$ as in Eq. (5) converges to 1, implying rank collapse. Furthermore, we show how setting the residual parameters α_1 and α_2 as dictated by Theorem 3.3, ensures that the correlation measure is dependent on the input in a non-trivial way even at infinite depth. To this end, we introduce the average correlation at layer ℓ:

$$\rho^{\ell} = \frac{1}{n(n-1)} \sum_{k \neq k'} \rho^{\ell}_{kk'}. \tag{14}$$

Note that $\rho^{\ell} = 1$ if and only if every pair of tokens is perfectly aligned. We are now ready to formalize the influence of the $1/\sqrt{L}$-scaling on the correlation between tokens' representations by stating Theorem 3.3.

Theorem 3.3. *Let the input tokens have the same norm, i.e.* $\|\mathbf{X}_k\| = \|\boldsymbol{x}\| \ \forall k \in [n]$ *for some* $\boldsymbol{x} \in \mathbb{R}^{d_v}$. *Under the depth scaling for the residual block parameters* $\alpha_1^2 = \frac{\tilde{\alpha}_1}{L}, \alpha_2^2 = \frac{\tilde{\alpha}_2}{L}$ *with* $\tilde{\alpha}_1, \tilde{\alpha}_2 \in \mathbb{R}$ *independent of L, we have that:*

$$\lim_{L \to \infty} \rho^{\ell} = \frac{n e^{\tilde{\alpha}_1} C(\mathbf{X})}{(n-1)[(e^{\tilde{\alpha}_1} - 1) C(\mathbf{X}) + n \|\mathbf{X}\|_F^2]} - \frac{1}{n-1}. \tag{15}$$

On the other hand, if $\alpha_1, \alpha_2 \neq 0$ *are some constants independent of L, we have that:*

$$\lim_{L \to \infty} \rho^{\ell} = 1. \tag{16}$$

The proof consists in noting that due to the symmetry of the problem at initialization, for a fixed layer the expected norm of each token is the same. Hence, by our definition of $\rho^{\ell}_{kk'}$, we can write $\mathbb{E}\langle \mathbf{X}^{\ell}_k, \mathbf{X}^{\ell}_{k'} \rangle = \rho^{\ell}_{kk'} \mathbb{E} \left\| \boldsymbol{x}^{\ell} \right\|^2$. By summing over the k, k' indexes, the resulting equation will depend on $\mathbb{E}[C(\mathbf{X}^{\ell})]$ and $\mathbb{E} \left\| \mathbf{X}^{\ell} \right\|^2$, which can be expanded using Lemma 3.2 and 3.3 respectively. The result is then given by solving for ρ^{ℓ}.

Note that under the $1/\sqrt{L}$-scaling, the correlation term is one if and only if $C(\mathbf{X}) = n \|\mathbf{X}\|^2$, which holds in the degenerate case where all the input tokens are perfectly aligned. In Appendix A.4, we give precise formulas for the expected correlations at any depth, showing that ρ^{ℓ} reaches values close to one even for relatively shallow networks when the $1/\sqrt{L}$-scaling is not adopted (see also Fig. 3 (left)). Additionally, in Fig. 4, we empirically show that in the presence of the $1/\sqrt{L}$-scaling, layer normalization (either PRE or POST) does not significantly affect the evolution of the correlations. On the other hand, without the residual scaling, PRE-LN seems to alleviate the rate of increase of $\rho^{\ell}_{kk'}$. It is intriguing that most deep Transformer models use this configuration [Brown et al., 2020]. We provide more extensive empirical results in Appendix B.

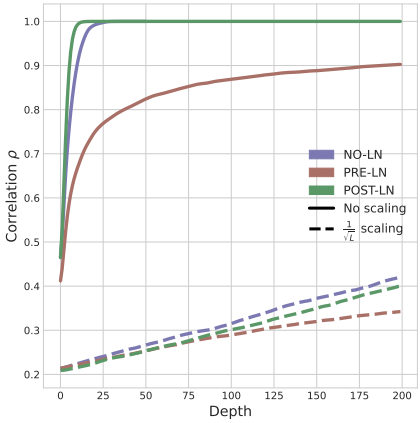

Figure 4: Evolution of Correlation in Transformers with (dashed lines) and without (solid lines) $1/\sqrt{L}$-scaling for PRE-LN, POST-LN and without layer normalization (No-LN).

Note that the $1/\sqrt{L}$ scaling for the residual branches has been previously studied in the context of stabilization of residual networks (see Section 4), here we extend these results to Transformers and provide new insights on its role in the context of rank preservation. Finally, note that by setting $\tilde{\alpha}_1, \tilde{\alpha}_2 = 0$, we recover the so called "ReZero" initialization [Bachlechner et al., 2021]. In this context, the $1/\sqrt{L}$ scaling extends this framework as it allows for wider range of values for $\tilde{\alpha}_1, \tilde{\alpha}_2$ while still guaranteeing stability.

ReLU extension. We mention here that extending these results from the linear activation to the ReLU case is known to be a hard problem, due to the technical difficulty of propagating the inner products across ReLU layers that are shared among the tokens (this is the case in the position-wise feed-forward layers in Transformers). Exact formulas can be found only in the case of one ReLU layer with Gaussian inputs in Cho and Saul [2009]. However, in the context of rank collapse analyzed here, the linear activation function provides a bound on the correlation with respect to the ReLU case. In fact, correlations are exactly preserved in expectation in the linear case, but increase in the ReLU case (for instance, see the contraction argument in Nachum et al. [2021] below Equation (2)). Hence, the perfect alignment (a.k.a rank collapse) that affects the linear case affects the ReLU case as well (in which case the rank collapses even faster with depth, as we show in Figure 10).

3.3 Dependence on the Angle between Tokens and the Input Norm

In this section, we drop the superscript ℓ as it is obvious from context and assume for simplicity that $d_k = d_v$. To gain a better intuition on the factors that affect the gradients and provide additional insights, we study the case in which every pair of distinct tokens are zero-mean Gaussian random variables, correlated in the same way, i.e $\rho^{\ell}_{ii'} = \rho$ for $i \neq i'$ or more precisely

$$\mathbb{E}\left[\mathbf{X}_{i,j}\mathbf{X}_{i',j'}\right] = \begin{cases} 0 & j \neq j' \text{ (independent dimensions)} \\ \sigma_x^2 & i = i', j = j' \\ \rho\sigma_x^2 & i \neq i', j = j' \end{cases} .$$

To see that this equation satisfies our definition of the correlation metric, note that $\mathbb{E}[\|\mathbf{X}_i\|^2] = d\sigma_x^2$ and $\mathbb{E}\langle \mathbf{X}_i, \mathbf{X}_{i'}\rangle = d\sigma_x^2\rho$, for $i \neq i'$. Then, the expected norm of the gradients for the values (Eq. (6)) simplifies to

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}}{\partial \mathbf{W}^V}\right\|_F^2 = \sigma_x^2 d^2\left(1 + \rho(n-1)\right). \tag{17}$$

By making the additional assumption that the norm and the correlation propagate independently, the respective norm for the queries — and symmetrically the keys — (Eq. (7)) reduces to:

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}}{\partial \mathbf{W}^Q}\right\|_F^2 = \sigma_x^6 \frac{(n-1)}{n}(1-\rho)^2 d(n+d). \tag{18}$$

In Appendix A.2 we provide a rigorous proof, that relies on Isserlis theorem [Isserlis, 1918] to compute higher-order moments. The above expressions reveal the different dependencies on four main actors, that we inspect separately here. The gradients of the queries depend via a cubic function on the *variance of the input,* σ_x^2, compared to a linear for the values. This provides an additional interpretation of the successful use of layer normalization, as in Xiong et al. [2020], either in the POST-LN or PRE-LN format, that standardizes the input variance σ_x^2 to the value 1.

Next, we emphasize the dependence on the *correlation between the tokens*, also illustrated in Fig. 3. Importantly, note how the queries/keys have opposite monotonic functional dependence with respect to ρ compared to the values. As revealed by Theorem 3.3 and Fig. 3 (center), inappropriate scaling of the residual branches can already lead to this phenomenon even in a relatively shallow network.

Finally, Eq. (17) and (18) reveal a different scaling in terms of the *embedding size* d and the *sequence length* n due to the self-attention operation itself. We hope that the identification of the different dependencies in the gradients of the parameters will inspire a new line of works aimed at solving some of the difficulties in training Transformers.

3.4 Connections to Optimization and Adaptive Methods

The existence of the discrepancy in the magnitude of the gradients with respect to the weights $\mathbf{W}^Q, \mathbf{W}^K$ and $\mathbf{W}^V$, might explain the success of adaptive optimization algorithms, as illustrated in Fig. 6, where we plot the effective learning rate computed by Adam [Kingma and Ba, 2014] in a toy encoder task (more details in Appendix C). Notice that the effective learning rate is increasingly larger (with depth) for the queries compared to the values — as postulated by our theory — and this difference is remarkably constant throughout training. Hence, we conjecture that the success of adaptive methods in Transformers' training can be partially explained by the need to fix this

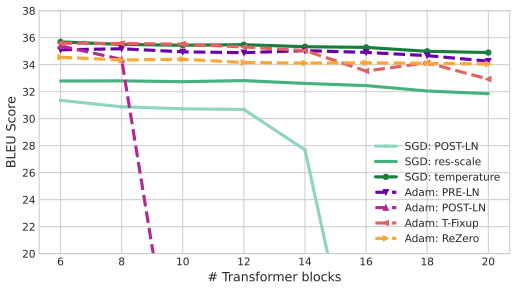

Figure 5: BLEU scores by increasing the number of transformers blocks. 'X' Transformer blocks implies in total 'X' encoder self-attention, 'X' decoder self-attention and 'X' decoder cross-attention layers.

Table 1: BLEU scores for the IWSLT14 German-to-English translation task. *SGD res-scale* refers to the training of SGD without layer normalization and initialization of the residual scaling $a_1 = a_2 = \frac{1}{\sqrt{L}}$. *SGD temperature* additionally employs an inverse temperature inside the softmax.

Method (6L-Encoder / 6L-Decoder)	BLEU ↑
SGD POST-LN	31.36
SGD res-scale	32.79
SGD temperature	**35.69**
Adam POST-LN [Vaswani et al., 2017]	35.39
Adam PRE-LN [Vaswani et al., 2017]	35.10
ReZero [Bachlechner et al., 2021]	34.55
T-Fixup Zhang et al. [2019]	35.59

disproportionate gradient's magnitude. To test this hypothesis, we propose a simple architectural modification, an inverse temperature scaling $\tau \in \mathbb{R}$ inside the softmax:

$$\mathbf{S}_\tau^\ell := \text{softmax}\left(\frac{\tau}{\sqrt{d_k}}\mathbf{X}^\ell\mathbf{W}^Q\left(\mathbf{X}^\ell\mathbf{W}^K\right)^\top\right)\mathbf{X}^\ell\mathbf{W}^V. \tag{19}$$

A direct consequence of our analysis is that τ allows controlling the magnitude of the gradients for the queries and keys' parameters. In Section C.2, we detail how one can choose τ such that the magnitude of the gradients as derived in Equation 17 and 18 is approximately matched at initialization.

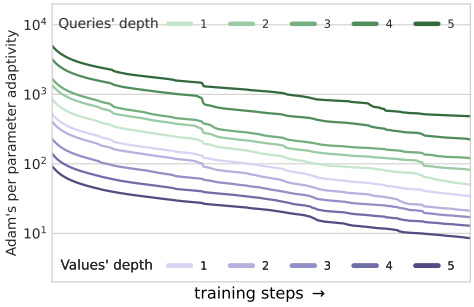

Figure 6: Adaptive learning rates computed by Adam in Transformers.

We evaluate our proposal, consisting of residual scaling and the aforementioned inverse temperature parameters, on the widely used IWSLT14 German-to-English (De-En) benchmark translation task. All details regarding the experimental setup and the choice of inverse temperature used are provided in Appendix C. We train a Transformer encoder-decoder of varying depth with stochastic gradient descent (SGD), after removing all normalization layers and adequately initializing the residual connections. For our training with SGD, we avoid using any learning rate warm-up, as commonly done for Adam, and instead use a step-scheduler to decrease the learning rate at 40% and 80% of training. We compare against the following methods that make use of Adam; POST-LN and PRE-LN refer to the aforementioned alternatives to apply layer normalization. We also compare against other successful techniques that rely on specific initializations to avoid layer normalization, such as ReZero [Bachlechner et al., 2021] and T-Fixup [Zhang et al., 2019]. We report the average BLEU score [Papineni et al., 2002] across 5 runs in Fig. 5 and Table 1.

Our proposed method considerably improves training with SGD, keeping up and in some cases surpassing any results achieved by the Adam optimizer. We are also able to train deeper networks without the use of layer normalization. We leave for future work to further investigate modifications or alternatives to the self-attention operation.

4 Related Work

Our work builds upon the rich literature on forward and backward signal propagation in random neural networks [Poole et al., 2016, Schoenholz et al., 2017, Xiao et al., 2018, Pennington et al., 2017, Orvieto et al., 2021, Noci et al., 2021, Zavatone-Veth and Pehlevan, 2021]. The $1/\sqrt{L}$ scaling scheme has been investigated in the literature for the stabilization of residual networks [Hanin and Rolnick, 2018, Arpit et al., 2019, Allen-Zhu et al., 2019, Hayou et al., 2021].

Our work draws inspiration from a series of recent works studying the rank of the representations of random feed-forward neural networks at initialization [Daneshmand et al., 2020, 2021]. In the context of Transformers, Dong et al. [2021] has recently identified the rank collapse issue object of study of the present work. Thanks to our analysis of the backward pass, we are able to demonstrate that rank collapse in Transformer architectures leads to vanishingly small gradients of queries and keys, thereby preventing effective training and allowing us to complete the analysis of [Dong et al., 2021].

Among the architectural components in Transformers, layer normalization is, arguably, one of the most important – and debated – ones [Chen et al., 2018, Wang et al., 2019, Nguyen et al., 2010, Xiong et al., 2020]. In the original architecture [Vaswani et al., 2017], layer normalization is used to stabilize the forward pass by reducing the variance of the inputs to the following sublayer. Our analysis of the forward pass shows that its inclusion is not strictly necessary for the purpose of controlling the norm of the representations. For a theoretical analysis of signal propagation in the presence of layer norm, we refer the reader to Xiong et al. [2020].

Additionally, our theoretical study of the backward pass provides a rigorous explanation of the empirically observed discrepancy between the magnitude of the gradients of the queries and the values, which Liu et al. [2020] hypothesize to be one of the causes of the success of adaptive methods in training Transformers [Liu et al., 2019, Zhang et al., 2020, Huang et al., 2020].

Finally, properly rescaled residual connections have been found to be beneficial for training Transformers by a number of recent research works [Zhang et al., 2019, Bachlechner et al., 2021, Wang et al., 2022]. However, none of these studies characterize the impact of skip connections on rank propagation, while our analysis suggests a theoretically-grounded way to stabilize it.

5 Conclusions and Future Work

In this paper, we showed how, at initialization, rank collapse and more generally high correlation in the tokens causes vanishing gradients of the queries and keys of a Transformer architecture. While residual connections help mitigate rank collapse at finite depth, we showed that they alone cannot prevent high alignments of the tokens' representations — unless properly scaled by a $1/\sqrt{L}$-factor. Finally, we have also discovered counter-intuitive dependencies on the variance of the input, embedding size, and sequence length, potentially causing large differences between the gradients of queries/keys compared to the values' parameters. Hence, we conclude that one of the strengths of Transformers lies in their carefully designed architecture together with an adequate initialization. Finally, we gave preliminary evidence that one of the factors contributing to the higher efficacy of Adam compared to SGD in training Transformers arises from the disproportionate magnitude of gradients as postulated by our theory. Nonetheless, other factors might further accentuate the difference between these two algorithms during training, leaving the door open for further research regarding the benefits of adaptive optimization methods with Transformers.

Acknowledgements

Sidak Pal Singh would like to acknowledge the financial support from Max Planck ETH Center for Learning Systems and the travel support from ELISE (GA no 951847).

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

Devansh Arpit, Víctor Campos, and Yoshua Bengio. How to initialize your network? robust initialization for weightnorm & resnets. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurelien Lucchi, and Giambattista Parascandolo. Neural symbolic regression that scales. In *International Conference on Machine Learning*, pages 936–945. PMLR, 2021.

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.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, et al. The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv:1804.09849*, 2018.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

Alexis Conneau and Guillaume Lample. *Cross-Lingual Language Model Pretraining*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Hadi Daneshmand, Jonas Kohler, Francis Bach, Thomas Hofmann, and Aurelien Lucchi. Batch normalization provably avoids ranks collapse for randomly initialised deep networks. *Advances in Neural Information Processing Systems*, 33:18387–18398, 2020.

Hadi Daneshmand, Amir Joudaki, and Francis Bach. Batch normalization orthogonalizes representations in deep random networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Jean Daunizeau. Semi-analytical approximations to statistical moments of sigmoid and softmax mappings of normal variables. *arXiv preprint arXiv:1703.00091*, 2017.

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. *arXiv preprint arXiv:2010.11929*, 2020.

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.

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? *Advances in neural information processing systems*, 31, 2018.

Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. *Advances in Neural Information Processing Systems*, 31, 2018.

Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable resnet. In *International Conference on Artificial Intelligence and Statistics*, pages 1324–1332. PMLR, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91(1), 1991.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2018.

Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. Improving transformer optimization through better initialization. In *International Conference on Machine Learning*, pages 4475–4483. PMLR, 2020.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond, 2019. URL https://arxiv.org/abs/1908.03265.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. *arXiv preprint arXiv:2004.08249*, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

Ido Nachum, Jan Hązła, Michael Gastpar, and Anatoly Khina. A johnson–lindenstrauss framework for randomly initialized cnns. *arXiv preprint arXiv:2111.02155*, 2021.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Lorenzo Noci, Gregor Bachmann, Kevin Roth, Sebastian Nowozin, and Thomas Hofmann. Precise characterization of the prior predictive distribution of deep relu networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Antonio Orvieto, Jonas Kohler, Dario Pavllo, Thomas Hofmann, and Aurelien Lucchi. Vanishing curvature and the power of adaptive methods in randomly initialized deep networks. *AISTATS 2022 (to appear)*, 2021.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/d9fc0cdb67638d50f411432d0d41d0ba-Paper.pdf.

Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. Formal mathematics statement curriculum learning. *arXiv preprint arXiv:2202.01344*, 2022.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29, 2016.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *ICLR*, 2017.

Alexander Shekhovtsov and Boris Flach. Feed-forward propagation in probabilistic neural networks with categorical and max layers. In *International conference on learning representations*, 2018.

Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic insights into structure and rank of neural network hessian maps. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=otDgw7LM7Nn.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. *arXiv preprint arXiv:2203.00555*, 2022.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5393–5402. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/xiao18a.html.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Jacob Zavatone-Veth and Cengiz Pehlevan. Exact marginal prior distributions of finite bayesian neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Hongyi Zhang, Yann N Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. *arXiv preprint arXiv:1901.09321*, 2019.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 3
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] Most of them in the Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code will be released upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Table of Contents

A Proof of Theorems **15**
- A.1 Backward Pass: Proofs of Lemma 3.1 and Theorem 3.2 . . . 16
- A.2 Gradient Analysis of Section 3.3 . . . 22
- A.3 Forward Pass: Proofs of Lemma 3.2 and 3.3 . . . 25
- A.4 Proof of Theorem 3.3: Correlations are Preserved under Residual Scaling . . . 28
- A.5 Motivation for Assumption 3.1 . . . 28

B Additional Results **30**
- B.1 On the Roles of the $1/\sqrt{L}$-Scaling of the Residuals and Layer Normalization . . . 30
- B.2 Further Empirical Assessment of Assumption 3.1 . . . 31
- B.3 Empirical Verification of the Gradient Analysis of Section 3.3 . . . 32

C Experimental Setup **33**
- C.1 Toy Example . . . 34
- C.2 Tempering the softmax . . . 34
- C.3 Translation Task . . . 36

A Proof of Theorems

Recall the defining equations of a Transformer:

$$\begin{aligned}\mathbf{Z}^\ell &= \alpha_1 \mathbf{S}^\ell + \mathbf{X}^\ell \\ \mathbf{Y}^\ell &= \sigma(\mathbf{Z}^\ell \mathbf{W}^{F,1})\mathbf{W}^{F,2} \\ \mathbf{X}^{\ell+1} &= \alpha_2 \mathbf{Y}^\ell + \mathbf{Z}^\ell,\end{aligned}$$

where the self attention layers are defined as follows:

$$\mathbf{S}^\ell := \mathbf{A}^\ell \mathbf{X}^\ell \mathbf{W}^V, \text{ where } \mathbf{A}^\ell = \text{softmax}\left(\frac{1}{\sqrt{d_k}} \mathbf{X}^\ell \mathbf{W}^Q \left(\mathbf{X}^\ell \mathbf{W}^K\right)^\top\right).$$

We remark that when it is clear from the context, we suppress the ℓ index to improve readability.

Initialization: Recall that we initialize our weights with the so called "Xavier" [Glorot and Bengio, 2010] or "He" [He et al., 2015] initialization: each weight is sampled independently from a distribution with zero-mean and variance $\sigma_v^2 = \frac{1}{d_v}$ for the values and feedforward weights and $\sigma_k^2 = \frac{1}{d_k}$ for the queries and keys.

Kronecker Delta: we introduce the Kronecker Delta notation

$$\delta_{ab} = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}$$

and, similarly: $\delta_{a \neq b} = 1 - \delta_{ab}$.

Notation: in this section, we also adopt the following shorthand notation for the argument of the softmax $\mathbf{M} := \frac{\mathbf{X}\mathbf{W}^Q \mathbf{W}^{K\top} \mathbf{X}^\top}{\sqrt{d_k}}$. We will first compute the gradients with respect to values, queries and input (note that the gradients of the keys have the form as the queries, hence we omit the derivation). Recall that for a matrix $\mathbf{X}$, we use $\mathbf{X}_k$ to indicate its k-th row. Finally, we indicate with $\mathbf{1}_{n\times n}$ the $n \times n$ matrix with all ones, with $\mathbf{1}_n$ the columns vector with all ones, and with $\mathbf{I}_n$ the n-dimensional identity matrix.

A.1 Backward Pass: Proofs of Lemma 3.1 and Theorem 3.2

In this section, we now look at the proofs of Lemma 3.1 and Theorem 3.2. We will first introduce our notation for the gradients, as well as some useful properties of the Kronecker product.

A.1.1 Preliminaries

For the gradients, we avoid directly working with tensors by vectorizing the matrices in a *row-wise* fashion (vec_r) and arranging the Jacobian in the numerator layout. More formally,

$$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}} := \frac{\partial\, \mathrm{vec}_r(\mathbf{Y})}{\partial\, \mathrm{vec}_r(\mathbf{X})^\top}\,.$$

Alongside this, we use the following rule ($\otimes$ is the Kronecker product):

$$\frac{\partial \mathbf{AWB}}{\partial \mathbf{W}} = \mathbf{A} \otimes \mathbf{B}^\top\,. \tag{20}$$

For the proof of this rule, we refer to Singh et al. [2021], and to Magnus and Neudecker [2019] for a complete introduction to matrix calculus. We will also use the following well-known properties of the Kronecker product.

Lemma A.1. *Given the matrices* $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathbf{C} \in \mathbb{R}^{m\times r}$, $\mathbf{D} \in \mathbb{R}^{q\times s}$, *then the following holds:*

$$\mathrm{tr}(\mathbf{A}\otimes\mathbf{B}) = \mathrm{tr}(\mathbf{A})\,\mathrm{tr}(\mathbf{B}), \tag{21}$$

and

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD}). \tag{22}$$

A.1.2 Proof of Lemma 3.1

In Lemma A.2 and Lemma A.3 we compute the gradients with respect to the queries, values and $\mathbf{X}$, respectively. Then we use these results to prove Lemma 3.1 by computing the expectation of the Frobenius norms.

Lemma A.2 (Gradients of Self Attention for parameter matrices)**.** *The gradients of the self attention layer defined in Eq. (1) have the following form:*

$$\frac{\partial \mathbf{S}}{\partial \mathbf{W}^V} = \mathrm{softmax}\left(\frac{\mathbf{X}\mathbf{W}^Q{\mathbf{W}^K}^\top\mathbf{X}^\top}{\sqrt{d_k}}\right)\mathbf{X}\otimes\mathbf{I}_{d_v}$$

$$\frac{\partial \mathbf{S}}{\partial \mathbf{W}^Q} = \left(\mathbf{I}_n\otimes{\mathbf{W}^V}^\top\mathbf{X}^\top\right)\frac{\partial \mathbf{A}}{\partial \mathbf{M}}\left(\frac{\mathbf{X}\otimes\mathbf{X}\mathbf{W}^K}{\sqrt{d_k}}\right),$$

where the gradients of the softmax with respect to its inputs are as follows:

$$\frac{\partial \mathbf{A}}{\partial \mathbf{M}} = \mathrm{blockdiag}\left(\frac{\partial \mathbf{A}_i}{\partial \mathbf{M}_i^\top}\right) \tag{23}$$

and where $\frac{\partial \mathbf{A}_i}{\partial \mathbf{M}_i^\top} = \mathrm{diag}(\mathbf{A}_i) - \mathbf{A}_i\mathbf{A}_i^\top$ *with* $\mathbf{A}_i$ *being the* i*-th row of* $\mathbf{A}$ *in column vector format. Finally, note that under the uniform-attention assumption, Eq. (23) simplifies to:*

$$\frac{\partial \mathbf{A}}{\partial \mathbf{M}} = \frac{1}{n}\mathbf{I}_n\otimes\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right). \tag{24}$$

Proof. Let's start with the simple case of the values' weights $\mathbf{W}^V$. Using the rule in Eq. (20), it is immediate that:

$$\frac{\partial \mathbf{S}}{\partial \mathbf{W}^V} = \mathrm{softmax}\left(\frac{\mathbf{X}\mathbf{W}^Q{\mathbf{W}^K}^\top\mathbf{X}^\top}{\sqrt{d_k}}\right)\mathbf{X}\otimes\mathbf{I}_{d_v} = \mathbf{AX}\otimes\mathbf{I}_{d_v}\,.$$

For the queries, a simple application of the chain rule and then again Eq. (20) gives:

$$\begin{aligned}\frac{\partial \mathbf{S}}{\partial \mathbf{W}^Q} &= \frac{\partial \mathbf{S}}{\partial \mathbf{A}}\frac{\partial \mathbf{A}}{\partial \mathbf{W}^Q} = \frac{\partial \mathbf{S}}{\partial \mathbf{A}}\frac{\partial \mathbf{A}}{\partial \mathbf{M}}\frac{\partial \mathbf{M}}{\partial \mathbf{W}^Q} \\ &= \left(\mathbf{I}_n \otimes \mathbf{W}^{V\top}\mathbf{X}^\top\right)\frac{\partial \mathbf{A}}{\partial \mathbf{M}}\left(\frac{\mathbf{X}\otimes \mathbf{X}\mathbf{W}^K}{\sqrt{d_k}}\right),\end{aligned}$$

which is the desired results. Finally, for the gradients of the softmax note that:

$$\frac{\partial \mathbf{A}_{pq}}{\partial \mathbf{M}_{ij}} = \frac{\partial}{\partial \mathbf{M}_{ij}}\frac{\exp(\mathbf{M}_{pq})}{\sum_k \exp(\mathbf{M}_{pk})} = \delta_{ip}\delta_{jq}\mathbf{A}_{ij} - \delta_{ip}\mathbf{A}_{iq}\mathbf{A}_{ij}.$$

By writing the above expression in the matrix notation described above, we obtain the desired result. More specifically, the block diagonal structure is given from the term δ_{ip} which stems from the fact that the softmax is applied row-wise. □

Lemma A.3 (Gradients of Self Attention with respect to the Embedding matrix)**.** *The gradients of the self attention layer with respect to the embedding matrix* $\mathbf{X}$ *defined in Eq.* (1) *have the following form*

$$\frac{\partial \mathbf{S}}{\partial \mathbf{X}} = \frac{1}{\sqrt{d_k}}(\mathbf{I}_n \otimes \mathbf{W}^{\mathbf{V}\top}\mathbf{X}^\top)\frac{\partial \mathbf{A}}{\partial \mathbf{M}}\left(\mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^K\mathbf{W}^{\mathbf{Q}\top} + \mathbf{K}_{nn}(\mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^Q\mathbf{W}^{\mathbf{K}\top})\right) + \mathbf{A}\otimes \mathbf{W}^{V\top}, \tag{25}$$

where the gradients of the softmax with respect to its inputs are denoted by $\frac{\partial \mathbf{A}}{\partial \mathbf{M}}$ *as before.*

Proof. Remember that we defined $\mathbf{S} = \mathrm{softmax}(\frac{1}{\sqrt{d_k}}\mathbf{X}\mathbf{W}^Q\mathbf{W}^{\mathbf{K}\top}\mathbf{X}^\top)\mathbf{X}\mathbf{W}^V$. Alongside with our previous shorthands $\mathbf{A}$, $\mathbf{M}$, let us define the remaining $\mathbf{X}\mathbf{W}^V$ as a matrix $\mathbf{T}$, so that $\mathbf{S} = \mathbf{A}\,\mathbf{T}$. Both $\mathbf{A}$ and $\mathbf{T}$ are functions of $\mathbf{X}$. So the matrix differential can be written as:

$$\begin{aligned}\frac{\partial \mathbf{S}}{\partial \mathbf{X}} &= \frac{\partial \mathbf{S}}{\partial \mathbf{A}}\frac{\partial \mathbf{A}}{\partial \mathbf{X}} + \frac{\partial \mathbf{S}}{\partial \mathbf{T}}\frac{\partial \mathbf{T}}{\partial \mathbf{X}} &(26)\\ &= \frac{\partial \mathbf{S}}{\partial \mathbf{A}}\frac{\partial \mathbf{A}}{\partial \mathbf{M}}\frac{\partial \mathbf{M}}{\partial \mathbf{X}} + \frac{\partial \mathbf{S}}{\partial \mathbf{T}}\frac{\partial \mathbf{T}}{\partial \mathbf{X}} &(27)\\ &= (\mathbf{I}_n \otimes \mathbf{W}^{\mathbf{V}\top}\mathbf{X}^\top)\frac{\partial \mathbf{A}}{\partial \mathbf{M}}\frac{\partial \mathbf{M}}{\partial \mathbf{X}} + (\mathbf{A}\otimes \mathbf{I}_d)(\mathbf{I}_n \otimes \mathbf{W}^{\mathbf{V}\top}) &(28)\\ &= (\mathbf{I}_n \otimes \mathbf{W}^{\mathbf{V}\top}\mathbf{X}^\top)\frac{\partial \mathbf{A}}{\partial \mathbf{M}}\frac{\partial \mathbf{M}}{\partial \mathbf{X}} + (\mathbf{A}\otimes \mathbf{W}^{\mathbf{V}\top}) &(29)\end{aligned}$$

Next, we use the matrix differential and then the identification theorem of matrix derivatives to compute the matrix gradient $\frac{\partial \mathbf{A}}{\partial \mathbf{X}}$

$$\mathrm{d}\mathbf{A} = \frac{1}{\sqrt{d_k}}\mathrm{d}(\mathbf{X})\,\mathbf{W}^Q\mathbf{W}^{\mathbf{K}\top}\mathbf{X}^\top + \frac{1}{\sqrt{d_k}}\mathbf{X}\mathbf{W}^Q\mathbf{W}^{\mathbf{K}\top}\,\mathrm{d}(\mathbf{X}^\top).$$

Vectorizing both sides:

$$\begin{aligned}\mathrm{d}\,\mathrm{vec}_r(\mathbf{A}) &= \frac{1}{\sqrt{d_k}}(\mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^K\mathbf{W}^{\mathbf{Q}\top})\mathrm{d}(\mathrm{vec}_r(\mathbf{X})) + \frac{1}{\sqrt{d_k}}(\mathbf{X}\mathbf{W}^Q\mathbf{W}^{\mathbf{K}\top}\otimes \mathbf{I}_n)\,\mathrm{d}(\mathrm{vec}_r(\mathbf{X}^\top))\\ &= \frac{1}{\sqrt{d_k}}(\mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^K\mathbf{W}^{\mathbf{Q}\top})\mathrm{d}(\mathrm{vec}_r(\mathbf{X})) + \frac{1}{\sqrt{d_k}}(\mathbf{X}\mathbf{W}^Q\mathbf{W}^{\mathbf{K}\top}\otimes \mathbf{I}_n)\mathbf{K}_{dn}\,\mathrm{d}(\mathrm{vec}_r(\mathbf{X})).\end{aligned}$$

Recall, for an arbitrary matrix $\mathbf{B}\in\mathbb{R}^{m\times n}$, the commutation matrix $\mathbf{K}_{mn}$ transforms columnwise vectorization into rowwise vectorization. More precisely,

$$\mathbf{K}_{mn}\,\mathrm{vec}_c(\mathbf{B}) = \mathrm{vec}_c(\mathbf{B}^\top)$$

and $\mathrm{vec}_c(\mathbf{B}) = \mathrm{vec}_r(\mathbf{B}^\top)$. Therefore, for rowwise vectorization, we have a similar result:

$$\begin{aligned}\mathbf{K}_{mn}\,\mathrm{vec}_r(\mathbf{B}^\top) &= \mathrm{vec}_r(\mathbf{B})\\ \mathrm{vec}_r(\mathbf{B}^\top) &= \mathbf{K}_{nm}\,\mathrm{vec}_r(\mathbf{B}),\end{aligned}$$

where in the last line we used the fact the commutation is a permutation matrix, so $\mathbf{K}_{mn}^{-1} = \mathbf{K}_{mn}^\top = \mathbf{K}_{nm}$. Thus, we get the required matrix derivative as follows:

$$\frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \frac{1}{\sqrt{d_k}} \mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^K \mathbf{W}^{\mathbf{Q}\top} + \frac{1}{\sqrt{d_k}} (\mathbf{X}\mathbf{W}^Q \mathbf{W}^{\mathbf{K}\top} \otimes \mathbf{I}_n) \mathbf{K}_{dn} \,.$$

Next, we will use a property of commutation matrix to make things simpler (Theorem 7.9, Magnus and Neudecker [2019]):

$$\frac{\partial \mathbf{A}}{\partial \mathbf{X}} = \frac{1}{\sqrt{d_k}} \mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^K \mathbf{W}^{\mathbf{Q}\top} + \frac{1}{\sqrt{d_k}} \mathbf{K}_{nn} (\mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^Q \mathbf{W}^{\mathbf{K}\top}).$$

Plugging this into the above Eq. (29), we get:

$$\frac{\partial \mathbf{S}}{\partial \mathbf{X}} = \frac{1}{\sqrt{d_k}} (\mathbf{I}_n \otimes \mathbf{W}^{\mathbf{V}\top} \mathbf{X}^\top) \frac{\partial \mathbf{A}}{\partial \mathbf{M}} \left(\mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^K \mathbf{W}^{\mathbf{Q}\top} + \mathbf{K}_{nn} (\mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^Q \mathbf{W}^{\mathbf{K}\top}) \right) + \mathbf{A} \otimes \mathbf{W}^{V\top}.$$

As a sanity check, we can calculate if the shapes of the matrices are consistent. LHS should be a $nd \times nd$ matrix, while the constituent matrices of the first term on RHS: $\mathbf{I}_n \otimes \mathbf{W}^{V\top} \mathbf{X}^\top \in \mathbb{R}^{nd \times n^2}$, $\frac{\partial \mathbf{A}}{\partial \mathbf{M}} \in \mathbb{R}^{n^2 \times n^2}$, the additive term next to it is a $n^2 \times nd$ matrix, and the second term on RHS is a Kronecker product of a $n \times n$ and a $d \times d$ matrix. □

Lemma 3.1. *Let $\mathbf{X}^\ell$ be the representations of the input sequence at the ℓ-th layer. Under the uniform-attention assumption, we have*

$$\mathbb{E} \left\| \frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^{V,\ell}} \right\|_F^2 = d_v n \mathbb{E} \|\bar{\boldsymbol{x}}^\ell\|^2 \,; \tag{6}$$

$$\mathbb{E} \left\| \frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 = \frac{\sigma_v^2 \sigma_k^2 d_v}{n^2} \cdot \mathbb{E} \left[\|\mathbf{X}^\ell\|_F^2 \cdot \|(\mathbf{X}^\ell)^\top \mathbf{X}^\ell - n \bar{\boldsymbol{x}}^\ell (\bar{\boldsymbol{x}}^\ell)^\top\|_F^2 \right] \,; \tag{7}$$

$$\mathbb{E} \left\| \frac{\partial \mathbf{S}^\ell}{\partial \mathbf{X}^\ell} \right\|_F^2 \leq \frac{8 \sigma_q^2 \sigma_k^2 \sigma_v^2 d_k d_v}{n} \cdot \mathbb{E} \left\| (\mathbf{X}^\ell)^\top \mathbf{X}^\ell - n \bar{\boldsymbol{x}}^\ell (\bar{\boldsymbol{x}}^\ell)^\top \right\|_F^2 + 2 d_v^2 \sigma_v^2 \,. \tag{8}$$

Proof. Here, we suppress the index ℓ.

Gradient with respect to the values matrix.

Recall that from Lemma A.2 we have that:

$$\frac{\partial \mathbf{S}}{\partial \mathbf{W}^V} = \mathrm{softmax} \left(\frac{\mathbf{X}\mathbf{W}^Q \mathbf{W}^{K\top} \mathbf{X}^\top}{\sqrt{d_k}} \right) \mathbf{X} \otimes \mathbf{I}_{d_v} \overset{\text{Ass. 3.1}}{=} \frac{1}{n} \mathbf{1}_{n \times n} \mathbf{X} \otimes \mathbf{I}_{d_v} \,.$$

By direct computation:

$$\begin{aligned}
\left\| \frac{\partial \mathbf{S}}{\partial \mathbf{W}^V} \right\|_F^2 = \mathrm{tr} \left(\frac{\partial \mathbf{S}}{\partial \mathbf{W}^V} \frac{\partial \mathbf{S}}{\partial \mathbf{W}^V}^\top \right) &\overset{(22)}{=} \frac{1}{n^2} \mathrm{tr}((\mathbf{1}_{n \times n} \mathbf{X}\mathbf{X}^\top \mathbf{1}_{n \times n}) \otimes \mathbf{I}_{d_v}) \\
&\overset{(21)}{=} \frac{1}{n^2} \mathrm{tr}(\mathbf{1}_{n \times n} \mathbf{X}\mathbf{X}^\top \mathbf{1}_{n \times n}) \, \mathrm{tr}(\mathbf{I}_{d_v}) \\
&= \frac{d}{n^2} \mathrm{tr}(\mathbf{X}\mathbf{X}^\top \mathbf{1}_{n \times n} \mathbf{1}_{n \times n}) \\
&= \frac{d_v}{n} \mathrm{tr}(\mathbf{X}\mathbf{X}^\top \mathbf{1}_{n \times n}) \\
&= \frac{d_v}{n} \mathrm{tr}(\mathbf{X}\mathbf{X}^\top \mathbf{1}_n \mathbf{1}_n^\top) \\
&= \frac{d_v}{n} \mathbf{1}_n^\top \mathbf{X}\mathbf{X}^\top \mathbf{1}_n \\
&= \frac{d_v}{n} \|\mathbf{X}^\top \mathbf{1}_n\|^2 = d_v n \, \|\bar{\boldsymbol{x}}\|^2 \,.
\end{aligned}$$

Gradients with respect to the queries/keys matrix.

First, recall the expression for the gradient of the softmax under the uniform-attention assumption (Eq. (24)):

$$\frac{\partial \mathbf{A}}{\partial \mathbf{M}} = \frac{1}{n}\mathbf{I}_n \otimes \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right).$$

Hence, we can rewrite the expression of Lemma A.2 for the gradients of the queries as:

$$\begin{aligned}
\frac{\partial \mathbf{S}}{\partial \mathbf{W}^Q} &= \left(\mathbf{I}_n \otimes \mathbf{W}^{V\top}\mathbf{X}^\top\right)\frac{\partial \mathbf{A}}{\partial \mathbf{M}}\left(\frac{\mathbf{X}\otimes \mathbf{X}\mathbf{W}^K}{\sqrt{d_k}}\right)\\
&= \frac{1}{\sqrt{d_k}n}\left(\mathbf{I}_n \otimes \mathbf{W}^{V\top}\mathbf{X}^\top\right)\left[\mathbf{I}_n \otimes \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right)\right]\left(\mathbf{X}\otimes \mathbf{X}\mathbf{W}^K\right)\\
&= \frac{1}{\sqrt{d_k}n}\mathbf{X}\otimes\left[\mathbf{W}^{V\top}\mathbf{X}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right)\mathbf{X}\mathbf{W}^K\right],
\end{aligned}$$

where in the last step we have used twice the property of the Kronecker product in Eq. (22) of Lemma A.1.

Hence,

$$\left\|\frac{\partial \mathbf{S}}{\partial \mathbf{W}^Q}\right\|_F^2 = \operatorname{tr}\left(\frac{\partial \mathbf{S}}{\partial \mathbf{W}^Q}\frac{\partial \mathbf{S}}{\partial \mathbf{W}^Q}^T\right) \overset{(21)}{=} \frac{1}{d_k n^2}\|\mathbf{X}\|_F^2 \cdot \left\|\mathbf{W}^{V\top}\mathbf{X}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right)\mathbf{X}\mathbf{W}^K\right\|_F^2,$$

where we have used the property on the trace of the Kronecker product (Lemma A.1, Eq. (21)). Note that if we are conditioning on $\mathbf{X}$, then we only have to take the expectation of the last term with respect to the weights $\mathbf{W}^K$ and $\mathbf{W}^V$. Let us call $\mathbf{L} := \mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}$ for notation simplicity.

Note: for a matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$ whose entries $w_{ij} \sim \mathcal{N}(0,\sigma^2)$, then $\mathbb{E}\mathbf{W}\mathbf{W}^\top = d\sigma^2\,\mathbf{I}_d$. Thus, exchanging the order of trace and expectation, we can write:

$$\begin{aligned}
\mathbb{E}\|\mathbf{W}^{\mathbf{V}\top}\mathbf{X}^\top\mathbf{L}\mathbf{X}\mathbf{W}^K\|_F^2 &= \mathbb{E}\operatorname{tr}(\mathbf{W}^{\mathbf{V}\top}\mathbf{X}^\top\mathbf{L}\mathbf{X}\mathbf{W}^K \cdot \mathbf{W}^{\mathbf{K}\top}\mathbf{X}^\top\mathbf{L}\mathbf{X}\mathbf{W}^V)\\
&= \operatorname{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}\mathbb{E}[\mathbf{W}^K\mathbf{W}^{\mathbf{K}\top}]\mathbf{X}^\top\mathbf{L}\mathbf{X}\mathbb{E}[\mathbf{W}^V\mathbf{W}^{\mathbf{V}\top}])\\
&= \sigma_v^2\sigma_k^2 d_k d_v \operatorname{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}\cdot\mathbf{X}^\top\mathbf{L}\mathbf{X})\\
&= \sigma_v^2\sigma_k^2 d_k d_v \|\mathbf{X}^\top\mathbf{L}\mathbf{X}\|_F^2\\
&= \sigma_v^2\sigma_k^2 d_k d_v \left\|\mathbf{X}^\top(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top)\mathbf{X}\right\|_F^2\\
&= \sigma_v^2\sigma_k^2 d_k d_v \|\mathbf{X}^\top\mathbf{X} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\|_F^2,
\end{aligned}$$

where, $\bar{\boldsymbol{x}} = \frac{1}{n}\mathbf{X}^\top\mathbf{1}_n \in \mathbb{R}^d$ is the mean embedding. Multiply this by $\frac{1}{d_k n^2}\|\mathbf{X}\|_F^2$ to get the final answer.

Gradient with respect to the input.

Plugging in the values of $\frac{\partial \mathbf{A}}{\partial \mathbf{M}}$ and $\mathbf{A}$ under the uniform-attention assumption into Eq. (25) gives rise to the following:

$$\begin{aligned}
\frac{\partial \mathbf{S}}{\partial \mathbf{X}} &= \frac{1}{n\sqrt{d_k}}\mathbf{I}_n \otimes \mathbf{W}^{\mathbf{V}\top}\mathbf{X}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)\mathbf{X}\mathbf{W}^K\mathbf{W}^{\mathbf{Q}\top}\\
&+ \frac{1}{n\sqrt{d_k}}\left[\mathbf{I}_n \otimes \mathbf{W}^{\mathbf{V}\top}\mathbf{X}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)\right]\mathbf{K}_{nn}(\mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^Q\mathbf{W}^{\mathbf{K}\top})\\
&+ \frac{1}{n}\mathbf{1}_{n\times n} \otimes \mathbf{W}^{\mathbf{V}\top}
\end{aligned}$$

Let's refer to the matrices on the right-hand side as $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ respectively. We compute the expected squared Frobenius norm of these as follows:

For $\mathbf{A}_3$:

$$\begin{aligned}
\mathbb{E}[\|\mathbf{A}_3\|_F^2] &= \frac{1}{n^2}\mathbb{E}[\operatorname{tr}(n\mathbf{1}_{n\times n} \otimes \mathbf{W}^{\mathbf{V}\top}\mathbf{W}^V)]\\
&\overset{(21)}{=} \frac{1}{n}\operatorname{tr}(\mathbf{1}_{n\times n})\operatorname{tr}(\mathbb{E}[\mathbf{W}^V\mathbf{W}^{\mathbf{V}\top}]) = d_v^2\sigma_v^2.
\end{aligned}$$

Similarly, for $\mathbf{A}_1$:

$$\mathbb{E}[\|\mathbf{A}_1\|_F^2] = \frac{\sigma_q^2\sigma_k^2\sigma_v^2 d_k d_v}{n^2}\ \mathrm{tr}(\mathbf{I}_n)\,\mathrm{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}\ \cdot\ \mathbf{X}^\top\mathbf{L}\mathbf{X}) \tag{30}$$

$$= \frac{\sigma_q^2\sigma_k^2\sigma_v^2 d_k d_v}{n}\ \left\|\mathbf{X}^\top\mathbf{X} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\right\|_F^2 \tag{31}$$

$$= \sigma_q^2\sigma_k^2\sigma_v^2 d_k d_v n\ \left\|\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\right\|_F^2. \tag{32}$$

Finally, for $\mathbf{A}_2$:

$$\mathbb{E}[\|\mathbf{A}_2\|_F^2] \tag{33}$$

$$= \frac{1}{n^2 d_k}\mathbb{E}\left[\mathrm{tr}(\left[\mathbf{I}_n \otimes \mathbf{W}^{\mathbf{V}\top}\mathbf{X}^\top\mathbf{L}\right]\mathbf{K}_{nn}(\mathbf{I}_n \otimes \mathbf{X}\mathbf{W}^Q\mathbf{W}^{\mathbf{K}\top}\mathbf{W}^K\mathbf{W}^{\mathbf{Q}\top}\mathbf{X}^\top)\mathbf{K}_{nn}\left[\mathbf{I}_n \otimes \mathbf{L}\mathbf{X}\mathbf{W}^V\right])\right] \tag{34}$$

$$= \frac{\sigma_q^2\sigma_k^2\sigma_v^2 d_k d_v}{n^2}\ \mathrm{tr}((\mathbf{I}_n \otimes \mathbf{X}\mathbf{X}^\top)[\mathbf{I}_n \otimes \mathbf{L}\mathbf{X}\mathbf{X}^\top\mathbf{L}]) \tag{35}$$

$$= \frac{\sigma_q^2\sigma_k^2\sigma_v^2 d_k d_v}{n}\ \mathrm{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}\ \cdot\ \mathbf{X}^\top\mathbf{L}\mathbf{X}) = \mathbb{E}[\|\mathbf{A}_1\|_F^2], \tag{36}$$

where in the second line we have taken the expectation inside and used the fact that $\mathbf{K}_{nn}$, being a commutation matrix, is orthogonal. Then, by simple properties of Kronecker product and cyclic property of trace, we have the result, which is the same as that for $\mathbf{A}_1$.

Finally, by the triangle inequality

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}}{\partial \mathbf{X}}\right\|^2 \leq 2\mathbb{E}\|\mathbf{A}_1 + \mathbf{A}_2\|^2 + 2\mathbb{E}\|\mathbf{A}_3\|^2 \tag{37}$$

$$\leq 4\mathbb{E}\|\mathbf{A}_1\|^2 + 4\mathbb{E}\|\mathbf{A}_2\|^2 + 2\mathbb{E}\|\mathbf{A}_3\|^2 \tag{38}$$

$$= 8\mathbb{E}\|\mathbf{A}_1\|^2 + 2\mathbb{E}\|\mathbf{A}_3\|^2 \tag{39}$$

$$= \frac{8\sigma_q^2\sigma_k^2\sigma_v^2 d_k d_v}{n}\ \left\|\mathbf{X}^\top\mathbf{X} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\right\|_F^2 + 2d_v^2\sigma_v^2. \tag{40}$$

This completes the proof. □

A.1.3 Proof of Theorem 3.2

Theorem 3.2 (Vanishing gradients under rank collapse)**.** *Suppose that the uniform-attention assumption holds. If additionally $\mathbf{X}^\ell$ for any $l \in [L]$ has rank-1, and there exists a vector $\boldsymbol{x} \in \mathbb{R}^d$ such that $\mathbf{X}^\ell = \mathbf{1}_n \boldsymbol{x}^T$, then:*

$$\mathbb{E}\left\|\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{Q,\ell}}\right\|_F^2 = 0, \quad \mathbb{E}\left\|\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{K,\ell}}\right\|_F^2 = 0, \tag{9}$$

where the expectation is taken over the weight matrices. This implies that these quantities are vanishing almost surely, due to the non-negativeness of the norm.

Before starting the proof, it is interesting to note that, even though the gradients of queries and keys vanish in the rank collapse regime (i.e. $\left\|\mathbf{X}^\top\mathbf{X} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\right\| = 0$), the gradient with respect to the values and the input does not (see Theorem 3.1). From this simple remark, we can conclude that, even in the rank collapse regime, information still propagates in the backward pass. In Section 3.4 (main paper), we show that even if gradients effectively propagate, the phenomenon studied in this theorem still greatly affects training.

Proof. By using the chain rule and the fact that for two matrixes $\mathbf{A}, \mathbf{B}$ we have that $\|\mathbf{AB}\|_F^2 \leq \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2$, we can upper bound the gradient as:

$$\begin{aligned}
\left\| \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 &\leq \prod_{i=\ell+1}^{L-1} \left\| \frac{\partial \mathbf{X}^{i+1}}{\partial \mathbf{X}^i} \right\|_F^2 \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{X}^L} \right\|_F^2 \left\| \frac{\partial \mathbf{X}^{\ell+1}}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 \\
&\leq \prod_{i=\ell+1}^{L-1} \left\| \frac{\partial \mathbf{X}^{i+1}}{\partial \mathbf{X}^i} \right\|_F^2 \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{X}^L} \right\|_F^2 \left\| \frac{\partial \mathbf{X}^{\ell+1}}{\partial \mathbf{Z}^\ell} \right\|_F^2 \left\| \frac{\partial \mathbf{Z}^\ell}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 \\
&\leq \prod_{i=\ell+1}^{L-1} \left\| \frac{\partial \mathbf{X}^{i+1}}{\partial \mathbf{X}^i} \right\|_F^2 \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{X}^L} \right\|_F^2 \left\| \frac{\partial \mathbf{X}^{\ell+1}}{\partial \mathbf{Z}^\ell} \right\|_F^2 \left(\left\| \frac{\partial \alpha_1 \mathbf{S}^\ell}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 + \underbrace{\left\| \frac{\partial \mathbf{X}^\ell}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2}_{=0} \right),
\end{aligned}$$

where we recall that $\mathbf{Z}^\ell = \alpha_1 \mathbf{S}^\ell + \mathbf{X}^\ell$ and in the last step we have used that $\mathbf{X}^\ell$ does not depend on $\mathbf{W}^{Q,\ell}$, hence the gradient vanishes. By taking expectation and using the tower property, we have that:

$$\mathbb{E} \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 \leq \mathbb{E} \left[\underbrace{\mathbb{E} \left[\prod_{i=\ell+1}^{L-1} \left\| \frac{\partial \mathbf{X}^{i+1}}{\partial \mathbf{X}^i} \right\|_F^2 \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{X}^L} \right\|_F^2 \left\| \frac{\partial \mathbf{X}^{\ell+1}}{\partial \mathbf{Z}^\ell} \right\|_F^2 \right]}_{=:G(\mathbf{X}^\ell)} \left\| \frac{\partial \alpha_1 \mathbf{S}^\ell}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 \right],$$

where the expectations are taken with respect to $\mathbf{X}^\ell$ for the outer one and conditioning on $\mathbf{X}^\ell$ for inner one. Indeed, the first three terms only depend on the network values after $\mathbf{X}^\ell$. Now, a repeated application of the tower property in $G(\mathbf{X}^\ell)$, together with the results on the gradients of Lemma 3.1, easily shows that $G(\mathbf{X}^\ell)$ stays bounded under our hypothesis. To see this one can also simply note that, since the softmax and its derivatives are almost surely bounded, the boundedness of $G(\mathbf{X}^\ell)$ is implied by an analogous statement for a vanilla linear MLP (i.e removing the softmax). In this setting, the random variable inside the expectation in $G(\mathbf{X}^\ell)$ is a finite linear combination of Gaussian products — which has bounded expectation.

All in all, we have that

$$\mathbb{E} \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 \leq \mathbb{E} \left[B_{\mathbf{X}^\ell} \left\| \frac{\partial \alpha_1 \mathbf{S}^\ell}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 \right],$$

where $B_{\mathbf{X}^\ell}$ is an almost-surely-bounded function of $\mathbf{X}^\ell$. Hence, to show that $\mathbb{E} \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 = 0$, we now just need to show that:

$$\mathbb{E} \left\| \frac{\partial \alpha_1 \mathbf{S}^\ell}{\partial \mathbf{W}^{Q,\ell}} \right\|_F^2 = 0$$

under the rank-1 hypothesis for $\mathbf{X}^\ell$. Let $\mathbf{X}_1^\ell, \ldots \mathbf{X}_n^\ell \in \mathbb{R}^{d_v}$ be the representations for the n tokens. Under the rank-1 assumption, each token can be written as a multiple of a single vector $\boldsymbol{x} \in \mathbb{R}^{d_v}$, and hence there exists $a_1, \ldots, a_n \in \mathbb{R}$ such that $\mathbf{X}_1 = a_1 \boldsymbol{x}, \ldots, \mathbf{X}_n = a_n \boldsymbol{x}$. From Lemma 3.1, we know that:

$$\mathbb{E} \left\| \frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^Q} \right\|_F^2 = \frac{\sigma_v^2 \sigma_k^2 d^2}{n^2} \cdot \mathbb{E} \left[\|\mathbf{X}^\ell\|_F^2 \cdot \|(\mathbf{X}^\ell)^\top \mathbf{X}^\ell - n \bar{\boldsymbol{x}}^\ell (\bar{\boldsymbol{x}}^\ell)^\top\|_F^2 \right].$$

The mean token simplifies to $\bar{\boldsymbol{x}}^l = \frac{\boldsymbol{x}}{n} \sum_k a_k$ and hence $\left(\bar{\boldsymbol{x}}^\ell (\bar{\boldsymbol{x}}^\ell)^\top\right)_{ij} = \frac{1}{n^2} (\sum_k a_k)^2 x_i x_j$. Similarly, $\left((\mathbf{X}^\ell)^\top \mathbf{X}^\ell\right)_{ij} = \sum_k a_k^2 x_i x_j$. If furthermore all the coefficients a_i are the same (which corresponds to the rank collapse assumption $\mathbf{X}^\ell = \mathbf{1}_n \boldsymbol{x}^T$ analyzed here), then it is easy to see that $\left((\mathbf{X}^\ell)^\top \mathbf{X}^\ell\right)_{ij} - n \left(\bar{\boldsymbol{x}}^\ell (\bar{\boldsymbol{x}}^\ell)^\top\right)_{ij} = 0 \; \forall i, j$ and hence $\|(\mathbf{X}^\ell)^\top \mathbf{X}^\ell - n \bar{\boldsymbol{x}}^\ell (\bar{\boldsymbol{x}}^\ell)^\top\|_F^2 = 0$.

□

A.2 Gradient Analysis of Section 3.3

Throughout this section we assume that between every pair of tokens, the same dimension is a zero-mean Gaussian random variable with the same correlation, meaning that

$$\mathbb{E}\left[\mathbf{X}_{i,j}\mathbf{X}_{i',j'}\right] = \begin{cases} 0 & j \neq j' \text{ (independent dimensions)} \\ \sigma_x^2 & i = i', j = j' \\ \rho\sigma_x^2 & i \neq i', j = j'. \end{cases}$$

As we will deal with the computation of 4-th order moments of correlated Gaussian random variables, we will make use of Isserlis theorem [Isserlis, 1918]:

Theorem A.1 (Isserlis). *Let $X_1, \dots X_m$ be m zero-mean Gaussian random variables. Then:*

$$\mathbb{E}[X_1 \cdots X_m] = \begin{cases} \sum_{p \in P_m^2} \prod_{(i,j)\in p} \mathbb{E}[X_i X_j] & m \text{ even} \\ 0 & m \text{ odd} \end{cases} \tag{41}$$

where P_m^2 is the set of all the possible pairings of the indexes $1, \dots, m$.

In particular, we will only need the 4-th order term, which reads:

$$\mathbb{E}[X_1X_2X_3X_4] = \mathbb{E}[X_1X_2]\mathbb{E}[X_3X_4] + \mathbb{E}[X_1X_3]\mathbb{E}[X_2X_4] + \mathbb{E}[X_1X_4]\mathbb{E}[X_2X_3].$$

Now we can prove Eq. (17) , which we re-state here:

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}}{\partial \mathbf{W}^V}\right\|_F^2 = \sigma_x^2 d^2 \left(1 + \rho(n-1)\right).$$

Also, from Eq. (6) we have that

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^{V,\ell}}\right\|_F^2 = dn\mathbb{E}\|\bar{\boldsymbol{x}}^\ell\|^2.$$

Now,

$$\mathbb{E}\|\bar{\boldsymbol{x}}^\ell\|^2 = \mathbb{E}\left(\sum_{i=1}^d (\bar{\boldsymbol{x}}_i^\ell)^2\right).$$

Each $\bar{\boldsymbol{x}}_i^\ell = \frac{1}{n}\sum_{k=1}^n \mathbf{X}_{ki}^\ell$ is equally distributed with mean

$$\mathbb{E}[\bar{\boldsymbol{x}}_i^\ell] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \mathbf{X}_{ki}^\ell\right] = 0$$

and variance

$$\text{Var}[\bar{\boldsymbol{x}}_i^\ell] = \text{Var}\left[\frac{1}{n}\sum_{k=1}^n \mathbf{X}_{ki}^\ell\right] = \frac{1}{n^2}\left(n\sigma_x^2 + n(n-1)\rho\sigma_x^2\right) = \frac{1}{n}\sigma_x^2(1+\rho(n-1)).$$

Finally we get

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^{V,\ell}}\right\|_F^2 = \sigma_x^2 d^2(1+\rho(n-1)).$$

We know prove Eq. (18), which reads:

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}}{\partial \mathbf{W}^Q}\right\|_F^2 = \sigma_x^6 \frac{(n-1)}{n}(1-\rho)^2 d(n+d).$$

For the queries (and the keys respectively), recall from Eq. (7) that

$$\mathbb{E}\left\|\frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^{Q,\ell}}\right\|_F^2 = \frac{\sigma_v^2\sigma_k^2 d^2}{dn^2}\cdot\mathbb{E}\left[\|\mathbf{X}^\ell\|_F^2\cdot\|(\mathbf{X}^\ell)^\top\mathbf{X}^\ell - n\bar{\boldsymbol{x}}^\ell(\bar{\boldsymbol{x}}^\ell)^\top\|_F^2\right].$$

To proceed, we drop the superscript ℓ and we make the additional assumption that $\|\mathbf{X}\|_F^2$ is uncorrelated from the correlation magnitude $\|\mathbf{X}^\top\mathbf{L}\mathbf{X}\|_F^2 = \|\mathbf{X}^\top\mathbf{X} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\|_F^2$.

Let us proceed with an expansion:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X}^\top\mathbf{L}\mathbf{X}\|_F^2\right] &= \sum_{i,j=1}^{d}\mathbb{E}\left[\left(\sum_{a,b=1}^{n}\mathbf{X}_{ai}\mathbf{L}_{ab}\mathbf{X}_{bj}\right)^2\right]\\ &= \sum_{i,j=1}^{d}\sum_{a,b,a',b'=1}^{n}\mathbf{L}_{ab}\mathbf{L}_{a'b'}\mathbb{E}\left[\mathbf{X}_{ai}\mathbf{X}_{a'i}\mathbf{X}_{bj}\mathbf{X}_{b'j}\right].\end{aligned}$$

Now, we have 2 cases: if $i \neq j$, which gives $d(d-1)$ equal terms, we need to compute

$$\mathrm{A} := \sum_{a,b,a',b'=1}^{n}\mathbf{L}_{ab}\mathbf{L}_{a'b'}\mathbb{E}\left[\mathbf{X}_{ai}\mathbf{X}_{a'i}\right]\cdot\mathbb{E}\left[\mathbf{X}_{bj}\mathbf{X}_{b'j}\right],$$

where (i,j) is any tuple with $i \neq j$ and we used uncorrelation of different dimensions. Otherwise, we get d each equal to

$$\mathrm{B} := \sum_{a,b,a',b'=1}^{n}\mathbf{L}_{ab}\mathbf{L}_{a'b'}\mathbb{E}\left[\mathbf{X}_{ai}\mathbf{X}_{a'i}\mathbf{X}_{bi}\mathbf{X}_{b'i}\right],$$

where i is any index.

Term A. Note that

$$\mathbb{E}\left[\mathbf{X}_{ai}\mathbf{X}_{a'i}\right]\cdot\mathbb{E}\left[\mathbf{X}_{bj}\mathbf{X}_{b'j}\right] = \begin{cases}\sigma_x^4 & a=a', b=b', \quad n^2 \text{ terms}\\ \rho\sigma_x^4 & a=a', b\neq b', \quad n^2(n-1) \text{ terms}\\ \rho\sigma_x^4 & a\neq a', b=b', \quad n^2(n-1) \text{ terms}\\ \rho^2\sigma_x^4 & a\neq a', b\neq b', \quad n^2(n-1)^2 \text{ terms}\end{cases}.$$

So basically A is the sum of 3 terms:

$$\begin{aligned}A_1 &:= \sigma_x^4\sum_{a,b}\mathbf{L}_{ab}^2 = (n-1)\sigma_x^4\\ A_2 &:= 2\rho\sigma_x^4\sum_{a,b}\sum_{b'\neq b}\mathbf{L}_{ab}\mathbf{L}_{ab'} = -2(n-1)\rho\sigma_x^4\\ A_3 &:= \rho^2\sigma_x^4\sum_{a,b}\sum_{a'\neq a}\sum_{b'\neq b}\mathbf{L}_{ab}\mathbf{L}_{a'b'} = \rho^2\sigma_x^4(n-1),\end{aligned}$$

where we leveraged the following direct calculations:

$$\begin{aligned}A_1 &= \sigma_x^4\sum_{a,b}\mathbf{L}_{ab}^2\\ &= \sigma_x^4\left(n\left(\frac{n-1}{n}\right)^2 + (n-1)n\frac{1}{n^2}\right)\\ &= \sigma_x^4\frac{(n-1)^2+(n-1)}{n}\\ &= \sigma_x^4(n-1).\end{aligned}$$

Next, we compute

$$
\begin{aligned}
A_2 &= 2\rho\sigma_x^4 \sum_{a,b} \mathbf{L}_{ab} \sum_{b'\neq b} \mathbf{L}_{ab'} \\
&= 2\rho\sigma_x^4 \left(\sum_a \mathbf{L}_{a,a} \sum_{b'\neq b} \mathbf{L}_{ab'} + \sum_a \sum_{b\neq a} \mathbf{L}_{ab} \sum_{b'\neq b} \mathbf{L}_{ab'} \right) \\
&= 2\rho\sigma_x^4 \left(\sum_a \frac{n-1}{n} \left[-(n-1)\frac{1}{n} \right] + \sum_a \sum_{b\neq a} \mathbf{L}_{ab} \left[\frac{n-1}{n} - \frac{n-2}{n} \right] \right) \\
&= 2\rho\sigma_x^4 \left(-\frac{(n-1)^2}{n} - \frac{1}{n}(n-1)n\frac{1}{n} \right) \\
&= -2\rho\sigma_x^4 (n-1).
\end{aligned}
$$

Finally, similar computations also lead to the last term. To follow the calculations, we invite the reader to draw the matrix $\mathbf{L}$ and to hide the columns over which summations are not performed:

$$
\begin{aligned}
A_3 &= \rho^2\sigma_x^4 \sum_{a,b} \mathbf{L}_{ab} \sum_{a'\neq a} \sum_{b'\neq b} \mathbf{L}_{a'b'} \\
&= \rho^2\sigma_x^4 \left(\sum_a \mathbf{L}_{aa} \sum_{a'\neq a} \sum_{b'\neq a} \mathbf{L}_{a'b'} + \sum_a \sum_{b\neq a} \mathbf{L}_{ab} \sum_{a'\neq a} \sum_{b'\neq b} \mathbf{L}_{a'b'} \right) \\
&= \rho^2\sigma_x^4 \left(\sum_a \mathbf{L}_{aa} (1 - \frac{1}{n} + \sum_a \sum_{b\neq a} \mathbf{L}_{ab} (-1\frac{1}{n}) \right) \\
&= \rho^2\sigma_x^4 \left(n(1-\frac{1}{n})(1-\frac{1}{n}) + n(n-1)(-\frac{1}{n})(-\frac{1}{n}) \right) \\
&= \rho^2\sigma_x^4 (n-1).
\end{aligned}
$$

All in all, we get:

$$\mathrm{A} = (n-1)(1-\rho)^2\sigma_x^4.$$

Term B. We make use of Isserlis theorem, stating that:

$$\mathbb{E}\left[\mathbf{X}_{ai}\mathbf{X}_{a'i}\mathbf{X}_{bi}\mathbf{X}_{b'i}\right] = \underbrace{\mathbb{E}\mathbf{X}_{ai}\mathbf{X}_{a'i}\mathbb{E}\mathbf{X}_{bi}\mathbf{X}_{b'i}}_{Q_1} + \underbrace{\mathbb{E}\mathbf{X}_{ai}\mathbf{X}_{bi}\mathbb{E}\mathbf{X}_{a'i}\mathbf{X}_{b'i}}_{Q_2} + \underbrace{\mathbb{E}\mathbf{X}_{ai}\mathbf{X}_{b'i}\mathbb{E}\mathbf{X}_{a'i}\mathbf{X}_{bi}}_{Q_3}.$$

By using our independence assumptions, we get:

$Q_1 = \sigma_x^4(\delta_{aa'} + \rho\delta_{a\neq a'})(\delta_{bb'} + \rho\delta_{b\neq b'}) = \sigma_x^4(\delta_{aa'}\delta_{bb'} + \rho\delta_{aa'}\delta_{b\neq b'} + \rho\delta_{a\neq a'}\delta_{bb'} + \rho^2\delta_{a\neq a'}\delta_{b\neq b'})$.

Similarly for Q_2 and Q_3:

$$Q_2 = \sigma_x^4(\delta_{ab}\delta_{a'b'} + \rho\delta_{ab}\delta_{a'\neq b'} + \rho\delta_{a\neq b}\delta_{a'b'} + \rho^2\delta_{a\neq b}\delta_{a'\neq b'})$$

and

$$Q_3 = \sigma_x^4(\delta_{ab'}\delta_{a'b} + \rho\delta_{ab'}\delta_{a'\neq b} + \rho\delta_{a\neq b'}\delta_{a'b} + \rho^2\delta_{a\neq b'}\delta_{a'\neq b}).$$

Hence,

$$B = \sum_{a,b,a',b'=1}^{n} \mathbf{L}_{ab}\mathbf{L}_{a'b'}\mathbb{E}\left[\mathbf{X}_{ai}\mathbf{X}_{a'i}\mathbf{X}_{bi}\mathbf{X}_{b'i}\right] = \sum_{a,b,a',b'=1}^{n} \mathbf{L}_{ab}\mathbf{L}_{a'b'}(Q_1 + Q_2 + Q_3).$$

Let's study it term by term. We will also use $\mathbf{L}_{ab} = (\delta_{ab} - \frac{1}{n})$, and so $\mathbf{L}_{ab}\mathbf{L}_{a'b'} = (\delta_{ab}\delta_{a'b'} - \frac{\delta_{ab}}{n} - \frac{\delta_{a'b'}}{n} + \frac{1}{n^2})$.

First term: we have that $\sigma_x^4 \sum_{aa'b'b'} \mathbf{L}_{ab}\mathbf{L}_{a'b'}Q_1$ which is equal to (omitting the constant σ_x^4):

$$
\begin{aligned}
&= \sum_{a,a',b,b'} (\delta_{ab}\delta_{a'b'} - \frac{\delta_{ab}}{n} - \frac{\delta_{a'b'}}{n} + \frac{1}{n^2})(\delta_{aa'}\delta_{bb'} + \rho\delta_{aa'}\delta_{b\neq b'} + \rho\delta_{a\neq a'}\delta_{bb'} + \rho^2\delta_{a\neq a'}\delta_{b\neq b'}) \\
&= \rho^2\left(n(n-1) - 2(n-1)(n-1) + (n-1)^2\right) + \rho(-4(n-1) + 2(n-1)) + n - 2 + 1 \\
&= \rho^2(n-1)\left(n - 2(n-1) + (n-1)\right) - 2\rho(n-1) + (n-1) \\
&= \rho^2(n-1) - 2\rho(n-1) + (n-1).
\end{aligned}
$$

Second term: we have that $\sigma_x^4 \sum_{aa'b'b'} \mathbf{L}_{ab}\mathbf{L}_{a'b'}Q_2$ which is equal to (omitting the constant σ_x^4):

$$
\begin{aligned}
&= \sum_{a,a',b,b'} (\delta_{ab}\delta_{a'b'} - \frac{\delta_{ab}}{n} - \frac{\delta_{a'b'}}{n} + \frac{1}{n^2})(\delta_{ab}\delta_{a'b'} + \rho\delta_{ab}\delta_{a'\neq b'} + \rho\delta_{a\neq b}\delta_{a'b'} + \rho^2\delta_{a\neq b}\delta_{a'\neq b'}) \\
&= \rho^2(n-1)^2 + \rho(-2n(n-1) + 2(n-1)) + n^2 - 2n + 1 \\
&= \rho^2(n-1)^2 - 2\rho(n-1)^2 + (n-1)^2.
\end{aligned}
$$

Third term: we have that $\sigma_x^4 \sum_{aa'b'b'} \mathbf{L}_{ab}\mathbf{L}_{a'b'}Q_3$ which is equal to (omitting the constant σ_x^4):

$$
\begin{aligned}
&= \sum_{a,a',b,b'} (\delta_{ab}\delta_{a'b'} - \frac{\delta_{ab}}{n} - \frac{\delta_{a'b'}}{n} + \frac{1}{n^2})(\delta_{ab'}\delta_{a'b} + \rho\delta_{ab'}\delta_{a'\neq b} + \rho\delta_{a\neq b'}\delta_{a'b} + \rho^2\delta_{a\neq b'}\delta_{a'\neq b}) \\
&= \rho^2\left(n(n-1) - 2(n-1)(n-1) + (n-1)^2\right) + \rho(-4(n-1) + 2(n-1)) + n - 2 + 1 \\
&= \rho^2(n-1) - 2\rho(n-1) + (n-1).
\end{aligned}
$$

Summing all the three terms, we get:

$$B = \sigma_x^4(n-1)\left[\rho^2(n+1) - 2(n+1)\rho + (n+1)\right] = \sigma_x^4(n-1)(n+1)(1-\rho)^2.$$

Plugging in the values of A and B we get:

$$\mathbb{E}[\|\mathbf{X}^\top\mathbf{L}\mathbf{X}\|_F^2] = d\cdot B + d(d-1)\cdot A = \sigma_x^4(1-\rho)^2 d(n-1)(n+d),$$

and finally assuming Xavier initialization

$$
\begin{aligned}
\mathbb{E}\left\|\frac{\partial \mathbf{S}^\ell}{\partial \mathbf{W}^{Q,\ell}}\right\|_F^2 &= \frac{\sigma_v^2\sigma_k^2 d^2}{dn^2}\cdot\mathbb{E}\left[\|\mathbf{X}^\ell\|_F^2\cdot\|(\mathbf{X}^\ell)^\top\mathbf{X}^\ell - n\bar{\boldsymbol{x}}^\ell(\bar{\boldsymbol{x}}^\ell)^\top\|_F^2\right] \\
&= \sigma_x^6\frac{n-1}{n}(1-\rho)^2 d(n+d).
\end{aligned}
$$

A.3 Forward Pass: Proofs of Lemma 3.2 and 3.3

First, we characterize the evolution of the correlations between tokens $\mathbf{X}_k, \mathbf{X}_{k'}$ with depth, under the assumptions of Theorem 3.3, namely uniform-attention assumption, and the adoption of a linear activation.

Lemma A.4 (Expectation of Linear Layers)**.** *Let $\mathbf{D} = \mathbf{X}\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{d\times d}$ is a random matrix with i.i.d random entries with variance $\sigma^2 = \frac{1}{d}$ and $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a fixed matrix:*

$$\mathbb{E}[\mathbf{D}_{kj}\mathbf{D}_{k'j}] = \frac{1}{d}\langle \mathbf{X}_k, \mathbf{X}_{k'}\rangle$$

Note that by summing over the indexes, Lemma A.4 implies:

$$
\begin{aligned}
\mathbb{E}\,\|\mathbf{D}\|_F^2 &= \mathbb{E}\,\|\mathbf{X}\|_F^2 \\
\mathbb{E}C(\mathbf{D}) &= \mathbb{E}C(\mathbf{X}).
\end{aligned}
$$

Proof.

$$\mathbb{E}[\mathbf{D}_{kj}\mathbf{D}_{k'j}] = \sum_{zz'}\mathbf{X}_{kz}\mathbf{X}_{k'z'}\mathbb{E}[\mathbf{W}_{zj}\mathbf{W}_{z'j}] = \sigma^2\sum_z \mathbf{X}_{kz}\mathbf{X}_{k'z} = \frac{1}{d}\langle \mathbf{X}_k, \mathbf{X}_{k'}\rangle.$$

□

Lemma A.5 (Expectation of skip connection)**.** *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times q}$*. Let* $\mathbf{D} := \alpha \mathbf{A} + \mathbf{B}$ *with* $\mathbb{E}[\mathbf{A}|\mathbf{B}] = \mathbf{0}$ *and* $\alpha \in \mathbb{R}$*. Then:*

$$\mathbb{E}\left[\mathbf{D}_{ij}\mathbf{D}_{i'j}\right] = \alpha^2 \mathbb{E}[\mathbf{A}_{ij}\mathbf{A}_{i'j}] + \mathbb{E}[\mathbf{B}_{ij}\mathbf{B}_{ij'}] \tag{42}$$

holds for all $i, i' \in [p], j \in [q]$*.*

Note that by summing over the indexes, Lemma A.5 implies:

$$\mathbb{E}\left\|\mathbf{D}\right\|_F^2 = \alpha^2 \mathbb{E}\left\|\mathbf{A}\right\|_F^2 + \mathbb{E}\left\|\mathbf{B}\right\|_F^2$$
$$\mathbb{E}C(\mathbf{D}) = \alpha^2 \mathbb{E}C(\mathbf{A}) + \mathbb{E}C(\mathbf{B}).$$

Proof.

$$\begin{aligned}\mathbb{E}[\mathbf{D}_{ij}\mathbf{D}_{i'j}] &= \mathbb{E}\left[(\alpha\mathbf{A}_{ij} + \mathbf{B}_{ij})(\alpha\mathbf{A}_{i'j} + \mathbf{B}_{i'j})\right] \\ &= \mathbb{E}\left[\alpha^2\mathbf{A}_{ij}\mathbf{A}_{i'j} + \alpha\mathbf{A}_{ij}\mathbf{B}_{i'j} + \alpha\mathbf{A}_{i'j}\mathbf{B}_{ij} + \mathbf{B}_{ij}\mathbf{B}_{i'j}\right] \\ &= \alpha^2\mathbb{E}\left[\mathbf{A}_{kj}\mathbf{A}_{i'j}\right] + \mathbb{E}\left[\mathbf{B}_{ij}\mathbf{B}_{i',j}\right],\end{aligned}$$

where using iterated expectations $\alpha\mathbb{E}[\mathbf{A}_{i'j}\mathbf{B}_{ij}] = \alpha\mathbb{E}[\mathbb{E}[\mathbf{A}_{i'j}|\mathbf{B}]\mathbf{B}_{ij}]] = 0$ and identically $\alpha\mathbb{E}[\mathbf{A}_{ij}\mathbf{B}_{i'j}] = 0$. □

Lemma A.6 (Expectation of Attention Layers)**.** *Under the uniform-attention assumption:*

$$\mathbb{E}[\mathbf{S}_{kj}\mathbf{S}_{k'j}] = \frac{1}{d_v n^2}\mathbb{E}C(\mathbf{X}).$$

In this case, by summing over the indexes we have that:

$$\mathbb{E}\left\|\mathbf{S}\right\|_F^2 = \frac{\mathbb{E}C(\mathbf{X})}{n}$$
$$\mathbb{E}C(\mathbf{S}) = \mathbb{E}C(\mathbf{X}).$$

Proof. Note that under the uniform-attention assumption:

$$\mathbf{S}_{kj} = \frac{1}{n}\left(\mathbf{1}_{n\times n}\mathbf{X}\mathbf{W}^V\right)_{kj} = \frac{1}{n}\sum_{zi}\mathbf{X}_{zi}\mathbf{W}^V_{ij}.$$

Hence, using the fact that the weights are i.i.d with variance $\sigma_v^2 = \frac{1}{d_v}$:

$$\mathbb{E}[\mathbf{S}_{kj}\mathbf{S}_{k'j}] = \frac{\sigma_v^2}{n^2}\sum_{z,z'}\sum_{i}\mathbb{E}[\mathbf{X}_{zi}\mathbf{X}_{z'i}] = \frac{1}{d_v n^2}\sum_{k,k'}\langle\mathbf{X}_z, \mathbf{X}_{z'}\rangle = \frac{1}{d_v n^2}\mathbb{E}C(\mathbf{X}).$$

□

Lemma 3.2 (Propagation of inner products)**.** *Let* $C(\mathbf{X}^\ell) = \sum_{k,k'}\langle\mathbf{X}^\ell_k, \mathbf{X}^\ell_{k'}\rangle$ *and* $\mathbf{X}$ *the input sequence. Under the Assumption 3.1 and if* σ *is the linear activation function, we have that:*

$$\mathbb{E}\left[C(\mathbf{X}^L)\right] = (\alpha_2^2 + 1)^L(\alpha_1^2 + 1)^L C(\mathbf{X}). \tag{10}$$

hence, under the depth scaling for the residual block parameters $\alpha_1^2 = \frac{\tilde{\alpha}_1}{L}, \alpha_2^2 = \frac{\tilde{\alpha}_2}{L}$ *with* $\tilde{\alpha}_1, \tilde{\alpha}_2 \in \mathbb{R}$ *independent of L, we have that:*

$$\lim_{L\to\infty}\mathbb{E}[C(\mathbf{X}^L)] = e^{\tilde{\alpha}_1 + \tilde{\alpha}_2}C(\mathbf{X}). \tag{11}$$

Proof. First, note that for the residual blocks we have that $\mathbb{E}[Y^\ell_{kj}|Z^\ell_{k'j}] = 0$ due to the independence assumption on the feedforward weights, and similarly $\mathbb{E}[S^\ell_{kj}|X^\ell_{k'j}] = 0$. Hence, we can use Lemma

A.5 in both the skip connections of the Transformer architecture. Therefore, using Lemma A.5 (skip), Lemma A.4 (linear) and Lemma A.6 (attention):

$$
\begin{aligned}
&\mathbb{E}[C(\mathbf{X}^{\ell+1})] \\
&\overset{\text{skip}}{=} \alpha_2^2 \mathbb{E}C(\mathbf{Y}^\ell) + \mathbb{E}C(\mathbf{Z}^\ell) \\
&\overset{\text{linear}}{=} \alpha_2^2 \mathbb{E}C(\mathbf{Z}^\ell) + \mathbb{E}C(\mathbf{Z}^\ell) \\
&= (\alpha_2^2 + 1)\mathbb{E}C(\mathbf{Z}^\ell) \\
&\overset{\text{skip}}{=} (\alpha_2^2 + 1)\left(\alpha_1^2 \mathbb{E}C(\mathbf{S}^\ell) + \mathbb{E}C(\mathbf{X}^\ell)\right) \\
&\overset{\text{attention}}{=} (\alpha_2^2 + 1)(\alpha_1^2 + 1)\mathbb{E}[C(\mathbf{X}^\ell)] \\
&\overset{\text{unroll recurs.}}{=} (\alpha_2^2 + 1)^{\ell+1}(\alpha_1^2 + 1)^{\ell+1} C(\mathbf{X}),
\end{aligned}
$$

where in the last step we have unrolled the recursion until the input layer.

For the limit as $L \to \infty$, simply note that:

$$
\lim_{L\to\infty} \left(\frac{\tilde{\alpha}_i}{L} + 1\right)^L = \mathrm{e}^{\tilde{\alpha}_i},
$$

with $i \in \{1, 2\}$. □

Now we are ready to re-state and prove Lemma 3.3.

Lemma 3.3 (Propagation of the norm)**.** *Let $\mathbf{X}^L$ be the representations of the input sequence at the final layer. Under the assumptions of Lemma 3.2, we have that:*

$$
\mathbb{E}\left\|\mathbf{X}^L\right\|_F^2 = n(\alpha_2^2 + 1)^L \alpha_1^2 \sum_{k=0}^{L-1} (\alpha_1^2 + 1)^k \left\|\bar{\boldsymbol{x}}\right\|^2 + (\alpha_2^2 + 1)^L ||\mathbf{X}||_F^2, \tag{12}
$$

hence, under the depth scaling for the residual block parameters $\alpha_1^2 = \frac{\tilde{\alpha}_1}{L}, \alpha_2^2 = \frac{\tilde{\alpha}_2}{L}$ with $\tilde{\alpha_1}, \tilde{\alpha_2} \in \mathbb{R}$ independent of L, we have that:

$$
\lim_{L\to\infty} \mathbb{E}\left\|\mathbf{X}^L\right\|_F^2 = n e^{\tilde{\alpha}_2}(e^{\tilde{\alpha}_1} - 1)\left\|\bar{\boldsymbol{x}}\right\|^2 + e^{\tilde{\alpha}_2}||\mathbf{X}||_F^2. \tag{13}
$$

Proof. The proof is in the same spirit as Lemma 3.2 but slightly more involved. Again, using Lemma A.5 in both the skip connections of the Transfomer architecture. Therefore, using Lemma A.5 (skip), Lemma A.4 (linear) and Lemma A.6 (attention):

$$
\begin{aligned}
\mathbb{E}[||\mathbf{X}^{\ell+1}||_F^2] &\overset{\text{skip}}{=} \alpha_2^2 \mathbb{E}||\mathbf{Y}^\ell||_F^2 + \mathbb{E}||\mathbf{Z}^\ell||_F^2 \\
&\overset{\text{linear}}{=} (\alpha_2^2 + 1)\mathbb{E}||\mathbf{Z}^\ell||_F^2 \\
&\overset{\text{skip}}{=} (\alpha_2^2 + 1)\left(\alpha_1^2 \mathbb{E}[||\mathbf{S}^\ell||_F^2] + \mathbb{E}[||\mathbf{X}^\ell||_F^2]\right) \\
&\overset{\text{softmax}}{=} (\alpha_2^2 + 1)\left(\frac{\alpha_1^2}{n}\mathbb{E}[C(\mathbf{X}^\ell)] + \mathbb{E}[||\mathbf{X}^\ell||_F^2]\right) \\
&= (\alpha_2^2 + 1)\frac{\alpha_1^2}{n}\mathbb{E}[C(\mathbf{X}^\ell)] + (\alpha_2^2 + 1)\mathbb{E}[||\mathbf{X}^\ell||_F^2] \\
&\overset{\text{unroll } C(\mathbf{X}^\ell)}{=} (\alpha_2^2 + 1)^{\ell+1}\alpha_1^2(\alpha_1^2 + 1)^\ell \frac{C(\mathbf{X})}{n} + (\alpha_2^2 + 1)\mathbb{E}[||\mathbf{X}^\ell||_F^2] \\
&\overset{\text{unroll } \left\|\mathbf{X}^\ell\right\|_F^2}{=} (\alpha_2^2 + 1)^{\ell+1}\alpha_1^2 \frac{C(\mathbf{X})}{n}\sum_{k=0}^{\ell}(\alpha_1^2 + 1)^k + (\alpha_2^2 + 1)^{\ell+1}||\mathbf{X}||_F^2,
\end{aligned}
$$

where in the second to last step we have used Lemma 3.2 and in the last step we have unrolled the recursion for $\left\|\mathbf{X}^\ell\right\|_F^2$ until the input layer.

For the second part, we now show that for a network of L layers, the choice $\alpha_1^2 = \frac{\tilde{\alpha}_1}{L}$ and $\alpha_2^2 = \frac{\tilde{\alpha}_2}{L}$ stabilizes the norm of the activations in the forward pass. Using the product law for the limits, we can study the converges of $(\alpha_2^2+1)^\ell$ and $\alpha_1 \sum_{k=0}^{\ell}(\alpha_1^2+1)^k$ separately.

Let $i \in \{1,2\}$. For the latter term we have that:

$$\begin{aligned}\lim_{L\to\infty} \frac{\tilde{\alpha}_i}{L}\sum_{l=0}^{L-1}\left(1+\frac{\tilde{\alpha}_i}{L}\right)^\ell &= \lim_{L\to\infty} \frac{\tilde{\alpha}_i}{L}\frac{1-\left(1+\frac{\tilde{\alpha}_i}{L}\right)^\ell}{1-1-\frac{\tilde{\alpha}_i}{L}} \\ &= \lim_{L\to\infty} -1+\left(1+\frac{\tilde{\alpha}_i}{L}\right)^\ell \\ &= \mathrm{e}^{\tilde{\alpha}_i}-1,\end{aligned}$$

while for the former term we have that $\lim_{L\to\infty}(\frac{\tilde{\alpha}_i}{L}+1)^\ell = \mathrm{e}^{\tilde{\alpha}_i}$. Hence, the norm of the representations converges to:

$$\lim_{L\to\infty} \mathbb{E}[||\mathbf{X}^\ell||_F^2] = \mathrm{e}^{\tilde{\alpha}_2}(\mathrm{e}^{\tilde{\alpha}_1}-1)\frac{C(\mathbf{X})}{n} + \mathrm{e}^{\tilde{\alpha}_2}||\mathbf{X}||_F^2.$$

The final results as stated in the theorem hold because of the following:

Remark: note that $C(\mathbf{X}) = \sum_{k,k'}\sum_j \mathbf{X}_{kj}\mathbf{X}_{kj'} = \sum_j(\sum_{k,k'}\mathbf{X}_{kj}\mathbf{X}_{k'j}) = \sum_j(\sum_k \mathbf{X}_{kj})^2 = n^2\left\|\bar{\boldsymbol{x}}\right\|^2$. □

A.4 Proof of Theorem 3.3: Correlations are Preserved under Residual Scaling

Theorem 3.3. *Let the input tokens have the same norm, i.e.* $\|\mathbf{X}_k\| = \|\boldsymbol{x}\|\ \ \forall k \in [n]$ *for some* $\boldsymbol{x} \in \mathbb{R}^{d_v}$. *Under the depth scaling for the residual block parameters* $\alpha_1^2 = \frac{\tilde{\alpha}_1}{L}, \alpha_2^2 = \frac{\tilde{\alpha}_2}{L}$ *with* $\tilde{\alpha}_1, \tilde{\alpha}_2 \in \mathbb{R}$ *independent of L, we have that:*

$$\lim_{L\to\infty}\rho^\ell = \frac{ne^{\tilde{\alpha}_1}C(\mathbf{X})}{(n-1)[(e^{\tilde{\alpha}_1}-1)C(\mathbf{X})+n\left\|\mathbf{X}\right\|_F^2]} - \frac{1}{n-1}. \tag{15}$$

On the other hand, if $\alpha_1, \alpha_2 \neq 0$ *are some constants independent of L, we have that:*

$$\lim_{L\to\infty}\rho^\ell = 1. \tag{16}$$

Proof. Due to the rotational symmetries of the Gaussian random matrices, if the input tokens have the same norm, then the expected norm at layer $\ell \in [L]$ is also the same across the token's representations. Hence, we can write $\mathbb{E}\left\|\mathbf{X}^\ell\right\|_F^2 = n\mathbb{E}\left\|\boldsymbol{x}^\ell\right\|^2$, where $\left\|\boldsymbol{x}^\ell\right\|^2$ is the norm of every token at layer ℓ. Furthermore, by definition of our correlation coefficient $\rho_{kk'}^l$, we have that $\mathbb{E}\langle\mathbf{X}_k^\ell, \mathbf{X}_{k'}^\ell\rangle = \rho_{kk'}^\ell\mathbb{E}\left\|\boldsymbol{x}^\ell\right\|^2$. By summing over the indexes k, k', we can expand the relation as:

$$\underbrace{\sum_{k,k'}\mathbb{E}\langle\mathbf{X}_k^\ell, \mathbf{X}_{k'}^\ell\rangle}_{\mathbb{E}C(\mathbf{X})} = \sum_{k,k'}\rho_{kk'}^\ell\mathbb{E}\left\|\boldsymbol{x}^\ell\right\|^2 = (n+\sum_{k\neq k'}\rho_{k,k'}^\ell)\mathbb{E}\left\|\boldsymbol{x}^\ell\right\|^2 = \underbrace{n\mathbb{E}\left\|\boldsymbol{x}^\ell\right\|^2}_{\mathbb{E}\|\mathbf{X}^\ell\|_F^2}(1+(n-1)\rho^\ell).$$

By solving for ρ^ℓ, we have that:

$$\rho^\ell = \frac{\mathbb{E}C(\mathbf{X}^\ell)}{(n-1)\mathbb{E}\left\|\mathbf{X}^\ell\right\|^2} - \frac{1}{n-1}.$$

Now we plug in the expressions for $\mathbb{E}C(\mathbf{X}^\ell)$ and $\mathbb{E}\left\|\mathbf{X}^\ell\right\|^2$ with the aid of Lemma 3.2 and Lemma 3.3, respectively. Finally, by taking the limits with respect to L, we get the desired result. □

A.5 Motivation for Assumption 3.1

We motivate here the following assumption, stated in the main paper. This assumption is crucial to compute expectations involving the softmax function.

Assumption 3.1 (Uniform attention)**.** *We assume that* $\mathbf{A}^\ell = \frac{1}{n}\mathbf{1}_{n\times n}$,

Theoretical analysis. We first show that this assumption holds when taking d_k to infinity, keeping d_v fixed.

Lemma A.7. *Consider initializing each entry of $\mathbf{W}^Q \in \mathbb{R}^{d_v \times d_k}$ and $\mathbf{W}^K \in \mathbb{R}^{d_v \times d_k}$ independently with variance $\sigma_k^2 = 2/(d_v + d_k)$ — i.e. Glorot initialization [Glorot and Bengio, 2010]. Let $\mathbf{M} = \frac{1}{\sqrt{d_k}} \mathbf{X}^\ell \mathbf{W}^{Q,\ell} \mathbf{W}^{K,\ell\top} \mathbf{X}^{\ell\top}$; for any $(i,j) \in [n] \times [n]$ we have*

$$\mathbb{E}[\mathbf{M}_{i,j} \mid \mathbf{X}] = 0, \qquad \mathbb{E}[\mathbf{M}_{i,j}^2 \mid \mathbf{X}] = \sigma_k^4 \cdot \|\mathbf{X}_{i,:}\|^2 \cdot \|\mathbf{X}_{j,:}\|^2. \tag{43}$$

While keeping $d_v < \infty$ fixed, taking d_k to infinity yields

$$\mathbb{E}[\mathbf{M}_{i,j}^2 \mid \mathbf{X}] = \mathcal{O}\left(\frac{1}{d_k^2}\right). \tag{44}$$

In other words, $\mathbf{M}$ converges to $\mathbf{0}_{n \times n}$ in L^2 as $d_k \to \infty$.

Proof. First, note that

$$\mathbf{M}_{i,j} = \frac{1}{\sqrt{d_k}} \sum_{a,c=1}^{d_v} \sum_{b=1}^{d_k} \mathbf{X}_{i,a} \mathbf{W}_{a,b}^Q \mathbf{W}_{c,b}^K \mathbf{X}_{j,c}.$$

Since $\mathbf{W}^Q$ is independent from $\mathbf{W}^K$ at initialization, $\mathbb{E}[\mathbf{M}_{i,j} \mid \mathbf{X}] = 0$. Next, we compute

$$\begin{aligned}
\mathbb{E}[\mathbf{M}_{i,j}^2] &= \frac{1}{d_k} \sum_{a,c,a',c'=1}^{d_v} \sum_{b,b'=1}^{d_k} \mathbf{X}_{i,a} \mathbf{X}_{i,a'} \mathbf{X}_{j,c} \mathbf{X}_{j,c'} \mathbb{E}\left[\mathbf{W}_{a,b}^Q \mathbf{W}_{a',b'}^Q \mathbf{W}_{c,b}^K \mathbf{W}_{c',b'}^K\right] \\
&= \frac{1}{d_k} \sum_{a,c,a',c'=1}^{d_v} \sum_{b,b'=1}^{d_k} \mathbf{X}_{i,a} \mathbf{X}_{i,a'} \mathbf{X}_{j,c} \mathbf{X}_{j,c'} \mathbb{E}\left[\mathbf{W}_{a,b}^Q \mathbf{W}_{a',b'}^Q\right] \mathbb{E}\left[\mathbf{W}_{c,b}^K \mathbf{W}_{c',b'}^K\right] \\
&= \frac{\sigma_k^4}{d_k} \sum_{a,c=1}^{d_v} \sum_{b=1}^{d_k} \mathbf{X}_{i,a}^2 \mathbf{X}_{j,c}^2 \\
&= \sigma_k^4 \|\mathbf{X}_{i,:}\|^2 \|\mathbf{X}_{j,:}\|^2.
\end{aligned}$$

This concludes the proof. □

The following classical result implies almost sure convergence of the softmax matrix as $d_k \to \infty$.

Lemma A.8 (Borel-Cantelli)**.** *Let (X_i) be a sequence of random variables. If for any $\epsilon > 0$*

$$\sum_{i=0}^{\infty} \mathbb{P}[|X_i - X| > \epsilon] < \infty,$$

then X_i converges to X almost surely[4].

Theorem A.2 (Almost-sure convergence)**.** *Consider initializing each entry of $\mathbf{W}^Q \in \mathbb{R}^{d_v \times d_k}$ and $\mathbf{W}^K \in \mathbb{R}^{d_v \times d_k}$ independently with variance $\sigma_k^2 = 2/(d_v + d_k)$ — i.e. Glorot initialization [Glorot and Bengio, 2010]. Let $d_v < \infty$ be fixed, as $d_k \to \infty$ we have that, for any $\mathbf{X}$,*

$$\mathbf{A} := \operatorname{softmax}\left(\frac{1}{\sqrt{d_k}} \mathbf{X} \mathbf{W}^Q \mathbf{W}^{K\top} \mathbf{X}^\top\right) \overset{a.s.}{\to} \frac{1}{n} \mathbf{1}_{n \times n}$$

and

$$\frac{\partial \mathbf{A}}{\partial \mathbf{M}} \overset{a.s.}{\to} \frac{1}{n} \mathbf{I}_n \otimes \left(\mathbf{I}_n - \frac{1}{n} \mathbf{1}_{n \times n}\right).$$

[4] That is, $\lim_{i \to \infty} X_i(\omega) = X(\omega)$ for almost every $\omega \in \Omega$ (i.e. with probability one).

Proof. Thanks to Lemma A.7 and Markov Inequality, we have fast convergence in probability: for any fixed $\mathbf{X}$,

$$\mathbb{P}[|\mathbf{M}_{i,j}| > \epsilon] \leq \frac{\mathbb{E}[\mathbf{M}_{i,j}^2]}{\epsilon^2} \leq \frac{C_\epsilon}{d_k^2}.$$

Borel Cantelli then directly yields almost sure convergence of $\mathbf{M}$ to $\mathbf{0}_{n\times n}$ as $d_k \to \infty$. Next, note that both $\mathbf{A}$ and $\frac{\partial \mathbf{A}}{\partial \mathbf{M}}$ are continuous functions of $\mathbf{A}$, hence we can apply standard continuity event-per-event. For almost every $\omega \in \Omega$,

$$\lim_{d_k\to\infty} \mathbf{A}(\mathbf{A}(\omega)) = \mathbf{A}\left(\lim_{d_k\to\infty} \mathbf{A}(\omega)\right) = \mathbf{A}(\mathbf{0}_{n\times n}) = \frac{1}{n}\mathbf{1}_{n\times n}.$$

Hence $\mathbf{A} \to \frac{1}{n}\mathbf{1}_{n\times n}$ almost surely. This can also be seen as a simple application of the continuous mapping theorem. The same reasoning yields almost sure convergence of

$$\frac{\partial \mathbf{A}}{\partial \mathbf{M}} = \text{blockdiag}\left(\text{diag}(\mathbf{A}_{i:}) - \mathbf{A}_{i:}\mathbf{A}_{i:}^\top\right),$$

to the corresponding limiting quantity. □

Empirical analysis. We empirically assess the validity of Assumption 3.1 and of its theoretical justification by performing the following experiments: for a range of increasing values of d_k, we compute $\mathbf{A}$ and we calculate $\frac{1}{n^2}||\mathbf{A}^\ell - \frac{1}{n}\mathbf{1}_{n\times n}||_F^2$, i.e. its average (entry-wise) distance from a uniform matrix with entries all equal to $1/n$. For each value of d_k, we repeat this calculation 200 times, each time with different random weight matrices. Fig. 7 displays how the $||\mathbf{A} - \frac{1}{n}\mathbf{1}_{n\times n}||_F^2$ averaged over 200 runs, tends to zero with a trend inversely proportional to d_k^2, as predicted by our theoretical analysis.

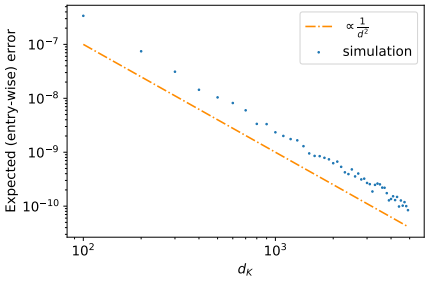

Figure 7: Evolution of $\frac{1}{n^2}||\mathbf{A}^\ell - \frac{1}{n}\mathbf{1}_{n\times n}||_F^2$ as a function of d_k for d_v fixed at 100.

B Additional Results

B.1 On the Roles of the $1/\sqrt{L}$-Scaling of the Residuals and Layer Normalization

We present some additional results on the propagation of the norm and the correlations in Figure 8. In particular, we empirically show that, with an adequate depth-dependent residual scaling, the norm and the correlation are stabilized, even for very deep networks. Furthermore, we demonstrate the propagation of the correlation and the gradient norms for the PRE-LN configuration in Figure 9. As also hinted in the main text, in Figure 4, the increase in correlation with depth for PRE-LN is much less wild. This also results in better stabilized gradients for the queries and keys' parameters. We also observe the opposite trend for the gradients of the values, in relation to the POST-LN case in Figure 3. We speculate that this different dependence, along with the better preserved correlation, is the main reason PRE-LN configured Transformers have been shown to scale better with depth. We plan to investigate this dependence more in future work. We provide the detailed experimental setup for Figures 3 and 9 in Table 2. Finally, as hinted in the main text, replacing the linear activation function with ReLU leads to an even faster increase in correlations (for instance, see the contraction argument in Nachum et al. [2021] below Equation (2)). We empirically evaluate that, by comparing the propagation of correlation (as in Figure 4) also for a network with ReLU activation functions. Results are shown in Figure 10.

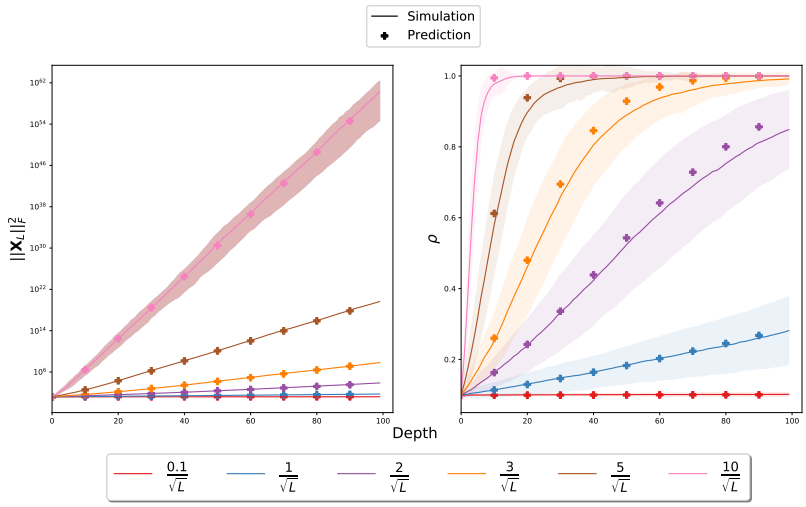

Figure 8: (Left) Propagation of the Frobenius norm of the input sequence; (Right) Propagation of the average token correlation. In both cases, no layer normalization layers are used and linear activations are employed.

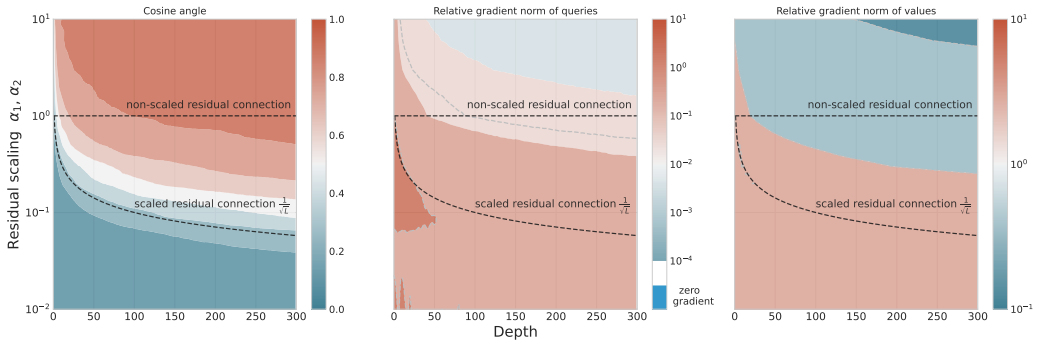

Figure 9: Same as Figure 3 but with a PRE-LN architecture. The correlation at depth 0 originates from the correlations in the randomly initialized tokens' embeddings and positional encodings.

Hyperparameters	Value
Embedding dimension	32
MLP dimension	128
Number of heads	4
Number of tokens	50
nonlinearity	ReLU

Table 2: Hyperparameters for Figures 3 and 9. Results are averaged across 50 runs.

B.2 Further Empirical Assessment of Assumption 3.1

Here, we empirically test the accuracy and limitations of the uniform-attention assumption.

For the empirical verification of Assumption 3.1 in the forward pass analysis, we plot the density of the norm of the representations for only-encoder Transformers of increasing depth. The results are shown in Fig 11. Note that when the standard deviation of the input is set to $1/\sqrt{d}$, then the uniform-attention assumption provide an excellent approximation to the common Xavier-initialization. On the contrary, we observe a deviation when the standard deviation of the input is increased. Also, note how as the depth increases, the distribution becomes more heavy-tailed. This heavy-tailedness

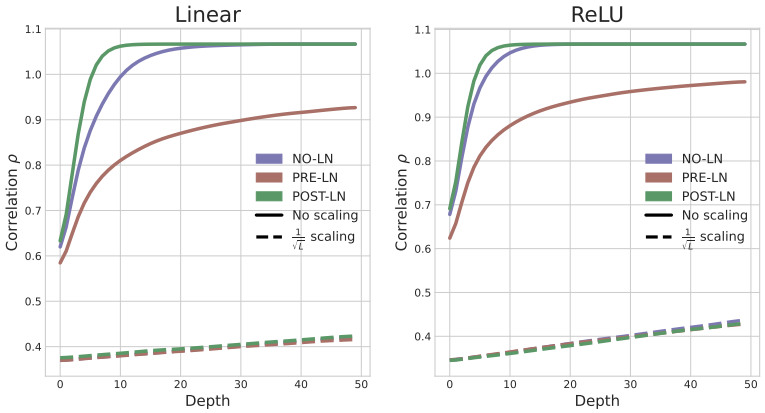

Figure 10: Increase in correlation at initialization for a network with linear (left) and ReLU (right) activation functions.

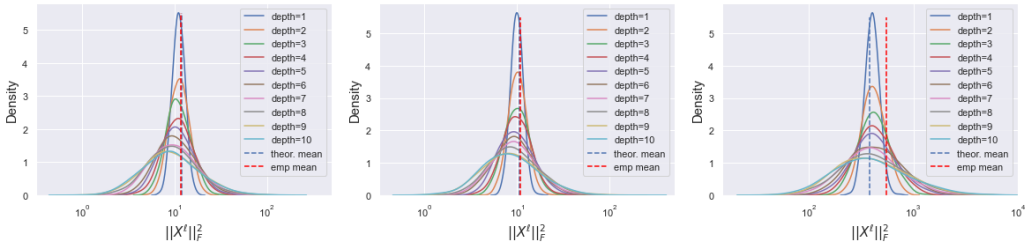

Figure 11: Density plots for $\left\|\mathbf{X}^\ell\right\|_F^2$ for Transformers of depths L from 1 to 10. The input $\mathbf{X}$ contains i.i.d Gaussian entries, simulating an embedding layer. We set $d := d_v = d_q = 30$. The empirical mean at $L = 10$ is highlighted in a vertical dashed red line, while the theoretical mean (Lemma 3.3) is a dashed blue line. The densities are estimated by sampling 1000 times the weights of the network. (Left): we adopt the uniform-attention. The standard deviation of the input is set to $1/\sqrt{d}$. (Center): Same, but removing the uniform-attention assumption. (Right): We remove the uniform-attention assumption, and set the standard deviation of the input to 1.

was recently formally shown for standard MLPs with and without ReLU activation [Noci et al., 2021, Zavatone-Veth and Pehlevan, 2021].

For the verification of the assumption in the backward pass, we additionally show in Fig. 12 how the norm of the gradients w.r.t queries and keys depends on the hidden dimension, the sequence length, the input correlation and the input variance. *Ground-truth* gradients are calculated with automatic differentiation, and they are compared with our theoretical results based on Assumption 3.1. As shown in Fig.12, our theoretical predictions show a very good agreement with the true gradients. Again, we notice that the smaller the values of the input standard deviation the tighter the agreement of the theory with the simulations. Intuitively, a higher input variance causes the argument of the softmax to have a large range of values. This in turn causes a deviation from the uniform distribution (i.e. maximum entropy), towards the distribution of minimum entropy (a Delta Dirac, corresponding to attending to only one token).

B.3 Empirical Verification of the Gradient Analysis of Section 3.3

Finally, in Figures 13 and 14 we show the dependence of the norm of the gradients for the keys and values based on the parameters of the architecture and the task-specific parameters. Figure 13 illustrates the true dependence and Figure 14 the one expected by the theory based on our assumptions. In short, the main takeaways are the following.

- As the correlation between the tokens increases (x-axis in the global plot), the norm of the gradients of the queries quickly diminishes compared to the one of the values.
- The dependence on the variance of the input σ_x^2 is different (y-axis in the global plot), being linear for the values and cubic for the queries. This highlights the importance of a stabilized forward pass and provides another explanation regarding the successful use of layer norm in Transformers.
- The dependence on n (x-axis in each subplot) and d (y-axis in each subplot) is more complicated, also being a function of the correlation ρ (compare the first column where $\rho = 0$ to the rest).

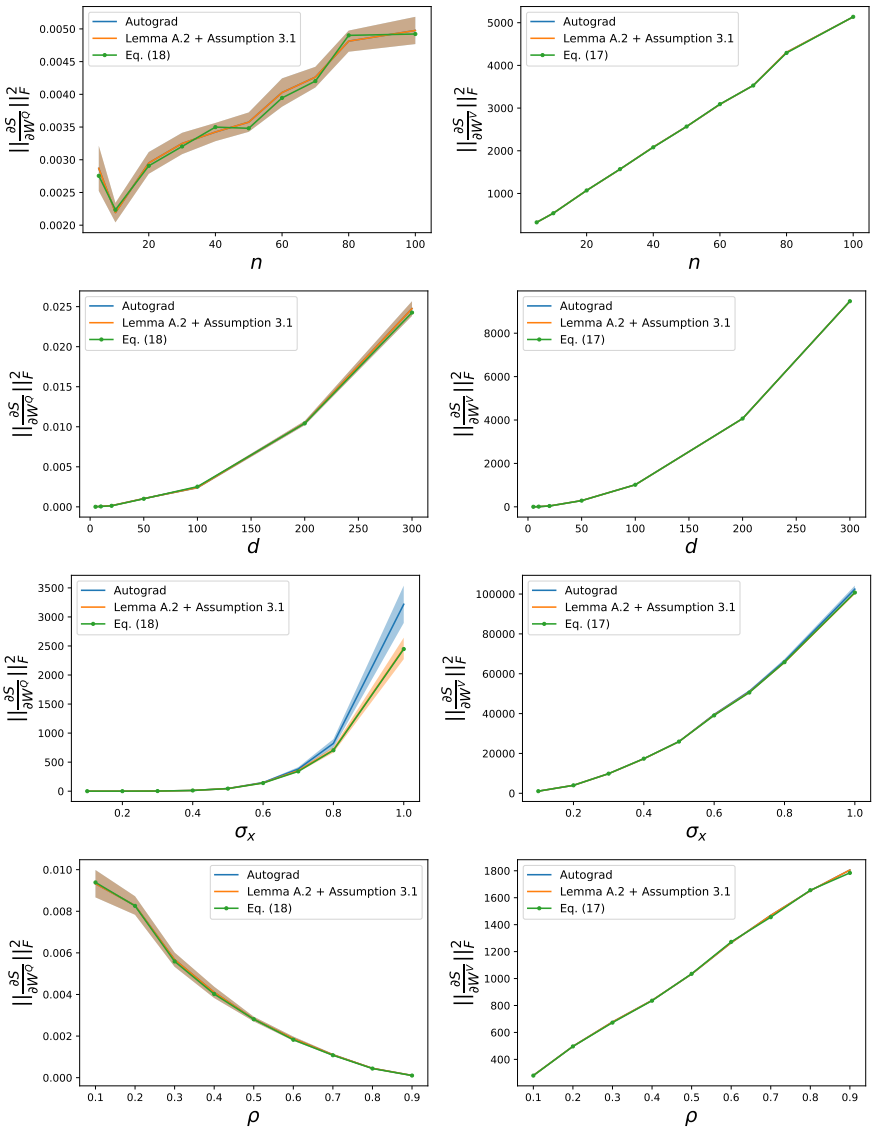

Figure 12: Empirical comparison of our theoretical findings. We sample, as aforementioned, the tokens according to a zero-mean Gaussian distribution, while varying the hidden dimension, sequence length, input correlation and input variance. Results are averaged over 20 runs.

C Experimental Setup

Here we provide more details regarding the experimental setup.

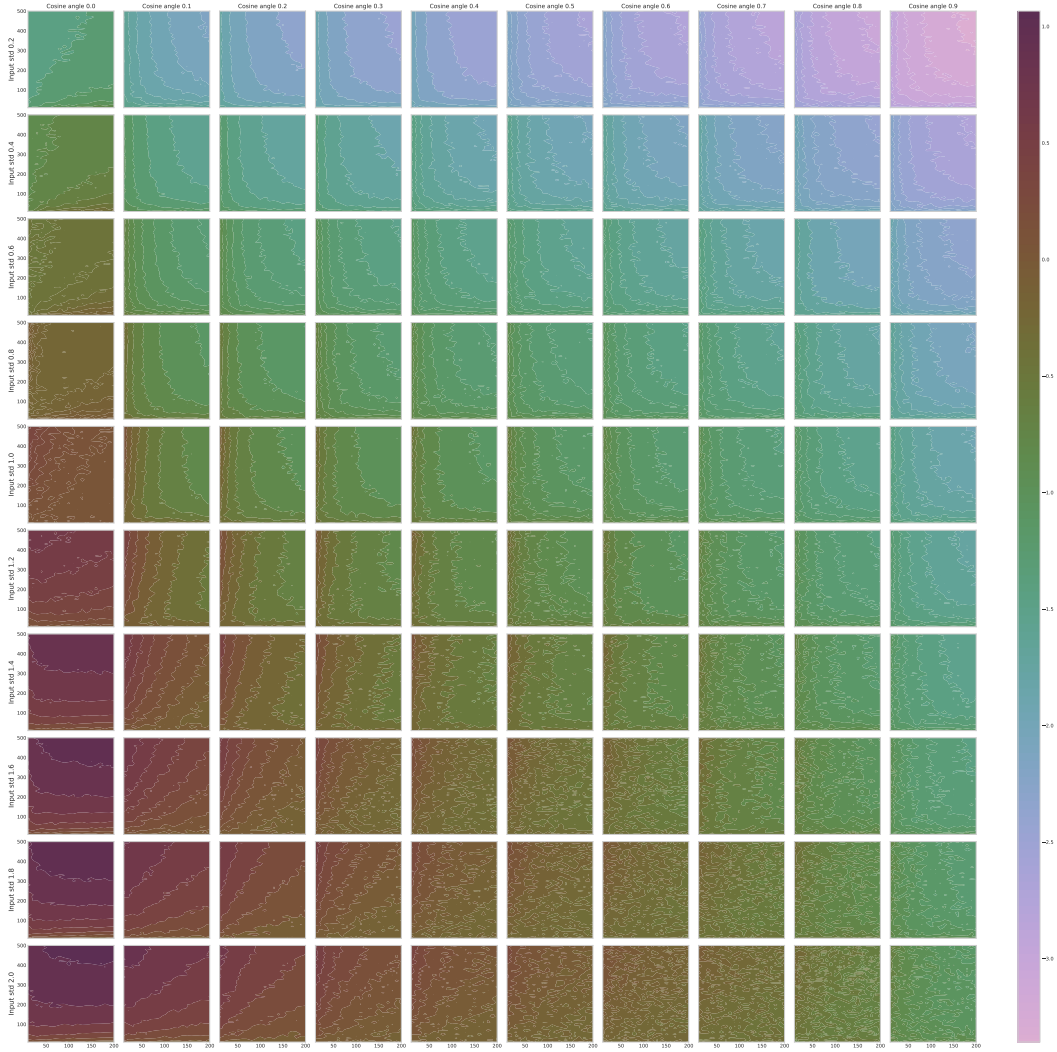

Figure 13: Log ratio of the norm of the gradients for the queries compared to those of the values for varying values of embedding dimension, sequence length, cosine of the tokens angle and standard deviation.

C.1 Toy Example

In Figure 6, we focus on a toy example where the task is to reverse a sequence of tokens. More specifically, given a sequence of 20 numbers in the range $0 - 9$, we predict the same tokens in the inverted order. We use an embedding layer of size 16, initializes with variance 1, and sinusoidal positional encodings to initially embed the input. We use a 5-layer POST-LN Transformer encoder model, with a single head attention operation and a two-layer feed-forward layer with a ReLU nonlinearity. We use residual scaling in this case equal to $\alpha_1 = \alpha_2 = 1$. We train using Adam with betas parameters $(0.9, 0.999)$, learning rate 0.01 and weight decay 0.

C.2 Tempering the softmax

The theory devised in Section 3.3 postulates that a different magnitude between the gradients of the queries/keys and the values should likely be observed. Here, we propose a simple remedy that consists in introducing an inverse temperature scaling τ inside the softmax that modifies the attention

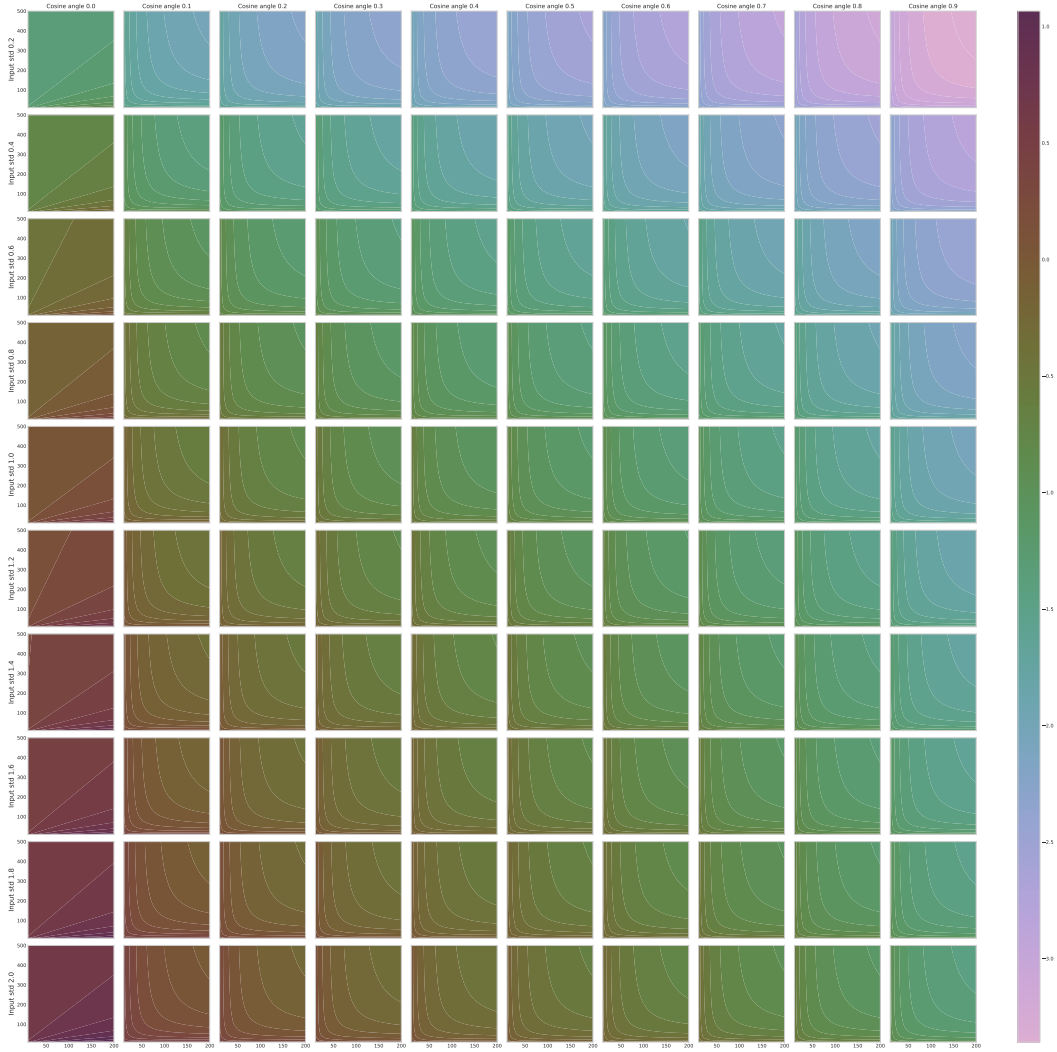

Figure 14: Log ratio of the norm of the gradients of the queries as expected by the theory, compared to those of the values for varying values of embedding dimension, sequence length, cosine of the tokens angle and standard deviation. We use Equations (17) and (18).

operation to

$$\mathbf{S}_\tau^\ell := \text{softmax}\left(\frac{\tau}{\sqrt{d_k}}\mathbf{X}^\ell\mathbf{W}^Q\left(\mathbf{X}^\ell\mathbf{W}^K\right)^\top\right)\mathbf{X}^\ell\mathbf{W}^V. \tag{45}$$

By computing the gradients of the queries/keys as in 3.3, Equation 18 becomes:

$$\mathbb{E}\left\|\frac{\partial\mathbf{S}_\tau}{\partial\mathbf{W}^Q}\right\|_F^2 = \tau^2\sigma_x^6\frac{(n-1)}{n}(1-\rho)^2 d(n+d). \tag{46}$$

Hence, the norm of gradients of the queries/keys scales linearly with τ. On the contrary, the temperature scaling inside the Softmax will not affect the gradients with respect to the values due to the fact that the Softmax normalizes the activations (see Lemma A.1 in Wang et al. [2022]). Hence τ can be heuristically chosen such that the magnitude of the gradients approximately matches:

$$\mathbb{E}\left\|\frac{\partial\mathbf{S}_\tau}{\partial\mathbf{W}^Q}\right\|_F^2 \overset{!}{=} \mathbb{E}\left\|\frac{\partial\mathbf{S}_\tau}{\partial\mathbf{W}^V}\right\|_F^2 \iff \tau^2 \approx \frac{dn(1+\rho(n-1))}{\sigma_x^4(1-\rho)^2(n+d)(n-1)}. \tag{47}$$

We stress that this requires a constant correlation ρ. In practice, this can be estimated as the mean correlation across all pairs of tokens (as we do in the computation of the correlations in Figure 1). Furthermore, both ρ and the variance σ_x^2 change across layers as our analysis in Section 3.2 predicts. Hence in practice a different temperature per layer should be adopted. Finally, note that in practice both ρ and σ_x^2 change during training, and is hard to study their the dynamics under SGD. We leave the time evolution of ρ and τ as an exciting future direction. Also, the value of n is set to be the average number of tokens per sentence. In this work, we set τ to a fixed value according to our analysis at initialization.

C.3 Translation Task

We now describe the experimental setup regarding the translation task on the IWSLT'14 De-En dataset. Using the ideas detailed in the previous section, we choose a temperature value of $\tau_{\text{final}} = 8.5$ to match the gradient norms of the values and queries as in Equations. (17) and (18). Doing so, we assume a constant small correlation between tokens (also empirically verified in Fig. 15) and set the sequence length n to the average found in our training dataset. Due to instabilities in training, we use warm-up on this temperature value. In short:

$$\tau = \tau_{\text{final}} \cdot \max(1, \frac{\text{step}}{\text{steps}_{\text{warmup}}}),$$

with '$\text{steps}_{\text{warmup}} = 1000$' and 'step' the current training step.

We base our implementation on fairseq [Ott et al., 2019]. For the hyperparameter configuration, we mostly rely on the extensive search already done in fairseq [Ott et al., 2019] and Liu et al. [2020]. The final used parameters are exhibited in Table 3. For the final evaluation, we use the best-performing model on the left-out validation set. We apply weight decay as in Loshchilov and Hutter [2017] for both SGD and Adam.

	Hyperparameters	Value
	Max tokens	4096
	Label smoothing	0.1
	clip-norm	0.0
	General Dropout	0.3
	Attention Dropout	0.1
	ReLU Dropout	0.1
	Hidden size	512
	FFN inner hidden size	2048
	Attention Heads	4
Adam	Learning rate	$7\epsilon^{-4}$
Adam	Learning rate scheduler	inverse sqrt
Adam	Warm-up updates	6000
Adam	Warm-up init learning rate	1e-7
Adam	Adam (β_1, β_2)	(0.9, 0.98)
Adam	Training updates	100K
Adam	Weight decay	0.0001
SGD	Learning rate	$2\epsilon^{-2}$
SGD	Learning rate scheduler	step
SGD	Step scheduler γ	0.1
SGD	Step scheduler update steps	[100K, 200K]
SGD	Training updates	250K
SGD	Weight decay	0.001

Table 3: Hyperparameters for the IWSLT'14 De-En translation task (Figure 1, 5).

Finally, in Figure 15 we display the evolution of correlations, residual scaling, and norm of the activations, with depth, for our best trained model. The residual scaling α_1, α_2 are trainable parameters. This enables them to weight differently the residual branches if deemed necessary. Although these values increase during training, the correlation between the tokens does not significantly increase, which as implied by our main results, allows efficient propagation of the gradients. The norm of the propagated forward signal tends to slightly increase with depth.

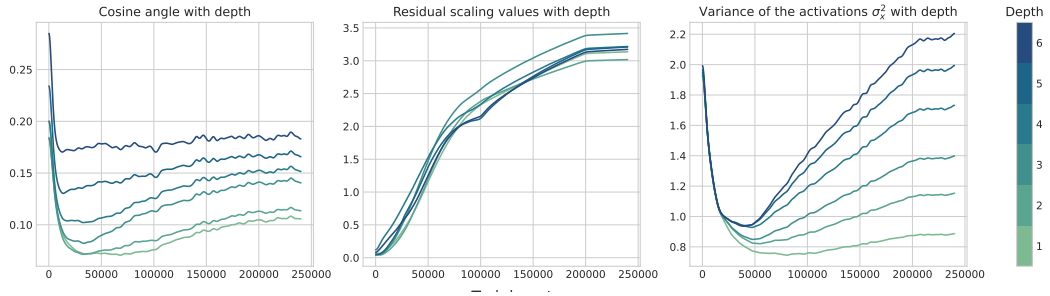

Figure 15: Evolution of the cosine of the angles, the trained residual α_1, α_2 and the activation norm throughout our training.