Weight decay induces low-rank attention layers

Anonymous Authors[1]

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 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 that the 2 losses become identical exponentially quickly during training. We thus explain in particular why optimizing $L2$-regularized objectives result in a low-rank inducing pressure on the matrix product very early during 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 modeling, indeed induce a significant reduction in the rank of $W_K^T W_Q$ and PW_V. We find that, in accordance with existing work, inducing low rank in attention layers damages language models, and see a performance improvement when omitting weight decay in attention layers.

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 & 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 & 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 & Wang, 2023; Arora et al., 2019; Li et al., 2021; Razin & 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 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

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

where ϕ is typically a softmax operation applied column-wise and M is typically the 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 & Hutter, 2022), and 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.

We will see in the following that optimizing neural network models with this particular parametrization in conjunction with the $L2$-regularization, i.e. optimizing the objective

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

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 & 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 BiT (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 regularization 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 setting, including when optimization with decoupled weight decay (Loshchilov & Hutter, 2019), on models ranging from deep linear networks to language models as well as Vision Transformers. 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 pretrained foundation models such as LLAMA 2 (Touvron et al., 2023) and Vision Transformer (Wu et al., 2020), by analyzing their pretrained weights.

2. Related Work

The setting we study is closely related to a setting extensively studied in the Matrix Completion literature (Srebro & Shraibman, 2005; Sun & Luo, 2016; Candes & 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 & 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 correctly recover the underlying low-rank matrix (Sun & Luo, 2016; Candes & 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 *linear* networks on the parametrized matrix. For example, small weight initialization in this setting was shown to induce low rank (Jacot et al., 2022; Arora et al., 2019; Li et al., 2021). 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 & 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

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

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 an 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

$$\begin{aligned}
\|W\|_* = \mathrm{Tr}(S) &= \mathrm{Tr}(U^\top AB^\top V) && (5)\\
&\leq \sqrt{\mathrm{Tr}(U^\top AA^\top U)\mathrm{Tr}(B^\top VV^\top B)} && (6)\\
&= \|A\|\|B\| && (7)\\
&\leq \frac{1}{2}(\|A\|^2 + \|B\|^2) && (8)
\end{aligned}$$

Thus, when using a factorized parametrization coupled with $L2$-regularization, the objective in equation 2 becomes an upper bound of the nuclear norm-regularized objective

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

3.2. Equivalence of optimization solution

In the following, we will first show that in fact, any objective of the form in equation 2 will coincide at any stationary point with the nuclear-norm regularized loss in equation 9, thus introducing a low-rank inducing bias in the solution found. For simplicity, we assume the argument A, B of $\mathcal{L}_{L2}$ to have number of rows greater or equal to number of columns, although our result holds in the general case.

We start by providing a sufficient condition under which the averaged Frobenius norm of 2 matrices would correspond to the nuclear norm of their product.

All proofs can be found in Appendix A.

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 2 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 our main result, which shows that this claim is in fact 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) $A^\top A = B^\top B$.*

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 inadvertedly) 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 2 parametrization may result in different optimization, and thus different solution, even if the loss landscape share the same local minima.

3.3. Optimization dynamic in the gradient flow limit

The above result give an equivalence of the local minima of the 2 losses. Our next result shows that the 2 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 $A^\top A - B^\top B$.
We provide in the appendix similar result when considering gradient flow with noise, as well as with momentum and decoupled weight decay.
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 & 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 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 B)

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

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

In particular, this implies, for any i,

$$\lambda \|a_i\|^2 = (s_i - \|a_i\|^2)\|a_i\|^2. \tag{18}$$

Clearly, if $\lambda \ge s_i$, 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. $d_2 \ge d_{1,3}$. In that

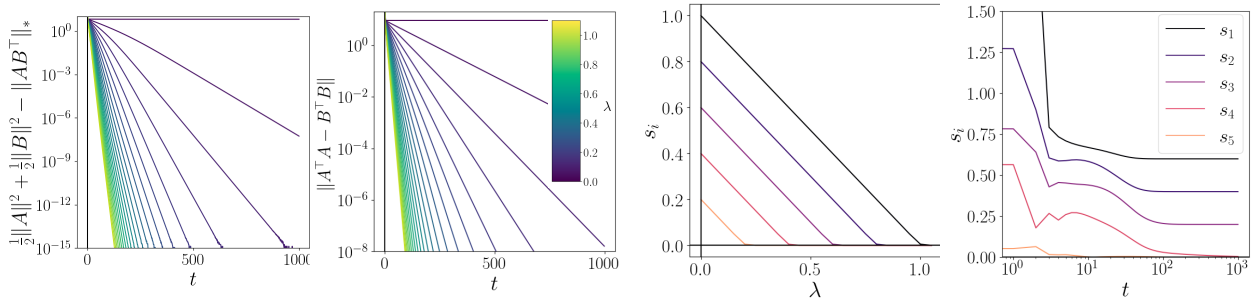

Figure 1. Optimization by gradient descent of 2 5-by-5 matrices A, B on the $L2$-regularized loss $\|AB^\top - D\|^2 + \frac{\lambda}{2}(\|A\|^2 + \|B\|^2)$ where $D = \text{diag}(0.2, 0.4, 0.6, 0.8, 1)$, with various regularization strength λ. t denotes the number of optimization steps. *Left*: difference 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 $\lambda = 0.4$.

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

$$W = \frac{1}{1+\lambda} S, \tag{19}$$

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 & 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 & 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 that under some simplifying assumptions, they coincide with those of a $L2$-regularized loss with a different regularization strength.

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. 2 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

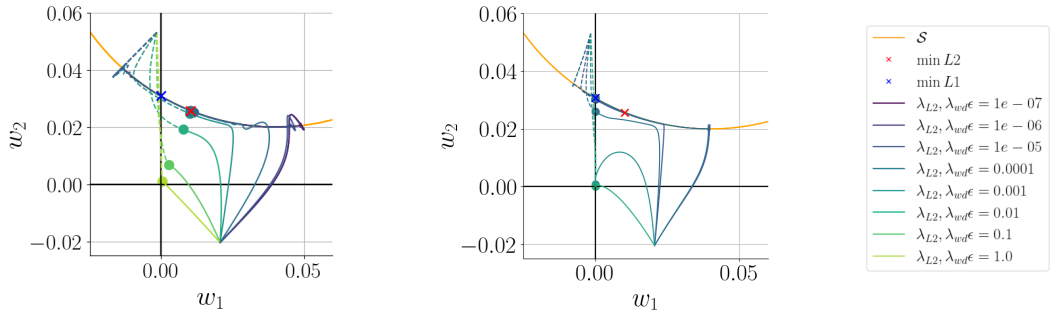

Figure 2. Trajectory of w_1, w_2 in the 2D plane when optimizing the underlying parameter for various hyperparameters. At every coordinate in the plane, the loss is defined as the squared distance to the surface $\mathcal{S}$ in orange. The red (resp. blue) cross represent the points on $\mathcal{S}$ 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 $\lambda_w d$ correspond 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 2 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 2 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.

that minimizes the nuclear norm, even when trained with AdamW.

4. Empirical results

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 foundation models. While we cannot train these models ourselves, we train a small-scale language model as well as Vision Transformer without changing common hyperparameters to demonstrate that their typical training is affected by 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.

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.

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 modeling (Fu et al., 2023; Poli et al., 2023).

We train a 2-layer self-attention only Transformer with AdamW optimizer on minibatch of size 128, for $N = 20$. To simulate additional noise, we perturb 5% of the labels with random labeling (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 equation 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.

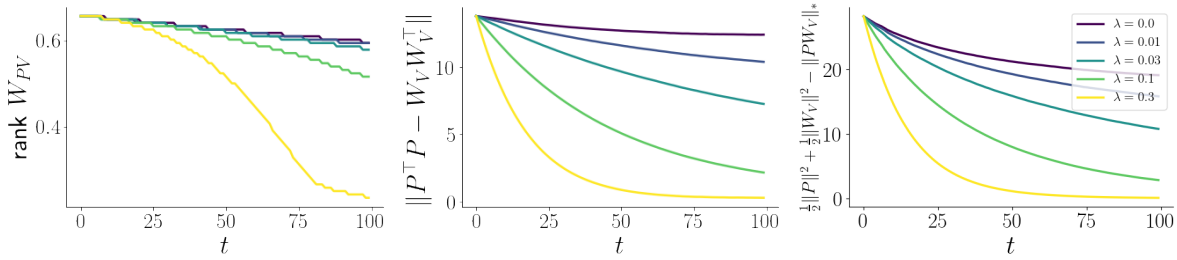

Figure 3. Left: The rank of weight matrix product PW_V of the first layer of a 2-layer Transformer trained on the associative recall task, during training, with AdamW, for various decay strength. To better account for the effect of weight decay on the attention layers, only the decay strength applied to attention layers is varied, while the strength for all other layers 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 correlated 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 $W_K^\top W_Q$.

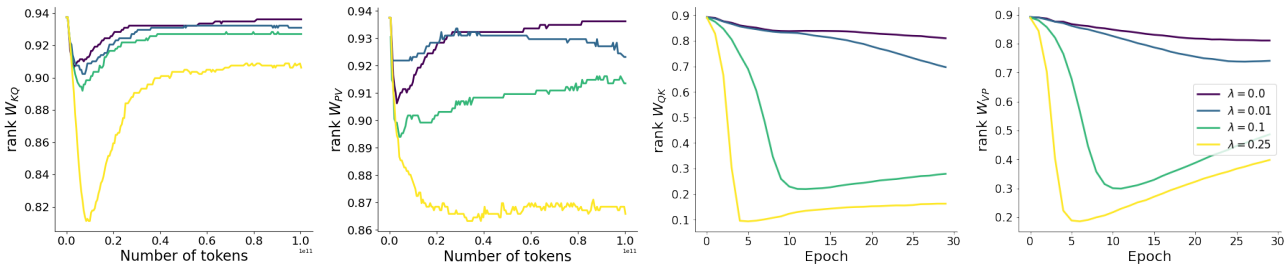

Figure 4. Left, center left: The rank of weight matrix products $W_K^\top W_Q$ and PW_V averaged across heads and layers within autoregressive Transformers trained on the Pile (Gao et al., 2020). *Center right, right*: The rank of weight matrix products $W_K^\top W_Q$ and PW_V averaged over all heads and all layers of a Vision Transformer trained following (Irandoust et al., 2022) on the ImageNet dataset (Deng et al., 2009). In both setting, 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.

4.2. 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 as well as 355 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, but we note that the weight decay strength is set to be 0.1 throughout the different setting considered in the following for all parameters of the MLPs, except when stated otherwise.

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 4).

Table 1. Test set perplexity of 125 million and 355 million parameter Transformer models trained on the Pile for 100 billion tokens with AdamW and different weight decay strengths λ.

Optimizer	125Mio	355Mio
AdamW, $\lambda = 0.25$	11.09	10.77
AdamW, $\lambda = 0.1$	10.95	10.35
AdamW, $\lambda = 0.01$	10.87	10.12
AdamW, $\lambda = 0.0$	10.87	9.87

Surprisingly, not applying weight decay to the attention layers improves test set accuracy for both model sizes, see Table 1, indicating that, while a weight decay strength of 0.1 does not significantly reduce the rank of the matrices, the pressure is enough to hurt the performance. Note that when during weight decay off completely i.e. removing it from

the weights of the MLPs as well, the model performance does not increase further but drops: We measure 10.92 and 10.21 test set perplexity for the 125Mio and 355Mio model respectively when turning off weight decay completely. We train all of our models for 'only' 100 billion tokens - running the models for longer does not seem to change relative difference between models.

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 hyperparameters fixed. We observe a similar effect of the decay strength on the ranks of the matrices W_{QK}, W_{VP} (c.f. Fig 4).

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 A.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 4.4 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 equation 5 is indeed mostly tight, consistent with a rank regularizing training.

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 $P W_V$.

We begin by establishing, theoretically, that at the stationary point of any arbitrary $L2$-regularized loss for a model employing factorized parametrization $W = AB^\top$, the Frobenius norm of matrices A and B converges precisely to the nuclear norm of the matrix W. Additionally, we demonstrate that, all local minima of such loss function regularized by the Frobenius norm of matrices A and B coincide with the local minima of the same loss function when regularized by the nuclear norm of matrix W. Finally, we show the discrepancy between the 2 losses disappear exponentially quickly during training, implying that the nuclear norm is co-optimized very early. We stress that our findings show that standard $L2$-regularization in combination with the discussed parametrization $W = AB^\top$ provides an intriguingly easy to optimize alternative to obtain a low rank matrix W, without introducing any suboptimal local minima.

We next validate these empirical results by studying the influence of weight decay in toy settings as well as when training language models and Vision Transformers, or analyzing large well-known pre-trained models hosted on huggingface. All of 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 are in fact in practice affected by the same regularization.

Furthermore, we find that turning off weight decay in the attention weights improves performance in our language modeling experiments while turning off the weight decay in the feedforward network part i.e MLPs was found on the other hand to hurt performance. 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 to understand 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.

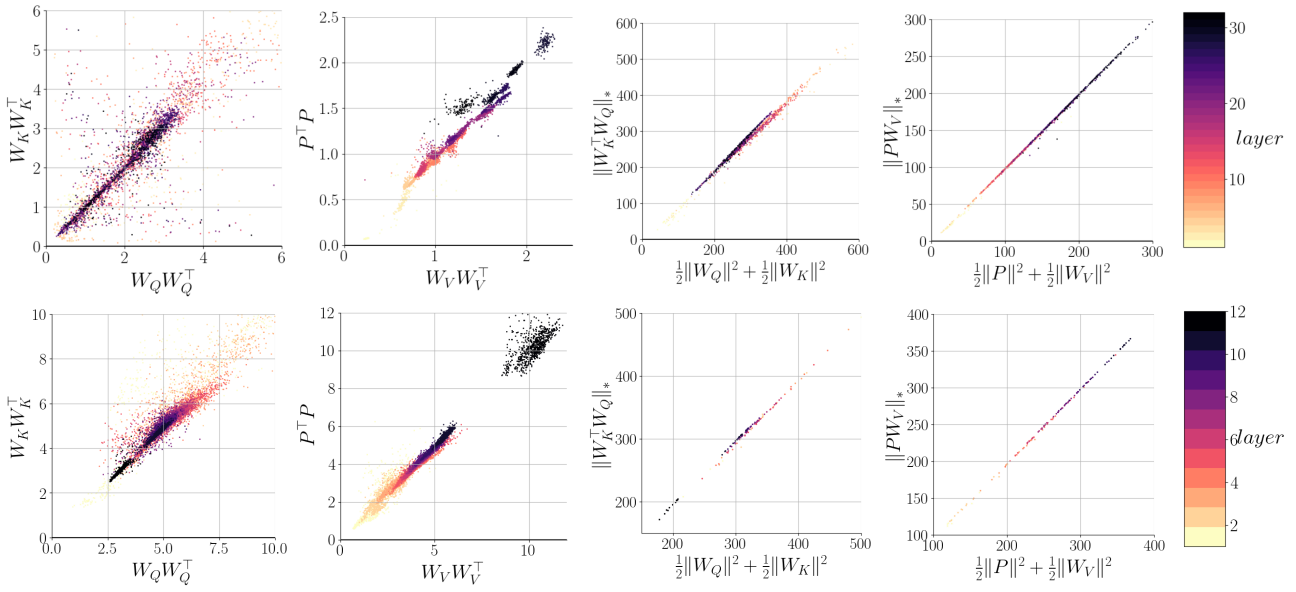

Figure 5. Top row: Analyses of attention layers in the pretrained LLAMA 2 model with 7 Billion parameters (Touvron et al., 2023). The leftmost (resp. center left) shows the squared norm of every row of W_Q (resp. W_V), for the first head of each layer, against the norm of the corresponding row of W_K (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 2-norms almost perfectly coincide. *Bottom row*: same analysis done on a pretrained Vision Transformer (Wu et al., 2020), available on huggingface under the id "google/vit-base-patch16-224-in21k".

References

Andriushchenko, M., D'Angelo, F., Varre, A., and Flammarion, N. Why do we need weight decay in modern deep learning?, 2023.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *CoRR*, abs/1905.13655, 2019. URL http://arxiv.org/abs/1905.13655.

Bhojanapalli, S., Yun, C., Rawat, A. S., Reddi, S. J., and Kumar, S. Low-rank bottleneck in multi-head attention models, 2020.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

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

Dai, Z., Karzand, M., and Srebro, N. Representation costs of linear neural networks: Analysis and design. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26884–26896. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e22cb9d6bbb4c290a94e4fff4d68a831-Paper.pdf.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. 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.

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

Galanti, T., Siegel, Z. S., Gupte, A., and Poggio, T. Characterizing the implicit bias of regularized sgd in rank minimization, 2023.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: an 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization, 2017.

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

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021a. URL https://arxiv.org/abs/2106.09685.

Hu, Z., Nie, F., Wang, R., and Li, X. 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.

Irandoust, S., Durand, T., Rakhmangulova, Y., Zi, W., and Hajimirsadeghi, H. Training a vision transformer from scratch in less than 24 hours with 1 gpu, 2022.

Jacot, A., Ged, F., Şimşek, B., Hongler, C., and Gabriel, F. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, 2022.

Khodak, M., Tenenholtz, N., Mackey, L., and Fusi, N. Initialization and regularization of factorized neural layers, 2022.

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

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

Li, Z., Luo, Y., and Lyu, K. 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.

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

Mackay, D. J. C. 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.

Phuong, M. and Hutter, M. Formal algorithms for transformers, 2022.

Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., and Ré, C. Hyena hierarchy: Towards larger convolutional language models, 2023.

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

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

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

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

Singh, A. K., Chan, S. C. Y., Moskovitz, T., Grant, E., Saxe, A. M., and Hill, F. The transient nature of emergent in-context learning in transformers, 2023.

Srebro, N. and Shraibman, A. Rank, trace-norm and max-norm. In Auer, P. and Meir, R. (eds.), *Learning Theory*, pp. 545–560, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31892-7.

Sun, R. and Luo, Z.-Q. 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.

Tibshirani, R. J. Equivalences between sparse models and neural networks. 2021. URL https://api.semanticscholar.org/CorpusID:233468306.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W.,

Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017.

von Oswald, J., Niklasson, E., Schlegel, M., Kobayashi, S., Zucchet, N., Scherrer, N., Miller, N., Sandler, M., y Arcas, B. A., Vladymyrov, M., Pascanu, R., and Sacramento, J. Uncovering mesa-optimization algorithms in transformers, 2023.

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

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., and Vajda, P. Visual transformers: Token-based image representation and processing for computer vision, 2020.

Xie, Z., zhiqiang xu, Zhang, J., Sato, I., and Sugiyama, M. 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.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. 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.

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

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features, 2021.

Ziyin, L. and Wang, Z. spred: Solving l1 penalty with SGD. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43407–43422. PMLR, 23–29 Jul 2023.

Ziyin, L., Li, B., and Meng, X. Exact solutions of a deep linear network. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=X6bp8ri8dV.

A. Proofs of theoretical results

A.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

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

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

□

A.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

$$\partial_A L = \left(\frac{\partial L}{\partial W}\Big|_{W = AB^\top} \right) B \tag{20}$$

$$\partial_B L = \left(\frac{\partial L}{\partial W}\Big|_{W = AB^\top} \right)^\top A \tag{21}$$

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

$$\lambda A = GB \tag{22}$$

$$\lambda B = G^\top A \tag{23}$$

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

□

A.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 is 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. □

A.4. Proof of Theorem 3.4

A.5. Exponential decay of $A^\top A - B^\top B$

We begin by showing the following result for the vailla gradient flow limit.

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

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:

$$\tau \dot{a}^i = Gb^i - \lambda a^i \tag{24}$$

$$\tau \dot{b}^i = G^\top a^i - \lambda b^i \tag{25}$$

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 $A^\top A - B^\top B = Qe^{-\frac{2\lambda}{\tau}}$, 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.

We start by stating a lemma:

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

We now state the main result:

Proposition A.3. *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,

$$H_t^A = \mu \int_0^t e^{-\mu(t-s)} G_s B_s ds + \sqrt{\frac{\mu\sigma^2}{2}} W^A_{1-e^{-2\mu t}} \tag{26}$$

$$H_t^B = \mu \int_0^t e^{-\mu(t-s)} G_s^\top A_s ds + \sqrt{\frac{\mu\sigma^2}{2}} W^B_{1-e^{-2\mu t}} \tag{27}$$

$$A_t^\top A_t - B_t^\top B_t = e^{-2\eta\lambda}(A_0^\top A_0 - B_0^\top B_0) - \eta \int_0^t e^{-2\eta\lambda(t-s)}(H_s^{A\top} A_s + A_s^\top H_s^A - H_s^{B\top} B_s - B_s^\top H_s^B)ds \tag{28}$$

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. Similarly,

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

We can rewrite $dA_t = -\eta\left(G_t B_t dt + \sigma dW_t^A - \frac{dH_t^A}{\mu} + \lambda A_t dt\right)$. If we now look at $A^\top A - B^\top B$, we get

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

which gives

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

□

Let's analyse the consequences of such a statement. First observe that, using lemma A.2, 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 $\sigma\sqrt{\mu n d \ln \frac{\varepsilon}{2nd}}$. If we assume A_t and B_t remain L2-bounded by $M > 0$, and that G_t remains L2-bounded by K (either using a locally Lipschitzian loss, or using clipping), then H^A and H^B will with probability $1-\varepsilon$ remain bounded by $KM + \sigma\sqrt{\mu n d \ln \frac{\varepsilon}{2nd}}$. 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 $4\frac{\eta M}{\lambda}\left(KM + +\sigma\sqrt{\mu n d \ln \frac{\varepsilon}{2nd}}\right)$. 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.

A.6. Upper bound of $\|AB^\top\|_*$ and $\frac{1}{2}(\|A\|^2 + \|B\|^2)$

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 A.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,

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

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.

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

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 Powers-Stormer inequality (Powers & Størmer, 1970), we have $\|\Delta\|_F^2 = \|V_B \Delta V_B^\top\|_F^2 \leqslant \|Q\|_*$.

From there, we rewrite A as

$$\begin{aligned} A &= U_A \Sigma_A V_A^\top V_B V_B^\top = U_A \begin{pmatrix} S_A V_A^\top V_B \\ 0 \end{pmatrix} V_B^\top \\ &= U_A \begin{pmatrix} V_A^\top V_B (S_B + \Delta) \\ 0 \end{pmatrix} V_B^\top \\ &= U_A \begin{pmatrix} V_A^\top V_B & 0 \\ 0 & I \end{pmatrix} \left(\Sigma_B + \begin{pmatrix} \Delta \\ 0 \end{pmatrix}\right) V_B^\top \end{aligned}$$

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

□

B. 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:

$$\lambda B^\top = A^\top(S - AB^\top) \tag{29}$$

$$\lambda A = (S - AB^\top)B \tag{30}$$

Clearly, equation 29 implies

$$\lambda B^\top B = A^\top(S - AB^\top)B \tag{31}$$

$$\lambda A^\top A = A^\top(S - AB^\top)B \tag{32}$$

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

Furthermore, equation 29 also implies

$$\lambda AB^\top = AA^\top(S - AB^\top) = AA^\top S - AA^\top AB^\top \tag{33}$$

$$\lambda AB^\top = (S - AB^\top)BB^\top = SBB^\top - AB^\top BB^\top \tag{34}$$

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, equation 33 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

$$\lambda a_i^\top b_i = \|a_i\|^2(s_i - a_i^\top b_i) \tag{35}$$

$$\lambda\|a_i\|^2 = (s_i - a_i^\top b_i)a_i^\top b_i \tag{36}$$

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

C. 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	1024
Vocabulary size	50257
Vocabulary dim	756
Optimizer	Adam (Kingma & Ba, 2015) with $\epsilon = 1e^{-8}, \beta_1 = 0.9, \beta_2 = 0.95$
Weight decay	See main text
Batchsize	256
Gradient clipping	Global norm of 1.
Positional encodings	We add standard positional encodings.
Dropout	We use embedding dropout of 0.1 right after adding positional encodings.
Architecture details 125Mio model	12 heads, key size 64, token size 756, no input- but output-embedding
Architecture details 355Mio model	16 heads, key size 64, token size 1024, 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 $3e^{-4}$ for the 125Mio model and $1e^{-4}$ for the 355Mio model in the first 8000 training steps, cosine annealing to $2e-4$ for the next 300 billion tokens
MLP size	Widening factor 4 i.e. hidden dimension $4 * 756$ / $4 * 1024$ with ReLU non-linearities (Hahnloser et al., 2000)