LAMBDA-SKIP CONNECTIONS: THE ARCHITECTURAL COMPONENT THAT PREVENTS RANK COLLAPSE

Federico Arangath Joseph
ETH Zurich
farangath@ethz.ch

Jerome Sieber
ETH Zurich
jsieber@ethz.ch

Melanie N. Zeilinger
ETH Zurich
mzeilinger@ethz.ch

Carmen Amo Alonso
Stanford University
camoalon@stanford.edu

ABSTRACT

Rank collapse, a phenomenon where embedding vectors in sequence models rapidly converge to a uniform token or equilibrium state, has recently gained attention in the deep learning literature. This phenomenon leads to reduced expressivity and potential training instabilities due to vanishing gradients. Empirical evidence suggests that architectural components like skip connections, LayerNorm, and MultiLayer Perceptrons (MLPs) play critical roles in mitigating rank collapse. While this issue is well-documented for transformers, alternative sequence models, such as State Space Models (SSMs), which have recently gained prominence, have not been thoroughly examined for similar vulnerabilities. This paper extends the theory of rank collapse from transformers to SSMs using a unifying framework that captures both architectures. We study how a parametrized version of the classic skip connection component, which we call *lambda-skip connections*, provides guarantees for rank collapse prevention. Through analytical results, we present a sufficient condition to guarantee prevention of rank collapse across all the aforementioned architectures. We also study the necessity of this condition via ablation studies and analytical examples. To our knowledge, this is the first study that provides a general guarantee to prevent rank collapse, and that investigates rank collapse in the context of SSMs, offering valuable understanding for both theoreticians and practitioners. Finally, we validate our findings with experiments demonstrating the crucial role of architectural components such as skip connections and gating mechanisms in preventing rank collapse.

1 INTRODUCTION

The phenomenon of rank collapse has been recently reported in the deep learning literature Dong et al. (2023); Noci et al. (2022); Shi et al. (2022); He et al. (2023); Geshkovski et al. (2024); Wu et al. (2024a;b). Rank collapse is a phenomenon by which embedding vectors exhibit a fast convergence rate into a *uniform token*, or equilibrium embedding vector. Various explanations why this phenomenon arises have been proposed: Dong et al. (2023) proposed a path decomposition argument and suggests that one of the reasons on why rank collapse happens is that the attention matrix is row stochastic. Geshkovski et al. (2024) instead proposes an analysis of rank collapse from a statistical physics perspective by treating transformers as mean-field interacting particle systems and observes that the resulting system converges to a metastable state corresponding to a Dirac delta measure. In practice, this convergence to a uniform token reduces model expressivity Daneshmand et al. (2020); Dong et al. (2023); Noci et al. (2022); Wu et al. (2024b) since a low rank in the output matrix means that the model is unable to capture complex relationships between tokens, also potentially causing vanishing gradients during training (Noci et al., 2022). Therefore, gaining a deeper understanding of this phenomenon—its origins, conditions for occurrence, and potential prevention strategies—is essential for designing models that are more robust, stable, and expressive.

Rank collapse has primarily been studied in the transformer architecture Vaswani et al. (2023). In particular, it has been reported that the convergence rate to the equilibrium embedding vector of self-

attention Vaswani et al. (2023) (corresponding to a rank-1 matrix) is doubly exponential (Dong et al. (2023)). Both empirical observations and theoretical results support that the presence of skip connections (He et al. (2015)) as well as the LayerNorm component (Ba et al. (2016)) in the architecture mitigate the issue of rank collapse Dong et al. (2023); Wu et al. (2024a). In particular, Dong et al. (2023) shows that by adding skip connections, there exist an infinite amount of parametrizations of the network for which rank collapse does not happen. Wu et al. (2024a) instead argues that adding LayerNorm slows down rank collapse by showing that in this setting the upper bound on the rank collapse measure is higher than for networks without these components. This implies that, in order to have convergence to 0, stricter conditions on input sequence and attention weights are needed.

Sequence models alternative to transformers have recently gained widespread attention. In particular, State Space Models (SSMs) such as S4 Gu et al. (2022b), S5 Smith et al. (2023), H3 Fu et al. (2023) and Mamba Dao and Gu (2024); Gu and Dao (2024) have been very prominent, where attention was replaced by SSM blocks, which operate in combination with other components such as Gating Mechanisms (Hochreiter and Schmidhuber (1997); Chung et al. (2014)), MLPs (Popescu et al. (2009) and skip connections (He et al. (2015)). Furthermore, comparisons between these architectures have been discussed in (Sieber et al., 2024; Hè and Kabic, 2023; Tiezzi et al., 2024), and a larger framework where both the transformer and SSMs are seen as different realizations of the same general model has recently been proposed (Ali et al. (2024), Dao and Gu (2024)). However, whether they suffer from similar weaknesses, and whether rank collapse happens more broadly in these architectures, is still an open question.

In this work, we focus on the specific issue of rank collapse and explore how slight modifications in the architecture, referred to as lambda-skip connections, can help to prevent it. Although parametrized versions of skip connections have been explored in the past (see e.g. He et al. (2016),Srivastava et al. (2015)), their impact in providing guarantees to prevent rank collapse has not been studied. Moreover, here we develop a general theory of rank collapse using the unifying framework for sequence models presented in Ali et al. (2024) and (Dao and Gu, 2024), comprising both transformers and SSM architectures. To the best of our knowledge, the rank collapse phenomenon in the SSM setting has not yet been studied and we are the first to provide a general lower bound for rank collapse which holds for both Transformers and SSMs. Additionally, we offer a mechanistic solution to prevent rank collapse by introducing a skip strength parameter, which comes with minimal computational overhead while enhancing the model's expressivity and stability. Since rank collapse can be a very serious issue for practitioners, investigating broadly when rank collapse occurs and designing robust architectures that prevent rank collapse is of central interest not only to theorists but also to experimentalists. Our contributions can be outlined as follows:

- We extend the theory on rank collapse originally developed for transfomers to SSMs by using the recently proposed framework unifying these two architecture types (Ali et al., 2024; Dao and Gu, 2024).
- We investigate how lambda-skip connections influence rank collapse. In particular, we give sufficient conditions that guarantee that rank collapse does not occur in the finite layers setting. Our conditions apply to both transformers and SSMs. We also study the necessity of this condition via ablation analysis and analytical examples.
- We provide experimental details in support of our theory, by showing how adding a parameter controlling the skip connection strength could be beneficial in preventing rank collapse.
- We also empirically show the role of gating mechanisms in preventing rank collapse. To the best of our knowledge, we are the first to make this connection between gating mechanisms, which were originally constructed for improving memory capabilities of the models (Hochreiter and Schmidhuber, 1997), and rank collapse.

2 Related Works

Rank Collapse. Rank collapse is the phenomenon in which the rank of the output layer of the network collapses to 1 (Dong et al. (2023)). To mathematically capture this, measuring the rank is not ideal since for every finite amount of layers, the output matrix will always be full rank with probability 1. This led to the development of an alternative metric to measure rank collapse Dong et al. (2023), which intuitively evaluates how "ill-conditioned" is the matrix of the output layer, i.e., how close it is to being rank 1. In Dong et al. (2023), this metric is defined by computing how far

the columns of the matrix are from the mean of all columns. It was shown in e.g. Dong et al. (2023); Wu et al. (2024a) that this metric quickly decays to 0 as the depth of the model increases. Hence, rank collapse is an intrinsic issue of deep-learning models. Rank collapse has mainly been studied for transformers (Vaswani et al. (2023)), Graph Neural Networks (Zhou et al. (2021)) and ReLU networks (Agarap (2019)), but not SSMs. The work in Daneshmand et al. (2020) studied rank collapse in randomly initialized linear and ReLU networks and proved that Batch Normalization can be helpful to mitigate the phenomenon. In Dong et al. (2023) the authors show that in self-attention-only transformers, rank collapse occurs doubly exponentially, i.e. with rate $O(a^{b^n})$ for some $a, b \in \mathbb{R}$. In the same paper, the authors propose for the first time that skip connections and MLPs (Popescu et al. (2009)) are helpful in preventing rank collapse from happening. Moreover, the work in Noci et al. (2022) shows that rank collapse is not only an issue during inference but also hinders training due to vanishing gradient problems at initialization. In Wu et al. (2024b) the authors show that rank collapse also happens in graph attention networks (in this setting the phenomenon is also known as oversmoothing), also causing the model to lose expressive power. Finally, the work in Wu et al. (2024a) studies the effect of LayerNorm on rank collapse in transformers, by proving that both of these components can help preventing or mitigating rank collapse.

Skip Connections. Skip connections were first introduced in ResNets He et al. (2015) with the purpose of easing training and optimization in deep neural networks (He et al. (2016); Veit et al. (2016); Balduzzi et al. (2018). More precisely, the introduction of skip connections addresses the vanishing gradient problem by decomposing back-propagation of gradients into multiple paths and thus allowing them to skip and bypass layers. As an alternative explanation, it was shown in Li et al. (2018) that adding skip connections also results in a loss function which is much smoother and hence it is harder for gradients to get stuck in local minima, facilitating the training procedure. In transformers and SSMs, the introduction of the skip connection followed the same motivation as for ResNets and other deep neural networks, i.e. to improve and facilitate optimization. The work in Dong et al. (2023) is the first to also connect the benefit of skip connections to avoiding rank collapse by studying the paths decomposition in transformers. Regarding modifications to the skip connection, Bachlechner et al. (2020) also considers a control parameter in residual networks. However, their approach differs in that it is used as a gating with the main network layer again with the purpose of stabilizing optimization, which allows for training of much deeper models. The work in Srivastava et al. (2015) proposes adding an extra feedforward layer to each residual layer. The output of each residual layer is then expressed as a convex combination of two terms: one gated by the residual connection and the other by the output of the primary network layer. The effect of these modifications on skip connection have not been analyzed. Lastly, He et al. (2016) studies the lambda-skip connection proposed here in terms of training efficiency and concludes that the standard choice of $\lambda = 1$ is usually beneficial.

LayerNorm. LayerNorm (Ba et al. (2016)) has become the standard choice of normalization layer for sequence models such as transformers (Vaswani et al. (2023)) and Mamba (Dao and Gu (2024)). In contrast to batch normalization (Ioffe and Szegedy (2015)), LayerNorm performs the shifting and the normalization token-wise and not sequence-wise. This is particularly beneficial as the embedding dimension is usually much smaller than the sequence dimension, resulting in a much cheaper computational overhead than BatchNorm. It has also been shown that LayerNorm is helpful to stabilize training (Xiong et al. (2020)) and to increase the expressivity of the attention layer (Brody et al. (2023)). Related instead to rank collapse, the work in Dong et al. (2023) was the first to analyze a potential connection between rank collapse and LayerNorm. However, they conjecture that LayerNorm has no impact on rank collapse. Later on, the work in Wu et al. (2024a) refutes this claim, showing both empirically and theoretically that LayerNorm helps slow down rank collapse.

3 Problem Setup

3.1 Sequence Model Deep-Learning Architecture

Rank collapse is defined as the convergence – in the limit of infinite depth – of the output matrix of the network's layers to a Rank-1 matrix. In this sense, rank collapse is intrinsic to deep-learning architectures. In what follows, we outline the different components on a *single-layer* of the architecture considered in this paper. We note that the presented architecture is general enough to capture both transformers and SSM-like architectures.

We start by introducing some notation. We index each layer of the architecture with $k \in \{1, ..., K\}$. We denote $X^{(k)} \in \mathbb{R}^{N \times d}$ and $Y^{(k)} \in \mathbb{R}^{N \times d}$ to be the input and output of the k-th layer, respectively, where N is the sequence length. In this setting, $Y^{(0)}$ is the input sequence to the model. Each layer k has the following components.

Main mechanism. The main mechanism differs between attention in transformers, and a recurrent block in SSMs. However, as shown in Ali et al. (2024) and Dao and Gu (2024), both models can be seen as part of a general framework

$$O^{(k)} = M^{(k)} V^{(k)}. \tag{1}$$

In the case of attention,

$$V^{(k)} = X^{(k)} W_V^{(k)} \in \mathbb{R}^{N \times d} \quad \text{and} \quad M^{(k)} = \text{softmax}\left(\frac{X^{(k)} W_Q^{(k)} (W_K^{(k)})^T (X^{(k)})^T}{\sqrt{d_{QK}}} \right) \in \mathbb{R}^{N \times N},$$

with $W_Q^{(k)} \in \mathbb{R}^{d \times d}$, $W_K^{(k)} \in \mathbb{R}^{d \times d}$ and $W_V^{(k)} \in \mathbb{R}^{d \times d}$ are the query, key and value matrices respectively.

In the case of the recurrent block, $V^{(k)} = X^{(k)}$, and $M^{(k)}$ is a lower triangular matrix with

$$M_{ji}^{(k)} = C_j^{(k)} \left(\prod_{l=0}^{j-i-1} A_{j-l}^{(k)} \right) B_i^{(k)},$$

where $A_t^{(k)} \in \mathbb{R}^{dH \times dH}, B_t^{(k)} \in \mathbb{R}^{dH \times 1}, C_t^{(k)} \in \mathbb{R}^{1 \times dH}, D_t^{(k)} \in \mathbb{R}^{1 \times 1}$. The first versions of the SSM blocks consisted of Linear Time Invariant (LTI) representations, i.e., $A_t = A$, $B_t = B$, $C_t = C$ and $D_t = D$ for all time steps t (Gu et al., 2022b;a; Smith et al., 2023). Recently, in order to improve the expressivity and the performance of such models, selective State Space Models have been proposed (Gu and Dao, 2024; Dao and Gu, 2024). In these models, matrices A_t, B_t, and C_t are input-dependent. We refer to Appendix A.1 for a detailed explanation of attention and the recurrent block.

Skip Connection. The skip connection sums the unmodified input to the output of the self-attention layer:

$$\tilde{Y}^{(k)} = X^{(k)} + O^{(k)}. \tag{2}$$

Although it was originally proposed to stabilize training by avoiding vanishing gradients (He et al. (2015)), it was recently shown that it also has a role in modulating rank collapse Dong et al. (2023). In SSM architectures, skip connections are often preceded by gating mechanisms, i.e., $O^{(k)}$ in equation 2 is replaced by $\tilde{O}^{(k)} = O^{(k)} \odot \sigma(W^{(k)} X^{(k)})$.[1]. However, we ignore these in the theoretical part of this paper for simplicity. In this work, we modify the skip connection by allowing it to have an additional parameter $\lambda^{(k)} \in \mathbb{R}$, that controls the strength of the skip connection, i.e.,

Definition 3.1. *Given a standard skip connection as per equation 2 and a parameter $\lambda^{(k)} \in \mathbb{R}$, a* lambda-skip connection *is defined as*

$$\tilde{Y}^{(k)} = \lambda^{(k)} X^{(k)} + O^{(k)}. \tag{3}$$

In the theoretical section, we will consider $\lambda^{(k)}$ to be fixed for all the layers for simplicity. However, in the experimental section, we will also consider models with learnable $\lambda^{(k)}$, i.e. it might vary across layers.

Layer Norm. The LayerNorm (Ba et al., 2016) computation shifts and normalizes the output $\tilde{Y}^{(k)}$ after the skip connection. Similar to the skip connection, the purpose of the LayerNorm is also to stabilize training. It was recently shown to also help to mitigate rank collapse (Wu et al., 2024a). Here, we consider a slightly simplified version of LayerNorm, similar to Wu et al. (2024a), where we only apply normalization and not shifting, namely:

$$Y^{(k)} = D^{(k)} \tilde{Y}^{(k)}, \tag{4}$$

where $D^{(k)} = \text{diag}(d_1^{(k)}, d_2^{(k)}, ..., d_N^{(k)}) \in \mathbb{R}^{N \times N}$ and $d_i^{(k)} = \frac{1}{||\tilde{Y}_{i,:}^{(k)}||_2}$.

[1] $\sigma(\cdot)$ is a non-linearity, often SILU, $W^{(k)}$ is a weight matrix and $\odot$ is the Hadamard product.

Other components. After the above-mentioned computations, the input Y passes through an MLP (Popescu et al., 2009), which can intuitively be seen as a non-linear feature extractor or a kernel. MLPs might play a role in mitigating rank collapse since they increase the Lipschitz constant of the network (Dong et al., 2023). In particular, the Lipschitz constant of the MLP layer appears in the double exponential decay term of the upper bounds of the rank collapse measure. This is beneficial since it allows for a wider range of choices for the parametrization of other components while maintaining a high-enough upper bound. Since the role of the MLP layer on rank collapse is well-understood, in this work we focus on the effect that LayerNorm and skip connections have on rank collapse. We anticipate that the addition of MLP will further improve the bound, and hence the rank collapse avoidance conditions.

3.2 Problem Statement

In mathematical terms, the phenomenon of rank collapse means that the columns of the final output matrix are proportional to each other. The representational capacity of the model is severely hindered by this phenomenon: since the final representations of each token are proportional to each other, there is limited diversity of tokens to be chosen from. Given that the rank of the output layer will be full rank with probability one (since rank-deficient matrices are a zero measure subspace of the space of all matrices), a different metric was introduced in Dong et al. (2023); Wu et al. (2024a) to assess rank collapse:

$$\mu(Y^{(k)}) = ||Y^{(k)} - \mathbf{1}\gamma_{Y^{(k)}}||_F, \tag{5}$$

where $Y^{(k)} \in \mathbb{R}^{N\times d}$ is the output matrix resulting from the k-th layer, $\gamma_{Y^{(k)}} := \frac{1}{N}\mathbf{1}^T Y^{(k)}$, and $||\cdot||_F$ is the Frobenius norm.

In order to study the impact of the architecture on the phenomenon of rank collapse, we make use of the equations describing the architecture: equation 1, equation 3, and equation 4. Moreover, we note that the output of layer $k-1$ is the input to layer k, i.e. $X^{(k+1)} = Y^{(k)}$. Combining these equations we obtain

$$Y^{(k)} = D^{(k)}(M^{(k-1)} + \lambda^{(k)} I)Y^{(k-1)}C_V^{(k-1)}, \tag{6}$$

where $C_V^{(k)} = W_V^{(k)}$ for the attention block and $C_V^{(k)} = I$ for the recurrent block.

The goal of this paper is to show how the introduction of $\lambda^{(k)}$ in the lambda-skip connection influences rank collapse. For this purpose, the main quantity of interest is $\mu(Y^{(K)})$ i.e., the value of the token-similarity metric on the output of the last layer. We provide a mathematical analysis of this metric in terms of the different architectural components in the next section.

4 Lambda-Skip Connection: Necessary and Sufficient to prevent rank collapse

Here, we present the analysis resulting from studying rank collapse with lambda-skip connections. In Section 4.1, we analyze the scenario where we replace the regular skip connection by a lambda-skip connection and show that, together with LayerNorm, an appropriate choice of λ in equation 3 is sufficient to prevent rank collapse in the finite layers setting. For simplicity, here we consider a fixed value of λ for all the layers, i.e. $\lambda^{(k)} = \lambda\ \forall k$. Specifically, we provide the value of $\lambda \in \mathbb{R}$ in the lambda-skip connection that guarantees the absence of rank collapse. This result holds for any arbitrary amount of layers and all the three architectures studied in the paper (transformers, LTI SSMs and selective SSMs). In Section 4.2, we study the necessity of this condition. First we show that in the ablated architecture without skip connection, rank collapse occurs exponentially, i.e., $\mu(Y^{(K)})$ converges to 0 exponentially. If, additionally, LayerNorm is also ablated, rank collapse occurs doubly exponentially. Second, we highlight via examples that the strength of the lambda-skip connection (parametrized through $\lambda \in \mathbb{R}$) plays a crucial role in determining the effect of skip connections on rank collapse.

4.1 Lambda-Skip Connection: sufficient to prevent rank collapse

Here we show the role that the parameter $\lambda \in \mathbb{R}$ plays in guaranteeing, under appropriate conditions, the prevention of rank collapse. To do this, we provide a λ-dependent lower bound for the rank

collapse metric $\mu(Y^{(K)})$ in equation 5. We start by defining the following quantity:

$$b := \frac{1}{a^K} \frac{2\lambda N S^2 C_M}{\lambda^2 c^2 - aS^2(C_M + \lambda)^2},$$

with $C_M := \sup_k ||M^{(k)}||_F$, $S := \sup_k \sigma_{\max}(C_V^{(k)})$, $c := \inf_k \sigma_{\min}(C_V^{(k)})$.[2] We also define the collapse rate as:

Definition 4.1. *The* collapse rate*, $a > 0$, is the rate at which the lower bound for the rank collapse metric $\mu(Y^{(K)})^2$ decays with the number of layers, i.e $\mu(Y^{(K)})^2 \geq a^K \mu(Y^{(0)})^2$.*

Theorem 4.1 (Lower Bound on Rank Collapse)**.** *Let the input sequence $Y^{(0)}$ be such that $\mu(Y^{(0)})^2 \geq b$. If the skip connection strength λ is chosen to satisfy*

$$\lambda^2 c^2 - aS^2(C_M + |\lambda|)^2 > 0, \tag{7}$$

then we can lower bound $\mu(Y^{(K)})$ by $\mu(Y^{(K)})^2 \geq a^K \mu(Y^{(0)})^2$ for all $K \in \mathbb{N}$.

Proof. The key part of the proof is to show the following relationship between the rank collapse measure at two consecutive layers: $\mu(Y^{(k+1)})^2 \geq \frac{1}{S^2(C_M+\lambda)^2}\left(\lambda^2 c^2 \mu(Y^{(k)})^2 - 2\lambda N S^2 C_M\right)$, which is done by algebraic manipulation and upper bounds on $||Y^{(k+1)}||_F$. The claim then follows by imposing the right-hand side above greater of equal to $a\mu(Y^{(k)})^2$, rearranging the terms and imposing the condition on λ to guarantee the feasibility of this. The full proof can be found in Appendix A.4. □

We note that this result is very general, since it holds for any model that can be expressed in the form of equation 6, independent of its specific implementation or parametrization. Moreover, the presence of LayerNorm implies that the bound provided in equation 7 does not depend on the input even when the matrix M is input-dependent. This is because the input is normalized before being fed through M. Specifically, the constant C_M is only dependent on the weights and on the sequence length [3]

Theorem 4.1 also provides an important intuition for the choice of λ. In particular, we note that to guarantee prevention of rank collapse, the ideal choice is $a = 1$. However, this choice will be mediated by the values of c and S in the different architectures. In particular, it is easy to see that the only way to guarantee a solution to 7 is by having $c^2 - aS^2 > 0$, for which λ needs to satisfy $|\lambda| > \frac{aC_M S^2 + \sqrt{ac^2 C_M^2 S^2}}{c^2 - aS^2}$.

Remark 4.1. *In the case of Mamba, $C_V = I$ and $c = S = 1$. This implies that the only possible choice for the collapse rate is $a < 1$. Although this does not guarantee $\mu(Y^{(K)})^2 \geq \mu(Y^{(0)})^2$, in practice a can be chosen to be very close to 1. For instance, one can set $a = 0.9999$. Given $K = 64$ (standard number of layers in Mamba), $a^K \approx 0.993$. Hence, in practice, if we choose λ appropriately, this choice still prevents rank collapse (see e.g. Figure 2).*

4.2 Lambda-Skip Connection: necessary to prevent rank collapse?

Next, we explore the role that the lambda-skip connection plays in terms of being necessary to prevent rank collapse. Although we do not provide a formal necessary condition, we explore this idea in two ways. First, we provide theoretical results on how rank collapse occurs when the (standard) skip connection is ablated. Then, we give two examples of how rank collapse may occur when the lambda-skip connection does not attain the bound in Theorem 4.1.

4.2.1 Rank collapse occurs without skip connections

Here we show how, for different architectures, rank collapse occurs in the absence of skip connections. Here we focus on transformers and selective SSMs. Additional results on LTI SSMs are presented in Appendix A.8.

[2] $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ denote the smallest and largest singular value of A.

[3] In transformers $C_M = \sqrt{N}$, in structured LTI SSMs $C_M \geq \sup_k ||A^{(k)}||\ ||B^{(k)}||\ ||C^{(k)}||_F$ and in selective SSMs $C_M \geq \sup_k ||W_B^{(k)}||\ ||W_C^{(k)}||_F$ (since in Selective SSMs it is common practice to choose $A_t^{(k)}$ with eigenvalues smaller than one).

Transformers. We start by reporting for completeness the result in Wu et al. (2024a) for transformers architectures with Self-Attention and LayerNorm only.

Theorem 4.2 (Corollary 1, Wu et al. (2024a) (Informal))**.** *Under certain assumptions on the self-attention weights and the input sequence, there exist $C > 0$, $\epsilon > 0$ and $r > 0$ such that $N\epsilon < 1$ and*

$$\mu(Y^{(K)}) \leq C(1-\epsilon^{2r})^{\frac{K}{2r}}, \quad \forall K \geq 0.$$

meaning that tokens converge to a common point on $\mathbb{S}^{d-1}$ exponentially.

We refer to Appendix A.5 for a rigorous presentation of Theorem 4.2. The main takeaway from this result is that, under appropriate conditions on the architecture and the input sequence, removing skip connections and only using attention layers and LayerNorm can cause an exponentially fast convergence to a rank-1 matrix. This highlights the importance of skip connections to mitigate rank collapse. It is also interesting to compare the result in Theorem 4.2 to the setting of self-attention only transformers, i.e. removing both LayerNorm and skip connections, reported in Dong et al. (2023). In particular, under certain assumptions on the attention matrix and the input sequence, rank collapse occurs at a doubly exponential rate. We report the full result for completeness in Theorem A.3 and Corollary A.3.1 (where we adapted the Theorem to the rank collapse measure provided in the problem statement).

Selective SSMs. According to Dao and Gu (2024), for selective SSMs where $A_t = \alpha_t I$ (e.g., Mamba-2), the matrix M can be compactly written as $M^{(k)} = 1\text{SS}(\alpha) \odot \left(Y^{(k)} W_C W_B^\top Y^{(k)^T}\right)$, where $1\text{SS}(\alpha)$ is a lower-triangular matrix 1-semiseparable matrix.[4] In what follows, we introduce the following assumption on selectivity:

Assumption 4.1. *Only the matrices $B_t = Y^{(k)}_{t,:} W_B$ and $C_t = Y^{(k)}_{t,:} W_C$ are input-dependent, while $A_t = A = \alpha I$ $\forall t$ (with $\alpha \leq 1$) is independent of the input.*

This choice is done for ease of exposition. Although it may seem restrictive in theory, we show experimentally that eliminating skip connections can cause rank collapse for selective SSMs where A_t is input dependent as well, such as Mamba (see e.g. the plot corresponding to $\lambda = 0$ in Figure 1). In the following derivation, we neglect gating mechanisms for the sake of the theoretical analysis–we will thoroughly analyze their effect on rank collapse in simulation (see e.g. Figure 3).

Theorem 4.3 (Rank Collapse for selective SSMs without skip connection)**.** *Let $\phi^{(k)} = \min_{i,j\in[N]} \langle Y^{(k)}_{i,:}, Y^{(k)}_{j,:} \rangle$, where $\langle \cdot, \cdot \rangle$ indicates the inner product. Under Assumption 4.1 and if $c \leq \phi^{(0)} < 1$ for some $c > 0 \in \mathbb{R}$, $\lambda_{min} > 0$, $\sum_j M_{ij} \geq 1$ $\forall i$, then it holds that:*

$$\mu(Y^{(K)}) \leq \sqrt{N}\left(1 - c^2 \lambda_{min}^2 \alpha^{2N}\right)^K \ \forall K \geq 0,$$

where λ_{min} and λ_{max} the minimum and maximum eigenvalues of $\frac{W_B W_C^\top + W_C W_B^\top}{2}$.

Proof. The proof adapts the derivation in Corollary 1 in Wu et al. (2024a) to selective SSMs. In particular, the key is to find a recursive relation between $1 - \phi^{(k)}$ and $1 - \phi^{(k+1)}$. The full proof can be found in Appendix A.7. □

In particular, when $\lambda_{\text{min}} \leq \frac{2}{c^2 \alpha^{2N}}$, the model suffers from exponential rank collapse. Note that the above result holds for a model with layers having the same parametrization.

In Appendix A.9 Theorem A.10, we show that if we instead have a selective SSM with ablated the skip connections and LayerNorms, then rank collapse can happen doubly exponentially under certain conditions, similarly to what we have in transformers architecture. In contrast with the LTI SSM case (Theorems A.7 and A.8), which exhibits an exponential decay, the matrix $M^{(k)}$

[4] Specifically, $1\text{SS}(\alpha) := \begin{bmatrix} 1 & & & & \\ \alpha_1 & 1 & & & \\ \alpha_2\alpha_1 & \alpha_2 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \alpha_{N-1}\dots\alpha_1 & \alpha_{N-1}\dots a_2 & \dots & \alpha_{N-1} & 1 \end{bmatrix}$.

depends quadratically on the input for selective SSMs: $||M^{(k)}||_F \leq \sqrt{N}||Y^{(k)}||_F^2||W_{BC}||_F$. where $W_{BC} = W_C W_B^\top$. This dependency is what causes the doubly exponential behavior.

4.2.2 Rank Collapse can still occur with skip connections under specific choices of λ: examples

Here we show analytical examples of how, for different architectures, rank collapse occurs in the presence of lambda-skip connections, if λ does not satisfy the condition in Theorem 4.1.

Structured LTI SSMs. Consider the system $A_t = A = \alpha I, B_t = B = I$ and $C_t = C = I$, and let $N = 2$ and $d = 2$. By using the relation $M_{ji} = C_j^\top A_j \dots A_{i+1} B_i$ whenever $j \geq i$ (Dao and Gu, 2024), we get that $M + \lambda I = \begin{pmatrix} 1+\lambda & 0 \\ \alpha & 1+\lambda \end{pmatrix}$. Let us denote this system, equipped with LayerNorm, as [Sys-1].

Proposition 4.3.1. *Given [Sys-1], if we choose $\alpha = 2$, i.e. $M = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$, $\lambda > -2$ and $Y^{(0)} = I$, then we have that $\mu(Y^{(k)}) \to 0$. Conversely, given the choices above with $\lambda < -2$ (with $\lambda \neq -1$, then rank collapse does not happen, i.e. $\mu(Y^{(k)}) \not\to 0$.*

Proof. The result follows by calculating $Y^{(k)}$ inductively and by showing that the corresponding $\mu(Y^{(k)})$ decays to 0. See Appendix A.10 for the full proof. □

Selective SSMs. Consider the system $A_t = A = \alpha I$ (for simplicity, we consider α not to be input dependent here), $B = YW_B$ and $C = YW_C$ input dependent (where y_t is the t-th row of Y) with LayerNorm applied to the output of each layer. Once again, we let $N = 2$ and $d = 2$. We have $M = W_A \odot \left(YW_C W_B^\top Y^\top\right)$ where $W_A = \begin{pmatrix} 1 & 0 \\ \alpha & 1 \end{pmatrix}$. By choosing $\alpha = 1$ and $W_B = W_C = I$, we get that $M = \begin{pmatrix} 1 & 0 \\ y_1^\top y_2 & 1 \end{pmatrix}$ since the input to the layer is normalized by LayerNorm. Let us denote this system as [Sys-2].

Proposition 4.3.2. *Given [Sys-2], if we then choose $\lambda > -\frac{3}{2}$, M as above and $Y^{(0)} = \begin{pmatrix} 1 & 0 \\ \frac{\alpha_0}{\sqrt{\alpha_0^2+\beta_0^2}} & \frac{\beta_0}{\sqrt{\alpha_0^2+\beta_0^2}} \end{pmatrix}$ for any α_0, β_0 with $\alpha_0 > 0$, then we have that $\mu(Y^{(k)}) \to 0$. On the other hand, if we choose $\lambda < -\frac{3}{2}$ and M and $Y^{(0)}$ as above, then rank collapse is avoided.*

Proof. Again, the result follows by calculating $Y^{(k)}$ inductively and by showing that the corresponding $\mu(Y^{(k)})$ decays to 0 for $\lambda > -\frac{3}{2}$ while this does not happen if $\lambda < -\frac{3}{2}$. We refer to Appendix A.11 for the full proof. □

5 Experiments

In this section, we aim to both explore the link between gating mechanisms and rank collapse and to empirically validate our theoretical findings. In Section 5.1, we empirically validate Theorem 4.1, demonstrating the importance of selecting the appropriate skip connection strength to mitigate rank collapse. In Section 5.2, we show that for the Mamba-2 architecture (Dao and Gu (2024)), gating mechanisms indeed play a crucial role in preventing rank collapse. Finally, we provide additional experiments on the S4 architecture (Gu et al., 2022b) in Appendix A.12.1.

5.1 Comparison of Rank Collapse for Mamba Architecture with different values of λ

We start by validating our main finding, i.e. that by controlling the value of λ in the skip connection, we can prevent rank collapse from happening. In order to do so, we consider a 2 billion parameters pre-trained Mamba-2 model available on HuggingFace and remove the gating mechanisms. Instead, we add an additive skip connection with varying λ. For simplicity, we keep the value of λ constant at every layer. Similar to Dong et al. (2023); Wu et al. (2024a), we sample 32 text excerpts from

Wikipedia using the Wikipedia API and tokenize them in sequences of at least 128 tokens using the BERT tokenizer Devlin et al. (2019). We then forward pass these inputs through the pre-trained model and evaluate the rank collapse measure μ (normalized by the norm of the layer output) for each of the 64 layers for different values of λ, as shown in Figure 1. We observe that when $|\lambda|$ is small, the model suffers from rank collapse. This happens even when $\lambda = 1$, which is usually the default design choice. After sufficiently increasing or decreasing λ, we note that we avoid rank collapse, consistent with our theoretical results, particularly Theorem 4.1. This suggests that rather than fixing the skip connection strength at $\lambda = 1$ (as is standard), it may be beneficial to treat λ as a learnable parameter or hyperparameter to improve model stability. In Figure 2, we analyze

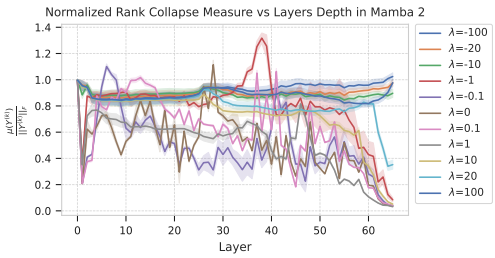

Figure 1: (Normalized) Rank Collapse measure plotted for different values of the skip connection's strength λ. Again, shaded areas represent one standard deviation from the mean calculated over the 32 examples.

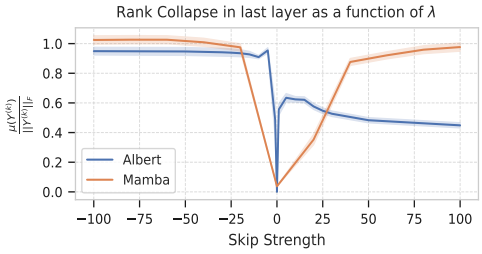

Figure 2: (Normalized) Rank Collapse measure at the last layer plotted for different values of λ for both Albert and Mamba. Shaded areas represent one standard deviation from the mean calculated over the 32 examples.

how the rank collapse measure at the last layer varies with different skip connection strengths for both the Albert's architecture (Lan et al. (2020)) and Mamba-2, keeping the full architecture (also LayerNorm). In both cases, we note the importance of tuning and choosing the correct value of λ to avoid rank collapse. We see how a value of λ very close to 0, as we expected, is detrimental and causes a very low value of μ, even in the presence of LayerNorm, which is consistent with our findings in Theorem 4.3. Additionally, we see that a negative λ leads to a higher μ. Intuitively, the reason for this could be that a negative λ can be interpreted as negative feedback in the system, hence helping stabilizing it. Note that even if in both cases our condition on λ in Theorem 4.1 is too conservative (e.g. for Mamba-2 with $a = 0.9$, λ needs to satisfy$|\lambda| > 937.924$), in practice much lower values of λ are good enough to prevent rank collapse.

To investigate whether using λ as a trainable parameter affects the learning performance, we train two transformer-based architectures and two SSM-based architectures on the image task of the long range arena (LRA) benchmark (Tay et al., 2020) and on a multi-query associative recall (MQAR) task (Arora et al., 2023). The accuracy of all four architectures on both tasks are reported in Table 1 and the details on the experimental setup are provided in Appendix A.13. From the achieved accuracy we note that learning λ does not affect the performance and even outperforms the models with fixed λ in some cases.

Task [%]	**Model**							
	Transformer		Lin. Transformer		Mamba		Mamba-2	
	$\lambda = 1$	var. λ	$\lambda = 1$	var. λ	$\lambda = 1$	var. λ	$\lambda = 1$	var. λ
`Image` (LRA)	32.64	32.85	34.10	32.80	63.04	62.92	42.28	38.92
MQAR ($L = 512$)	99.6	98.9	1.6	1.55	81.5	85.3	97.3	99.1

Table 1: Model performance in terms of test accuracy on the `Image` task of the LRA benchmark and the MQAR task $\{L = 512, \text{KV-pairs} = 64\}$.

5.2 Comparison of Rank Collapse for Mamba Architecture with and without Gating Mechanisms and LayerNorm

We also analyze the effect of gating mechanisms and LayerNorms in the Mamba architecture. Because gating mechanisms can be viewed as a multiplicative form of skip connections, we aim to empirically explore whether they, like skip connections, contribute to stability and prevent rank collapse. We repeat the same procedure as above, i.e., forward passing 32 excerpts from Wikipedia to the pre-trained model and calculating the rank collapse measure across all layers. We replicate this for all considered cases, namely gating plus LayerNorm, only gating, only LayerNorm, and neither gating nor LayerNorm.

As shown in Figure 3, when removing the gating mechanism, we either observe the rank collapse measure approaching zero especially for later layers if the LayerNorm is present, or we observe an initial drop in the rank collapse measure followed by a dramatic increase if we remove the LayerNorm. This highlights the importance of having LayerNorm to keep the model stable, as was also observed in Dao and Gu (2024). Indeed, even when the model contains a gating mechanism, although we prevent rank collapse, removing LayerNorm leads to a sudden increase in the rank collapse measure, indicating instability. Interestingly, we find that gating mechanisms, originally designed to enhance memory capabilities, also play a crucial role in preventing rank collapse. To the best of our knowledge, this is the first time this connection has been made.

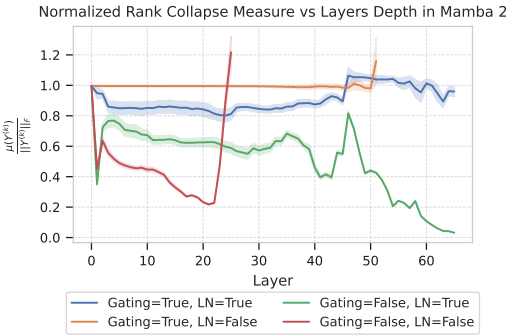

Figure 3: (Normalized) Rank Collapse measure of the Mamba-2 model plotted as a function of layer depth. The shaded areas represent one standard deviation from the mean calculated over the 32 examples. Gating=True/False indicates we use a Mamba2 architecture with/without gating whereas LN=True/False indicate we use the architecture with/without LayerNorm

6 Conclusion

This paper studies the phenomenon of rank collapse across general sequence model architectures, using the unifying framework from Ali et al. (2024) and Dao and Gu (2024). Our main contribution is the introduction of a lambda-skip connection, which adds a regulatory term to the standard skip connection component to control its strength. We provide sufficient conditions for this parameter under which rank collapse does not occur. We do this via a lower bound on the rank collapse measure, which holds for transformers, LTI SSMs, selective SSMs and more generally, all architectures that can be written in the framework from Ali et al. (2024) and Dao and Gu (2024). We further study the necessity of the lambda-skip connection by showing that when this component is ablated, selective SSMs also experience rank collapse even in the presence of LayerNorm. Moreover, we show that similarly to Transformers, selective SSMs can also be affected by doubly exponential rank collapse when both skip connections and LayerNorm are ablated. We suggest that the primary cause of this doubly exponential rank collapse is the input dependence of the $M^{(k)}$ matrix. We validate our theoretical analysis with simulation experiments and further analyze other components such as the Gating Mechanism, which we have found also play a role in rank collapse.

Limitations and Future Work. A limitation of this work is the exclusion of gating mechanisms from the theoretical analysis. Since gating mechanisms help to mitigate the effect of rank collapse, we expect their inclusion in the derivations would result in tighter upper bounds on the rank collapse measure for Mamba. Furthermore, one other line worth exploring is to consider the effect on lower bounds of MLPs on a theoretical standpoint instead of only considering skip connections and LayerNorm, which have not been considered in this paper. Finally, deriving a tighter lower bound that matches the results in our experiments is also a very promising avenue of research. Finally, Remark 4.1 offers intriguing insights from a control-theoretic perspective, as it suggests that the λ parameter can be interpreted as a control gain. We believe that further theoretical analysis of the stability and performance associated with λ-skip connections could lead to a more principled design of learning architectures.

REFERENCES

A. F. Agarap. Deep learning using rectified linear units (relu), 2019. URL https://arxiv.org/abs/1803.08375.

A. Ali, I. Zimerman, and L. Wolf. The hidden attention of mamba models, 2024. URL https://arxiv.org/abs/2403.01590.

S. Arora, S. Eyuboglu, A. Timalsina, I. Johnson, M. Poli, J. Zou, A. Rudra, and C. Ré. Zoology: Measuring and improving recall in efficient language models, 2023. URL https://arxiv.org/abs/2312.04927.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

T. Bachlechner, B. P. Majumder, H. H. Mao, G. W. Cottrell, and J. McAuley. Rezero is all you need: Fast convergence at large depth, 2020. URL https://arxiv.org/abs/2003.04887.

D. Balduzzi, M. Frean, L. Leary, J. Lewis, K. W.-D. Ma, and B. McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question?, 2018. URL https://arxiv.org/abs/1702.08591.

S. Brody, U. Alon, and E. Yahav. On the expressivity role of layernorm in transformers' attention. *arXiv preprint arXiv:2305.02582*, 2023.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. URL https://arxiv.org/abs/1412.3555.

H. Daneshmand, J. Kohler, F. Bach, T. Hofmann, and A. Lucchi. Batch normalization provably avoids rank collapse for randomly initialised deep networks, 2020. URL https://arxiv.org/abs/2003.01652.

T. Dao and A. Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024. URL https://arxiv.org/abs/2405.21060.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Y. Dong, J.-B. Cordonnier, and A. Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2023. URL https://arxiv.org/abs/2103.03404.

Y. Fang, K. Loparo, and X. Feng. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994. doi: 10.1109/9.362841.

D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré. Hungry hungry hippos: Towards language modeling with state space models, 2023. URL https://arxiv.org/abs/2212.14052.

B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. A mathematical perspective on transformers, 2024. URL https://arxiv.org/abs/2312.10794.

A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Re. Hippo: Recurrent memory with optimal polynomial projections, 2020. URL https://arxiv.org/abs/2008.07669.

A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré. Combining recurrent, convolutional, and continuous-time models with linear state-space layers, 2021. URL https://arxiv.org/abs/2110.13985.

A. Gu, K. Goel, A. Gupta, and C. Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022a.

A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*, 2022b.

B. He, J. Martens, G. Zhang, A. Botev, A. Brock, S. L. Smith, and Y. W. Teh. Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation, 2023. URL https://arxiv.org/abs/2302.10322.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks, 2016. URL https://arxiv.org/abs/1603.05027.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

H. Hè and M. Kabic. A unified view of long-sequence models towards modeling million-scale dependencies, 2023. URL https://arxiv.org/abs/2302.06218.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. URL https://arxiv.org/abs/1502.03167.

A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020. URL https://arxiv.org/abs/1909.11942.

H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets, 2018. URL https://arxiv.org/abs/1712.09913.

I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation, 2015. URL https://arxiv.org/abs/1508.04025.

L. Noci, S. Anagnostidis, L. Biggio, A. Orvieto, S. P. Singh, and A. Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse, 2022. URL https://arxiv.org/abs/2206.03126.

A. Orvieto, S. L. Smith, A. Gu, A. Fernando, C. Gulcehre, R. Pascanu, and S. De. Resurrecting recurrent neural networks for long sequences, 2023. URL https://arxiv.org/abs/2303.06349.

M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis. Multilayer perceptron and neural networks. *WSEAS Trans. Cir. and Sys.*, 8(7):579–588, July 2009. ISSN 1109-2734.

H. Shi, J. Gao, H. Xu, X. Liang, Z. Li, L. Kong, S. M. S. Lee, and J. T. Kwok. Revisiting oversmoothing in bert from the perspective of graph, 2022. URL https://arxiv.org/abs/2202.08625.

J. Sieber, C. A. Alonso, A. Didier, M. N. Zeilinger, and A. Orvieto. Understanding the differences in foundation models: Attention, state space models, and recurrent neural networks, 2024. URL https://arxiv.org/abs/2405.15731.

J. T. H. Smith, A. Warrington, and S. W. Linderman. Simplified state space layers for sequence modeling, 2023. URL https://arxiv.org/abs/2208.04933.

S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data, 2016. URL https://arxiv.org/abs/1611.06067.

R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks, 2015. URL https://arxiv.org/abs/1505.00387.

M. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber. Deep networks with internal selective attention through feedback connections, 2014. URL https://arxiv.org/abs/1407.3068.

Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler. Long range arena: A benchmark for efficient transformers, 2020. URL https://arxiv.org/abs/2011.04006.

M. Tiezzi, M. Casoni, A. Betti, M. Gori, and S. Melacci. State-space modeling in long sequence processing: A survey on recurrence in the transformer era, 2024. URL https://arxiv.org/abs/2406.09062.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

A. Veit, M. Wilber, and S. Belongie. Residual networks behave like ensembles of relatively shallow networks, 2016. URL https://arxiv.org/abs/1605.06431.

A. Voelker, I. Kajić, and C. Eliasmith. Legendre memory units: Continuous-time representation in recurrent neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/952285b9b7e7a1be5aa7849f32ffff05-Paper.pdf.

X. Wu, A. Ajorlou, Y. Wang, S. Jegelka, and A. Jadbabaie. On the role of attention masks and layernorm in transformers, 2024a. URL https://arxiv.org/abs/2405.18781.

X. Wu, A. Ajorlou, Z. Wu, and A. Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks, 2024b. URL https://arxiv.org/abs/2305.16102.

R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu. On layer normalization in the transformer architecture. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/xiong20b.html.

J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications, 2021. URL https://arxiv.org/abs/1812.08434.

A Appendix

A.1 Preliminaries

We present a brief description of the sequence models considered in this work, namely attention (Stollenga et al. (2014); Luong et al. (2015); Song et al. (2016); Vaswani et al. (2023)) and recurrent blocks in State Space Models (Voelker et al. (2019); Gu et al. (2020; 2021; 2022b); Fu et al. (2023), and their corresponding architectural components (from Transformers Vaswani et al. (2023) and Mamba Gu and Dao (2024)).

A.2 Attention

Attention is considered to be the backbone of the Transformer architecture. This is thanks to its ability to capture correlations and dependencies between tokens in the sequence. In particular, the attention block can be expressed as follows:

$$O = \text{softmax}\left(\frac{XW_QW_K^TX^T}{\sqrt{d_{QK}}}\right)XW_v \in \mathbb{R}^{N\times d}, \tag{8}$$

where the softmax operation is applied row-wise and $\sqrt{d_{QK}} \in \mathbb{R}$ acts as a sort of temperature parameter to modulate the strength of the interactions between tokens. $W_Q \in \mathbb{R}^{d\times d}$, $W_K \in \mathbb{R}^{d\times d}$ and $W_V \in \mathbb{R}^{d\times d}$ are the query, key and value matrices respectively, and $X \in R^{N\times d}$ is the input sequence, where d is the embedding dimension and N is the length of the sequence. In preparation for later sections, we also provide an alternative form for the attention block, as proposed by Dao and Gu (2024):

$$O = MV \tag{9}$$

with $V = XW_v \in \mathbb{R}^{N\times d}$ and $M = \text{softmax}(\frac{XW_QW_K^TX^T}{\sqrt{d_{QK}}}) \in \mathbb{R}^{N\times N}$ for attention.

A.3 Recurrent Block

SSMs lie their foundations in linear (often input-dependent) dynamical systems, which can be seen as linear RNNs (Orvieto et al. (2023)). In particular, the classic recurrent block for a generic SSM takes the following form:

$$\begin{cases} h_k = A_kh_{k-1} + B_kx_k \\ o_k = C_kh_k + D_kx_k \end{cases} \tag{10}$$

where $x_k \in R^d$ is the input at time k, where $k \in \{1, 2, ..., N\}$, $h_k \in R^H$ is the hidden state and $o_k \in R^d$ is the output of the SSM layer at time k. We note the connection with the Transformer's notation $x_k := X_{k,:}$, and similarly denote $O \in R^{N\times d}$ to be the output matrix of the SSM block, i.e. $o_k = O_{k,:}$. Here, given a matrix A, we indicate $A_{k,:}$ to be the k-th row.

One fundamental difference between attention and the recurrent block is that the attention mechanism does not have the capacity for memory, so the whole sequence of previous inputs X has to be fed every time step. In contrast, the recurrent block in equation 10 is able to capture previous inputs' information, so only the last token has to be fed in. Yet, this expression also allows for a convolutional representation, where equation 10 is unrolled in time as:

$$o_k = \sum_{j=0}^{k} C_k\left(\prod_{l=0}^{k-j-1} A_{k-l}\right)B_jx_j + D_kx_k, \tag{11}$$

where $A_k \in \mathbb{R}^{H\times H}, B_k \in \mathbb{R}^{H\times d}, C_k \in \mathbb{R}^{d\times H}, D_k \in \mathbb{R}^{d\times d}$. The first versions of the SSM blocks consisted of Linear Time Invariant (LTI) representations, i.e., $A_k = A$, $B_k = B$, $C_k = C$ and $D_k = D$ for all time steps k. Recently, in order to improve the expressivity and the performance of such models, selective State Space Models have been proposed Gu and Dao (2024); Dao and Gu (2024). In these models, matrices A_k, B_k, and C_k are input-dependent.

As shown in Dao and Gu (2024), the SSM update rule in equation 11 can be equivalently expressed as in equation 9, where M is a lower triangular matrix with $M_{ji} = C_j \left(\prod_{l=0}^{j-i-1} A_{j-l}\right) B_i$ and $V = X$. In the reminder of the paper, this formulation will be used as it covers both the attention and the recurrent block.

A.4 Proof of Lower Bound of Rank Collapse Measure

Lemma A.1. *Let $A \in R^{m\times n}$ and $B \in R^{n\times p}$ two matrices. Then it holds that $||AB||_F \geq \sigma_{min}(B)||A||_F$, where $\sigma_{min}(B)$ is the smallest singular value of B.*

Proof. From Fang et al. (1994), it holds that $||CD||_F \geq \sigma_{\min}(C)||D||_F$.
By the properties of the Frobenius norm, we have that $||AB||_F = ||B^\top A^\top||_F$. Hence, we have $||B^\top A^\top||_F \geq \sigma_{\min}(B^\top)||A^\top||_F$. The result then simply follows from the fact that $\sigma_{\min}(B^\top) = \sigma_{\min}(B)$. □

Theorem 4.1 (Lower Bound on Rank Collapse)**.** *Let the input sequence $Y^{(0)}$ be such that $\mu(Y^{(0)})^2 \geq b$. If the skip connection strength λ is chosen to satisfy*

$$\lambda^2 c^2 - aS^2(C_M + |\lambda|)^2 > 0, \tag{7}$$

then we can lower bound $\mu(Y^{(K)})$ by $\mu(Y^{(K)})^2 \geq a^K \mu(Y^{(0)})^2$ for all $K \in \mathbb{N}$.

Proof. In the following, we denote $V_i^{(k)} := V_{:,i}^{(k)}$ and $\mu^{(k)} = \mu(Y^{(k)})$ for brevity.

$$\begin{aligned}
\tilde{\mu}^{(k+1)^2} &= \left|\left|\tilde{Y}^{(k+1)} - \frac{\mathbf{1}\mathbf{1}^T}{N}\tilde{Y}^{(k+1)}\right|\right|_F^2 = \left|\left|(M^{(k)} + \lambda I)V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N}(M^{(k)} + \lambda I)V^{(k)}\right|\right|_F^2 \\
&= \lambda^2\left|\left|V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N}V^{(k)}\right|\right|_F^2 + \sum_{i=1}^{d}\left|\left|M^{(k)}V_i^{(k)} - \mathbf{1}\mathbf{1}^{\boldsymbol{T}}M^{(k)}\frac{V_i^{(k)}}{N}\right|\right|_2^2 + \\
&+ \sum_{i=1}^{d} 2\lambda\left(M^{(k)}V_i^{(k)} - \mathbf{1}\mathbf{1}^{\boldsymbol{T}}M^{(k)}\frac{V_i^{(k)}}{N}\right)^T\left(V_i^{(k)} - \mathbf{1}\mathbf{1}^{\boldsymbol{T}}\frac{V_i^{(k)}}{N}\right) \\
&= \lambda^2\left|\left|V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N}V^{(k)}\right|\right|_F^2 + \sum_{i=1}^{d}\left(M^{(k)}V_i^{(k)} - \mathbf{1}\mathbf{1}^{\boldsymbol{T}}M^{(k)}\frac{V_i^{(k)}}{N}\right)^T\left(M^{(k)}V_i^{(k)} - \right. \\
&\left. - \mathbf{1}\mathbf{1}^{\boldsymbol{T}}M^{(k)}\frac{V_i^{(k)}}{N} + 2\lambda V_i^{(k)} - 2\lambda\mathbf{1}\mathbf{1}^{\boldsymbol{T}}\frac{V_i^{(k)}}{N}\right) \\
&= \lambda^2\left|\left|V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N}V^{(k)}\right|\right|_F^2 + \sum_{i=1}^{d}\left(M^{(k)}V_i^{(k)} - \mathbf{1}\mathbf{1}^{\boldsymbol{T}}M^{(k)}\frac{V_i^{(k)}}{N}\right)^T\left((M^{(k)} + 2\lambda I)V_i^{(k)} - \right. \\
&\left. \mathbf{1}\mathbf{1}^{\boldsymbol{T}}(M^{(k)} + 2\lambda I)\frac{V_i^{(k)}}{N}\right) \\
&= \lambda^2\left|\left|V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N}V^{(k)}\right|\right|_F^2 + \sum_{i=1}^{d}\left(V_i^{(k)^T}M^{(k)^T}(M^{(k)} + 2\lambda I)V_i^{(k)} + \right. \\
&+ \frac{V_i^{(k)^T}}{N}M^{(k)^T}\mathbf{1}\mathbf{1}^{\boldsymbol{T}}(\mathbf{1}\mathbf{1}^{\boldsymbol{T}})^T M^{(k)}\frac{V_i^{(k)}}{N} + \frac{V_i^{(k)^T}}{N}M^{(k)^T}\mathbf{1}\mathbf{1}^{\boldsymbol{T}}(M^{(k)} + 2\lambda I)V_i^{(k)} - \\
&\left. - V_i^{(k)^T}M^{(k)^T}\mathbf{1}\mathbf{1}^{\boldsymbol{T}}(M^{(k)} + 2\lambda I)\frac{V_i^{(k)}}{N}\right) \\
&= \lambda^2\left|\left|V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N}V^{(k)}\right|\right|_F^2 + \sum_{i=1}^{d}V_i^{(k)^T}\left[M^{(k)^T}(M^{(k)} + 2\lambda I) + \frac{1}{N}M^{(k)^T}\mathbf{1}\mathbf{1}^{\boldsymbol{T}}M^{(k)} - \right.
\end{aligned}$$

$$
\begin{aligned}
&- \frac{2}{N} M^{(k)^T} \mathbf{1}\mathbf{1}^{\boldsymbol{T}} (M^{(k)} + 2\lambda I) \Bigg] V_i^{(k)} \\
&= \lambda^2 \left|\left| V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N} V^{(k)} \right|\right|_F^2 + \sum_{i=1}^{d} V_i^{(k)^T} \Bigg[M^{(k)^T} \left(I - \frac{1}{N} \mathbf{1}\mathbf{1}^{\boldsymbol{T}} \right) M^{(k)} + \\
&+ 2\lambda M^{(k)^T} \left(I - \frac{2}{N} \mathbf{1}\mathbf{1}^{\boldsymbol{T}} \right) \Bigg] V_i^{(k)} \\
&\geq \lambda^2 \left|\left| V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N} V^{(k)} \right|\right|_F^2 + \sum_{i=1}^{d} V_i^{(k)^T} \left[2\lambda M^{(k)^T} \left(I - \frac{2}{N} \mathbf{1}\mathbf{1}^{\boldsymbol{T}} \right) \right] V_i^{(k)}
\end{aligned}
$$

where the last inequality follows from the fact that $M^{(k)^T} \left(I - \frac{1}{N} \mathbf{1}\mathbf{1}^{\boldsymbol{T}} \right) M^{(k)}$ is Symmetric and Positive Semi-Definite.
Continuing from here, we get:

$$
\begin{aligned}
\tilde{\mu}^{(k+1)^2} &\geq \lambda^2 \left|\left| V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N} V^{(k)} \right|\right|_F^2 - 2\lambda \sum_{i=1}^{d} ||V_i^{(k)}||^2 \, ||M^{(k)}|| \left|\left| \left(I - \frac{2}{N} \mathbf{1}\mathbf{1}^{\boldsymbol{T}} \right) \right|\right| \\
&\geq \lambda^2 \left|\left| V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N} V^{(k)} \right|\right|_F^2 - 2\lambda \sum_{i=1}^{d} ||V_i^{(k)}||^2 \, ||M^{(k)}|| \\
&\geq \lambda^2 \left|\left| V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N} V^{(k)} \right|\right|_F^2 - 2\lambda \sum_{i=1}^{d} ||V_i^{(k)}||^2 \, ||M^{(k)}||_F \\
&= \lambda^2 \left|\left| V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N} V^{(k)} \right|\right|_F^2 - 2\lambda ||V^{(k)}||_F^2 \, ||M^{(k)}||_F \\
&\geq \lambda^2 \left|\left| V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N} V^{(k)} \right|\right|_F^2 - 2\lambda N \sigma_{\max}(C_V^{(k)})^2 ||M^{(k)}||_F
\end{aligned}
$$

where the last inequality follows since $||V^{(k)}||_F \leq ||Y^{(k)}||_F \sigma_{\max}(C_V^{(k)})$ and since we are using LayerNorm and hence every row of $Y^{(k)}$ has norm 1.

Furthermore, from Lemma A.1, we have that:

$$
\left|\left| V^{(k)} - \frac{\mathbf{1}\mathbf{1}^T}{N} V^{(k)} \right|\right|_F^2 \geq \sigma_{\min}(C_V^{(k)})^2 \mu^{(k)^2}
$$

Hence, we have that:

$$
\tilde{\mu}^{(k+1)^2} \geq c^2 \lambda^2 \mu^{(k)^2} - 2\lambda N S^2 ||M^{(k)}||_F
$$

We now need to analyze the relationship between $\mu^{(k+1)}$ and $\tilde{\mu}^{(k+1)}$.

$$
\begin{aligned}
\mu^{(k+1)^2} &= \left|\left| D^{(k+1)} \tilde{Y}^{(k+1)} - \frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{N} D^{(k+1)} \tilde{Y}^{(k+1)} \right|\right|_F^2 = \left|\left| \left(I - \frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{N} \right) D^{(k+1)} \tilde{Y}^{(k+1)} \right|\right|_F^2 \\
&= \sum_{j=1}^{N} \sum_{i=1}^{d} \left(\sum_{l=1}^{N} \left(I - \frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{N} \right)_{jl} \frac{\tilde{Y}_{il}^{(k+1)}}{||\tilde{Y}_{i,:}^{(k+1)}||_2} \right)^2 \\
&\geq \sum_{j=1}^{N} \sum_{i=1}^{d} \left(\sum_{l=1}^{N} \left(I - \frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{N} \right)_{jl} \frac{\tilde{Y}_{il}^{(k+1)}}{\max_{i \in [N]} ||\tilde{Y}_{i,:}^{(k+1)}||_2} \right)^2
\end{aligned}
$$

$$\geq \frac{1}{\max_{i\in[N]} ||\tilde{Y}_{i,:}^{(k+1)}||_2^2} \sum_{j=1}^{N}\sum_{i=1}^{d}\left(\sum_{l=1}^{N}\left(I-\frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{N}\right)_{jl}\tilde{Y}_{il}^{(k+1)}\right)^2$$
$$= \frac{1}{\max_{i\in[N]} ||\tilde{Y}_{i,:}^{(k+1)}||_2^2}\tilde{\mu}^{(k+1)^2}$$

Combining the previous two results, we then get:

$$\mu^{(k+1)^2} \geq \frac{1}{\max_{i\in[N]} ||\tilde{Y}_{i,:}^{(k+1)}||_2^2}\left(c^2\lambda^2\mu^{(k)^2} - 2\lambda N S^2 ||M^{(k)}||_F\right)$$

We will now proceed in upper bounding $\max_{i\in[N]} ||\tilde{Y}_{i,:}^{(k+1)}||_2^2$.

$$\max_{i\in[N]} ||\tilde{Y}_{i,:}^{(k+1)}||_2^2 = \max_{i\in[N]}\sum_j \tilde{Y}_{ij}^{(k+1)^2} = \max_{i\in[N]}\sum_j\left(\sum_l\left(M^{(k)}+\lambda I\right)_{il}V_{lj}^{(k)}\right)^2$$
$$\leq \max_{i\in[N]}\sum_l\left(M^{(k)}+\lambda I\right)_{il}^2\sum_j\sum_m \underbrace{Y_{lm}^{(k)^2}}_{\leq 1} C_{V_{mj}}^2$$
$$\leq S^2(||M^{(k)}||_F + |\lambda|)^2$$

Using the fact that $||M^{(k)}||_F \leq C_M$, we get:

$$\mu^{(k+1)^2} \geq \frac{1}{S^2(C_M+|\lambda|)^2}\left(\lambda^2c^2\mu^{(k)^2} - 2\lambda N S^2 C_M\right)$$

To guarantee that $\mu^{(k+1)^2} \geq a\mu^{(k)^2}$ for some $a < 1$, we can simply lower bound the right hand side. Hence, we want that $\frac{1}{S^2(C_M+|\lambda|)^2}\left(\lambda^2c^2\mu^{(k)^2} - 2\lambda N S^2 C_M\right) \geq a\mu^{(k)^2}$.
Hence, we need:

$$\left(\lambda^2c^2 - aS^2(C_M+|\lambda|)^2\right)\mu^{(k)^2} \geq 2\lambda N C_M$$

To do so, we need to guarantee that $\lambda^2c^2 - aS^2(C_M+|\lambda|)^2 > 0$. Hence, if λ satisfies the condition above and we have that $\mu^{(k)^2} \geq \frac{2\lambda N S^2 C_M}{\lambda^2c^2 - aS^2(C_M+|\lambda|)^2}$, then $\mu^{(k+1)^2} \geq a\mu^{(k)^2}$. Note that in order to satisfy the condition on $\mu^{(k)^2}$ up until time K, we need to choose $\mu^{(0)^2} \geq \frac{1}{a^K}\frac{2\lambda N S^2 C_M}{\lambda^2c^2 - aS^2(C_M+|\lambda|)^2}$, which concludes the proof.

□

A.5 Theorem on Rank Collapse in Self-Attention Only Transformers with LayerNorm

We first define some notation that is used in the Theorem. First, we denote $\phi^{(t)} = Y_{i,:}^{(t)}, Y_{j,:}^{(t)}\rangle$ where $\langle,\cdot,\cdot,\rangle$ is the inner product. Then, $\mathcal{G}$ represents the graph induced by the attention mask. In particular, if there is a directed edge from j to i, this means that token i attends to token j. A Quasi-Strongly Connected graph is a graph $\mathcal{G}$ in which there exists a node from which every other node in the directed graph $\mathcal{G}$ is reachable (i.e. there exist a directed path connecting the two nodes). Finally, the authors consider the following assumptions:

Assumption A.1. *$\mathcal{G}$ contains self-loops, i.e. every token in the sequence attends to itself*

Assumption A.2. *There exists a constant $C \in \mathbb{R}$ such that* $\max_{t\in\mathbb{N}}\left\{||W_Q^{(t)}||_2, ||W_K^{(t)}||_2\right\} \leq C$

Theorem A.2 (Corollary 1, Wu et al. (2024a)). *Consider the architecture defined in 6 by choosing $\lambda = 0$, i.e. a Self-Attention Network with LayerNorm and no skip connection. Let $\mathcal{G}$ be a quasi-strongly connected graph. Under A.1 and A.2, if $W_V^{(t)}$ is orthogonal for all $t \geq 0$ and $\phi^{(0)} \geq 0$, there exist $C > 0$ and $\epsilon > 0$ such that $N\epsilon < 1$ and*

$$\mu(Y^{(K)}) \leq C(1 - \epsilon^{2r})^{\frac{K}{2r}}, \quad \forall K \geq 0.$$

where r is the diameter of $\mathcal{G}$, meaning that tokens converge to a common point on $\mathbb{S}^{d-1}$ exponentially.

A.6 Theorem on Rank Collapse in Self-Attention only Transformers without LayerNorm

Let $\text{res}(Y) = Y - \frac{\mathbf{1}\mathbf{1}^\top Y}{N}$. Note that with this notation, we can redefine $\mu(Y) = ||\text{res}(Y)||_F$. Then, it holds that:

Theorem A.3 (Rank Collapse Transformers Dong et al. (2023)). *For any single-head Self-Attention network consisting of K layers with $\left\|\boldsymbol{W}_Q^{(k)} W_K^{(k)^\top}\right\|_1 \left\|\boldsymbol{W}_V^{(k)}\right\|_{1,\infty} \leq \beta$ and for a term γ that depends on the attention entries, we have that*

$$\|\operatorname{res}(Y^{(K)})\|_{1,\infty} \leq \left(\frac{4\gamma\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^K-1}{2}} \|\operatorname{res}(Y^{(0)})\|_{1,\infty}^{3^K}$$

which amounts to a doubly exponential convergence to a rank-1 matrix.

Corollary A.3.1. *Under the same conditions of Theorem A.3, it holds that*

$$||\operatorname{res}(Y^{(K)})||_F \leq \left(\frac{4\gamma\sqrt{\min(N,d)}\beta}{\sqrt{d_{qk}}}\right)^{\frac{3^K-1}{2}} ||\operatorname{res}(Y^{(0)})||_F^3.$$

Proof. Follows straightforward from the relations between Frobenius norm and the 1- and ∞-norms, i.e. given a matrix we have $A \in R^{N\times d}$ $||A||_F \geq ||A||_1$, $||A||_F \geq ||A||_\infty$, $||A||_F \leq \min(N,d)||A||_1$, $||A||_F \leq \min(N,d)||A||_\infty$. □

A.7 Proof of Rank Collapse for Selective SSMs with LayerNorm but without Skip Connections

Lemma A.4. *If $\alpha \leq 1$, We have that $M_{ij}^{(k)} \geq \lambda_{min} \cdot \phi^{(k)}\alpha^N$ for $i \geq j$*

Proof. If $i \geq j$, we have that:

$$M_{ij}^{(k)} = \alpha^{i-j} Y_{i,:}^{(k)^\top} W_C W_B^\top Y_{j,:}^{(k)} \geq \alpha^{i-j}\lambda_{\min}\langle Y_{i,:}^{(k)}, Y_{j,:}^{(k)}\rangle \geq \alpha^{j-i}\lambda_{\min} \cdot \phi^{(k)}$$

Note that the worst case is when $i - j = N - 1$, which leads to the desired lower bound by lower bounding $\alpha^{i-j} \geq \alpha^{N-1} \geq \alpha^N$ since $a \leq 1$. □

Lemma A.5. *We have that $D_{i,i}^{(k)} \geq \frac{1}{N^{\frac{3}{2}}\lambda_{max}}$ $\forall k$*

Proof. From the definition of LayerNorm, we have $D_{i,i}^{(k)} = \frac{1}{||\tilde{Y}_{i,:}^{(k)}||_2}$. Furthermore, we have that $||\tilde{Y}_{i,:}^{(k)}||_2 = ||\sum_{j=1}^N M_{ij}^{(k-1)} Y_{j,:}^{(k-1)}||_2 \leq \sum_{j=1}^N M_{ij}^{(k-1)} \underbrace{||Y_{j,:}^{(k-1)}||_2}_{=1} \leq N\lambda_{\max}$. □

Theorem 4.3 (Rank Collapse for selective SSMs without skip connection). *Let $\phi^{(k)} = \min_{i,j\in[N]}\langle Y_{i,:}^{(k)}, Y_{j,:}^{(k)}\rangle$, where $\langle\cdot,\cdot\rangle$ indicates the inner product. Under Assumption 4.1 and if $c \leq \phi^{(0)} < 1$ for some $c > 0 \in \mathbb{R}$, $\lambda_{min} > 0$, $\sum_j M_{ij} \geq 1$ $\forall i$, then it holds that:*

$$\mu(Y^{(K)}) \leq \sqrt{N}\left(1 - c^2\lambda_{min}^2\alpha^{2N}\right)^K \ \forall K \geq 0,$$

where λ_{min} and λ_{max} the minimum and maximum eigenvalues of $\frac{W_B W_C^\top + W_C W_B^\top}{2}$.

Proof. For the fist part of the proof, we will follow the proof of Corollary 1 in Wu et al. (2024a):

$$\begin{aligned}
\phi^{(k+1)} &= \min_{i,j\in[N]} \langle Y^{(k+1)}_{i,:}, Y^{(k+1)}_{j,:}\rangle = \langle Y^{(k+1)}_{i^{(k+1)},:}, Y^{(k+1)}_{j^{(k+1)},:}\rangle \\
&= \left\langle \sum_{k_1=1}^{N} \left(D^{(k)}M^{(k)}\right)_{i^{(k+1)}k_1} Y^{(k)}_{k_1,:}, \sum_{l_1=1}^{N} \left(D^{(k)}M^{(k)}\right)_{j^{(k+1)}l_1} Y^{(k)}_{l_1,:} \right\rangle \\
&\geq \frac{1}{N\lambda_{\max}} \left\langle \sum_{k_1=1}^{N} M^{(k)}_{i^{(k+1)}k_1} Y^{(k)}_{k_1,:}, \sum_{l_1=1}^{N} M^{(k)}_{j^{(k+1)}l_1} Y^{(k)}_{l_1,:} \right\rangle
\end{aligned}$$

where the last inequality follows from Lemma A.5.

From here, using Lemma A.4 and the assumption that $\lambda_{\max} \leq \frac{1}{N}$ we have that:

$$\begin{aligned}
\phi^{(k+1)} &\geq \sum_{k_1=1}^{N}\sum_{l_1=1}^{N} M^{(k)}_{i^{(k+1)}k_1} M^{(k)}_{j^{(k+1)}l_1} \langle Y^{(k)}_{k_1,:}, Y^{(k)}_{l_1,:}\rangle \\
&= \sum_{m=1}^{N} M^{(k)}_{i^{(k+1)}m} M^{(k)}_{j^{(k+1)}m} \underbrace{\langle Y^{(k)}_{m,:}, Y^{(k)}_{m,:}\rangle}_{=1} + \sum_{k_1=1}^{N} \sum_{l_1=1, l_1\neq k_1}^{N} M^{(k)}_{i^{(k+1)}k_1} M^{(k)}_{j^{(k+1)}l_1} \underbrace{\langle Y^{(k)}_{k_1,:}, Y^{(k)}_{l_1,:}\rangle}_{\geq \phi^{(k)}} \\
&\geq \lambda^2_{\min} {\phi^{(k)}}^2 \alpha^{2N} + (1 - \lambda^2_{\min}\alpha^{2N}{\phi^{(k)}}^2)\phi^{(k)} \\
&\geq c^2\lambda^2_{\min}\alpha^{2N} + \left(1 - c^2\lambda^2_{\min}\alpha^{2N}\right)\phi^{(k)}
\end{aligned}$$

where the last two inequalities follow since $\sum_j M_{ij} \geq 1\ \forall i$ and $\phi^{(k)} \geq c$ respectively. Note also that by iterating the above recurrence starting from $t = 1$, it is easy to show by induction that since $\phi^{(0)} \geq c$, then $\phi^{(k)} \geq c\ \forall k$. By rearranging the terms, we have that:

$$1 - \phi^{(k+1)} \leq \left(1 - c^2\lambda^2_{\min}\alpha^{2N}\right)\left(1 - \phi^{(k)}\right)$$

By recursively unrolling the relationship, we get:

$$\begin{aligned}
1 - \phi^{(K)} &\leq \left(1 - c^2\lambda^2_{\min}\alpha^{2N}\right)^K \left(1 - \phi^{(0)}\right) \\
&\leq \left(1 - c^2\lambda^2_{\min}\alpha^{2N}\right)^K
\end{aligned}$$

since $0 < c \leq \phi^{(0)} < 1$

Since by definition, $1 - \phi^{(k)} \geq 1 - \langle Y^{(k)}_{i,:}, Y^{(k)}_{j,:}\rangle \geq \|Y^{(k)}_{i,:} - Y^{(k)}_{j,:}\|_2^2/2$ it follows that:

$$\begin{aligned}
\mu(Y^{(K)}) = \|Y^{(K)} - \mathbf{1}\mathbf{1}^\top Y^{(K)}/N\|_F &= \sqrt{\sum_{i=1}^{N} \|Y^{(K)}_{i,:} - \mathbf{1}^\top Y^{(K)}/N\|_2^2} \\
&= \sqrt{\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} \|Y^{(K)}_{i,:} - Y^{(K)}_{j,:}\|_2^2} \\
&\leq \sqrt{N}\left(1 - c^2\lambda^2_{\min}\alpha^{2N}\right)^K
\end{aligned}$$

□

A.8 Rank Collapse for LTI SSMs with no Skip Connections and LayerNorm

Lemma A.6.

$$||Y^{(k)}||_F \leq ||W||^k ||Y^0||_F$$

Proof. We have that $||Y^{(k+1)}||_F = ||M^{(k)}Y^{(k)}||_F \leq ||W|| ||Y^{(k)}||_F$. By simply solving the recurrence we get the desired result. □

We now analyze rank collapse for Structured LTI SSMs. For simplicity of exposition, here we neglect the gating mechanism and focus on the recurrent block. In particular, we will analyze a network consisting of K layers of stacked Structured LTI SSM blocks. We recall that for structured LTI SSM blocks, which satisfy with $A_t = a_t I$, we have that the matrix M takes the form $M^{(k)} = W_A^{(k)} \odot (W_C^{(k)} {W_B^{(k)}}^T) := W^{(k)}$, where $W_A = \begin{bmatrix} 1 & & & & \\ a_1 & 1 & & & \\ a_2 a_1 & a_2 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ a_{N-1} \dots a_1 & a_{N-1} \dots a_2 & \dots & a_{N-1} & 1 \end{bmatrix}$.
We first consider the case where all the layers have the same parameters, i.e. it holds that $W^{(k)} = W \; \forall k$ for some matrix W. We then have the following result:

Theorem A.7 (Rank Collapse Structured LTI SSMs)**.** *For a K layer network of only Structured LTI SSM blocks where all layers have the same parametrization, it holds that $\mu(Y^{(k)}) \leq ||W||^K \; ||Y^{(0)}||_F$*

Proof. We have that:

$$M^{(k)}Y^{(k)} - 1\gamma_{k+1} = \left(I - \frac{11^\top}{N}\right) WY^{(k)}$$

From here, by using Lemma A.6, we get:

$$\begin{aligned} \mu(Y^{(k+1)}) &= ||M^{(k)}Y^{(k)} - 1\gamma_{k+1}||_F \\ &\leq \left|\left|I - \frac{11^\top}{N}\right|\right| \left|\left|WY^{(k)}\right|\right|_F \leq ||Y^{(k+1)}||_F \\ &\leq ||W||^{k+1} ||Y^{(0)}||_F \end{aligned}$$

□

Hence, whenever $||W||_F < 1$, we have rank collapse for the case of structured LTI-SSMs. Note that however here the overall dynamics and behavior is different: the convergence is exponential rather than doubly exponential like in the Attention case. This is due to the fact that the matrix M is not input dependent, which is what causes the double exponential convergence. We will see that for selective SSMs, which also have a similar input dependence to attention (apart from the softmax) for the matrix M, the convergence will again be doubly exponential. Furthermore, note that here the initial input sequence does not have any influence on rank collapse, whereas in the attention case this was not the case (since there was also a doubly exponential dependence on $||\text{res}(Y^{(0)})||_F$). Again, we will see a similar behavior in the selective SSM setting.
For the case where we have different parametrizations for different layers, we then have:

Theorem A.8 (Rank Collapse Structured LTI SSMs General)**.** *For a K layer network of only Structured LTI SSM blocks it holds that $\mu(Y^{(K)}) \leq \prod_{k=1}^K ||W^{(k)}|| \; ||Y^{(0)}||_F$*

Proof. From the proof of Lemma A.6, we have that $||Y^{(k+1)}||_F = ||M^{(k)}Y^{(k)}||_F \leq ||W^{(k)}|| \; ||Y^{(k)}||_F$. Applying this recursively gives us $||Y^{(K)}||_F \leq \prod_{k=1}^K ||W^{(k)}|| \; ||Y^{(0)}||_F$. Then, similarly to the proof of TheoremA.7, we have:

$$
\begin{aligned}
\mu(Y^{(K+1)}) &= ||M^{(K)}Y^{(K)} - 1\gamma_{K+1}||_F \\
&\leq \left|\left| I - \frac{11^\top}{N} \right|\right| \left|\left| W^{(K)}Y^{(K)} \right|\right|_F \leq \sqrt{N}||Y^{(K+1)}||_F \\
&\leq \prod_{k=1}^{K+1} ||W^{(k)}|| \; ||Y^{(0)}||_F
\end{aligned}
$$

□

In this case, one simple sufficient condition to have exponential convergence of the rank collapse measure is that $||W^{(k)}||_F < 1 \; \forall k$.

A.9 Proof of Upper Bound of Rank Collapse Measure for Selective SSMs without LayerNorm and Skip Connections

Lemma A.9. *For a selective SSM, it holds that:*

$$
||Y^{(k)}||_F \leq \left(\sqrt{N}||W_{BC}||_F\right)^{\frac{3^k-1}{2}} ||Y^{(0)}||_F^{3^k}
$$

where $W_{BC} = W_C W_B^\top$

Proof. We have that:

$$
\begin{aligned}
||Y^{(k+1)}||_F &= ||M^{(k)}Y^{(k)}||_F \leq ||M^{(k)}||_F \; ||Y^{(k)}||_F \\
&= ||1\text{SS}(\alpha) \odot \left(Y^{(k)} W_C W_B^\top Y^{(k)^T}\right)||_F \; ||Y^{(k)}||_F \\
&\leq ||1\text{SS}(\alpha)||_F ||Y^{(k)} W_C W_B^\top Y^{(k)^T}||_F \; ||Y^{(k)}||_F \\
&\leq \sqrt{N}||W_{BC}||_F ||Y^{(k)}||_F^3
\end{aligned}
$$

By simply solving the recurrence we get the desired result. □

Theorem A.10 (Rank Collapse for Selective SSMs)**.** *Given a selective SSM with same parametrization for each layer with no skip connection, it holds that:*

$$
\mu(Y^{(k)}) \leq \left(\sqrt{N}||W_{BC}||_F\right)^{\frac{3^{k-1}+1}{2}} ||Y^{(0)}||_F^{3^k}.
$$

Proof. We start by observing that:

$$
\begin{aligned}
\mu(Y^{(k)}) &= \left|\left| \left(I - \frac{11^\top}{N}\right) M^{(k-1)}Y^{(k-1)} \right|\right|_F \leq \sqrt{N}||M^{(k-1)}||_F \; ||Y^{(k-1)}||_F \\
&\leq \sqrt{N}||W_{BC}|| \; ||Y^{(k-1)}||_F^3
\end{aligned}
$$

By using Lemma A.9, we get that:

$$
\mu(Y^{(k)}) \leq \left(\sqrt{N}||W_{BC}||_F\right)^{\frac{3^{k-1}+1}{2}} ||Y^{(0)}||_F^{3^k}
$$

proving the result. □

A.10 Proof of Counterexample for LTI SSMs

Proposition 4.3.1. *Given [Sys-1], if we choose $\alpha = 2$, i.e. $M = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$, $\lambda > -2$ and $Y^{(0)} = I$, then we have that $\mu(Y^{(k)}) \to 0$. Conversely, given the choices above with $\lambda < -2$ (with $\lambda \neq -1$, then rank collapse does not happen, i.e. $\mu(Y^{(k)}) \not\to 0$.*

Proof. We will prove the theorem by induction.
In particular, we will prove that $Y^{(k)}$ will take the following form:

$$Y^{(k)} = \begin{pmatrix} 1 & 0 \\ \sqrt{\frac{\alpha_k - 1}{\alpha_k}} & \frac{1}{\sqrt{\alpha_k}} \end{pmatrix} \tag{12}$$

where $\alpha_0 = 1$ and α_k satisfy the recurrence: $\alpha_{k+1} = \alpha_k \left(1 + \frac{4}{(1+\lambda)^2}\right) + \frac{4}{1+\lambda}\sqrt{\alpha_k - 1}\sqrt{\alpha_k}$. For $\lambda > -1$, we simply have that $\alpha_{k+1} \geq \alpha_k \left(1 + \frac{4}{(1+\lambda)^2}\right)$, implying that $\alpha_k \geq \alpha_0 \left(1 + \frac{4}{(1+\lambda)^2}\right)^k$, which causes α_k to diverge. Similarly, if $-2 < \lambda < -1$, we have that $\alpha_{k+1} \geq \alpha_k \left(1 + \frac{2}{1+\lambda}\right)^2$, i.e. $\alpha_k \geq \alpha_0 \left(\frac{3+\lambda}{1+\lambda}\right)^{2k}$, which also diverges. This implies that $Y^{(k)} \to \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$, which is a rank one matrix, hence also implying that $\mu(Y^{(k)}) \to 0$.

Let's consider now the case where $\lambda < -2$. We will prove that the sequence α_k remains bounded from above. Note that this automatically proves that rank collapse does not happen, since a necessary condition for this is that $\alpha_k \to \infty$. Let define $r = \frac{1}{1+\lambda}$ and $\beta = \frac{1}{1-r^2} > 1$ for $\lambda < -2$. We prove by induction that $\alpha_k \leq \beta \forall k$. Since $\beta > 1$, the statement clearly holds for $k = 0$. Suppose now that the statement holds for k. We prove that the statement holds for $k+1$. We have that $\alpha_{k+1} \leq \beta(1 + 4r^2) + 4r\sqrt{\beta}\sqrt{\beta - 1}$. We want the quantity on the right hand side to be smaller than β. This is equivalent to saying that $r^2\beta \leq -r\sqrt{\beta}\sqrt{\beta - 1}$. By squaring both sides and solving the inequality we can see that $\beta = \frac{1}{1-r^2}$ satisfies this inequality.

We now go on to prove the above statement in Equation 12 about the structure of $Y^{(k)}$ using induction. Clearly, the statement is true for $k = 0$. Now, let's assume the statement holds for k. We will prove that then the statement holds for $k+1$. We have that:

$$\tilde{Y}^{(k+1)} = (M + \lambda I)Y^{(k)} = \begin{pmatrix} 1+\lambda & 0 \\ 2 & 1+\lambda \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \sqrt{\frac{\alpha_k - 1}{\alpha_k}} & \frac{1}{\sqrt{\alpha_k}} \end{pmatrix} = \begin{pmatrix} 1+\lambda & 0 \\ \frac{(1+\lambda)\sqrt{\alpha_k - 1} + 2\sqrt{\alpha_k}}{\sqrt{\alpha_k}} & \frac{1+\lambda}{\sqrt{\alpha_k}} \end{pmatrix}$$

Hence, after we apply LayerNorm to $\tilde{Y}^{(k+1)}$, by letting $\gamma = \left(4 + (1+\lambda)^2\right)\alpha_k + 4(1+\lambda)\sqrt{\alpha_k}\sqrt{\alpha_k - 1}$, we get that:

$$Y^{(k+1)} = \begin{pmatrix} 1 & 0 \\ \frac{(1+\lambda)\sqrt{\alpha_k - 1} + 2\sqrt{\alpha_k}}{\sqrt{\gamma}} & \frac{1+\lambda}{\sqrt{\gamma}} \end{pmatrix}$$

Choosing $\alpha_{k+1} = \left(1 + \frac{4}{(1+\lambda)^2}\right)\alpha_k + \frac{4}{1+\lambda}\sqrt{\alpha_k - 1}\sqrt{\alpha_k}$ clearly results to $k+1$ being of the form

$$Y^{(k+1)} = \begin{pmatrix} 1 & 0 \\ \sqrt{\frac{\alpha_{k+1} - 1}{\alpha_{k+1}}} & \frac{1}{\sqrt{\alpha_{k+1}}} \end{pmatrix}$$

which proves the induction hypothesis. □

A.11 Proof of Counterexample for Selective SSMs

Proposition 4.3.2. *Given [Sys-2], if we then choose $\lambda > -\frac{3}{2}$, M as above and $Y^{(0)} = \begin{pmatrix} 1 & 0 \\ \frac{\alpha_0}{\sqrt{\alpha_0^2+\beta_0^2}} & \frac{\beta_0}{\sqrt{\alpha_0^2+\beta_0^2}} \end{pmatrix}$ for any α_0, β_0 with $\alpha_0 > 0$, then we have that $\mu(Y^{(k)}) \to 0$. On the other hand, if we choose $\lambda < -\frac{3}{2}$ and M and $Y^{(0)}$ as above, then rank collapse is avoided.*

Proof. Again, we proceed by induction. In particular, we will show that under the assumptions of the theorem, $Y^{(k)}$ takes the following form:

$$Y^{(k)} = \begin{pmatrix} 1 & 0 \\ \frac{(2+\lambda)^k\alpha_0}{\sqrt{(2+\lambda)^{2k}\alpha_0^2+(1+\lambda)^{2k}\beta_0^2}} & \frac{(1+\lambda)^k\beta_0}{\sqrt{(2+\lambda)^{2k}\alpha_0^2+(1+\lambda)^{2k}\beta_0^2}} \end{pmatrix}$$

Note that if this result holds, as long as $\alpha_0 > 0$ and $|2+\lambda| > |1+\lambda|$ (which happens iff $\lambda > -\frac{3}{2}$), we have that $Y^{(k)} \to \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$, which is a rank one matrix, hence also implying that $\mu(Y^{(k)}) \to 0$. On the other hand, if $\lambda < -\frac{3}{2}$ we have that $Y^{(k)}$, as k increases, tends to oscillate between the matrices $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. Although in this case the limit of $Y^{(k)}$ does not exist, we have that both the matrices above have full rank and hence rank collapse does not happen.

We now proceed to prove the statement above. Clearly, the statement is true for $k = 0$. Let's suppose that the statement holds for k. We will prove that it holds for $k+1$. We have that under the choice of M and λ for the theorem, we have that:

$$\begin{aligned} \tilde{Y}^{(k+1)} &= \begin{pmatrix} 1+\lambda & 0 \\ \frac{(2+\lambda)^k\alpha_0}{\sqrt{(2+\lambda)^{2k}\alpha_0^2+(1+\lambda)^{2k}\beta_0^2}} & 1+\lambda \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{(2+\lambda)^k\alpha_0}{\sqrt{(2+\lambda)^{2k}\alpha_0^2+(1+\lambda)^{2k}\beta_0^2}} & \frac{(1+\lambda)^k\beta_0}{\sqrt{(2+\lambda)^{2k}\alpha_0^2+(1+\lambda)^{2k}\beta_0^2}} \end{pmatrix} \\ &= \begin{pmatrix} 1+\lambda & 0 \\ \frac{(2+\lambda)^{k+1}\alpha_0}{\sqrt{(2+\lambda)^{2k}\alpha_0^2+(1+\lambda)^{2k}\beta_0^2}} & \frac{(1+\lambda)^{k+1}\beta_0}{\sqrt{(2+\lambda)^{2k}\alpha_0^2+(1+\lambda)^{2k}\beta_0^2}} \end{pmatrix} \end{aligned}$$

Hence, after applying LayerNorm to $\tilde{Y}^{(k+1)}$, we get that:

$$Y^{(k)} = \begin{pmatrix} 1 & 0 \\ \frac{(2+\lambda)^{k+1}\alpha_0}{\sqrt{(2+\lambda)^{2k+2}\alpha_0^2+(1+\lambda)^{2k+2}\beta_0^2}} & \frac{(1+\lambda)^{k+1}\beta_0}{\sqrt{(2+\lambda)^{2k+2}\alpha_0^2+(1+\lambda)^{2k+2}\beta_0^2}} \end{pmatrix}$$

which proves the claim and finishes the proof. □

A.12 Additional Experiments

A.12.1 Comparison of Rank Collapse for the S4 Architecture with and without skip connection and LayerNorm

In the following, we consider a different SSM model, i.e. the S4 architecture Gu et al. (2022b), to analyze the effect of skip connections and LayerNorm on a different model. In particular, we use the S4 variant with diagonal A, i.e. S4D Gu et al. (2022a). We train the S4D architecture with 32 layers and 1.6 million parameters on the Cifar10 dataset. We then sample 32 images from this dataset and forward pass them through the trained model. Again, we calculate the normalized rank collapse measure μ at each layer. We repeat this for all considered cases, namely skip connections plus LayerNorm, only skip connections, only LayerNorm, and neither skip connections or LayerNorm.

Figure 4 shows that for S4D, even when removing the skip connection and the LayerNorm from the model, rank collapse is not observed. This can be explained by the fact that the condition on the Frobenius norm of the weight matrix being smaller than 1 to guarantee rank collapse is

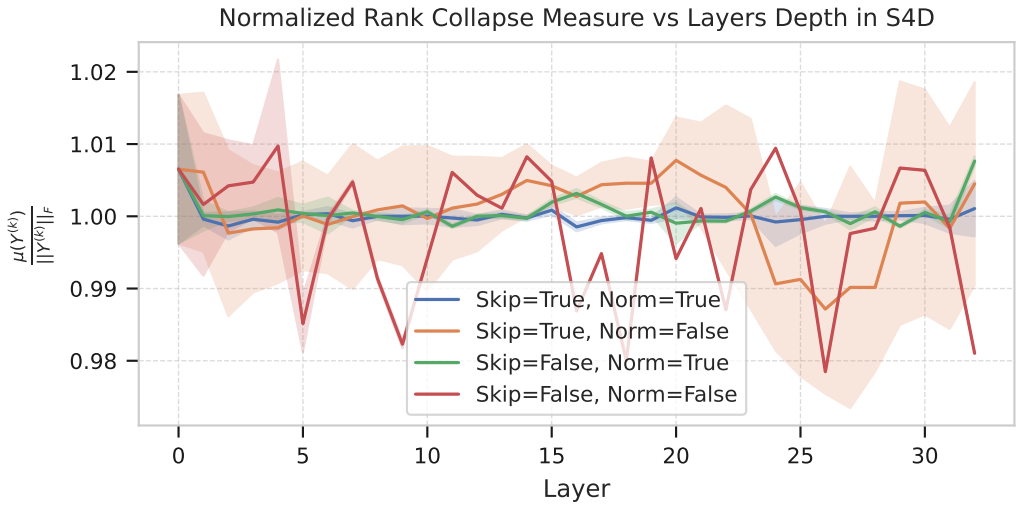

Figure 4: (Normalized) Rank Collapse measure as a function of layer depth for the S4D model. The shaded areas represent one standard deviation from the mean calculated over the 32 examples. Skip=True/False indicates we use a S4D architecture with/without skip connection whereas LayerNorm=True/False indicate we use a S4D architecture with/without LayerNorm

quite stringent in practice and hence it is not met by the model. Additionally, note that while for Transformers and for Mamba-2 the observation of rank collapse depended (at least theoretically) on both the values of the weight matrices and the inputs values, for standard SSMs (e.g. S4) rank collapse only depends on the weight matrix. This renders the SSM model more "robust" to rank collapse, in the sense that to observe rank collapse one needs to find a precise parametrization of the model and cannot simply find an "adversarial" input sequence causing rank collapse (which instead could be done for every parametrization of Transformers and selective SSMs).

A.13 Experimental Details

For both the LRA image task and the MQAR task we use the standard code bases provided online.[5] We use the standard transformer architecture, the linear attention model proposed in Katharopoulos et al. (2020), the Mamba architecture (Gu and Dao, 2024), and the Mamba-2 architecture (Dao and Gu, 2024). For the MQAR experiments we use the following training protocol:

- **Optimizer and schedule:** Weight decay of 0.1, linear warmup with duration of 10%, AdamW optimizer (Loshchilov and Hutter, 2019). For each run, we sweep the learning rates in `np.logspace`$(-4, -2, 4)$ and train for 64 epochs. This is the same setup as in (Arora et al., 2023).
- **Initialization:** For all models we use their standard initialization and initialize $\lambda = -1$ in each layer.
- **Training duration:** We use a global batch size of 64.
- **Width and depth:** For all runs, we use two layers (each with a sequence model and a MLP, interleaved with layer normalization). The model dimensions $d = 128$, state dimension $n = 64$, sequence length $L = 512$, and number of KV pairs $= 64$ are kept constant for all four architectures.
- **Position information:** Positional embeddings (Brown et al., 2020) are used for the attention architectures, but not for the SSM architecture classes. This is the same setup as in (Arora et al., 2023).

[5] `https://github.com/HazyResearch/zoology` and `https://github.com/google-research/long-range-arena`

- **Data:** Each model is trained on 100,000 datapoints and evaluated on 3,000 datapoints. The data and its order are constant for all runs. This is the same setup as in (Arora et al., 2023).

For the LRA image experiments we use the following training protocol:

- **Optimizer and schedule:** Linear warmup with duration of 10%, AdamW optimizer (Loshchilov and Hutter, 2019). For attention-based models we use weight decay 0.0 and learning rate $5e-4$ and for SSM-based models we use weight decay 0.01 and learning rate $2e-4$.
- **Initialization:** For all models we use their standard initialization and initialize $\lambda = -1$ in each layer.
- **Training duration:** We use a global batch size of 64.
- **Width and depth:** For all runs, we use four layers (each with a sequence model and a MLP, interleaved with layer normalization). The model dimensions $d = 256$ and state dimension $n = 64$ are kept constant for all four architectures.
- **Data:** Each model is trained on 35,200 datapoints and evaluated on 7,850 datapoints. The data and its order are constant for all runs.

Since we keep the model sizes constant among all architectures and do not optimize the hyperparameters, the accuracies reported in Table 1 are generally lower than in the literature. However, the main point of these experiments is to investigate if learning λ affects performance.