Weight decay induces low-rank attention layers

Seijin Kobayashi[*1,2], **Yassir Akram**[*1]
Johannes von Oswald[2]

[1]Department of Computer Science, ETH Zürich
[2]Google, Paradigms of Intelligence Team
{seijink, jvoswald}@google.com, yakram@ethz.ch

Abstract

The effect of regularizers such as weight decay when training deep neural networks is not well understood. We study the influence of weight decay as well as $L2$-regularization when training neural network models in which parameter matrices interact multiplicatively. This combination is of particular interest as this parametrization is common in attention layers, the workhorse of transformers. Here, key-query, as well as value-projection parameter matrices, are multiplied directly with each other: $W_K^T W_Q$ and PW_V. We extend previous results and show on one hand that any local minimum of a $L2$-regularized loss of the form $L(AB^\top) + \lambda(\|A\|^2 + \|B\|^2)$ coincides with a minimum of the nuclear norm-regularized loss $L(AB^\top) + \lambda\|AB^\top\|_*$, and on the other hand that the 2 losses become identical exponentially quickly during training. We thus complement existing works linking $L2$-regularization with low-rank regularization, and in particular, explain why such regularization on the matrix product affects early stages of training. Based on these theoretical insights, we verify empirically that the key-query and value-projection matrix products $W_K^T W_Q, PW_V$ within attention layers, when optimized with weight decay, as usually done in vision tasks and language modelling, indeed induce a significant reduction in the rank of $W_K^T W_Q$ and PW_V, even in fully online training. We find that, in accordance with existing work, inducing low rank in attention matrix products can damage language model performance, and observe advantages when decoupling weight decay in attention layers from the rest of the parameters.

1 Introduction

The influence of $L2$-regularization, as well as *weight decay* regularization when training deep neural network models remains poorly understood and is still a subject of active research [van Laarhoven, 2017, Zhang et al., 2021, 2019, Loshchilov and Hutter, 2019, Zhang et al., 2021, Xie et al., 2023, Andriushchenko et al., 2023]. Given a model parametrized by matrix W, the standard motivation of adding $\frac{\lambda}{2}\|W\|^2$ to the optimization loss $L(W)$ comes from framing learning the model weights W as maximum a posteriori (MAP) estimation and choosing a Gaussian prior with zero mean [Mackay, 1995, Krogh and Hertz, 1991].

Previous works have studied the effect of regularization on the rank of weight matrices when training a model with gradient-based optimization [Ziyin and Wang, 2023, Arora et al., 2019, Li et al., 2021, Razin and Cohen, 2020, Gunasekar et al., 2017]. Here, we focus on the effect of $L2$-regularization on models using a *factorized* parametrization, where some weight matrices are parametrized as products

[*]Equal contribution, order determined randomly

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0146

of (often lower rank) matrices, $W = AB^\top$. This parametrization is used heavily in attention layers inside transformers [Vaswani et al., 2017] which we will focus on in the following.

Indeed, at the heart of the Transformer architecture is the attention operation which updates the T *tokens* concatenated into a matrix $E \in \mathbb{R}^{d_m \times T}$ inside the network according to

$$E \leftarrow E + PW_V E \phi\big((E^\top W_K^\top W_Q E) \odot M\big) \tag{1}$$

where ϕ is typically a softmax operation applied column-wise and M is typically a causal mask. The matrices $W_V, W_K, W_Q \in \mathbb{R}^{d_k \times d_m}$ are respectively the value, key, and query matrices that linearly transform E into some typically smaller space of dimension d_k [Phuong and Hutter, 2022], which can potentially subsume bias terms by appending a constant 1 to the tokens. The weight matrix $P \in \mathbb{R}^{d_m \times d_k}$ projects the weighted sum of value vectors back into the original token dimension. Therefore (multi-head) attention layers indeed consist of parameter matrix products i.e. $W_{QK} = W_K^\top W_Q$ as well as $W_{VP} = PW_V$, regardless of the choice of ϕ, or the presence or absence of causal masks.

When optimizing neural network models with this particular parametrization in conjunction with $L2$-regularization, and for any such two weight matrices A and B (e.g. the P and W_V for a given layer and a given head), we can rewrite the loss as:

$$\mathcal{L}_{L2}(A, B, \theta) \coloneqq L(AB^\top, \theta) + \frac{\lambda}{2}(\|A\|^2 + \|B\|^2), \tag{2}$$

where θ accounts for all the remaining parameters. We will see in the following that optimizing such losses has in practice implications on regularizing the rank of $W = AB^\top$. In fact, while it is classically known that the summed Frobenius norm $\frac{1}{2}(\|A\|^2 + \|B\|^2)$ is a tight upper bound on the *nuclear norm* $\|AB^\top\|_*$ [Srebro and Shraibman, 2005, Tibshirani, 2021], we theoretically show in the following that gradient-based optimization of the above objective result in the upper bound becoming tight exponentially quickly, for arbitrary loss, and thus directly optimizes for the nuclear norm which is known to induce low rank.

We highlight the relevance of this study since high weight decay is commonly used when training Transformer models. For example, GPT-3 [Brown et al., 2020], LLaMa [Touvron et al., 2023], LLaMa 2 [Touvron et al., 2023] and ViT [Dosovitskiy et al., 2021] report a weight decay strength of $\lambda = 0.1$. Interestingly, this is even true when fine-tuning, for example with low-rank adaptation (LoRA) [Hu et al., 2021a].

We summarize our contributions below:

- We show that for models with factorized parametrization, all local minima of any loss regularized by the Frobenius norm of A, B coincide with local minima of the same loss regularized by the nuclear norm of W. We further show theoretically that the discrepancy between the 2 regularizations vanishes exponentially quickly during training, thus implying that training such models with weight regularization can be subjected to low rank inducing pressure long before convergence.
- We empirically validate our result on various experimental settings, including when optimization with decoupled weight decay [Loshchilov and Hutter, 2019], on models ranging from deep linear networks to language models as well as Vision Transformers Dosovitskiy et al. [2021]. Intriguingly, we observe that this inductive bias of factorized parametrization with weight decay seems to hurt the performance on some tasks, raising the question of whether it is a feature or a bug.
- We provide evidence suggesting that this rank-regularizing effect in fact seems to affect the pretraining of popular pre-trained foundation models such as LLAMA 2 [Touvron et al., 2023] and Vision Transformer [Wu et al., 2020], by analyzing their pre-trained weights.

2 Related Work

The setting we study is closely related to a setting extensively studied in the Matrix Completion literature [Srebro and Shraibman, 2005, Sun and Luo, 2016, Candes and Tao, 2009], where the goal is to recover an unknown low-rank matrix for which only a subset of its entries are specified. Nuclear norm regularization is often used as a convex relaxation of the problem [Hu et al., 2021b], and

its equivalence at the global optimum with the $L2$-regularization on factorized matrix [Srebro and Shraibman, 2005], which has the advantage of being differentiable everywhere, has been exploited as a popular approach for large-scale matrix completion. Extensive prior work has focused in this setting on the theoretical guarantee of the factorization formulation to recover the underlying low-rank matrix correctly [Sun and Luo, 2016, Candes and Tao, 2009]. Similarly, similar loss landscape analyses were performed in the context of unconstrained features models [Zhu et al., 2021]. In contrast, our analysis does not rely on assumptions about the data, the loss (other than its differentiability) or convergence.

In a different line of work, recent efforts have focused on the effect of gradient-based optimization of deep networks on the parametrized matrix. For example, small weight initialization in this setting was shown to induce low rank in deep linear networks [Jacot et al., 2022, Arora et al., 2019, Li et al., 2021]. More recently, [Jacot, 2023] has shown the representation cost of deep networks with homogeneous nonlinearity converges to a notion of rank over nonlinear functions. More related to our work, equivalence between $L2$ regularization applied on factorized matrices and a low-rank inducing L_p-Schatten norm on the matrix they parametrize has been shown in several prior works [Dai et al., 2021, Tibshirani, 2021]. This is particularly relevant as $L2$ regularization can be applied explicitly or implicitly, such as when training deep networks with homogeneous activation coupled with e.g. the cross entropy loss [Jacot et al., 2022, Arora et al., 2019]. Crucially, however, these existing works characterize the low-rank inducing bias on neural networks that globally minimize $L2$ regularization while fitting training data.

Recently, [Galanti et al., 2023] have studied the effect of SGD with L2-regularization on a general architecture. Similarly to our work, they consider a general differentiable loss, but bound the rank of matrices at sufficiently large training steps, employing a theoretical argument that crucially does not leverage low-rank inducing norms due in part to the generality of the architecture they consider. [Wang and Jacot, 2023] have studied the same effect in the context of deep fully connected linear networks, showing that SGD strengthens the already existing low-rank bias induced by L2-regularization, albeit on matrix completion problems. Similarly to our work, they draw for the first time, to the best of our knowledge, an equivalence between the critical points of L2-regularized loss on the factorized matrix and Nuclear norm regularized loss on the parametrized matrix.

In contrast to these past works, we show both theoretically and empirically that for any arbitrary differentiable loss, the two regularizations become exponentially quickly identical during gradient-based optimization, and thus, that the low-rank inducing effect comes into play very early in during training. This brings a theoretical understanding to empirical observations made in previous works [Khodak et al., 2022], and is particularly relevant for many practical settings, in which learning does not converge, such as foundation model trained online, as is commonly done for large language models (LLMs) and large vision models.

Finally, given the significance of self-attention models, there has been work trying to understand the implicit inductive biases of some of their design choices. [Bhojanapalli et al., 2020] shows, in particular, the head size heuristic commonly used causes a low-rank bottleneck and limits the expressive power of the multi-head attention layer. Recent work has shown indeed that reducing the rank of attention matrices post-training of LLMs can hurt downstream performance [Sharma et al., 2023]. Our empirical work complements these observations and sheds light on the potentially damaging effect of the implicit rank-reducing effect of weight decay in the context of Attention layers, an unintended side effect contrary to the matrix completion setting.

3 Theoretical results

3.1 Preliminaries

We begin by reviewing the definition of the nuclear norm of a matrix and its upper bound when applied to a factorized matrix. We denote by $\|\cdot\|$ the Frobenius norm when applied on matrices.

3.1.1 Nuclear norm

The nuclear norm (also known as trace norm) of a real-valued matrix W, denoted by $\|W\|_*$, is defined as

$$\|W\|_* = \mathrm{Tr}(\sqrt{WW^\top}) \tag{3}$$

https://doi.org/10.52202/079017-0146

When using the singular value decomposition (SVD) of W, $W = USV^\top$, denoting $(s_i)_i$ the singular values, we can see that

$$\|W\|_* = \mathrm{Tr}(\sqrt{USV^\top VSU^\top}) = \mathrm{Tr}(S) = \sum_i s_i \tag{4}$$

i.e. the nuclear norm is the sum of the singular values of W.

The nuclear norm is often used in the low-rank regularization literature [Hu et al., 2021b] as it intuitively is a convex relaxation of the rank, and regularizing it typically induces low rank by injecting sparsity in the singular values.

3.1.2 Upper bound of the nuclear norm of a factorized matrix

Let two matrices A, B such that $W = AB^\top$. Then, using the Cauchy-Schwarz inequality, we have that

$$\|W\|_* = \mathrm{Tr}(S) = \mathrm{Tr}(U^\top AB^\top V) \tag{5}$$

$$\leq \sqrt{\mathrm{Tr}(U^\top AA^\top U)\,\mathrm{Tr}(B^\top VV^\top B)} \tag{6}$$

$$= \|A\|\|B\| \leq \frac{1}{2}(\|A\|^2 + \|B\|^2) \tag{7}$$

3.1.3 Considered losses

We will consider $L2$ losses of the format

$$\mathcal{L}_{L2}(A, B) \coloneqq L(AB^\top) + \frac{\lambda}{2}(\|A\|^2 + \|B\|^2), \tag{8}$$

and their $L_\star$ counterpart

$$\mathcal{L}_*(AB^\top) \coloneqq L(AB^\top) + \lambda\|AB^\top\|_*. \tag{9}$$

As a consequence of the above inequality, the $L2$-regularized objective (8) is an upper bound of the nuclear norm-regularized objective.

The meticulous reader should spot that those objectives don't account for the remaining parameters θ as in (2), while those parameters also evolve through learning. In fact, one can convince oneself that this can be safely ignored without loss of generality. The reader is referred to appendix D for more details about this point.

3.2 Equivalence of optimization solution

In the following, we will first show that in fact, any objective of the form in (8) will coincide at any stationary point with the nuclear-norm regularized loss in (9), thus introducing a low-rank inducing bias in the solution found. We assume A, B to have a bottleneck, i.e. to have the number of rows greater or equal to the number of columns, as is usual in attention layers. All proofs can be found in Appendix B.

We start by providing a sufficient condition under which the averaged Frobenius norm of two matrices would correspond to the nuclear norm of their product.

Proposition 3.1. *Let A, B be matrices such that $A^\top A = B^\top B$. Then, denoting $AB^\top = USV^\top$ the SVD of $AB^\top$, there exist an orthogonal matrix O such that $A = U \begin{pmatrix} \sqrt{S} \\ 0 \end{pmatrix} O^\top$ and $B = V \begin{pmatrix} \sqrt{S} \\ 0 \end{pmatrix} O^\top$. In particular, $\|AB^\top\|_* = \frac{1}{2}(\|A\|^2 + \|B\|^2)$.*

This condition states that the scalar product of any two columns of A should match the scalar product of corresponding columns of B. We will show next that at any stationary point of the objective $\mathcal{L}_{L2}$, that condition is fulfilled. We assume the loss L is differentiable and $\lambda > 0$.

Lemma 3.2. *At any stationary point A, B of $\mathcal{L}_{L2}$ we have that $A^\top A = B^\top B$.*

The above Lemma, together with Proposition 3.1, implies that at a stationary point A, B, $\mathcal{L}_{L2}(A, B)$ and $\mathcal{L}_*(AB^\top)$ coincide. However, this is not enough to claim that finding a (local) minimum of $\mathcal{L}_{L2}$ will in fact find a (local) minimum of $\mathcal{L}_*$. We now provide a result which shows that this claim is true.

Theorem 3.3. *A, B is a local minimum of $\mathcal{L}_{L2}$, if and only if 1) $W = AB^\top$ is a local minimum of $\mathcal{L}_*$, constrained to matrices of rank r where r is the maximum rank achievable by $AB^\top$, and 2) $A^\top A = B^\top B$.*

A more general formulation of the above results, albeit without a bottleneck dimension, was recently shown in [Wang and Jacot, 2023] (c.f. Theorem 3.1) where it was applied in the matrix completion context. We restate and prove it here for completion in the context of self-attention and transformer models.

The Theorem states that there is in fact a one-to-one mapping between the local minima of $\mathcal{L}_{L2}(A, B)$ and (the equivalence class of) local minima of $\mathcal{L}_*(AB^\top)$, for a general unregularized loss L.

In particular, if one wishes to optimize $\mathcal{L}_*$ for some matrix W, potentially under rank constraint, one can reparametrize W as a product of two matrices A, B and optimize the differentiable objective $\mathcal{L}_{L2}$ on A, B without introducing bad minima, and obtain rank-regularized solutions. In principle, one can still converge to a bad minimum for a general loss, but this is not due to the reparametrization.

On the other hand, the theorem shows that naively optimizing the $L2$-regularized loss with a factorized parametrization will (often inadvertently) result in actually finding solutions that exactly minimize the nuclear-norm regularized loss, introducing unintended low-rank inducing bias to the solution.

Note that however, the two parametrizations may result in different optimization, and thus different solutions, even if the loss landscape shares the same local minima.

3.3 Optimization dynamic in the gradient flow limit

The above result establishes equivalence of the local minima of the two losses. Our next result shows that the two losses will in fact coincide exponentially quickly during training.

Theorem 3.4. *Consider the gradient flow limit over the loss $\mathcal{L}_{L2}$. If $\|A\|, \|B\|$ remain bounded during training, then we have that $\left|\mathcal{L}_{L2}(A, B) - \mathcal{L}_*(AB^\top)\right|$ converges exponentially to 0.*

In order to prove the theorem, we first show that during gradient flow optimization, the condition from Proposition 3.1 becomes true exponentially quickly. This is then followed by a new bound bounding the gap between $\|AB^\top\|_*$ and $\frac{1}{2}(\|A\|^2 + \|B\|^2)$ by the norm of $A^\top A - B^\top B$. For completeness, we also provide a general result bounding the analogous gap for a L-layer deep linear network.

We provide in the appendix a similar result when considering gradient flow with noise, as well as with momentum and decoupled weight decay. We furthermore provide in appendix B.5 a discussion about the soundness condition.

The above result complements Theorem 3.3 by showing that optimizing $\mathcal{L}_{L2}$ will result in co-optimizing $\mathcal{L}_*$ very quickly during training, long before stationary points are found. The theorem also confirms previous empirical observations [Khodak et al., 2022].

3.4 Case study: 2-layer linear network

To illustrate the low-rank inducing bias of the factorized parametrization coupled with weight decay, we will study in the following the optimization within a 2-layer linear network and characterize the network at equilibrium. Such a network corresponds in fact to a drastically simplified softmax attention layer with $T = 1$. The derivations are similar to those used when studying deep linear networks [Ziyin et al., 2022, Saxe et al., 2013] and the redundant parameterization studied in [Ziyin and Wang, 2023].

Consider the following model

$$f(AB^\top) : x \to AB^\top x \tag{10}$$

where $B^\top \in \mathbb{R}^{d_2 \times d_1}$, $A \in \mathbb{R}^{d_3 \times d_2}$. For simplicity of presentation, we assume $d_3 = d_1 = d_{1,3}$, but the result can be easily extended to the general case. Given D data points $(x_i, y_i)_{1 \le i \le D}$, in matrix

https://doi.org/10.52202/079017-0146

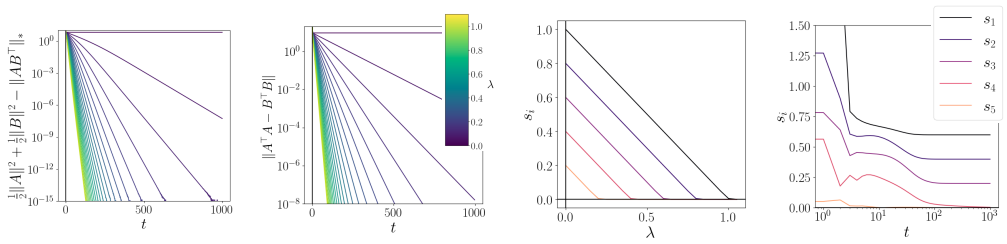

Figure 1: Optimization by gradient descent of two 5-by-5 matrices A, B on the $L2$-regularized loss $\|AB^\top - D\|^2 + \frac{\lambda}{2}(\|A\|^2 + \|B\|^2)$ where $D = \text{diag}(0.2, 0.4, 0.6, 0.8, 1)$, with various regularization strength λ. t denotes the number of optimization steps. *Left*: difference between the nuclear norm $\|AB^\top\|_*$ with the Frobenius norm $\frac{1}{2}\|A\|^2 + \frac{1}{2}\|B\|^2$ throughout optimization. For all cases, other than $\lambda = 0$, the trajectory converges exponentially quickly to 0 as predicted by our theory. *Center left*: Norm of the discrepancy between $A^\top A$ and $B^\top B$ over training steps. As predicted the discrepancy exponentially vanishes, with a time constant proportional to the λ. *Center right*: Singular values of the matrix $AB^\top$ at $t = 1000$, for various regularization strength λ. As predicted, s_i decays linearly with λ, until $\lambda \geq s_i$, at which point the singular value vanishes. *Right*: Singular values of the matrix $AB^\top$ during optimization, for $\lambda = 0.4$.

form, the $L2$-regularized mean squared error can be expressed as

$$\mathcal{L} = \frac{1}{2}\|Y - AB^\top X\|^2 + \frac{\lambda}{2}(\|B\|^2 + \|A\|^2) \tag{11}$$

where $X = (x_i)_i \in \mathbb{R}^{d_1 \times D}$, $Y = (y_i)_i \in \mathbb{R}^{d_3 \times D}$, and $\lambda > 0$.

Using full batch gradient flow, the differential equation governing the parameter dynamic becomes

$$\tau \dot{B}^\top = A^\top(\Sigma_{YX} - AB^\top \Sigma_{XX}) - \lambda B^\top \tag{12}$$

$$\tau \dot{A} = (\Sigma_{YX} - AB^\top \Sigma_{XX})B - \lambda A \tag{13}$$

where $\Sigma_{YX} = YX^\top$, $\Sigma_{XX} = XX^\top$, and τ is a constant controlling the learning rate.

To further simplify the above equations, we follow [Saxe et al., 2013] and assume $\Sigma_{XX} = I$, an assumption which holds exactly for whitened input data. Finally, without loss of generality, we perform a change of basis such that $\Sigma_{YX} = S$ where S is the diagonal matrix which diagonal consists of the singular values $(s_i)_{i \in [1..d_{1,3}]}$ of $YX^\top$.

At equilibrium, we thus have the following set of equations

$$\lambda B^\top = A^\top(S - AB^\top) \tag{14}$$

$$\lambda A = (S - AB^\top)B. \tag{15}$$

Denoting by a_i, b_i the i-th row of A, B, and assuming the $(s_i)_{i \in [1..d_{1,3}]}$ are all non-zero and distinct, we have the following conditions at equilibrium (cf Appendix C)

$$\forall i \in [1..d_{1,3}],\ a_i = b_i \tag{16}$$

$$\forall i, j \in [1..d_{1,3}]^2 \text{ s.t. } i \neq j,\ a_i^\top b_j = 0. \tag{17}$$

In particular, this implies, for any i, $\lambda\|a_i\|^2 = (s_i - \|a_i\|^2)\|a_i\|^2$. Clearly, if $\lambda \geq s_i$, then the equation can only be true if $a_i = 0$. If on the other hand $\lambda < s_i$, either $a_i = 0$ or $\|a_i\|^2 = s_i - \lambda$ satisfy the equilibrium condition, with the former being an unstable equilibrium point if the number of hidden units d_2 is greater than the number of elements in $\{i \in [1..d_{1,3}] \mid s_i > \lambda\}$.

To highlight the result, let us consider the case where the hidden layer has enough capacity, i.e. $d_2 \geq d_{1,3}$. In that case, the result tells us that at a stable equilibrium, $AB^\top$ will drop all singular values s that are less than λ, while keeping those that are larger. In other words, it performs a sort of low rank approximation of the input-output correlation matrix where the rank is controlled by λ. A related result was already obtained in the analyses of [Saxe et al., 2013] who studied the exact

solutions of 11 without the regularization term but introducing a bottleneck in the hidden layer, i.e. $d_2 < d_{1,3}$. Remarkably here, regularization achieves a similar effect even in an overcomplete network, where increasing λ gradually *prunes* the hidden neurons to ignore the smallest variations of the data, i.e. reducing d_2 adaptively. We confirm these results empirically in Figure 1.

Importantly, this result is only obtained because the regularization is applied to the parametrization involving a matrix multiplication. If $AB^\top$ were replaced by a single matrix $W \in \mathbb{R}^{d_{1,3} \times d_{1,3}}$, then the equilibrium condition would be $W = \frac{1}{1+\lambda} S$, whose rank remains constant w.r.t. the regularization strength.

3.5 Weight decay with Adam optimizer

While the regularized loss is a convenient setting for studying what happens to the parameters at equilibrium, in the vast majority of practical settings, decoupled weight decay [Loshchilov and Hutter, 2019], simply referred to as weight decay in the following, is used instead optimizing a regularized loss. A popular choice of optimizer for deep neural networks, including those with self-attention layers, is AdamW [Loshchilov and Hutter, 2019], which update the weights by using the Adam optimizer on the non-regularized loss while simultaneously applying weight decay.

While it is non-trivial to analyze the equilibrium points of AdamW in general, we show in Appendix E that under some simplifying assumptions, they coincide with those of a $L2$-regularized loss with a different regularization strength.

4 Empirical results

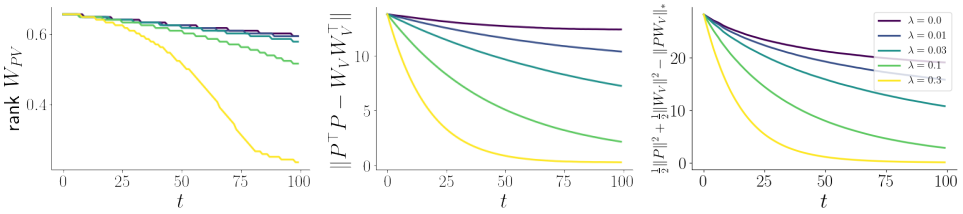

Figure 2: *Left*: The rank of weight matrix product PW_V of the first layer of a 2-layer Transformer trained on the associative recall task, during training, with AdamW, for various decay strengths. To better account for the effect of weight decay on the attention layers, only the decay strength applied to attention layers is varied, while the strength for all other layers is fixed at 0.1. We observe that rank reduction correlates strongly with weight decay strength. *Center*: Norm of the discrepancy between $P^\top P$ and $W_V W_V^\top$, during training. As predicted, the difference seems to converge to 0 when $\lambda > 0$ towards the end of training. While for AdamW we no longer have the guarantee of an exponential decay, we see that the discrepancy nonetheless vanishes quickly, with a time constant which perfectly correlates with the decay strength. *Right*: The difference of the nuclear norm of W_{VP} with the Frobenius norm upper bounding it. As the discrepancy between $P^\top P$ and $W_V W_V^\top$ decreases, the difference approaches 0, and thus the bound becomes tight. The optimization of $\mathcal{L}_{L2}$ thus gradually switches to that of $\mathcal{L}_*$, explaining the rank regularization. Qualitative findings are identical when studying $W_K^\top W_Q$.

The primary objective of our experimental analysis is to empirically validate the theoretical findings in more practical settings. Specifically, we aim to investigate the effect of decoupled weight decay, adaptive optimizers, as well as noisy gradient and lack of exact convergence to stationary points on the theoretical findings.

The second objective is to establish that the theory is relevant in the training of large neural network models. Due to the large computational costs we chose to avoid re-training large scale models but trained small-scale language models as well as a Vision Transformer without changing common hyperparameters. We aim to demonstrate that their typical training is affected by the rank-regularizing effect predicted by our theory. Finally, we investigate pre-trained weights of the relevant foundation models to show that they are consistent with rank-regularizing training.

https://doi.org/10.52202/079017-0146

To quantitatively measure the rank of matrices in the context of our experiments with attention layers, we use the following definition of *pseudo rank*: Let W be a weight matrix with singular values $\sigma_1, \sigma_2, ..., \sigma_n$, ordered such that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n$. The pseudo rank (referred to simply as rank in the following) of W is defined as $\frac{k}{n}$ where k is the smallest number such that:

$$\frac{\sum_{i=1}^{k} \sigma_i}{\sum_{i=1}^{n} \sigma_i} \geq 0.95.$$

In simpler terms, it represents the fraction of the largest singular values required to capture at least 95% of the sum of all singular values of the matrix W.

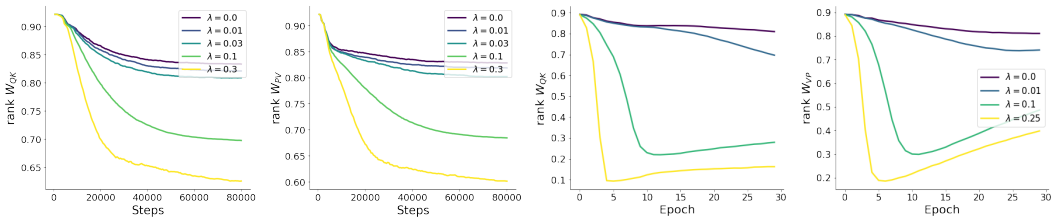

Figure 3: *Left, center left*: The rank of weight matrix products $W_K^\top W_Q$ and PW_V averaged across heads of layer 5 of an autoregressive transformers trained on the Pile [Gao et al., 2020]. *Center right, right*: The rank of weight matrix products $W_K^\top W_Q$ and PW_V averaged over all heads and all layers of a Vision Transformer trained following [Irandoust et al., 2022] on the ImageNet dataset [Deng et al., 2009]. In both settings, the decay strength applied to attention layers is varied, while keeping the strength for all other layers fixed. In all cases, we observe again that rank reduction correlates strongly with weight decay strength when optimizing with AdamW. The weight decay strength of 0.1 commonly used to pretrain some known large foundation models in fact noticeably reduces the rank of the generated matrices compared to when weight decay is turned off.

4.1 Associative recall task

In this simple memory task, a model is presented with a sequence of paired tokens $[x_1, y_1, \ldots, x_T, y_T, x_{T+1}]$. Specifically, the task is parameterized by an integer N, representing the number of unique tokens that can be mapped to N corresponding tokens. The sequence presented to the model therefore consists of $2N + 1$ tokens (with $T = N$), and the final token is repeated and appears in the sequence before, i.e. $x_{T+1} = x_j$ for some $j \in [0, \ldots, T]$. The model is trained to remember the correct association observed in-context and predict y_j. This task has been attributed and proposed as a proxy for language modelling [Fu et al., 2023, Poli et al., 2023].

We train a 2-layer self-attention only Transformer with AdamW optimizer on minibatches of size 128, for $N = 20$. To simulate additional noise, we perturb 5% of the labels with random labelling [Zhang et al., 2021].

Figure 4 shows that even in this setting, the stationary condition of a $L2$-regularized loss in Lemma 3.2 is approached, and the gap between the nuclear norm and the Frobenius norm in (5) vanishes, thus confirming that AdamW in fact also optimizes for the nuclear norm. Furthermore, the convergence speed is perfectly correlated with the weight decay strength. The results furthermore show that AdamW leads indeed to a consistent decrease in the rank in both parameter weight products as the decay strength increases. This aligns with the effect of optimizing the nuclear norm of these matrices.

4.2 Language Modelling

In order to validate our theoretical findings in larger scale experiments, we now present results when training standard small scale Transformer models, with 125 million parameters, on the Pile [Gao et al., 2020] - a common language modeling dataset. All design decisions such as the Transformer architecture as well as the optimizer and training schedule are identical to the ones proposed in the GPT-3 paper [Brown et al., 2020], which are now used in various other studies e.g. [Fu et al., 2023, von Oswald et al., 2023]. Details can be found in the Appendix G.

First, we confirm again that increasing weight decay with AdamW drastically reduces the rank of $W_K^T W_Q$ as well as PW_V, on average across depth and heads, of the trained models (c.f. Figure 3.

Table 1: Test set perplexity of 125 million Transformer models trained on the Pile for 10 billion tokens with AdamW and different weight decays λ for the self-attention (SA) and the feed-forward (MLP) weights. $\pm$ standard error of the mean computed over 5 seeds.

	SA-$\lambda = 0.0$	SA-$\lambda = 0.01$	SA-$\lambda = 0.025$	SA-$\lambda = 0.1$	SA-$\lambda = 0.25$
MLP-$\lambda = 0.0$	$12.00^{\pm 0.03}$	$12.01^{\pm 0.05}$	$11.98^{\pm 0.00}$	$11.92^{\pm 0.02}$	$12.02^{\pm 0.01}$
MLP-$\lambda = 0.01$	$11.94^{\pm 0.02}$	$11.95^{\pm 0.01}$	$11.94^{\pm 0.03}$	$11.89^{\pm 0.02}$	$11.97^{\pm 0.03}$
MLP-$\lambda = 0.025$	$11.89^{\pm 0.01}$	$11.90^{\pm 0.04}$	$11.90^{\pm 0.04}$	$11.80^{\pm 0.03}$	$11.92^{\pm 0.02}$
MLP-$\lambda = 0.1$	$11.72^{\pm 0.02}$	$11.71^{\pm 0.03}$	$11.68^{\pm 0.03}$	$11.67^{\pm 0.03}$	$11.70^{\pm 0.02}$
MLP-$\lambda = 0.25$	$11.63^{\pm 0.02}$	$11.65^{\pm 0.04}$	$11.62^{\pm 0.04}$	$11.52^{\pm 0.03}$	$11.58^{\pm 0.03}$

Appendix Figure 7). Furthermore, we observe that while increasing the weight decay strength of MLP beyond 0.1 is generally beneficial, see Table 1, doing the same for attention matrices starts slightly hurting performance. Results are averaged over 3 seeds. We observe a sweet spot around weight decay strength of 0.1 applied to self-attention weights, indicating that some rank regularization is beneficial for this task. Nevertheless, too much weight decay and therefore rank regularization seems to be detrimental. Finally, applying weight decay to the MLP weights seems to more important with a generally higher effect on performance. We leave a more nuanced investigation of decoupling the weight decay strength of matrices affected by our theory from the rest of the parameters for future research.

4.3 Vision Transformers

Next, we focus on computer vision tasks and train a Vision Transformer on the ImageNet dataset [Deng et al., 2009] for 24 hours, following the exact training protocol of [Irandoust et al., 2022]. We follow the previous section and vary the decay strength only in the attention layers, while keeping every other hyperparameter fixed. We observe a similar effect of the decay strength on the ranks of the matrices W_{QK}, W_{VP} (c.f. Fig 3).

4.4 Pretrained foundation models

Finally, we turn to pre-trained foundation models, and provide some evidence that their training is also impacted by the rank-regularizing effect of weight decay. Specifically, following Proposition B.4, it is sufficient to observe that the matrices $W_Q W_Q^\top$ resp. $P^\top P$ are close to $W_K W_K^\top$ resp. $W_V W_V^\top$. Because the matrices $W_Q, W_K, W_V, P^\top$ are typically wide rectangular matrices, the off-diagonal elements of $W_Q W_Q^\top$, etc, are mostly 0. For $\mathcal{L}_{L2}$ to approximately correspond to $\mathcal{L}_*$, it thus suffices that the diagonal elements of $W_Q W_Q^\top$ resp. $P^\top P$ are close to those of $W_K W_K^\top$ resp. $W_V W_V^\top$.

Figure 4, 6 shows that this is mostly the case, for all layers of the model. For each layer and head, we further find that the gap from (5) is indeed mostly tight, consistent with a rank regularizing training.

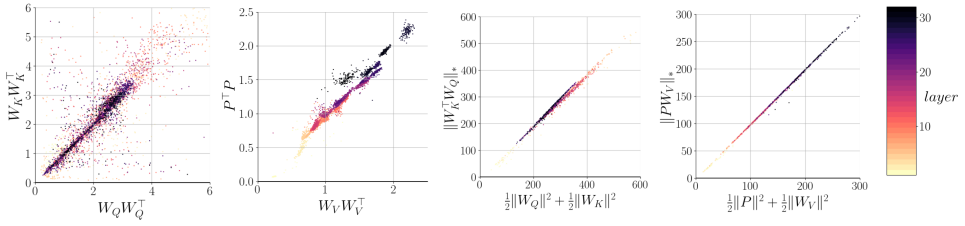

Figure 4: Analyses of attention layers in the pretrained LLAMA 2 model with 7 Billion parameters [Touvron et al., 2023]. The leftmost (resp. center left) shows the squared norm of every row of W_Q (resp. W_V), for the first head of each layer, against the norm of the corresponding row of W_K (resp. column of P). The condition $W_K W_K^\top = W_Q W_Q^\top$ would require these norms to be equal, which in fact is mostly true. While the model has not reached a stationary point, this indicates the optimization has advanced enough for this sufficient condition for $\mathcal{L}_*$ to be identical to $\mathcal{L}_{L2}$ to emerge. In fact, the center right (resp. rightmost) plot show the scatter plot mapping the Frobenius norm against the nuclear norm for all heads across all layers. The two norms almost perfectly coincide.

https://doi.org/10.52202/079017-0146

5 Discussion

Our results provide further insights into the interplay between $L2$-regularization and weight decay regularization and the optimization of models that consist of parameter matrix products. This is of particular interest since attention layers in transformer exhibit this parametrization as key-query, as well as value-projection parameter matrices, are multiplied directly with each other: $W_K^T W_Q$ and PW_V. Our empirical findings strongly support our theoretical predictions about the impact of weight decay on the rank of attention layers and clearly show a rank-regularizing effect even without convergence. We provide evidence that the training of some foundation models such as Llama are in fact in practice affected by the same regularization.

Furthermore, we find that decoupling weight decay in the attention weights and tuning its weight decay strength can improve performance, for example in our language modelling experiments. These findings complement the recent observation that reducing the rank of language model MLP matrices post-training improves their reasoning performance, while doing the same for attention layer matrices mostly hurt it [Sharma et al., 2023]. In particular, our findings suggest that the conventional practice of applying uniform regularization strategies across all layers may not be optimal for other deep learning architectures as well. This finding opens up new avenues for model- or layer-specific regularization strategies that could significantly enhance the performance of these models.

Our findings once more highlight the complexity of understanding optimization techniques in conjunction with particular neural network models, particularly transformers. For example, the difficulty of understanding the effect when varying regularization strengths on different components of these models underscores the need for a more nuanced theoretical understanding of layer-specific regularization. We are particularly excited about further research that aims to disentangle the role of weight decay in in-weight vs. in-context learning within MLPs and self-attention layers, building on [Singh et al., 2023]. In conclusion, while our findings mark a step forward in understanding and improving the usage of weight decay when training deep neural networks, in particular transformers, our study shed light on the intricate interplay of neural network regularization and its parametrization.

Acknowledgments

Seijin Kobayashi, Yassir Akram and Johannes von Oswald deeply thank João Sacramento and Angelika Steger for their support and guidance. The authors also thank Robert Obryk, Moritz Firsching, Luca Versari, Nicolas Zucchet, Alexander Meulemans, Simon Schug, Blaise Agüera y Arcas, Alexander Mordvintsev, Ettore Randazzo and Eyvind Niklasson for fruitful discussions.

References

Twan van Laarhoven. L2 regularization versus batch and weight normalization, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776.

Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1lz-3Rct7.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Zeke Xie, zhiqiang xu, Jingzhao Zhang, Issei Sato, and Masashi Sugiyama. On the overlooked pitfalls of weight decay and how to mitigate them: A gradient-norm perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vnGcubtzR1.

Maksym Andriushchenko, Francesco D'Angelo, Aditya Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning?, 2023.

David J C Mackay. Probable networks and plausible predictions — a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Neural Information Processing Systems*, 1991. URL https://api.semanticscholar.org/CorpusID:10137788.

Liu Ziyin and Zihao Wang. spred: Solving l1 penalty with SGD. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43407–43422. PMLR, 23–29 Jul 2023.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *CoRR*, abs/1905.13655, 2019. URL http://arxiv.org/abs/1905.13655.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=AHOs7Sm5H7R.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms, 2020.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization, 2017.

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

Mary Phuong and Marcus Hutter. Formal algorithms for transformers, 2022.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In Peter Auer and Ron Meir, editors, *Learning Theory*, pages 545–560, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31892-7.

Ryan J Tibshirani. Equivalences between sparse models and neural networks. *Working Notes. URL https://www. stat. cmu. edu/ryantibs/papers/sparsitynn. pdf*, 2021.

 https://doi.org/10.52202/079017-0146

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021a. URL https://arxiv.org/abs/2106.09685.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, November 2016. ISSN 1557-9654. doi: 10.1109/tit.2016.2598574. URL http://dx.doi.org/10.1109/TIT.2016.2598574.

Emmanuel J. Candes and Terence Tao. The power of convex relaxation: Near-optimal matrix completion, 2009.

Zhanxuan Hu, Feiping Nie, Rong Wang, and Xuelong Li. Low rank regularization: A review. *Neural Networks*, 136:218–232, 2021b. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2020.09.021. URL https://www.sciencedirect.com/science/article/pii/S089360802030352X.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features, 2021.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, 2022.

Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions, 2023.

Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26884–26896. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e22cb9d6bbb4c290a94e4fff4d68a831-Paper.pdf.

Tomer Galanti, Zachary S. Siegel, Aparna Gupte, and Tomaso Poggio. Characterizing the implicit bias of regularized sgd in rank minimization, 2023.

Zihan Wang and Arthur Jacot. Implicit bias of sgd in l_2-regularized linear dnns: One-way jumps from high to low rank, 2023.

Mikhail Khodak, Neil Tenenholtz, Lester Mackey, and Nicolò Fusi. Initialization and regularization of factorized neural layers, 2022.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models, 2020.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. The truth is in there: Improving reasoning in language models with layer-selective rank reduction, 2023.

Liu Ziyin, Botao Li, and Xiangming Meng. Exact solutions of a deep linear network. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=X6bp8ri8dV`.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, 2013. URL `https://arxiv.org/abs/1312.6120`.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: an 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Saghar Irandoust, Thibaut Durand, Yunduz Rakhmangulova, Wenjie Zi, and Hossein Hajimirsadeghi. Training a vision transformer from scratch in less than 24 hours with 1 gpu, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models, 2023.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models, 2023.

Johannes von Oswald, Eyvind Niklasson, Maximilian Schlegel, Seijin Kobayashi, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Blaise Agüera y Arcas, Max Vladymyrov, Razvan Pascanu, and João Sacramento. Uncovering mesa-optimization algorithms in transformers, 2023.

Aaditya K. Singh, Stephanie C. Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers, 2023.

Paul Lévy. Sur certains processus stochastiques homogènes. *Compositio Mathematica*, 7:283–339, 1940. URL `http://www.numdam.org/item/CM_1940__7__283_0/`.

Robert T Powers and Erling Størmer. Free states of the canonical anticommutation relations. *Communications in Mathematical Physics*, 16(1):1–33, 1970.

Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Richard H. R. Hahnloser, Rahul Sarpeshkar, Misha A. Mahowald, Rodney J. Douglas, and H. Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 2000.

 https://doi.org/10.52202/079017-0146

A Compute budget

We estimate the total compute budget to 4 Nvidia RTX 4090 for two months. The LLMs were punctually trained on a cluster of 16xA100 GPUs for 4 days.

B Proofs of theoretical results

B.1 Proof of Proposition 3.1

Proof. Using singular value decomposition, we write $A = U_A \Sigma_A V_A^\top$ and $B = U_B \Sigma_B V_B^\top$, with $\Sigma_A = \begin{pmatrix} S_A \\ 0 \end{pmatrix}$ and $\Sigma_B = \begin{pmatrix} S_B \\ 0 \end{pmatrix}$. Substituting in the equation $A^\top A = B^\top B$, we get

$$V_A S_A^2 V_A^\top = V_B S_B^2 V_B^\top$$

By positivity and uniqueness of singular values, necessarily, $\Sigma_A = \Sigma_B = \begin{pmatrix} S \\ 0 \end{pmatrix}$. Furthermore, by rearranging the above equation, we get $S^2 V_A^\top V_B = V_A^\top V_B S^2$, i.e. that $V_A^\top V_B$ commutes with S. We rewrite A as

$$\begin{aligned} A &= U_A \Sigma_A V_A^\top V_B V_B^\top = U_A \begin{pmatrix} S V_A^\top V_B \\ 0 \end{pmatrix} V_B^\top \\ &= U_A \begin{pmatrix} V_A^\top V_B S \\ 0 \end{pmatrix} V_B^\top = U_A \begin{pmatrix} V_A^\top V_B & 0 \\ 0 & I \end{pmatrix} \Sigma_A V_B^\top \end{aligned}$$

Redefining U_A as $U_A \begin{pmatrix} V_A^\top V_B & 0 \\ 0 & I \end{pmatrix}$, setting $\Sigma = \Sigma_A$, and $V = V_B$, we can write $A = U_A \Sigma V^\top$ and $B = U_B \Sigma V^\top$.

In particular, $AB^\top = U_A \Sigma \Sigma^\top U_B^\top$, which is a valid SVD of $AB^\top$. It remains to show that, if $AB^\top$ is diagonal, then there exists an orthogonal matrix O such that $A = \Sigma O^\top$ and $B = \Sigma O^\top$.

Let us assume the diagonality, i.e. $AB^\top = \Sigma \Sigma^\top$. Then, we have

$$(AB^\top)^2 = U_A \Sigma \Sigma^\top \Sigma \Sigma^\top U_A^\top = \Sigma \Sigma^\top \Sigma \Sigma^\top = U_B \Sigma \Sigma^\top \Sigma \Sigma^\top U_B^\top$$

i.e. that U_A, U_B commute with $\Sigma \Sigma^\top$, and thus that they are block diagonal. Furthermore, $\Sigma \Sigma^\top U_A U_B^\top = \Sigma \Sigma^\top$. They can then be written as

$$U_A = \begin{pmatrix} U & 0 \\ 0 & U'_A \end{pmatrix}$$

$$U_B = \begin{pmatrix} U & 0 \\ 0 & U'_B \end{pmatrix}$$

where U, U'_A, U'_B are orthogonal matrices, and the block of U corresponds to the non zero singular values of $\Sigma \Sigma^\top$.

We can then rewrite A, B as $A = \Sigma \begin{pmatrix} U & 0 \\ 0 & I \end{pmatrix} V^\top$ and $B = \Sigma \begin{pmatrix} U & 0 \\ 0 & I \end{pmatrix} V^\top$, which conclude the proof by setting $O = \begin{pmatrix} U & 0 \\ 0 & I \end{pmatrix} V^\top$.

Finally, $AB^\top = U_A \Sigma \Sigma^\top U_B^\top$, and therefore $\|AB^\top\|_* = \|U_A \Sigma \Sigma^\top U_B^\top\|_* = \mathrm{Tr}(\Sigma \Sigma^\top) = \frac{1}{2}(\|A\|^2 + \|B\|^2)$. □

B.2 Proof of Lemma 3.2

Proof. Let A, B a stationary point of the unregularized loss L in $\mathcal{L}_{L2}$. One can show that the gradient of $L(W = AB^\top)$ with respect to A (resp. B) is

$$\partial_A L = \left(\frac{\partial L}{\partial W}\big|_{W=AB^\top} \right) B \tag{18}$$

$$\partial_B L = \left(\frac{\partial L}{\partial W}\big|_{W=AB^\top} \right)^\top A \tag{19}$$

where $\frac{\partial L}{\partial W}\big|_{W=AB^\top}$ is a matrix, which we denote by $-G$. Differentiating L, at the stationary point, the following equations must then be satisfied

$$\lambda A = GB \tag{20}$$

$$\lambda B = G^\top A \tag{21}$$

In particular, $A^\top A = \frac{1}{\lambda} A^\top GB = \frac{1}{\lambda}(G^\top A)^\top B = B^\top B$.

□

B.3 Proof of Theorem 3.3

Proof. ($\Leftarrow$) We start by proving the backward implication, by contradiction. Let M a local minimum of $\mathcal{L}_*$, and A, B such that $M = AB^\top$ and $A^\top A = B^\top B$. Then by Proposition 3.1, $\mathcal{L}_{L2}(A, B) = \mathcal{L}_*(M)$. Assume A, B is not a local minimum of $\mathcal{L}_{L2}$, i.e. there exists an infinitesimally perturbed matrices A', B' such that $\mathcal{L}_{L2}(A', B') < \mathcal{L}_{L2}(A, B)$. By continuity of matrix multiplication, $M' = A'B'^\top$ is an infinitesimally perturbed matrix M. Since $\mathcal{L}_*(M') \leq \mathcal{L}_{L2}(A', B') < \mathcal{L}_{L2}(A, B) = \mathcal{L}_*(M)$, we get a contradiction.

($\Rightarrow$) Assume now that A, B is a local minimum of $\mathcal{L}_{L2}$, and that $W = AB^\top$ is not a local minimum of $\mathcal{L}_*$ constrained to rank r matrices. Then, we can construct a sequence $(W_n)_n$ of rank r matrices such that $\lim_{n\to\infty} W_n = W$, and for all n, $\mathcal{L}_*(W_n) < \mathcal{L}_*(W)$. For all n, let $W_n = U_n S_n V_n^\top$ the SVD of W_n. By continuity of the mapping from a matrix to its singular values, $\lim_{n\to\infty} S_n = S$, where S are the singular values of W. Because the set of orthogonal matrices is compact, there exists a subsequence of $((U_n, V_n))_n$ which converges to some orthogonal matrices (U, V). Without loss of generality, we redefine the sequence to this converging subsequence. By continuity of matrix multiplication, necessarily $USV^\top = W$. $USV^\top$ is a valid SVD of W. Since by local minimality of A, B, following Lemma 3.2 and Proposition 3.1, we get that $A = U\Sigma O^\top$ and $B = V\Sigma O^\top$ where $\Sigma = \begin{pmatrix} \sqrt{S} \\ 0 \end{pmatrix}$ and O is some orthogonal matrix. Let for all n, $A_n = U_n \Sigma_n O^\top$ and $B_n = V_n \Sigma_n O^\top$, where $\Sigma_n = \begin{pmatrix} \sqrt{S_n} \\ 0 \end{pmatrix}$. Then, $\lim_{n\to\infty}(A_n, B_n) = (A, B)$ and yet, because for all n, $A_n^\top A_n = B_n^\top B_n$ and $A_n B_n^\top = W_n$, we have $\mathcal{L}_{L2}(A_n, B_n) = \mathcal{L}_*(W_n) < \mathcal{L}_*(W) = \mathcal{L}_{L2}(A, B)$. This is a contradiction. □

B.4 Proof of Theorem 3.4

In order to prove the theorem, we first show that during optimization, the condition from Proposition 3.1 becomes true exponentially quickly. This is then followed by a new bound bounding the gap between $\|AB^\top\|_*$ and $\frac{1}{2}(\|A\|^2 + \|B\|^2)$ by the norm of $A^\top A - B^\top B$.

B.4.1 Exponential decay of $A^\top A - B^\top B$

We provide the result for the vanilla gradient flow limit, but also provide an alternative proof for the stochastic gradient flow with momentum and decoupled weight decay, to illustrate that the exponential decay would hold in many practical settings. We note that the gradient flow limit is a good approximation for a small learning rate in the discrete dynamic.

Lemma B.1. *In the gradient flow limit over the loss $\mathcal{L}_{L2}$, $A^\top A - B^\top B$ will converge exponentially to* 0.

https://doi.org/10.52202/079017-0146

Proof. For any i, we denote by a^i, b^i the i-th column of A and B. The columns follow the following differential equations:

$$\tau \dot{a}^i = Gb^i - \lambda a^i \tag{22}$$

$$\tau \dot{b}^i = G^\top a^i - \lambda b^i \tag{23}$$

where $G = -\frac{\partial L}{\partial W}\big|_{W=AB^\top}$, and τ is some time constant controlling the learning rate. Given a pair i, j, we can now look at the dynamic of $a^{i\top}a^j - b^{i\top}b^j$:

$$\begin{aligned}
\tau \frac{d}{dt}(a^{i\top}a^j - b^{i\top}b^j) &= \tau(a^{j\top}\dot{a}^i + a^{i\top}\dot{a}^j - \dot{b}^{j\top}b^i - \dot{b}^{i\top}b^j) \\
&= a^{j\top}Gb^i - \lambda a^{j\top}a^i \\
&+ a^{i\top}Gb^j - \lambda a^{i\top}a^j \\
&- (a^{j\top}Gb^i - \lambda b^{j\top}b^i) \\
&- (a^{i\top}Gb^j - \lambda b^{i\top}b^j) \\
&= -2\lambda(a^{i\top}a^j - b^{i\top}b^j)
\end{aligned}$$

Therefore, we have $A^\top A - B^\top B = Qe^{-\frac{2\lambda t}{\tau}}$, where Q is $A^\top A - B^\top B$ at initialization, and in particular, every entry of $A^\top A - B^\top B$ converge to 0 exponentially. □

We now provide a similar result, in the gradient flow regime but with momentum, as well as decoupled weight decay - a tractable approximation to AdamW as shows the following proposition:

Proposition B.2. *We consider the following dynamics approximating stochastic gradient flow with weight decay:*

$$\begin{aligned}
dH_t^A &= \mu(G_t B_t dt + \sigma dW_t^A - H_t^A dt) \\
dH_t^B &= \mu(G_t^\top A_t dt + \sigma dW_t^B - H_t^B dt) \\
dA_t &= -\eta(H_t^A + \lambda A_t)dt \\
dB_t &= -\eta(H_t^B + \lambda B_t)dt
\end{aligned}$$

where $\mu, \eta, \sigma > 0$ and W^A and W^B are independent matrix Wiener processes. Initial condition are $H_0^A = 0$ and $H_0^B = 0$. H^A (resp. H^B) is the momentum gradient with respect to A (resp. B). Then,

$$H_t^A = \mu \int_0^t e^{-\mu(t-s)} G_s B_s ds + \sqrt{\frac{\mu\sigma^2}{2}} W^A_{1-e^{-2\mu t}} \tag{24}$$

$$H_t^B = \mu \int_0^t e^{-\mu(t-s)} G_s^\top A_s ds + \sqrt{\frac{\mu\sigma^2}{2}} W^B_{1-e^{-2\mu t}} \tag{25}$$

$$A_t = e^{-\eta\lambda t} A_0 - \eta \int_0^t e^{-\eta\lambda(t-s)} H_s^A ds \tag{26}$$

$$B_t = e^{-\eta\lambda t} B_0 - \eta \int_0^t e^{-\eta\lambda(t-s)} H_s^B ds \tag{27}$$

$$A_t^\top A_t - B_t^\top B_t = e^{-2\eta\lambda t}(A_0^\top A_0 - B_0^\top B_0) \tag{28}$$

$$- \eta \int_0^t e^{-2\eta\lambda(t-s)}(H_s^{A\top} A_s + A_s^\top H_s^A - H_s^{B\top} B_s - B_s^\top H_s^B) ds. \tag{29}$$

Proof. We have

$$\begin{aligned}
d(e^{\mu t} H_t^A) &= e^{\mu t}(\mu H_t^A dt + dH_t^A) \\
&= \mu e^{\mu t}(G_t B_t dt + \sigma dW_t^A)
\end{aligned}$$

such that

$$\begin{aligned}
H_t^A &= e^{-\mu t}\left(H_0^A + \mu \int_0^t e^{\mu s}(G_s B_s ds + \sigma dW_s^A)\right) \\
&= \mu \int_0^t e^{-\mu(t-s)} G_s B_s ds + \sqrt{\frac{\mu\sigma^2}{2}} W^A_{1-e^{-2\mu t}}
\end{aligned}$$

Note that the second term is an abuse of notation. The derivation of the integral form of H_t^B, A_t and B_t follows the same logic.

For the last two equations, we get

$$\begin{aligned}
d(A_t^\top A_t - B_t^\top B_t) =& dA_t^\top A_t + A_t^\top dA_t - dB_t^\top B_t - B_t^\top dB_t \\
=& -\eta\left(\left(H_t^A + \lambda A_t\right)^\top A_t + A_t^\top\left(H_t^A + \lambda A_t\right)\right) \\
&+\eta\left(\left(H_t^B + \lambda B_t\right)^\top B_t + B_t^\top\left(H_t^B + \lambda B_t\right)\right) \\
=& -2\eta\lambda(A_t^\top A_t - B_t^\top B_t) - \eta(H_t^{A\top} A_t + A_t^\top H_t^A - H_t^{B\top} B_t - B_t^\top H_t^B)dt \\
\coloneqq& -2\eta\lambda(A_t^\top A_t - B_t^\top B_t) + Q_t dt
\end{aligned}$$

which gives

$$A_t^\top A_t - B_t^\top B_t = e^{-2\eta\lambda}(A_0^\top A_0 - B_0^\top B_0) + \int_0^t e^{-2\eta\lambda(t-s)} Q_s ds$$

□

Before analyzing the implications of this proposition, let us state a lemma that will allow us to bound the probability of a Brownian motion diverging:

Lemma B.3. *[Lévy, 1940] Let (B_t) be a 1D Wiener process. Then, for $t, L > 0$, $\mathbb{P}[\max_{s\in[0,t]} B_s > L] = 2\mathbb{P}[B_t > L]$.*

Let us now analyse the consequences of Proposition B.2. We will assume that A_t and B_t remain L2-bounded by $M > 0$ (which is true for all converging dynamic modulo steady state noise), and that G_t remains L2-bounded by K (either using a Lipschitzian loss, or using clipping).

First observe that, using lemma B.3, for $\varepsilon > 0$, with probability $1 - \varepsilon$, the term $\sqrt{\frac{\mu}{2}}\sigma W^A_{1-e^{-2\mu t}}$ and the correspond B will remain bounded by $\sigma\sqrt{\mu n d \ln\frac{4nd}{\varepsilon}}$.

This way, H^A and H^B will with probability $1 - \varepsilon$ remain bounded by $KM + \sigma\sqrt{\mu n d \ln\frac{4nd}{\varepsilon}}$. With that same probability, the term

$$\eta\int_0^t e^{-2\eta\lambda(t-s)}(H_s^{A\top} A_s + A_s^\top H_s^A - H_s^{B\top} B_s - B_s^\top H_s^B)ds$$

will remain bounded by $4\frac{\eta M}{\lambda}\left(KM + +\sigma\sqrt{\mu n d \ln\frac{4nd}{\varepsilon}}\right)$. This is the same order of magnitude as the stochastic term. Until $A^\top A - B^\top B$ is of that order, it exponentially decays.

B.4.2 Upper bound of $\left|\|AB^\top\|_* - \frac{1}{2}(\|A\|^2 + \|B\|^2)\right|$

Finally, we provide the following general result, bounding the gap between $\|AB^\top\|_*$ and $\frac{1}{2}(\|A\|^2 + \|B\|^2)$ by the norm of $A^\top A - B^\top B$.

Proposition B.4. *For any matrices A, B, we have*

$$\left|\|AB^\top\|_* - \|A\|_F^2\right| \leqslant \sqrt{\|A^\top A - B^\top B\|_*}\|A\|_*.$$

In particular,

$$\begin{aligned}
&\left|\|AB^\top\|_* - \frac{\|A\|_F^2 + \|B\|_F^2}{2}\right| \\
&\qquad\leqslant \sqrt{\|A^\top A - B^\top B\|_*}\frac{\|A\|_* + \|B\|_*}{2}.
\end{aligned}$$

https://doi.org/10.52202/079017-0146

Proof. Let $Q \coloneqq A^\top A - B^\top B$. Using singular value decomposition, we write $A = U_A \Sigma_A V_A^\top$ and $B = U_B \Sigma_B V_B^\top$, with $\Sigma_A = \begin{pmatrix} S_A \\ 0 \end{pmatrix}$ and $\Sigma_B = \begin{pmatrix} S_B \\ 0 \end{pmatrix}$. Substituting in the previous equation, we get

$$V_A S_A^2 V_A^\top = V_B S_B^2 V_B^\top + Q$$

i.e.

$$V_A S_A V_A^\top = \sqrt{V_B S_B^2 V_B^\top + Q} = V_B (S_B + \Delta) V_B^\top \tag{30}$$

where $V_B \Delta V_B^\top \coloneqq \sqrt{V_B S_B^2 V_B^\top + Q} - \sqrt{V_B S_B^2 V_B^\top}$. By the Powers-Stormer inequality [Powers and Størmer, 1970], we have $\|\Delta\|_F^2 = \|V_B \Delta V_B^\top\|_F^2 \leqslant \|Q\|_*$.

From there, we rewrite A as

$$A = U_A \Sigma_A V_A^\top V_B V_B^\top = U_A \begin{pmatrix} S_A V_A^\top V_B \\ 0 \end{pmatrix} V_B^\top \tag{31}$$

$$\overset{(30)}{=} U_A \begin{pmatrix} V_A^\top V_B (S_B + \Delta) \\ 0 \end{pmatrix} V_B^\top \tag{32}$$

$$= U_A \begin{pmatrix} V_A^\top V_B & 0 \\ 0 & I \end{pmatrix} \left(\Sigma_B + \begin{pmatrix} \Delta \\ 0 \end{pmatrix} \right) V_B^\top \tag{33}$$

Consequently, $\|AB^\top\|_* = \|S_B^2 + \Delta S_B\|_*$ and

$$\begin{aligned} \left| \|AB^\top\|_* - \|B\|_F^2 \right| &\leqslant \|\Delta S_B\|_* \leqslant \|\Delta\| \|S_B\|_* \leqslant \|\Delta\|_F \|S_B\|_* \\ &\leqslant \sqrt{\|Q\|_*} \|B\|_*. \end{aligned}$$

Similarily, we have $\left| \|AB^\top\|_* - \|A\|_F^2 \right| \leqslant \sqrt{\|Q\|_*} \|A\|_*$.

The second inequality is obtained by using the triangle inequality.

□

B.4.3 Upper bound of $\left\| \|\prod_l A_l\|_{2/L}^{2/L} - \frac{1}{L} \sum \|A_l\|_F^2 \right\|$

We here provide a more general version of Proposition 3.1.

Proposition B.5. *Let $q \geq r > 0$, $A_1 \in \mathcal{M}_{q,r}$, $A_l \in \mathcal{M}_{r,r}$ for $l \in [2..L-1]$, $A_L \in \mathcal{M}_{q,r}$ and $A = \prod_{l=1}^L A_l$. Let $\varepsilon > 0$. We assume that the sequence $(A_l)_{l \in [1..L]}$ is ε-balanced, i.e. that for $l \in [1..L-1]$,*

$$\left\| A_l^\top A_l - A_{l+1} A_{l+1}^\top \right\|_* \leqslant \varepsilon.$$

Furthermore, assume that

$$\varepsilon \leq \frac{1}{L^4} \min_l \|A_l\|_*^L.$$

Then for $k \in [1..L]$, we have

$$\left| \left\| \prod_{l=1}^L A_l \right\|_{2/L}^{2/L} - \|A_k\|_F^2 \right| \leq r \|A_k\|_*^{L-1} e^{2/L} L^{4/L} \varepsilon^{1/L}$$

In particular,

$$\left| \left\| \prod_{l=1}^L A_l \right\|_{2/L}^{2/L} - \frac{1}{L} \sum_{l=1}^L \|A_l\|_F^2 \right| \leq \frac{r}{L} \sum_l \|A_l\|_*^{L-1} e^{2/L} L^{4/L} \varepsilon^{1/L}$$

Proof. We will assume for this proof that the A_l are square matrices for all $l \in [1..L]$. This proof holds as it is for the more general case, with more cumbersome notations, and with the trick used in equation 33. For $l \in [1..L]$, we denote by $A_l = U_l \Sigma_l {V_l}^\top$ be the SVD of A_l. Let $k \in [1..L]$. We first show by recurrence that we can write

$$A_l = O_{l-1}(\Sigma_k + \Delta_l){O_l}^\top \qquad \text{for } l \in [1..L]$$

for an appropriate choice of orthogonal matrices $(O_l)_{l\in[0..L]}$ and symmetric matrices $(\Delta_l)_{l\in[1..L]}$, and such that : $O_{k-1} = U_k$, $O_k = V_k$ and $\|\Delta_l\|_F \leqslant \sqrt{\varepsilon}|l-k|$ for $l \in [1..k]$. The recurrence is symmetric for $l \leqslant k$ and for $l \geqslant k$. We will prove it in the latter case. For $l = k$, the statement holds. Let $l \in [k..L-1]$. We assume we can write $A_l = O_{l-1}(\Sigma_k + \Delta_l){O_l}^\top$, with $\|\Delta_l\|_F \leqslant \sqrt{\varepsilon}(l-k)$. Let $Q \coloneqq A_l^\top A_l - A_{l+1}A_{l+1}^\top$. We have

$$O_l(\Sigma_k + \Delta_l)^2 {O_l}^\top = U_{l+1}\Sigma_{l+1}^2 U_{l+1}^\top + Q.$$

Similar to the previous proof, we can write

$$O_l(\Sigma_k + \Delta_l + \Delta){O_l}^\top = U_{l+1}\Sigma_{l+1}U_{l+1}^\top$$

with Δ symmetric verifying

$$\|\Delta\|_F^2 \leqslant \|Q\|_* \leqslant \varepsilon.$$

We can rewrite

$$\begin{aligned} A_{l+1} &= U_{l+1}\Sigma_{l+1}V_{l+1}^\top \\ &= O_l {O_l}^\top U_{l+1}\Sigma_{l+1}V_{l+1}^\top \\ &= O_l(\Sigma_k + \Delta_l + \Delta){O_l}^\top U_{l+1}V_{l+1}^\top. \end{aligned}$$

We set $O_{l+1} = V_{l+1}U_{l+1}^\top O_l$, $\Delta_{l+1} = \Delta_l + \Delta$ to get the desired result. We verify that $\|\Delta_{l+1}\|_F \leqslant \sqrt{\varepsilon}(l-k+1)$.

Now that we have proven our lemma, let us observe that

$$\prod_{l=1}^{L} A_l = O_0 \left(\prod_l (\Sigma_k + \Delta_l)\right) {O_L}^\top$$

with

$$\left\|\prod_{l=1}^{L} A_l\right\|_{2/L} = \left\|\prod_l (\Sigma_k + \Delta_l)\right\|_{2/L}.$$

Notice that $\left\|\Sigma_k^L\right\|_{2/L}^{2/L} = \|A_k\|_F^2$. Using the triangular inequality of $A \to \|A\|_p^p$ for $0 < p \leq 1$, we have

$$\begin{aligned} \left| \left\|\prod_{l=1}^{L} A_l\right\|_{2/L}^{2/L} - \|A_k\|_F^2 \right| &= \left| \left\|\prod_{l=1}^{L} (\Sigma_k + \Delta_l)\right\|_{2/L}^{2/L} - \left\|\Sigma_k^L\right\|_{2/L}^{2/L} \right| \\ &\leq \left\|\prod_{l=1}^{L} (\Sigma_k + \Delta_l) - \Sigma_k^L\right\|_{2/L}^{2/L}. \end{aligned}$$

https://doi.org/10.52202/079017-0146

Furthermore, using the fact that $\|AB\|_{2/L} \leq \|A\|_{2/L}\|B\|_F$, we have

$$\begin{aligned}
\left\|\prod_{l=1}^{L}(\Sigma_k + \Delta_l) - \Sigma_k^L\right\|_{2/L} \leq & \sum_{l=1}^{L}\binom{L}{l}\|\Sigma_k^{L-l}\|_{2/L} L^l \varepsilon^{l/2} \\
\leq & \sum_{l=1}^{L}\binom{L}{l} r^{L/2}\|\Sigma_k^{L-l}\|_*^{L/2} L^l \varepsilon^{l/2} \\
\leq & \sum_{l=1}^{L}\binom{L}{l} r^{L/2}\|\Sigma_k\|_*^{L/2(L-l)} L^l \varepsilon^{l/2} \\
\leq & r^{L/2}(\|\Sigma_k\|_*^{L/2} + L\varepsilon^{1/2})^L - \|\Sigma_k\|_*^{L^2/2}) \\
\leq & r^{L/2}\|\Sigma_k\|_*^{L^2/2}\left(\left(1 + L\sqrt{\frac{\varepsilon}{\|\Sigma_k\|_*^L}}\right)^L - 1\right) \\
\leq & r^{L/2}\|\Sigma_k\|_*^{L^2/2} e L^2 \sqrt{\frac{\varepsilon}{\|\Sigma_k\|_*^L}} \\
\leq & r^{L/2}\|\Sigma_k\|_*^{L(L-1)/2} e L^2 \varepsilon^{1/2}
\end{aligned}$$

where in the penultimate line, we used $(1+x)^n - 1 \leq enx$ whenever $x < \frac{1}{n}$, and using the assumption of the proposition.

Ultimately, we thus obtain

$$\left|\left\|\prod_{l=1}^{L} A_l\right\|_{2/L}^{2/L} - \|A_k\|_F^2\right| \leqslant r\|A_k\|_*^{L-1} e^{2/L} L^{4/L} \varepsilon^{1/L}$$

which concludes the proof for the first inequality. The second inequality is obtained by using the triangular inequality. □

Note that the above proposition can be used to argue that optimizing a deep linear network of depth L will very quickly co-optimize the L_p-Schatten norm of the product, with an exponential time constant λ/L. In fact, even in a deep linear network, it can be shown that the matrices become exponentially balanced using the same proof as in Lemma B.1. For the assumption $\varepsilon \leq \frac{1}{L^4}\min_l \|A_l\|_*^L$ to hold, it suffices that the matrix norm remains lower-bounded by a strictly positive value, a reasonable assumption for loss functions of interest. Note however that this assumption is used only to obtain the convenient upper bound expression, but the exponential decay is trivially true even without it.

B.5 On the boundness condition of A and B

We now examine the assumption that both A and B are bounded. It can be shown that, under stochastic dynamics with momentum and for certain loss types, A and B will remain bounded with high probability. Below, we present two examples of sufficient conditions on the loss function to ensure this boundedness. Although one could formulate an expanding set of such sufficient conditions to cover a broader class of losses that lead to bounded parameters, there are still practical scenarios where certain losses might not satisfy these conditions. In these cases, practitioners often achieve stable dynamics through hyperparameter tuning, and our theorem remains applicable, as it is designed to accommodate these scenarios rather than exclude them. Therefore, the boundedness assumption is broadly relevant in practice, particularly in training setups that use weight decay, and we leverage this assumption to discuss the rank-regularization effect that impacts such training processes.

B.5.1 Sufficient conditions

Henceforth, we will refer by $\theta = (A, B)$.

Sufficient condition 1: Gradient flow with lower bounded loss function We consider a the following gradient flow dynamics with weight decay:

$$\dot{\theta} = -\eta(\nabla_\theta L + \lambda\theta)$$

η is the learning rate hyperparameter, and λ the weight decay strength.

A sufficient condition on the loss is that it is lower bounded, which is the case for most common losses.

Indeed, the above dynamic is the gradient flow dynamic of the loss $L'(\theta) := L(\theta) + \lambda\|\theta\|^2$. Given that L' is also lower bounded, and that $L'(\theta_t)$ is a monotonically decreasing function of time, $L'(\theta_t)$ must converge to a constant real value, i.e. $L(\theta_t) + \lambda\|\theta_t\|^2 \to_{t\to\infty} c$ for some c. If $\|\theta\|$ is unbounded from above, then necessarily L is unbounded from below, which is a contradiction.

Sufficient condition 2: Gradient flow with momentum with Lipschitz gradient We consider a the following gradient flow dynamics with momentum and decoupled weight decay:

$$\dot{G} = \mu(\nabla_\theta L - G)$$
$$\dot{\theta} = -\eta(G + \lambda\theta)$$

where G is the exponential average of the gradient of θ. μ is the momentum hyperparameter, η the learning rate, and λ the weight decay strength.

A sufficient condition on the gradient is to be $\min(1, \frac{\eta\lambda}{\mu})$-Lipschitz with respect to the parameters θ sufficiently far, i.e. for $\|\theta\| > P$ for a given P:

The momentum makes the analysis of the dynamics more complicated. However, defining $F = \begin{bmatrix} \theta \\ G \end{bmatrix}$, and $M = \begin{pmatrix} \eta\lambda & \eta \\ 0 & \mu \end{pmatrix}$ and $U = \begin{bmatrix} 0 \\ \nabla_\theta L \end{bmatrix}$ one can rewrite the equations as:

$$\dot{F} = -MF + \mu U$$

The derivative of the squared norm of F verifies:

$$\begin{aligned}
\frac{d}{dt}\|F\|^2 &= \operatorname{Tr} \dot{F}F^T \\
&= -\operatorname{Tr} MFF^\top + \mu \operatorname{Tr} U^\top F \\
&\leqslant -\min(\eta\lambda, \mu)\|F\|^2 + \mu \operatorname{Tr} U^\top F \\
&\leqslant -\min(\eta\lambda, \mu)\|F\|^2 + \mu\|\nabla_\theta L\|\|F\| \\
&= \mu\|F\|\left(\|\nabla_\theta L\| - \min(1, \frac{\eta\lambda}{\mu})\|F\|\right) \\
&\leqslant \mu\|F\|\left(\|\nabla_\theta L\| - \min(1, \frac{\eta\lambda}{\mu})\|\theta\|\right)
\end{aligned}$$

The dynamics of F are flow dynamics. Whenever $\|F\|$ reaches P, its norm is decreasing. $\|F\|$ can thus never exceed P. As $\|F\|$ is an upperbound on $\|\theta\|$, the same holds for $\|\theta\|$. We also observe that the Lipschitz condition doesn't need to hold for all $\theta > P$. In fact, it suffices that it holds for any borderless submanifold of codim -1 (for example the sphere of radius M) that contains the initialization point.

Note on stochasticity To deal with stochasticity, we consider similar equations:

$$dG = \mu(\nabla_\theta L dt + dW - G dt)$$
$$d\theta = -\eta(G + \lambda\theta)dt$$

https://doi.org/10.52202/079017-0146

which become:

$$dF = -MFdt + \mu Udt + \mu dW$$

The integral form is:

$$F = F(0) + \mu e^{-tM} \int e^{sM} U ds + \mu e^{-tM} \int e^{sM} dW$$

The process can diverge because of the stochasticity. However, similarly to proposition B.2, we can fix for any $\delta > 0$ an upper bound on W that holds with probability at least $1 - \delta$; this allows to bound the contribution of the stochasticity. We deal with the gradient component similar to the non-stochastic proof.

B.5.2 Pathological examples where the boundedness does not hold

An example of a loss for which the parameters will diverge is when we allow losses to be negative, and diverge to minus infinity "stronger" than the weight regularization term.

An obvious, albeit constructed such loss is $L(AB^\top) = -\|AB^\top\|^2$. Then, the gradient of L w.r.t. A (resp. B) is $\nabla_A L = -AB^\top B$ (resp. $\nabla_B L = -BA^\top A$). Even with the decay term, one can see that if A, B are initialized to be e.g. orthogonal matrices scaled by some $\alpha > \lambda$, both A, B will diverge to infinity.

Such negatively unbounded objective functions to be minimized may be found in e.g. the reinforcement learning setting, when using undiscounted returns.

C Equilibrium condition of 2-layer Linear network

We assume $\lambda > 0$ and that the singular values of $Y^\top X$ are all non-zero and distinct.

We start with the following set of equations:

$$\lambda B^\top = A^\top (S - AB^\top) \tag{34}$$

$$\lambda A = (S - AB^\top) B \tag{35}$$

Clearly, (34) implies

$$\lambda B^\top B = A^\top (S - AB^\top) B \tag{36}$$

$$\lambda A^\top A = A^\top (S - AB^\top) B \tag{37}$$

and thus that $B^\top B = A^\top A$.

Furthermore, (34) also implies

$$\lambda AB^\top = AA^\top (S - AB^\top) = AA^\top S - AA^\top AB^\top \tag{38}$$

$$\lambda AB^\top = (S - AB^\top) BB^\top = SBB^\top - AB^\top BB^\top \tag{39}$$

Using $B^\top B = A^\top A$, we get that $AA^\top S = SBB^\top$.

Denoting by a_i, b_i the i-th row of A, B, for any $i, j \in [1..d_{1,3}]^2$, we have $s_i a_i^\top a_j = s_j b_i^\top b_j$ and $s_j a_j^\top a_i = s_i b_j^\top b_i$. Thus, $\|a_i\|^2 = \|b_i\|^2$, and $a_i^\top a_j = b_i^\top b_j = 0$, since s_i, s_j are distinct and positive. Taken together, $AA^\top$ is a diagonal matrix, which we denote by D. We have $D = diag((\|a_i\|^2)_{i \in [1..d_{1,3}]})$.

In particular, (38) implies $\lambda AB^\top = D(S - AB^\top)$, i.e. $(\lambda I + D) AB^\top = DS$. Because the entries of D are positive, $(\lambda I + D)$ is invertible, and thus $AB^\top = (\lambda I + D)^{-1} DS$. In other words, the off-diagonal entries of $AB^\top$ are zero, i.e. $a_i^\top b_j = 0$ for all $i \neq j$, $i, j \in [1..d_{1,3}]^2$.

In particular, for a given i, we have

$$\lambda a_i^\top b_i = \|a_i\|^2 (s_i - a_i^\top b_i) \tag{40}$$

$$\lambda \|a_i\|^2 = (s_i - a_i^\top b_i) a_i^\top b_i \tag{41}$$

Using the positivity of s_i and λ, one can see that necessarily, $a_i = b_i$.

D On the link between the full loss and the restricted loss

In the theoretical section, we study a pruned version of the total losses:

$$\mathcal{L}_{L2}(A, B, \theta) \coloneqq L(AB^\top,) + \frac{\lambda}{2}(\|A\|^2 + \|B\|^2),$$
$$\mathcal{L}_*(AB^\top, \theta) \coloneqq L(AB^\top) + \lambda\|AB^\top\|_*,$$

where the remaining parameters are not accounted for. This in fact still accounts for the general case. Indeed:

- stationary points of $\mathcal{L}_{L2}$ will be also stationary in (A, B), hence Lemma 3.2 still holds. In fact, the latter condition suffices.
- For theorem 3.3, the same proof B.3 shows that (A, B, θ) is a local minima of $\mathcal{L}_{L2}$ iff 1) $(W = AB^\top.\theta)$ is a local minima of $\mathcal{L}_*$ constrained to matrices W of rank r and 2) $A^\top A = B^\top B$
- in the proof B.4.1 of theorem 3.4, the gradient G of the first lemma now hides a dependence in θ, but the proof still hold as is. In the second lemma, the gradient G_t depends on the time, and is indirectly hiding a dependence on θ.

E Link between solutions of AdamW and $L2$-regularization

Consider the following dynamic induced by AdamW, with $\lambda > 0$:

$$\begin{aligned}
G_t &\leftarrow \beta_1 \cdot G_{t-1} + (1-\beta_1) \cdot \nabla_W \mathcal{L}(W_t) \\
B_t &\leftarrow \beta_2 \cdot B_{t-1} + (1-\beta_2) \cdot \nabla_W \mathcal{L}(W_t)^2 \\
\hat{G}_t &\leftarrow G_t/(1-\beta_1^t) \\
\hat{B}_t &\leftarrow B_t/(1-\beta_2^t) \\
W_{t+1} &\leftarrow W_t - \eta \cdot \left(\hat{G}_t / \left(\sqrt{\hat{B}_t} + \varepsilon \right) + \lambda W_t \right)
\end{aligned}$$

where η represents the learning rate and $\beta_1, \beta_2, \varepsilon$ are the common hyperparameters of Adam, W_t is the parameter at time t, and where the various operations are applied element-wise.

Note that the term $-\lambda W_t$ stems from weight decay.

If the dynamic converges, then necessarily, $G_\infty = \nabla_W \mathcal{L}(W_\infty)$, $B_\infty = (\nabla_W \mathcal{L}(W_\infty))^2$, and thus $\lambda W_\infty = \frac{-\nabla_W \mathcal{L}(W_\infty)}{|\nabla_W \mathcal{L}(W_\infty)| + \epsilon}$. Clearly, this implies that $\lambda |W_\infty| < 1$. If we further assume that $\lambda |W_\infty| \ll 1$, then the condition becomes $\epsilon \lambda W_\infty \approx -\nabla_W \mathcal{L}(W_\infty)$, which is the equilibrium point of a $L2$-regularized loss with regularization strength $\frac{\epsilon\lambda}{2}$. Thus, the stationary points of the AdamW optimizer can in practice correspond to stationary points of $L2$-regularized loss, and thus the same low-rank inducing solutions can be found.

We show in Fig. 5 a toy experiments illustrating the equivalence in the solutions found by AdamW with decay strength λ_{WD} and hyperparameter ϵ, with those found by Adam with $L2$-regularization with regularization strength $\lambda_{L2} = \lambda_{WD}\epsilon$. In particular, we illustrate how a factorized parametrization in this setting will still result in solutions that minimize the nuclear norm, even when trained with AdamW.

F Pretrained foundation models: ViT

We provide in Figure 6 the ViT counterpart of Figure 4.

https://doi.org/10.52202/079017-0146

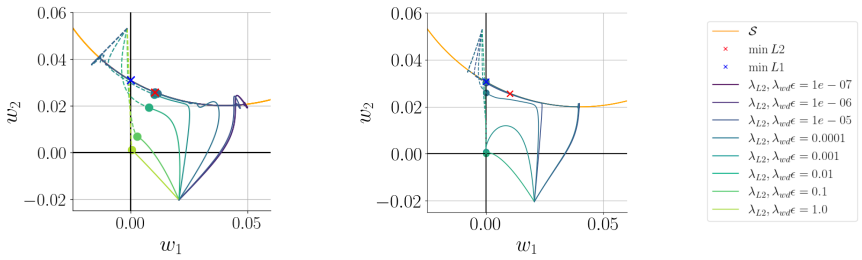

Figure 5: Trajectory of w_1, w_2 in the 2D plane when optimizing the underlying parameter for various hyperparameters. At every coordinate in the plane, the loss is defined as the squared distance to the surface $\mathcal{S}$ in orange. The red (resp. blue) cross represents the points on $\mathcal{S}$ minimizing the $L2$-norm (resp. $L1$-norm). *Left*: w_1, w_2 are directly parametrized and optimized by AdamW with decoupled weight decay (in solid line) or Adam with $L2$-regularization (in dotted line). As conjectured, the convergence point of AdamW given the hyperparameter ϵ and decay strength $\lambda_w d$ corresponds to that of the equilibrium point of the $L2$-regularized loss with regularization strength $\lambda_{L2} = \lambda_w d\epsilon$. *Right*: w_1, w_2 are parameterized as a product of two scalars, i.e. $w_1 = a_1 b_1, w_2 = a_2 b_2$, where a_1, b_1, a_2, b_2 are now optimized by AdamW or Adam with $L2$ regularization. Again, the two optimizers find the same convergence point for equivalent hyperparameters. However, the solution found now corresponds to those of the loss regularized by the $L1$-norm of w_1, w_2, (corresponding to the nuclear norm for scalars) as predicted.

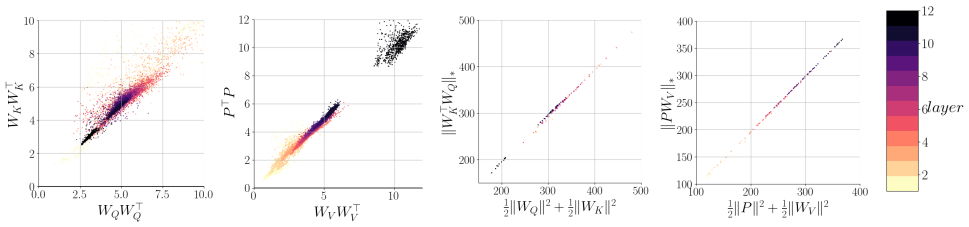

Figure 6: Analyses of attention layers in the pre-trained Vision Transformer [Wu et al., 2020], available on huggingface under the id "google/vit-base-patch16-224-in21k". The leftmost (resp. center left) shows the squared norm of every row of W_Q (resp. W_V), for the first head of each layer, against the norm of the corresponding row of W_K (resp. column of P). The condition $W_K W_K^\top = W_Q W_Q^\top$ would require these norms to be equal, which in fact is mostly true. While the model has not reached a stationary point, this indicates the optimization has advanced enough for this sufficient condition for $\mathcal{L}_*$ to be identical to $\mathcal{L}_{L2}$ to emerge. In fact, the centre right (resp. rightmost) plot shows the scatter plot mapping the Frobenius norm against the nuclear norm for all heads across all layers. The two norms almost perfectly coincide.

G Language modelling experimental details

Here, we present details of our language modeling experiments, employing standardized values from the literature and consistent, untuned hyperparameters across all trials. Unless specified otherwise, we utilize the conventional GPT-2 transformer architecture with LayerNorm (Ba et al., 2016), incorporating MLPs between self-attention layers and applying skip-connections after each layer. Training is conducted using a standard (autoregressively) masked cross-entropy loss, omitting an input embedding layer but incorporating an output projection before computing logits. Further details can be found in Table 2.

Table 2: Hyperparameters for language modelling experiments.

Hyperparameter	Value
Dataset	The pile [Gao et al., 2020]
Tokenizer	GPT-2 tokenizer - we append a special "EOS" token between every sequence
Context size	756
Vocabulary size	50257
Vocabulary dim	756
Optimizer	Adam [Kingma and Ba, 2015] with $\epsilon = 1e^{-8}, \beta_1 = 0.9, \beta_2 = 0.95$
Weight decay	See main text
Batchsize	128
Gradient clipping	Global norm of 1.
Positional encodings	We add standard positional encodings.
Dropout	We use an embedding dropout of 0.1 right after adding positional encodings.
Architecture details	12 layers, 12 heads, key size 64, token size 756, no input- but output-embedding
Weight init	$W \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.02$ and bias parameter to zero. We scale all weight matrices before a skip connection with $\frac{1}{2\sqrt{N}}$ with N the number of layers.
Learning rate scheduler	Linear warm-up starting from $1e^{-6}$ to $1e^{-3}$ in the first 8000 training steps, cosine annealing to 10% of the learning rate after warm-up for the end of training
MLP size	Widening factor 4 i.e. hidden dimension $4 * 756$ with ReLU non-linearities [Hahnloser et al., 2000]

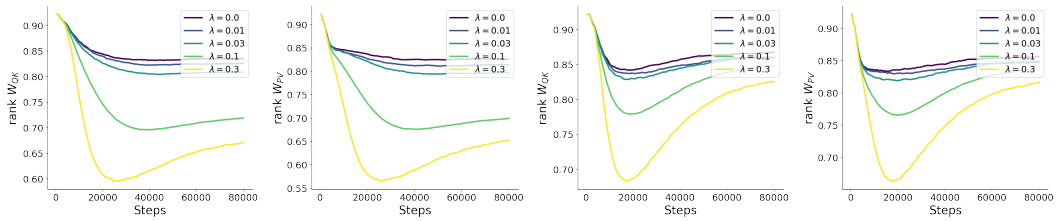

Figure 7: The rank of weight matrix products $W_K^\top W_Q$ and PW_V averaged across heads of layer 7 (left and outer left) and layer 9 (right and outer right) of an autoregressive transformers trained on the Pile [Gao et al., 2020]. For both layers, the decay strength applied to attention layers is varied, while keeping the strength for all other layers fixed. In all cases, we observe again that rank reduction correlates strongly with weight decay strength when optimizing with AdamW.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide both theoretical and empirical results supporting the main claim of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.

https://doi.org/10.52202/079017-0146

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The discussion highlights some of the limitation of this work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All statements are provided with proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.

- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We hope to provide all hyperparameters and experimental details in the appendix and provide code to reproduce most of the experiments.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We aim to collect the code as soon as possible in a Git repository.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all experimental details in the main text, as well as the appendix.

https://doi.org/10.52202/079017-0146

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer:[No]

   Justification: Many of the experiments were performed on large scale models or foundation models, rendering the computation of multiple seeds unrealistic.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide an estimate of compute used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper should be considered a theory and/or conceptual paper. We discussed implication for robust machine learning in the main text, and can not anticipate that the presented results can not conform in any aspect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper shows the importance of choosing a proper weight decay while training transformer.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No data and models realese.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

 https://doi.org/10.52202/079017-0146

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.