Batch Normalization Provably Avoids Rank Collapse for Randomly Initialised Deep Networks

Hadi Daneshmand[*]
INRIA Paris, ETH Zurich
`seyed.daneshmand@inria.fr`

Jonas Kohler[*]
Department of Computer Science, ETH Zurich
`jonas.kohler@inf.ethz.ch`

Francis Bach
INRIA-ENS-PSL, Paris
`francis.bach@inria.fr`

Thomas Hofmann
Department of Computer Science, ETH Zurich
`thomas.hofmann@inf.ethz.ch`

Aurelien Lucchi
Department of Computer Science, ETH Zurich
`aurelien.lucchi@inf.ethz.ch`

Abstract

Randomly initialized neural networks are known to become harder to train with increasing depth, unless architectural enhancements like residual connections and batch normalization are used. We here investigate this phenomenon by revisiting the connection between random initialization in deep networks and spectral instabilities in products of random matrices. Given the rich literature on random matrices, it is not surprising to find that the rank of the intermediate representations in unnormalized networks collapses quickly with depth. In this work we highlight the fact that batch normalization is an effective strategy to avoid rank collapse for both linear and ReLU networks. Leveraging tools from Markov chain theory, we derive a meaningful lower rank bound in deep linear networks. Empirically, we also demonstrate that this rank robustness generalizes to ReLU nets. Finally, we conduct an extensive set of experiments on real-world data sets, which confirm that rank stability is indeed a crucial condition for training modern-day deep neural architectures.

1 Introduction and related work

Depth is known to play an important role in the expressive power of neural networks [28]. Yet, increased depth typically leads to a drastic slow down of learning with gradient-based methods, which is commonly attributed to unstable gradient norms in deep networks [15]. One key aspect of the training process concerns the way the layer weights are initialized. When training contemporary neural networks, both practitioners and theoreticians advocate the use of randomly initialized layer weights with i.i.d. entries from a zero mean (Gaussian or uniform) distribution. This initialization strategy is commonly scaled such that the variance of the layer activation stays constant across layers [13, 14]. However, this approach can not avoid spectral instabilities as the depth of the network increases. For example, [26] observes that for linear neural networks, such initialization lets all but one singular values of the last layers activation collapse towards zero as the depth increases.

Nevertheless, recent advances in neural architectures have allowed the training of very deep neural networks with standard i.i.d. initialization schemes *despite* the above mentioned shortcomings.

---

[*] Shared first authorship

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Among these, both residual connections and normalization layers have proven particularly effective and are thus in widespread use (see [17, 24, 14] to name just a few). Our goal here is to bridge the explanatory gap between these two observations by studying the effect of architectural enhancements on the spectral properties of randomly initialized neural networks. We also provide evidence for a strong link of the latter with the performance of gradient-based optimization algorithms.

One particularly interesting architectural component of modern day neural networks is Batch Normalization (BN) [17]. This simple heuristics that normalizes the pre-activation of hidden units across a mini-batch, has proven tremendously effective when training deep neural networks with gradient-based methods. Yet, despite of its ubiquitous use and strong empirical benefits, the research community has not yet reached a broad consensus, when it comes to a theoretical explanation for its practical success. Recently, several alternatives to the original "internal covariate shift" hypothesis [17] have appeared in the literature: decoupling optimization of direction and length of the parameters [20], auto-tuning of the learning rate for stochastic gradient descent [3], widening the learning rate range [7], alleviating sharpness of the Fisher information matrix [18], and smoothing the optimization landscape [25]. Yet, most of these candidate justifications are still actively debated within the community. For example, [25] first made a strong empirical case against the original internal covariate shift hypothesis. Secondly, they argued that batch normalization simplifies optimization by smoothing the loss landscape. However, their analysis is on a per-layer basis and treats only the largest eigenvalue. Furthermore, even more recent empirical studies again dispute these findings, by observing the exact opposite behaviour of BN on a ResNet20 network [34].

1.1 On random initialization and gradient based training

In light of the above discussion, we take a step back – namely to the beginning of training – to find an interesting property that is provably present in batch normalized networks and can serve as a solid basis for a more complete theoretical understanding.

The difficulty of training randomly initialized, un-normalized deep networks with gradient methods is a long-known fact, that is commonly attributed to the so-called vanishing gradient effect, i.e., a decreasing gradient norm as the networks grow in depth (see, e.g., [27]). A more recent line of research tries to explain this effect by the condition number of the input-output Jacobian (see, e.g., [32, 33, 23, 7]). Here, we study the spectral properties of the above introduced initialization with a particular focus on the rank of the hidden layer activations over a batch of samples. The question at hand is whether or not the network preserves a diverse data representation which is necessary to disentangle the input in the final classification layer.

As a motivation, consider the results of Fig. 1, which plots accuracy and output rank when training batch-normalized and un-normalized neural networks of growing depth on the Fashion-MNIST dataset [31]. As can be seen, the rank in the last hidden layer of the vanilla networks collapses with depth and they are essentially unable to learn (in a limited number of epochs) as soon as the number of layers is above 10. The rank collapse indicates that the direction of the output vector has become independent of the actual input. In other words, the randomly initialized network no longer preserves information about the input. Batch-normalized networks, however, preserve a high rank across all network sizes and their training accuracy drops only very mildly as the networks reach depth 32.

The above example shows that both rank and optimization of even moderately-sized, unnormalized networks scale poorly with depth. Batch-normalization, however, stabilizes the rank in this setting and the obvious question is whether this effect is just a slow-down or even simply a numerical phenomenon, or whether it actually generalizes to networks of infinite depth.

In this work we make a strong case for the latter option by showing a remarkable stationarity aspect of BN. Consider for example the case of passing N samples $x_i \in \mathbb{R}^d$ arranged column-wise in an input matrix $X \in \mathbb{R}^{d \times N}$ through a very deep network with fully-connected layers. Ideally, from an information propagation perspective, the network should be able to differentiate between individual samples, regardless of its depth [27]. However, as can be seen in Fig. 2, the hidden representation of X collapses to a rank one matrix in vanilla networks, thus mapping all x_i to the same line in $\mathbb{R}^d$. Hence, the hidden layer activations and along with it the individual gradient directions become

---

[3] Computed using $torch.matrix\_rank()$, which regards singular values below $\sigma_{\max} \times d \times 10^{-7}$ as zero. This is consistent with both Matlab and Numpy.

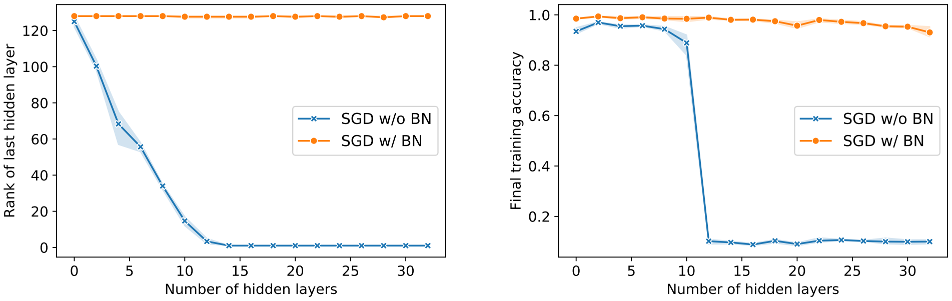


Figure 1: **Effect of depth on rank and learning**, on the Fashion-MNIST dataset with ReLU multilayer perceptrons (MLPs) of depth 1-32 and width 128 hidden units. Left: Rank[3] after random initialization as in PyTorch [22]. Right: Training accuracy after training 75 epochs with SGD, batch size 128 and grid-searched learning rate. Mean and 95% confidence interval of 5 independent runs.

independent from the input x_i as depth goes to infinity. We call this effect "directional" gradient vanishing (see Section 3 for a more thorough explanation).

Interestingly this effect does not happen in batch-normalized networks, which yield – as we shall prove in Theorem 2 – a stable rank for *any* depth, thereby preserving a disentangled representation of the input and hence allowing the training of very deep networks. These results substantiate earlier empirical observations made by [7] for random BN-nets, and also validates the claim that BN helps with *deep information propagation* [27].

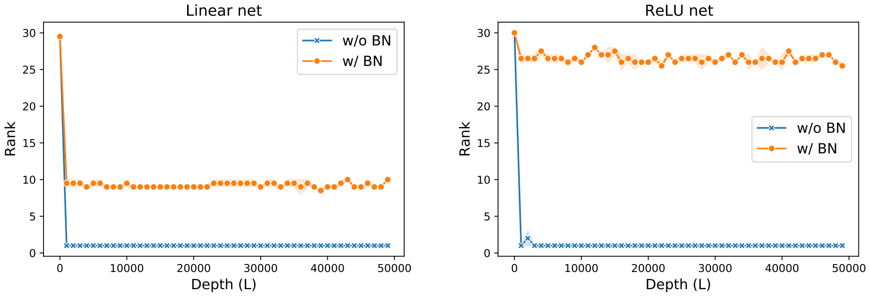


Figure 2: **Rank comparison of last hidden activation**: Log(rank) of the last hidden layer's activation over total number of layers (blue for BN- and orange for vanilla-networks) for Gaussian inputs. Networks are MLPs of width $d = 32$. (Left) Linear activations, (Right) ReLU activations. Mean and 95% confidence interval of 10 independent runs. While the rank quickly drops in depth for both networks, BN stabilizes the rank above $\sqrt{d}$.

1.2 Contributions

In summary, the work at hand makes the following two key contributions:

(i) We theoretically prove that BN indeed avoids rank collapse for deep linear neural nets under standard initialization and for any depth. In particular, we show that BN can be seen as a computationally cheap rank preservation operator, which may not yield hidden matrices with full rank but still preserves sufficient modes of variation in the data to achieve a scaling of the rank with $\Omega(\sqrt{d})$, where d is the width of the network. Subsequently, we leverage existing results from random matrix theory [9] to complete the picture with a simple proof of the above observed rank collapse for linear vanilla networks, which interestingly holds regardless of the presence of residual connections (Lemma 3). Finally, we connect the rank to difficulties in gradient based training of deep nets by showing that rank collapse makes the directional component of the gradients independent of the input.

(ii) We empirically show that the rank is indeed a crucial quantity for gradient-based learning. In particular, we show that both the rank and the final training accuracy quickly diminish in depth unless

BN layers are incorporated in both simple feed-forward and convolutional neural nets. To take this reasoning beyond mere correlations, we actively intervene with the rank of networks before training and show that (a) one can break the training stability of BN by initializing in a way that reduces its rank-preserving properties, and (b) a rank-increasing pre-training procedure for vanilla networks can recover their training ability even for large depth. Interestingly, our pre-training method allows vanilla SGD to outperform BN on very deep MLPs. In all of our experiments, we find that SGD updates preserve the order of the initial rank throughout optimization, which underscores the importance of the rank at initialization for the entire convergence behavior.

2 Background and Preliminaries

Network description. We consider a given input $X \in \mathbb{R}^{d\times N}$ containing N samples in $\mathbb{R}^d$. Let $\mathbf{1}_k \in \mathbb{R}^k$ denote the k-dimensional all one vector and $H_\ell^{(\gamma)}$ denote the hidden representation of X in layer ℓ of a BN-network with residual connections. The following recurrence summarizes the network mapping

$$H_{\ell+1}^{(\gamma)} = \text{BN}_{0,\mathbf{1}_d}(H_\ell^{(\gamma)} + \gamma W_\ell H_\ell^{(\gamma)}), \quad H_0^{(\gamma)} = X, \tag{1}$$

where $W_\ell \in \mathbb{R}^{d\times d}$ and γ regulates the skip connection strength (in the limit, $\gamma = \infty$ recovers a network without skip connection)[4]. Throughout this work, we consider the network weights W_ℓ to be initialized as follows.

Definition 1 (Standard weight initialization)**.** *The elements of weight matrices W_ℓ are i.i.d. samples from a distribution $\mathcal{P}$ that has zero-mean, unit-variance, and its density is symmetric around zero*[5]*. We use the notation μ for the probability distribution of the weight matrices.*

We define the BN operator $\text{BN}_{\alpha,\beta}$ as in the original paper [17], namely

$$\text{BN}_{\alpha,\beta}(H) = \beta \circ \left(\text{diag}\left(M(H)\right)\right)^{-1/2} H + \alpha \mathbf{1}_N^\top, M(H) := \frac{1}{N} H H^\top, \tag{2}$$

where $\circ$ is a row-wise product. Both $\alpha \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$ are trainable parameters. Throughout this work we assume the initialization $\alpha = 0$ and $\beta = \mathbf{1}_d$, and also omit corrections of the mean activity. As demonstrated empirically in Fig. 5, and theoretically in App. C this simplification does not change the performance of BN in our settings.

Rank notions. To circumvent numerical issues involved in rank computations we introduce a soft notion of the rank denoted by $\text{rank}_\tau(H)$ (soft rank). Specifically, let $\sigma_1, \dots, \sigma_d$ be the singular values of H. Then, given a $\tau > 0$, we define $\text{rank}_\tau(H)$ as

$$\text{rank}_\tau(H) = \sum_{i=1}^d \mathbf{1}(\sigma_i^2/N \geq \tau). \tag{3}$$

Intuitively, $\text{rank}_\tau(H)$ indicates the number of singular values whose absolute values are greater than $\sqrt{N\tau}$. It is clear that $\text{rank}_\tau(H)$ is less or equal to $\text{rank}(H)$ for all matrices H. For analysis purposes, we need an analytic measure of the collinearity of the columns and rows of H. Inspired by the so-called stable rank (see, e.g., [29]), we thus introduce the following quantity

$$r(H) = \text{Tr}(M(H))^2/\|M(H)\|_F^2, \quad M(H) = HH^\top/N. \tag{4}$$

In contrast to the algebraic rank, $r(H)$ is differentiable with respect to H. Furthermore, the next lemma proves that the above quantity lower-bounds the soft-rank for the hidden representations.

Lemma 1. *For an arbitrary matrix $H \in \mathbb{R}^{d\times d}$, rank$(H) \geq r(H)$. For the sequence $\{H_\ell^{(\gamma)}\}_{\ell=1}^\infty$ defined in Eq.* (2)*, rank$_\tau(H_\ell^{(\gamma)}) \geq (1-\tau)^2 r(H_\ell^{(\gamma)})$ holds for $\tau \in [0,1]$.*

---

[4] For the sake of simplicity, we here assume that the numbers of hidden units is equal across layers. In App. E we show how our results extend to nets with varying numbers of hidden units.

[5] Two popular choices for $\mathcal{P}$ are the Gaussian distribution $\mathcal{N}(0,1)$ and the uniform distribution $\mathcal{U}([-1,1])$. The variance can be scaled with the choice of γ to match the prominent initializations from [14] and [13]. Note that the symmetry implies that the law of each element $[W_\ell]_{ij}$ equates the law of $-[W_\ell]_{ij}$.

3 Batch normalization provably prevents rank collapse

Since our empirical observations hold equally for both non-linear and linear networks, we here focus on improving the theoretical understanding in the linear case, which constitutes a growing area of research [26, 19, 6, 2]. First, inspired by [10] and leveraging tools from Markov Chain theory, our main result proves that the rank of linear batch-normalized networks scales with their width as $\Omega(\sqrt{\text{width}})$. Secondly, we leverage results from random matrix theory [8] to contrast our main result to unnormalized linear networks which we show to provably collapse to rank one, even in the presence of residual connections.

3.1 Main result

In the following we state our main result which proves that batch normalization indeed prevents the rank of all hidden layer activations from collapsing to one. Please see Appendix E for the more formal version of this theorem statement.

Theorem 2. *[Informal] Suppose that the rank*$(X) = d$ *and that the weights* W_ℓ *are initialized in a standard i.i.d. zero-mean fashion (see Def. 1). Then, the following limits exist such that*

$$\lim_{L\to\infty} \frac{1}{L}\sum_{\ell=1}^{L} rank_\tau(H_\ell^{(\gamma)}) \geq \lim_{L\to\infty} \frac{(1-\tau)^2}{L}\sum_{\ell=1}^{L} r(H_\ell^{(\gamma)}) = \Omega((1-\tau)^2\sqrt{d}) \tag{5}$$

holds almost surely for a sufficiently small γ *(independent of* ℓ*) and any* $\tau \in [0, 1)$*, under some additional technical assumptions. Please see Theorem 14 in the Appendix for the formal statement.*

Theorem 2 yields a non trivial width-dependency. Namely, by setting for example $\tau := 1/2$, the result states that the average number of singular values with absolute value greater than $\sqrt{N/2}$ is at least $\Omega(\sqrt{d})$ on average. To put this into context: If one were to replace $\text{diag}(M)^{-1/2}$ by the *full* inverse $(M)^{-1/2}$ in Eq. (2), then BN would effectively constitute a classical whitening operation such that all $\{H_\ell^{(\gamma)}\}_{\ell=1}^{L}$ would be full rank (equal to d). However, as noted in the original BN paper [17], whitening is obviously expensive to compute and furthermore prohibitively costly to incorporate in back-propagation. As such, BN can be seen as a computationally inexpensive approximation of whitening, which does not yield full rank hidden matrices but still preserves sufficient variation in the data to provide a rank scaling as $\Omega(\sqrt{d})$. Although the lower-bound in Thm. 2 is established on the average over infinite depth (i.e., $L \to \infty$), Corollary 15 (in App. E) proves that the same bound holds for all $\text{rank}(H_\ell)$ and $\text{rank}_\tau(H_\ell)$.

Necessary assumptions. The above result relies on two key assumptions: (i) First, the input X needs to be full rank. (ii) Second, the weights have to be drawn according to the standard initialization scheme. We believe that both assumptions are indeed necessary for BN to yield a robust rank.

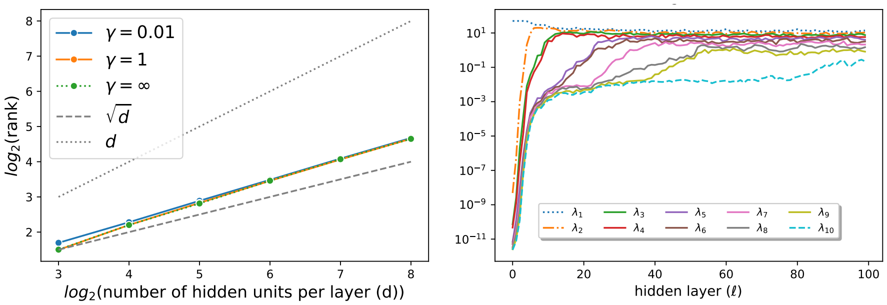
Regarding (i), we consider a high input rank a natural condition since linear neural nets cannot possibly *increase* the rank when propagating information through their layers. Of course, full rank is easily achieved by an appropriate data pre-processing. Yet, even when the matrix is close to low rank we find that BN is actually able to amplify small variations in the data (see Fig. 3.b).[6] Notably, we observed that hidden representations remain full rank if $H_0^{(\gamma)}$ is full-rank and $N = O(\sqrt{d})$. Regarding (ii), we derive – based on our theoretical insights – an adversarial initialization strategy that corrupts both the rank robustness and optimization performance of batch-normalized networks, thus suggesting that the success of BN indeed relies heavily on the standard i.i.d. zero-mean initialization.

Experimental validation. In order to underline the validity of Theorem 2 we run multiple simulations by feeding Gaussian data of dimensionality $d = N$ into networks of growing size and with different residual strengths. For each network, we compute the mean and standard deviation of the soft rank rank_τ with $\tau = 0.5$. As depicted in Fig. 3, the curves clearly indicate a $\Omega(\sqrt{d})$ dependency for $\lim_{L\to\infty}\sum_{\ell=1}^{L}\text{rank}_\tau(H_\ell)/L$, just as predicted in the Theorem. Although the established guarantee requires the weight on the parametric branch (i.e., γ) to be small, the results of Fig. 3 indicate that

---

[6] Intuitively this means that even if two data points are very close to each other in the input space, their hidden presentation can still be disentangled in batch-normalized networks (see Appendix E for more details)

the established lower bound holds for a much wider range including the case where no residual connections are used at all ($\gamma = \infty$).



a) b)

Figure 3: a) Result of Theorem 2 for different values of γ, where $\gamma = \infty$ depicts networks *without* skip connections. Each point is the average rank$_{1/2}$ over depth ($L = 10^6$) of nets of width $d \in \{8, 16, .., 256\}$ an on x-axis. b) Top 10 singular values of $H_\ell^{(\gamma)}$ for increasing values of ℓ given nearly collinear inputs. As can be seen, BN quickly amplifies smaller variations in the data while reducing the largest one.

3.2 Comparison with unnormalized networks

In order to stress the importance of the above result, we now compare the predicted rank of H_ℓ with the rank of unnormalized linear networks, which essentially constitute a linear mapping in the form of a product of random matrices. The spectral distribution of products of random matrices with i.i.d. standard Gaussian elements has been studied extensively [7, 12, 21]. Interestingly, one can show that the gap between the top and the second largest singular value increases with the number of products (i.e., ℓ) at an exponential rate[7] [12, 21]. Hence, the matrix converges to a rank one matrix after normalizing by the norm. In the following, we extend this result to products of random matrices with a residual branch that is obtained by adding the identity matrices. Particularly, we consider the hidden states $\widehat{H}_\ell$ of the following linear residual network:

$$\widehat{H}_\ell = \mathbf{B}_\ell X, \quad \mathbf{B}_\ell := \prod_{k=1}^{\ell} (I + \gamma W_k). \tag{6}$$

Since the norm of $\widehat{H}_\ell$ is not necessarily bounded, we normalize as $\widetilde{H}_\ell = B_\ell X / \|B_\ell\|$. The next lemma characterizes the limit behaviour of $\{\widetilde{H}_\ell\}$.

Lemma 3. *Suppose that $\gamma \in (0, 1)$ and assume the weights W_ℓ to be initialized as in Def. 1 with element-wise distribution $\mathcal{P}$. Then we have for linear networks, which follow recursion* (6)*, that:*

a. If $\mathcal{P}$ is standard Gaussian, then the sequence $\{\widetilde{H}_\ell\}$ converges to a rank one matrix.

b. If $\mathcal{P}$ is uniform$[-\sqrt{3}, \sqrt{3}]$, then there exists a monotonically increasing sequence of integers $\ell_1 < \ell_2, \ldots$ such that the sequence $\{\widetilde{H}_{\ell_k}\}$ converges to a rank one matrix.

This results stands in striking contrast to the result of Theorem 2 established for batch-normalized networks.[8] Interestingly, even residual skip connections cannot avoid rank collapse for very deep neural networks, unless one is willing to incorporate a depth dependent down-scaling of the parametric branch as for example done in [1], who set $\gamma = O(\frac{1}{L})$. Remarkably, Theorem 2 shows that BN layers provably avoid rank collapse *without* requiring the networks to become closer and closer to identity. Remarkably, the remaining direction after rank collapse depends exclusively on the random weights and it is independent of the input.

---

[7] The growth-rate of the i-th singular value is determined by the i-th Lyapunov exponent of the product of random matrices. We refer the reader to [12] for more details on Lyapunov exponents.

[8] According to the observations in Fig. 2, the result of part b holds for the usual sequence of indices $\{\ell_k = k\}$, which indicates that $\{\widetilde{H}_k\}$ converges to a rank one matrix even in the case of uniform initialization.

Implications of rank collapse on gradient based learning. In order to explain the severe consequence of rank collapse on optimization performance reported in Fig. 1, we study the effect of rank one hidden-layer representations on the gradient of the training loss for distinct input samples. Let $\mathcal{L}_i$ denote the training loss for datapoint i on a vanilla network as in Eq. (6). Furthermore, let the final classification layer be parametrized by $W_{L+1} \in \mathbb{R}^{d_{out} \times d}$. Then, given that the hidden presentation at the last hidden layer L is rank one, the normalized gradients of the loss with respect to weights of individual neurons $k \in 1, ..., d_{out}$ in the classification layer (denoted by $\nabla_{W_{L+1,k}}\mathcal{L}_i$, where $\|\nabla_{W_{L+1,k}}\mathcal{L}_i\| = 1$) are collinear for any two datapoints i and j, i.e. $\nabla_{W_{L+1,k}}\mathcal{L}_i = \mp\nabla_{W_{L+1,k}}\mathcal{L}_j$. A formal statement is presented in Prop. 19 in the Appendix alongside empirical validations on a VGG19 network (Fig. 10). This result implies that the commonly accepted vanishing gradient *norm* hypothesis is not descriptive enough since SGD does not take small steps into the *right* direction, but into a random direction that is *independent* from the input. In other words, deep neural networks are prone to *directional gradient vanishing* after initialization, which is caused by the collapse of the last hidden layer activations to a very small subspace (one line in $\mathbb{R}^d$ in the extreme case of rank one activations).

4 The important role of the rank

The preceding sections highlight that the rank of the hidden representations is a key difference between random vanilla and BN networks. We now provide three experimental findings that substantiate the particular importance of the rank at the beginning of training: First, we find that an unsupervised, rank-increasing pre-training allows SGD on vanilla networks to outperform BN networks. Second, we show that the performance of BN-networks is closely tied to a high rank at initialization. Third, we report that SGD updates preserve the initial rank magnitude throughout the optimization process.

Outperforming BN using a pre-training step. As discussed above, batch normalization layers are very effective at avoiding rank collapse. Yet, this is of course not the only way to preserve rank. Based upon our theoretical insights, we leverage the lower bound established in Eq. (4) to design a pre-training step that not only avoids rank collapse but also accelerates the convergence of SGD. Our proposed procedure is both simple and computationally cheap. Specifically, we *maximize* the lower-bound $r(H_\ell)$ (in Eq. (4)) on the rank of the hidden presentation H_ℓ in each layer ℓ. Since this function is differentiable with respect to its input, it can be optimized sufficiently by just a few steps of (stochastic) gradient ascent (see Section G in the Appendix for more details).

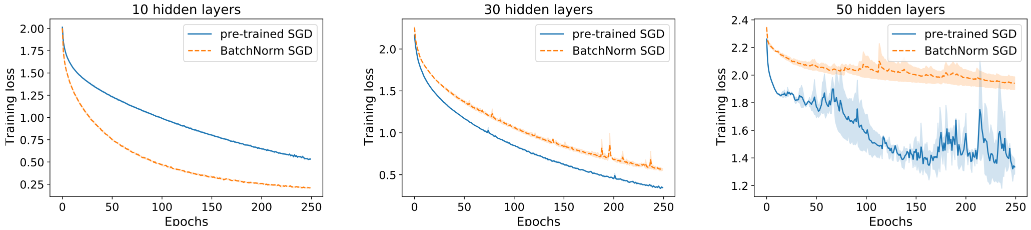


Figure 4: **Pre-training versus BN:** Loss over epochs on CIFAR-10 for MLPs of increasing depth with 128 hidden units and ReLU activation. Trained with SGD (batchsize 64) and grid-searched stepsize. See Fig. 11 for the corresponding test loss and accuracy as well as Fig. 12 for FashionMNIST results.

Fig. 4 compares the convergence rate of SGD on pre-trained vanilla networks and BN-networks. As can be seen, the slow down in depth is much less severe for the pre-trained networks. This improvement is, also, reflected both in terms of training accuracy and test loss (see Fig. 11 in Appendix). Interestingly, the pre-training is not only faster than BN on deep networks, but it is also straight-forward to use in settings where the application of BN is rather cumbersome such as for very small batch sizes or on unseen data [16, 30].

Breaking batch normalization. Some scholars hypothesize that the effectiveness of BN stems from a global landscape smoothing [25] or a certain learning rate tuning [3], that are thought to be induced by the normalization. Under these hypotheses, one would expect that SGD converges fast on BN-nets *regardless* of the initialization. Yet, we here show that the way that networks are initialized does play a crucial role for the subsequent optimization performance of BN-nets.

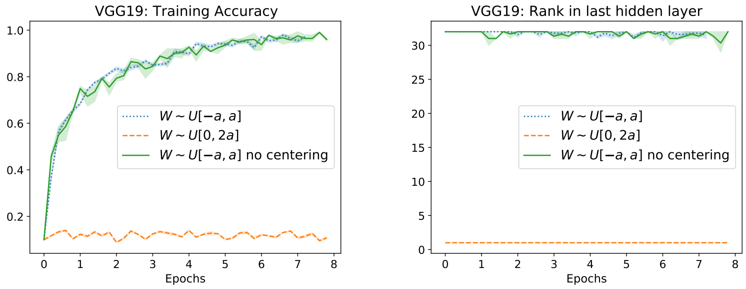


Figure 5: **Breaking Batchnorm:** CIFAR-10 on VGG19 with standard PyTorch initialization as well as a uniform initialization of same variance. (Left) training accuracy, (Right) Rank of last hidden layer computed using $torch.matrix\_rank()$. Plot also shows results for standard initialization and BN *without* mean deduction. Avg. and 95% CI of 5 independent runs. (See Fig. 13 in Appendix for similar results on ResNet-50).

Particularly, we train two MLPs with batchnorm, but change the initialization for the second net from the standard PyTorch way $W_{l,i,j} \sim \text{uniform}\left[-1/\sqrt{d_l}, 1/\sqrt{d_l}\right]$ [22, 13] to $W_{l,i,j} \sim \text{uniform}\left[0, +2/\sqrt{d_l}\right]$, where d_l is the layer size. As can be seen to the right, this small change reduces the rank preserving quality of BN significantly, which is reflected in much slower learning behaviour. Even sophisticated modern day architectures such as VGG and ResNet networks are unable to fit the CIFAR-10 dataset after changing the initialization in this way (see Fig. 5).

Rank through the optimization process. The theoretical result of Theorem 2 considers the rank at random initialization. To conclude, we perform two further experiments which confirm that the initial rank strongly influences the speed of SGD throughout the entire optimization process. In this regard, Fig. 6 reports that SGD preserves the initial magnitude of the rank to a large extent, regardless of the specific network type. This is particularly obvious when comparing the two BN initializations. A further noteworthy aspect is the clear correlation between the level of pre-training and optimization performance on vanilla nets. Interestingly, this result does again not only hold on simple MLPs but also generalizes to modern day networks such as the VGG-19 (see Fig. 5) and ResNet50 architecture (see Appendix I).

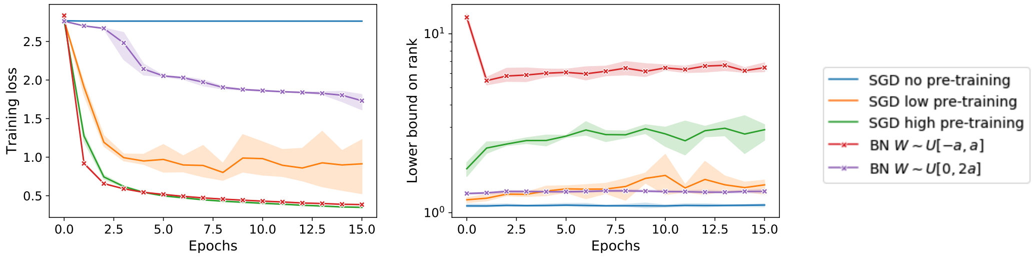


Figure 6: **Pretraining:** Fashion-MNIST on MLPs of depth 32 and width 128. (Left) Training accuracy, (Right) Lower bound on rank. Blue line is a ReLU network with standard initialization. Other solid lines are pre-trained layer-wise with 25 (orange) and 75 (green) iterations to increase the rank. Dashed lines are batchnorm networks with standard and asymmetric initialization. Average and 95% confidence interval of 5 independent runs.

5 Discussions

In summary, our work highlights a key difference between random vanilla- and BN networks. While the rank of the hidden representations quickly collapses to one as the depth of vanilla networks increases, BN is robust against such rank collapse. This intriguing property arises due to the standard initialization of weights and also it is preserved through the optimization process. Notably, our theoretical analysis proves this striking difference for linear MLPs and holds empirically across a wide range of data sets and network architectures. Our experiments further highlight the determining role of the rank quantity in the training speed. Inspired by these observations, we develop a novel pre-training method that allows previously un-trainable very deep vanilla networks to learn, sometimes even faster than BN-MLPs of the same size. Thereby our study reveals a key requirement for a proper initialization of deep neural networks, opening doors to the development of effective initialization schemes for modern-day architectures.

We thus consider our work a relevant step towards a better understanding of optimization for deep neural networks. Furthermore, our findings give rise to several interesting follow-up questions: (i) Can one generalize the analysis of Theorem 2 to ReLU and other non-linear nets to prove the observed rank robustness (e.g. Fig. 2)? (ii) is it possible to rigorously prove that SGD updates preserve the rank magnitude throughout optimization, as observed in Fig. 6)? (iii) Is it possible to use the develop a similarly effective pre-training for convolution and recurrent networks? (iv) How can one theoretically characterize the connection between the convergence of SGD and the rank quantity (a follow-up on directional gradient vanishing)? (v) Does rank robustness explain the success of related architectures such as layer normalization [4], weight normalization [24]) and modern initialization techniques such as fix up initialization [35]? We believe that these questions will spark an interesting line of future research towards the goal of fully understanding optimization in deep neural networks.

Broader impact

As we only contribute to a better understanding of neural network training in general, we consider our work fundamental research without any specific application. Hence a broader impact discussion is not applicable.

Funding

This project is fully supported by ETH-Zurich fellowships.

Acknowledgement

We thank Dr. Aran Raoufi and Dr. Olivier Ledoit as well as Gary Becigneul for their helpful discussions.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[2] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

[3] Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. *arXiv preprint arXiv:1812.03981*, 2018.

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

[5] Boaz Barak and David Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. *Course notes: http://www. sumofsquares. org/public/index. html*, 2016.

[6] Peter L. Bartlett, David P. Helmbold, and Philip M. Long. Gradient descent with identity initialization efficiently learns positive-definite linear transformations by deep residual networks. *Neural computation*, 31(3):477–502, 2019.

[7] Nils Bjorck, Carla P. Gomes, Bart Selman, and Kilian Q. Weinberger. Understanding batch normalization, 2018.

[8] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[9] Philippe Bougerol. *Products of Random Matrices with Applications to Schrödinger Operators*, volume 8. Springer Science & Business Media, 2012.

[10] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *arXiv preprint arXiv:1707.06386*, 2017.

[11] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer, 2018.

[12] Peter J. Forrester. Lyapunov exponents for products of complex Gaussian random matrices. *Journal of Statistical Physics*, 151(5):796–808, 2013.

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

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

[15] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[16] Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Advances in neural information processing systems*, pages 1945–1953, 2017.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[18] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. The normalization method for alleviating pathological sharpness in wide neural networks. In *Advances in Neural Information Processing Systems*, pages 6403–6413, 2019.

[19] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.

[20] Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Ming Zhou, Klaus Neymeyr, and Thomas Hofmann. Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization. *arXiv preprint arXiv:1805.10694*, 2018.

[21] Dang-Zheng Liu, Dong Wang, and Lun Zhang. Bulk and soft-edge universality for singular values of products of ginibre random matrices. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 1734–1762. Institut Henri Poincaré, 2016.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[23] Jeffrey Pennington, Samuel S Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. *arXiv preprint arXiv:1802.09979*, 2018.

[24] Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 901–909, 2016.

[25] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization?(no, it is not about internal covariate shift). *arXiv preprint arXiv:1805.11604*, 2018.

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

[27] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

[28] Matus Telgarsky. Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485*, 2016.

[29] Joel A. Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[30] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[31] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[32] Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In *Advances in neural information processing systems*, pages 7103–7114, 2017.

[33] Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. A mean field theory of batch normalization. *arXiv preprint arXiv:1902.08129*, 2019.

[34] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. PyHessian: Neural networks through the lens of the Hessian. *arXiv preprint arXiv:1912.07145*, 2019.

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

Appendix

A Preliminaries

Recall that $H_\ell^{(\gamma)}$ denotes the hidden representations in layer ℓ. These matrices make a Markov chain that obeys the recurrence of Eq. (1), which we restate here

$$H_{\ell+1}^{(\gamma)} = \text{BN}(H_\ell^{(\gamma)} + \gamma W_\ell H_\ell^{(\gamma)}), \quad H_0^\gamma = X, \tag{7}$$

where we use the compact notation BN for $\text{BN}_{0,\mathbf{1}_d}$. Let $M_\ell^{(\gamma)}$ be second moment matrix of the hidden representations $H_\ell^{(\gamma)}$, i.e. $M_\ell^{(\gamma)} := H_\ell^{(\gamma)} \left(H_\ell^{(\gamma)}\right)^\top /N$. Batch normalization ensures that the rows of H_ℓ have the same norm $\sqrt{N}$ for $\ell > 0$ –where N is the size of mini-batch. Let $\mathcal{H}$ be space of $d \times d$-matrices that obey this propery. This property enforces two key characteristics on $M_\ell^{(\gamma)}$:

(p.1) its diagonal elements are one (8)

(p.2) the absolute value of its off-diagonal elements is less than one (9)

Property (p.1) directly yields that the trace of $M_\ell^{(\gamma)}$ (and hence the sum of its eigenvalues) is equal to d. We will repeatedly use these properties in our analysis.

Furthermore, the sequence $\{H_\ell^{(\gamma)}\}_{\ell=1}^\infty$ constitute a Markov chain. Under mild assumptions, this chain admits an invariant distribution that is defined bellow[11].

Definition 2. *Distribution ν is an invariant distribution of the hidden representations $\{H_\ell^{(\gamma)}\}_{\ell=1}^\infty$ if it obeys*

$$\int \text{BN}(H + \gamma W H)\mu(dW)\nu(dH) = \int \text{BN}(H)\nu(dH) \tag{10}$$

where μ denotes the probability measure of random weights.

Later, we will see that the above invariance property allows us to determine the commutative behaviour of the sequence of hidden presentations.

B Lower bounds on the (soft) rank

Recall that we introduced the ratio $r(H) = \text{Tr}(M(H))^2/\|M(H)\|_F^2$ in Eq. (4) as a lower bound on both the rank(H) as well as the soft rank rank$_\tau(H)$ (stated in Lemma 1). This section establishes these lower bounds.

Proof of Lemma 1 (part 1). We first prove that rank$(H) \geq r(H)$. Let $M := M(H) = HH^\top/N$. Since the eigenvalues of H are obtained by a constant scaling factor of the squared singular values of H, these two matrices have the same rank. We now establish a lower bound on rank(M). Let $\lambda \in \mathbb{R}^d$ contain the eigenvalues of matrix M, hence $\|\lambda\|_1 = \text{Tr}(M)$ and $\|\lambda\|_2^2 = \|M\|_F^2$. Given λ, we define the vector $w \in \mathbb{R}^d$ as

$$w_i = \begin{cases} 1/\|\lambda\|_0 & : \lambda_i \neq 0 \\ 0 & : \lambda_i = 0. \end{cases} \tag{11}$$

To proof the assertion, we make use if a straightforward application of Cauchy-Schwartz

$$|\langle \lambda, w\rangle| \leq \|\lambda\|_2 \|w\|_2 \tag{12}$$

$$\implies \|\lambda\|_1/\|\lambda\|_0 \leq \|\lambda\|_2/\|\lambda\|_0^{1/2} \tag{13}$$

$$\implies \frac{\|\lambda\|_1}{\|\lambda\|_2} \leq \|\lambda\|_0^{1/2}. \tag{14}$$

Replacing $\|\lambda\|_2 = \|M\|_F$ and $\|\lambda\|_1 = \text{Tr}(M)$ into the above equation concludes the result. Note that the above proof technique has been used in the planted sparse vector problem [5]. □

Proof of Lemma 1 (part 2). Now, we prove that $\text{rank}_\tau(H_\ell^{(\gamma)}) \geq (1-\tau)^2 r(H_\ell^{(\gamma)})$. Let $\lambda \in \mathbb{R}_+^d$ be a vector containing the eigenvalues of the matrix $M_\ell^{(\gamma)} = M(H_\ell^{(\gamma)})$. Let $\sigma \in \mathbb{R}_+^d$ contain the singular values of H. Then, one can readily check that $\sigma_i^2/N = \lambda_i$. Furthermore, $\|\lambda\|_1 = d$ due to (p.1) in Eq. (8). Furthermore, we have by definition that

$$\text{rank}_\tau(H_\ell^{(\gamma)}) = h_\tau(\lambda) := \sum_{i=1}^d \mathbf{1}(\sigma_i^2/N \geq \tau) = \sum_{i=1}^d \mathbf{1}(\lambda_i \geq \tau). \tag{15}$$

Let us now define a vector $w \in \mathbb{R}^d$ with entries

$$w_i = \begin{cases} 1/h_\tau(\lambda) & : \lambda_i \geq \tau \\ 0 & : \text{otherwise.} \end{cases} \tag{16}$$

Then, we use Cauchy-Schwartz to get

$$|\langle \lambda, w\rangle| \leq \|\lambda\|_2 \|w\|_2. \tag{17}$$

It is easy to check that $\|w\|_2 = h_\tau(\lambda)^{-1/2}$ holds. Furthermore,

$$\begin{aligned} h_\tau(\lambda)|\langle w, \lambda\rangle| &= \sum_{|\lambda_i|\geq\tau}^d |\lambda_i| && (18)\\ &\geq \|\lambda\|_1 - d\tau && (19)\\ &\geq (1-\tau)\|\lambda\|_1, && (20)\end{aligned}$$

where we used the fact that $\|\lambda\|_1 = d$ in the last inequality. Replacing this result into the bound of Eq. (17) yields

$$\text{rank}_\tau(H_\ell^{(\gamma)}) = h_\tau(\lambda) \geq (1-\tau)^2 \|\lambda\|_1^2/\|\lambda\|_2^2 = (1-\tau)^2 r(H_\ell^{(\gamma)}), \tag{21}$$

which conludes the proof. ☐

C Initialization consequences

The particular weight initialization scheme consider through out this work (recall Def. 1), imposes an interesting structure in the invariant distribution of the sequence of hidden presentations (defined in Def. 2).

Lemma 4. *Suppose that the chain $\{H_\ell^{(\gamma)}\}_{\ell=1}^\infty$ (defined in Eq. 7) admits a unique invariant distribution ν_γ and H is drawn from ν_γ; then, the law of $H_{i:}$ equates the law of $-H_{i:}$ where $H_{i:}$ denotes the ith row of matrix H.*

Proof. Let S be a sign filliping matrix: it is diagonal and its diagonal elements are in $\{+1, -1\}$. Then $SW \overset{d}{=} W$ holds for a random matrix W whose elements are drawn i.i.d. from a symmetric distribution. Let H be drawn from the invariant distribution of the chain denoted by ν_γ; Leveraging the invariance property, we get

$$H \overset{d}{=} H_+ \overset{d}{=} \left(\text{diag}(H_{1/2}H_{1/2}^\top/N)\right)^{-1/2} H_{1/2}, \quad H_{1/2} := H + \gamma SWSH \tag{22}$$

By multiplying both sides with S, we get

$$SH \overset{d}{=} SH_+ \overset{d}{=} \left(\text{diag}\left(H_{1/2}H_{1/2}^\top/N\right)\right)^{-1/2} \widetilde{H}_{1/2}, \quad \widetilde{H}_{1/2} := SH + \gamma WSH \tag{23}$$

Note that we use the fact that diagonal matrices commute in the above derivation. According to the definition, $S^2 = I$ holds. Considering this fact, we get

$$\begin{aligned} \text{diag}\left(H_{1/2}H_{1/2}^\top\right) &= \text{diag}\left((H + \gamma SWSH)(H + \gamma SWSH)^\top\right) && (24)\\ &= \text{diag}\left((SSH + \gamma SWSH)(SSH + \gamma SWSH)^\top\right) && (25)\\ &= \text{diag}\left(S(SH + \gamma WSH)(SH + \gamma WSH)^\top S\right) && (26)\\ &= \text{diag}\left((SH + \gamma WSH)(SH + \gamma WSH)^\top\right) && (27)\\ &= \widetilde{H}_{1/2}\widetilde{H}_{1/2}^\top && (28)\end{aligned}$$

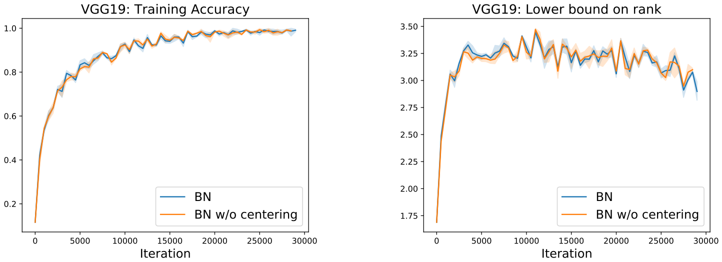
Replacing the above result into Eq. (29) yields

$$SH \overset{d}{=} SH_+ \overset{d}{=} \text{diag}^{-1/2}\left(\widetilde{H}_{1/2}\widetilde{H}_{1/2}^\top/N\right)\widetilde{H}_{1/2}, \quad \widetilde{H}_{1/2} := SH + \gamma WSH. \tag{29}$$

Hence the law of SH is invariant too. Since the invariant distribution is assumed to be unique, $SH \overset{d}{=} H$ holds and thus $H_{i:} \overset{d}{=} -H_{i:}$.

$\square$

Comment on BN-centering Let ν_γ be the unique invariant distribution associated with Markov chain $\{H_\ell^{(\gamma)}\}$. A straightforward implication of last Lemma is $\mathbb{E}[H_i] = 0$ for $H \sim \nu_\gamma$, hence the rows of $H_\ell^{(\gamma)}$ are mean-zero, hence their average is close to zero [9] and the mean-zero operation in BN is redundant. Although this theoretical argument is established for linear networks, we empirically observed that BN without centering also works well on modern neural architectures. For example, Fig. 7 shows that the centering does not affect the performance of BN on a VGG net when training CIFAR-10.



Training Accuracy $\qquad$ $r(H_L)$

Figure 7: Centering for BN. The experiment is conducted on a VGG network. The blue line indicates the original BN network and the orange line is BN without mean adaption. The vertical axis in the left plot is training accuracy. In the right plot it is $r(H_L)$, where H_L is the data representation in the last hidden layer L. The horizontal axis indicates the number of iterations.

D Main Theorem: warm-up analysis

As a warm-up analysis, the next lemma proves that rank$(H_\ell^{(\gamma)}) \geq 2$ holds. Later, we will prove a stronger result. Yet, this initial results provides valuable insights into our proof technique. Furthermore, we will use the following result in the next steps.

Lemma 5. *Suppose that each element of the weight matrices is independently drawn from distribution $\mathcal{P}$ that is zero-mean, unit-variance, and its support lies in interval $[-B, B]$. If the Markov chain $\{H_\ell\}_{\ell\geq 1}$ admits a unique invariant distribution, then*

$$\textit{rank}(H_\ell^{(\gamma)}) \geq 2 \tag{30}$$

holds almost surely for all integers ℓ and $\gamma \leq 1/(8d)$.

Proof. Let the weights $\{W_\ell\}$ be drawn from the distribution μ, defined in Def. 1. Such a distribution obeys an important property: element-wise symmetricity. That is, $[W_\ell]_{ij}$ is distributed as $-[W_\ell]_{ij}$. Such an initialization enforces an interesting structural property for the invariant distribution ν_γ that is stated in Lemma 4. It is easy to check that this implies

$$\mathbb{E}\left[[M(H_\ell^{(\gamma)})]_{ij}\right] = -\mathbb{E}\left[[M(H_\ell^{(\gamma)})]_{ij}\right] = 0, \tag{31}$$

for any $i \neq j$. Recall, $M(H) = HH^\top/N$. The above property enforces $[M(H)]_{ij}^2$ to be small and hence $\|M_\ell^{(\gamma)}\|_F^2$ is small as well. Now, as rank$(H_\ell^{(\gamma)})$ is proportional to $1/\|M_\ell^{(\gamma)}\|_F^2$ (compare

---

[9] When d is sufficiently large and assuming that coordinates in one row are weakly dependent, the central limit theorem implies that the empirical average of the rows converges to zero.

Eq. (4)), it must consequently stay large. The rest of the proof is based on this intuition. Given the uniqueness of the invariant distribution, we can invoke Birkhoff's Ergodic Theorem for Markov Chains (Theorem 5.2.1 and 5.2.6 [11]) which yields

$$\lim_{L\to\infty} \frac{1}{L}\sum_{\ell=1}^{L} [M_\ell^{(\gamma)}]_{ij} = \mathbb{E}_{H\sim\nu_\gamma}\left[[M(H)]_{ij}\right]. \tag{32}$$

This allows us to conclude the proof by a simple contradiction. Assume that rank$(H_k^{(\gamma)})$ is indeed one. Then, as established in the following Lemma, in the limit all entries of $M(H_\ell^{(\gamma)})$ are constant and either -1 or 1.

Lemma 6. *Suppose the assumptions of Lemma 5 hold. If rank*$(H_k^{(\gamma)}) = 1$ *for an integer* k*, then* $M(H_\ell^{(\gamma)}) = M(H_k^{(\gamma)})$ *holds for all* $\ell > k$*. Furthermore, all elements of all matrices* $\{M(H_\ell^{(\gamma)})\}_{\ell\geq k}$ *have absolute value one, hence*

$$\lim_{L\to\infty} \frac{1}{L}\sum_{\ell=1}^{L} [M(H_\ell^{(\gamma)})]_{ij} \in \{1, -1\} \tag{33}$$

holds.

As a result, leveraging the ergodicity established in (61), we get that then

$$\mathbb{E}_{H\sim\nu_\gamma}\left[[M(H)]_{ij}\right] \in \{+1, -1\} \tag{34}$$

must also hold. However, this contradicts the consequence of the symmetricity (Eq. (31)) which states that for any $j \neq i$ we have $\mathbb{E}_{H\sim\nu_\gamma}\left[[M(H)]_{ij}\right] = -\mathbb{E}_{H\sim\nu_\gamma}\left[[M(H)]_{ij}\right] = 0$. Thus, the rank one assumption cannot hold, which proves the assertion. ☐

To complete the proof of the last theorem, we prove Lemma 6.

Proof of Lemma 6. For the sake of simplicity, we omit all superscripts (γ) throughout the proof. Suppose that rank$(H_k) = 1$, then rank$(H_\ell) = 1$ for all $\ell \geq k$ as the sequence $\{\text{rank}(H_\ell)\}$ is non-increasing [10]. Invoking the established rank bound from Lemma 1, we get

$$r(H_\ell) = \frac{\text{Tr}(M_\ell)^2}{\|M_\ell\|_F^2} \leq \text{rank}(H_\ell) = 1. \tag{35}$$

The above inequality together with properties (p.1) and (p.2) (presented in Eqs 8 and 9) yield Tr$(M_\ell) = d$. Replacing this into the above equation gives that $\|M_\ell\|_F^2 \geq d^2$ must hold for the rank of H_ℓ to be one. Yet, recalling property (p.2), this can only be the case if $[M_\ell]_{ij} \in \{+1, -1\}$ for all i, j. Replacing the definition $M(H) = HH^\top/N$ into updates of hidden presentation in Eq. 1 obtains

$$M_{\ell+1} = \text{diag}\left(M_{\ell+\frac{1}{2}}\right)^{-1/2}\left(M_{\ell+\frac{1}{2}}\right)\text{diag}\left(M_{\ell+\frac{1}{2}}\right)^{-1/2} \tag{36}$$

where

$$M_{\ell+\frac{1}{2}} = M_\ell + \Delta M_\ell, \quad \Delta M_\ell := \gamma W_\ell M_\ell + \gamma M_\ell W_\ell^\top + \gamma^2 W_\ell M_\ell W_\ell^\top \tag{37}$$

We now prove that the sign of $[M_\ell]_{ij}$ and $[M_{\ell+1}]_{ij}$ are the same for $[M_\ell]_{ij} \in \{+1, -1\}$. The above update formula implies that the sign of $[M_{\ell+1}]_{ij}$ equates that of $[M_{\ell+1/2}]_{ij}$. Furthermore, it is easy to check that $|[\Delta M_\ell]_{ij}| \leq 4\gamma B$. For $\gamma \leq 1/(8Bd)$, this bound yields $|[\Delta M_\ell]_{ij}| \leq \frac{1}{2}$. Therefore, the sign of $[M_{\ell+1/2}]_{ij}$ is equal to the one of $[M_\ell]_{ij}$. Since furthermore $[M_{\ell+1}]_{ij} \in \{1, -1\}$ holds, we conclude that all elements of M_ℓ remain constant for all $\ell \geq k$, which yields the limit stated in Eq. 33 . ☐

---

[10] Recall that the updates in Eq. (1) is obtained by matrix multiplications, hence it does not increase the rank.

E Main theorem: Proof

In this section, we prove that BN yields an $\Omega(\sqrt{d})$-rank for hidden representation.

Proof sketch for Thm. 2. The proof is based on an application of ergodic theory (as detailed for example in Section 5 of [11]). In fact, the chain of hidden representations, denoted by $H_\ell^{(\gamma)}$ (1), constitutes a Markov chain in a compact space. This chain admits at least one invariant distribution ν for which the following holds

$$\int g(\mathrm{BN}_{0,\mathbf{1}_d}(H + \gamma W H))\mu(dW)\nu(dH) = \int g(H)\nu(dH), \tag{38}$$

for every bounded Borel function $g : \mathbb{R}^{d\times d} \to \mathbb{R}^d$. The above invariance property provides an interesting characterization of the invariant measure ν. Particularly, we show in Lemma 13 that

$$\int r(H)\nu(dH) = \Omega(\sqrt{d}) \tag{39}$$

holds, where $r(H)$ is the established lower-bound on the rank (see Lemma 1). Under weak assumptions, the chain obey Birkhoff's Ergodicity, which yields that the average behaviour of the hidden representations is determined by the invariant measure ν:

$$\lim_{L\to\infty} \frac{1}{L}\sum_{i=\ell} r(H_\ell^{(\gamma)}) = \int r(H)\nu(dH) \overset{(39)}{=} \Omega(\sqrt{d}). \tag{40}$$

Finally, the established lower bound in Lemma 1 allows us to directly extend this result to a lower bound on the soft rank itself. □

Characterizing the change in Frobenius norm Recall the established lower bound on the rank denoted by $r(H)$, for which

$$r(H_\ell) = \frac{\mathrm{Tr}(M_\ell)^2}{\|M_\ell\|_F^2} = \frac{d^2}{\|M_\ell\|_F^2} \tag{41}$$

holds for all H_ℓ defined in Eq. 1.[11] Therefore, $\|M_\ell\|_F^2$ directly influences $\mathrm{rank}_\tau(H_\ell)$ (and also $\mathrm{rank}(H_\ell)$) according to Lemma 1. Here, we characterize the change in $\|M(H)\|_F^2$ after applying one step of the recurrence in Eq. 7 to H, i.e. passing it trough one hidden layer. This yields

$$H_+ = (\mathrm{diag}(M(H_\gamma(W)))^{-1/2} H_\gamma(W), \quad H_\gamma(W) = (I + \gamma W)H. \tag{42}$$

Let $M = M(H)$ and $M_+ = M(H_+)$ for simplicity. The next lemma estimates the expectation (taken over the randomness of W) of the difference between the Frobenius norms of M_+ and M.

Lemma 7. *If $W \sim \mu$ (defined in Def. 1), then*

$$\left(\mathbb{E}_W\|M_+\|_F^2 - \|M\|_F^2\right)/(\gamma^2) = \underbrace{2d^2 - 2\|M\|_F^2 - 8Tr(M^3) + 8Tr(diag(M^2)^2)}_{\delta_F(M)} + O(\gamma) \tag{43}$$

holds as long as the support of distribution $\mathcal{P}$ (in Def. 1) lies in a finite interval $[-B, B]$.

The proof of the above lemma is based on a Taylor expansion of the BN non-linear operator. We postpone the detailed proof to the end of this section. While the above equation seems complicated at first glance, it provides some interesting insights.

Interlude: Intuition behind Lemma 7. In order to gain more understanding of the implications of the result derived in Lemma 7, we make the simplifying assumption that all the rows of matrix M have the same norm. We emphasize that this assumption is purely for intuition purposes and is not necessary for the proof of our main theorem. Under such an assumption, the next proposition shows that the change in the Frobenius norm directly relates to the spectral properties of matrix M.

---

[11] Recall $\mathrm{Tr}(M_\ell) = d$ holds due to property (p.2) in Eq. 9

Proposition 8. *Suppose that all the rows of matrix M have the same norm. Let $\lambda \in \mathbb{R}^d$ contain the eigenvalues of matrix M. Then,*

$$Tr(M^3) = \|\lambda\|_3^3, \quad Tr(diag(M^2))^2 = \|\lambda\|^4/d, \quad \|M\|_F^2 = \|\lambda\|_2^2 \tag{44}$$

holds and hence

$$\delta_F(M) = \delta_F(\lambda) := 2d^2 - 2\|\lambda\|_2^2 - 8\|\lambda\|_3^3 + 8\|\lambda\|^4/d. \tag{45}$$

We postpone the proof to the end of this section. This proposition re-expresses the polynomial of Lemma 7 in terms of the eigenspectrum of M.

Based on the above proposition, we can make sense of interesting empirical observation reported in Figure 3.b. This figure plots the evolution of the eigenvalues of $M(H_\ell^{(\gamma)})$ after starting from a matrix $M(H_0)$ whose leading eigenvalue is large and all other eigenvalues are very small. We observe that a certain fraction of the small eigenvalues of $M(H_\ell^{(\gamma)})$ grow quickly with ℓ, while the leading eigenvalue is decreases in magnitude. In the next example, we show that the result of the last proposition actually predicts this observation.

Example 9. *Suppose that M is a matrix whose rows have the same norm. Let $\lambda_1 \geq \lambda_2, \dots, \lambda_d$ be the eigenvalues associated with the matrix M such that $\lambda_d = \lambda_{d-1} = \lambda_2 = \gamma^2$ and $\lambda_1 = d - \gamma^2(d-1)$. In this setting, Prop. 8 implies that $\mathbb{E}_W \|M_+\|_F^2 < \|M\|_F^2 - \gamma^4 d^2$ for a sufficiently small γ. This change has two consequences in expectation:(i.) the leading eigenvalue of M_+ is $O(-\gamma^4 d)$ smaller than the leading eigenvalue of M, and (ii.) some small eigenvalues of M_+ are greater than those of M (see Fig. 3.b).*

We provide a more detailed justification for the above statement at the end of this section. This example illustrates that the change in Frobenius norm (characterized in Lemma 7) can predict the change in the eigenvalues of $M(H_\ell^{(\gamma)})$ (singular values of $H_\ell^{(\gamma)}$) and hence the desired rank. Inspired by this, we base the proof of Theorem 2 on leveraging the invariance property of the unique invariant distribution with respect to Frobenius norm – i.e. setting $g(H) = \|M(H)\|_F^2$ in Def. 2.

An observation: regularity of the invariant distribution We now return to the result derived in Lemma 7 that characterizes the change in Frobenius norm of $M(H)$ after the recurrence of Eq. (7). We show how such a result can be used to leverage the invariance property with respect to the Frobenius norm. First, we observe that the term $\mathrm{Tr}(M(H)^3)$ in the expansion can be shown to dominate the term $\mathrm{Tr}(\mathrm{diag}(M(H)^2)^2)$ in expectation. The next definition states this dominance formally.

Definition 3. *(Regularity constant α) Let ν be a distribution over $H \in \mathcal{H}$. Then the regularity constant associated with ν is defined as the following ratio:*

$$\alpha = \mathbb{E}_{H\sim\nu}\left[Tr\left(diag(M(H)^2)^2\right)\right] / \left(\mathbb{E}_{H\sim\nu}\left[Tr\left(M(H)^3\right)\right]\right). \tag{46}$$

The next lemma states that the regularity constant α associated with the invariant distribution ν_γ is always less than one. Our analysis will in fact directly rely on $\alpha < 1$.

Lemma 10. *Suppose that the chain $\{H_\ell^{(\gamma)}\}$ admits the unique invariant distribution ν_γ (in Def. 2). Then, the regularity constant of ν_γ (in Def. 3) is less than one for a sufficiently small γ.*

Proof. We use a proof by contradiction where we suppose that the regularity constant of distribution ν_γ is greater than one. In this case, we prove that the distribution cannot be invariant with respect to the Frobenius norm.

If the regularity constant α is greater than one, then

$$\mathbb{E}_{H\sim\nu_\gamma}\left[-\mathrm{Tr}(M(H)^3) + \mathrm{Tr}(\mathrm{diag}(M(H)^2)^2)\right] \geq 0 \tag{47}$$

holds. According to Theorem 5, the rank of $M(H)$ is at least 2. Since the sum of the eigenvalues is constant d, the leading eigenvalue is less than d. This leads to

$$\|M(H)\|_F^2 = \sum_i \lambda_i^2 \leq \max_i \lambda_i \left(\sum_j \lambda_j\right) \leq d \max_i \lambda_i < d^2.$$

Plugging the above inequality together with inequality 47 into the established bound in Lemma 7 yields

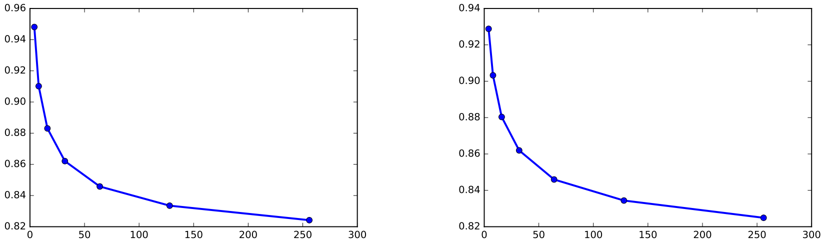
$$\mathbb{E}_{W,H\sim\nu_\gamma}\left[\|M(H_+)\|_F^2 - \|M(H)\|_F^2\right] > 0 \tag{48}$$

for a sufficiently small γ. Therefore, ν_γ does not obey the invariance property for $g(H) = \|M(H)\|_F^2$ in Def. 2. $\square$

We can experimentally estimate the regularity constant α using the Ergodicity of the chain. Assuming that the chain is Ergodic[12],

$$\lim_{L\to\infty} \frac{1}{L}\sum_{\ell=1}^{L} g(H_\ell^{(\gamma)}) = \mathbb{E}_{H\sim\nu_\gamma}\left[g(H)\right] \tag{49}$$

holds almost surely for every Borel bounded function $g: \mathcal{H} \to \mathbb{R}$. By setting $g_1(H) = \text{Tr}(M(H)^3)$ and $g_2(H) = \text{Tr}(\text{diag}(M(H)^2)^2)$, we can estimate $\mathbb{E}_{H\sim\nu_\gamma}\left[g_i(H)\right]$ for $i = 1$, and 2. Given these estimates, α can be estimated. Our experiments in Fig. 8 show that the regularity constant of the invariant distribution ν_γ is less than 0.9 for $d > 10$.



$\gamma = 1$ $\gamma = 0.1$

Figure 8: Regularity constant of the invariant distribution. The vertical axis is the estimated regularity constant α and the horizontal axis is d. We use $L = 10^5$ (in Eq. (49)).

Interlude: intuition behind the regularity We highlight the regularity constant does by itself not yield the desired rank property in Theorem 2. This is illustrated in the next example that shows how the regularity constant relates to the spectral properties of $M(H)$.

Example 11. *Suppose that the support of distribution ν contains only matrices $H \in \mathcal{H}$ for which all rows of $M(H)$ have the same norm. If the regularity constant of ν is greater than or equal to one, then all non-zero eigenvalues of matrix $M(H)$ are equal.*

A detailed justification of the above statement is presented at the end of this section. This example shows that the regularity constant does not necessarily relate to the rank of H, but instead it is determined by how much non-zero eigenvalues are close to each other. We believe that a sufficient variation in non-zero eigenvalues of $M(H)$ imposes the regularity of the law of H with a constant less than one (i.e. $\alpha < 1$ in Def. 3). The next example demonstrates this.

Example 12. *Suppose the support of distribution ν contains matrices $H \in \mathcal{H}$ for which all rows of $M(H)$ have the same norm. Let $\lambda \in \mathbb{R}^d$ contain sorted eigenvalues of $M(H)$. If $\lambda_1 = \Theta(d^\beta)$ and $\lambda_i = o(d^\beta)$ for $i > 1$ and $\beta < 1$,[13] then the regularity constant α associated with ν is less than 0.9 for sufficiently large d.*

We later provide further details about this example.

Invariance consequence The next lemma establishes a key result on the invariant distribution ν_γ.

Lemma 13. *Suppose that the chain $\{H_\ell^{(\gamma)}\}$ (see Eq. 7) admits the invariant distribution ν_γ (see Def. 2). If the regularity constant associated with ν_γ is $\alpha < 1$ (defined in Def. 3), then*

$$\mathbb{E}_{H\sim\nu_\gamma}\left[\|M(H)\|_F^2\right] \le d^{3/2}/\sqrt{1-\alpha} \tag{50}$$

holds for a sufficiently small γ.

---

[12] The uniqueness of the invariant distribution implies Ergodicity (see Theorem 5.2.1 and 5.2.6 [11]).

[13] According to definition, $\lim_{d\to\infty} o(d^\beta)/\Theta(d^\beta) = 0$

Proof. Leveraging invariance property in Def. 2,

$$\mathbb{E}_{W,H\sim\nu_\gamma}\left[\|M(H_+)\|_F^2 - \|M(H)\|_F^2\right] = 0 \tag{51}$$

holds where the expectation is taken with respect to the randomness of W and ν_γ.[14] Invoking the result of Lemma 7, we get

$$\mathbb{E}_{H\sim\nu_\gamma}\left[2d^2 - 2\|M(H)\|_F^2 - 8\text{Tr}(M(H)^3) + 8\text{Tr}(\text{diag}(M(H)^2)^2)\right] + \text{O}(\gamma) = 0. \tag{52}$$

Having a regularity constant less than one for ν_γ implies

$$0 \leq 2d^2 - \mathbb{E}_{H\sim\nu_\gamma}\left[2\|M(H)\|_F^2 - 8(1-\alpha)\text{Tr}(M(H)^3)\right] \tag{53}$$

holds for sufficiently small γ. Let $\lambda \in \mathbb{R}^d$ be a random vector containing the eigenvalues of the random matrix $M(H)$.[15] The eigenvalues of M^3 are λ^3, hence the invariance result can be written alternatively as

$$0 \leq 2d^2 - \mathbb{E}\left[2\|\lambda\|_2^2 - 8(1-\alpha)\|\lambda\|_3^3\right]. \tag{54}$$

The above equation leads to the following interesting spectral property:

$$\mathbb{E}\|\lambda\|_3^3 \leq d^2/(1-\alpha). \tag{55}$$

A straightforward application of Cauchy-schwarz yields:

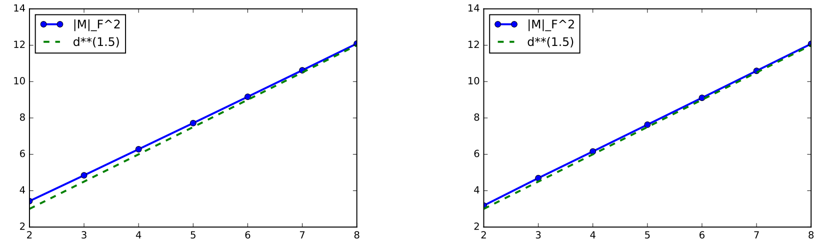
$$\|\lambda\|_2^2 = \sum_i \lambda_i^2 = \sum_i \lambda_i^{1/2}\lambda_i^{3/2} \leq \sqrt{\sum_i \lambda_i \sum_j \lambda_i^3} \leq \sqrt{d\|\lambda\|_3^3} \tag{56}$$

Given (i) the above bound, (ii) an application of Jensen's inequality, (iii) and the result of Eq. (55), we conclude with the desired result:

$$\mathbb{E}_{H\sim\nu_\gamma}\left[M(H)\right] = \mathbb{E}\left[\|\lambda\|_2^2\right] \overset{\text{(i)}}{\leq} \mathbb{E}\sqrt{d\|\lambda\|_3^3} \overset{\text{(ii)}}{\leq} \sqrt{d\mathbb{E}\|\lambda\|_3^3} \overset{\text{(iii)}}{\leq} d^{3/2}/\sqrt{1-\alpha} \tag{57}$$

$\square$

Notably, the invariant distribution is observed to have a regularity constant less than 0.9 (in Fig. 8) for sufficiently large d. This implies that an upper-bound $\text{O}\left(d^{3/2}\right)$ is achievable on the Frobenius norm. Leveraging Ergodicity (with respect to Frobenius norm in Eq. (49)), we experimentally validate the result of the last lemma in Fig. 9.



$\gamma = 1$ $\gamma = 0.1$

Figure 9: Dependency of $\mathbb{E}_{\nu_\gamma}\|M(H)\|_F^2$ on d. The horizontal axis is $\log_2(d)$ and the vertical axis shows $\log_2(\frac{1}{L}\sum_{\ell=1}^L \|M(H_\ell^{(\gamma)})\|_F^2)$ for $L = 10^5$. The green dashed-line plots $\log_2(d^{1.5})$.

Proof of the Main Theorem Here, we give a formal statement of the main Theorem that contains all required additional details (which we omitted for simplicity in the original statement).

---

[14] This result is obtained by setting $g(H) = \|M(H)\|_F^2$ in Def. 2.

[15] Note that $H \in \mathcal{H}$ is a random matrix whose law is ν_γ, hence $\lambda \in \mathbb{R}^d$ is also a random vector.

Theorem 14 (Formal statement of Theorem 2)**.** *Suppose that rank$(X) = d$, γ is sufficiently small, and all elements of the weight matrices $\{W_\ell\}$ are drawn i.i.d. from a zero-mean, unit variance distribution whose support lies in $[-B, B]$ and its law is symmertic around zero. Furthermore, assume that the Markov chain $\{H_\ell^{(\gamma)}\}$ (defined in Eq. 1) admits a unique invariant distribution. Then, the regularity constant $\alpha > 0$ associated with ν_γ (see Def. 3) is less than one and the following limits exist such that*

$$\lim_{L\to\infty} \frac{1}{L}\sum_{\ell=1}^{L} \mathit{rank}_\tau(H_\ell^{(\gamma)}) \geq \lim_{L\to\infty} \frac{(1-\tau)^2}{L}\sum_{\ell=1}^{L} r(H_\ell^{(\gamma)}) \geq (1-\tau)^2(1-\alpha)^{1/2}\sqrt{d} \tag{58}$$

holds almost surely for all $\tau \in [0, 1]$. Assuming that the regularity constant α does not increase with respect to d, the above lower-bound is proportional to $(1-\alpha)^{1/2}\sqrt{d} = \Omega(\sqrt{d})$.

Remarkably, we experimentally observed (in Fig. 8) that the regularity constant α is decreasing with respect to d. Examples 11 and 12 provide insights about the regularity constant. We believe that it is possible to prove that the constant α is non-increasing with respect to d.

Proof of Theorem 2 . Lemma 10 proves that the regularity constant α is less than one for the unique invariant distribution. Suppose that $H \in \mathcal{H}$ is a random matrix whose law is the one of the unique invariant distribution of the chain. For $H \in \mathcal{H}$, we get $\mathrm{Tr}(M(H)) = d$. A straightforward application of Jensen's inequality yields the following lower bound on the expectation of $r(H)$ (i.e. the lower bound on the rank):

$$\mathbb{E}\left[r(H)\right] = \mathbb{E}\left[\mathrm{Tr}(M(H))^2/\|M(H)\|_F^2\right] = \mathbb{E}\left[d^2/\|M(H)\|_F^2\right] \geq d^2/\mathbb{E}\left[\|M(H)\|_F^2\right] \tag{59}$$

where the expectation is taken over the randomness of H (i.e. the invariant distribution). Invoking the result of Lemma 13, we get an upper-bound on the expectation of the Frobenius norm – in the right-side of the above equation. Therefore,

$$\mathbb{E}\left[r(H)\right] \geq \sqrt{(1-\alpha)d} \tag{60}$$

holds. The uniqueness of the invariant distribution allows us to invoke Birkhoff's Ergodic Theorem for Markov Chains (Theorem 5.2.1 and 5.2.6 [11]) to get

$$\lim_{L\to\infty} \frac{1}{L}\sum_{\ell=1}^{L} r(H_\ell^{(\gamma)}) = \mathbb{E}\left[r(H)\right] \geq \sqrt{(1-\alpha)d}. \tag{61}$$

The established lower bound on $\mathrm{rank}_\tau(H_\ell^{(\gamma)})$ –in terms of $r(H_\ell^{(\gamma)})$– in Lemma 1 concludes

$$\lim_{L\to\infty} \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{rank}_\tau(H_\ell^{(\gamma)}) \geq \lim_{L\to\infty} \frac{(1-\tau)^2}{L}\sum_{\ell=1}^{L} r(H_\ell^{(\gamma)}) \geq (1-\tau)^2\sqrt{(1-\alpha)d}. \tag{62}$$

□

As shown in the following corollary, one can extend the result of Theorem 14 for any finite ℓ.

Corollary 15. *Under the setting of Thm. 14, rank$(H_\ell) = \Omega(\sqrt{d})$ holds almost surely for all finite integer ℓ. Assuming that $\{$rank$_\tau(H_\ell)\}$ is a monotonically no-increasing sequence, then rank$_\tau(H_\ell) = \Omega((1-\tau)^2\sqrt{d})$ holds almost surely for all finite ℓ.*

Proof. The proof is based on the no-increasing property of the rank[16]. Next lemma presents a straightforward implication of this property.

Lemma 16. *Consider a sequence of non-increasing bounded finite integers $\{y_k\}_{k=1}^{\infty}$. If $\lim_{N\to\infty}\sum_{k=1}^{N} y_k/N$ exists and is greater than α, then $y_k \geq \alpha$ for all finite k.*

The proof of the last lemma is provided at the end of this section. Replacing the result of Thm. 14 into the above lemma concludes the proof of the corollary.

□

---

[16] Recall that the rank does not increases in updates of Eq. (7)

A remark on the number of hidden units. The focus of our analysis was networks with the same number of hidden units in each layer. Yet, this result extends to more general architectures. Most of modern neural architectures consists of blocks in which the number of hidden units are constant. For example, VGG19-Nets and ResNets are consist of blocks convolutional layers with 64, 128, 256, and 512 channels where the number channels are equal in each block. An analogy of such an architecture is an MLP with different blocks of hidden layers where the numbers of hidden units are the same in each block. According to Cor. 15, the rank preservation property holds in each block after applying BN. In this way, one can extend the established results of Thm. 14 and Cor. 15 to a general family of architectures with varying number of hidden units.

Postponed proofs.

Proof of Lemma 7. The proof is based on a Taylor expansion of the BN non-linear recurrence function, which we restate here for simplicity:

$$H_+ = (\operatorname{diag}(M(H_\gamma)))^{-1/2} H_\gamma, \quad H_\gamma = (I + \gamma W)H \tag{63}$$

Consider the covariance matrices $M = M(H)$ and $M_+ = M(H_+)$ which obey

$$M_\gamma := M(H_\gamma) = M + \Delta M, \quad \Delta M := \gamma W M + \gamma M W^\top + \gamma^2 W M W^\top \tag{64}$$

$$[M_+]^2_{ij} = g_{ij}(M_\gamma) = [M_\gamma]^2_{ij} / [M_\gamma]_{ii} [M_\gamma]_{jj} \tag{65}$$

For the sake of simplicity, we use the compact notation $g := g_{ij}$ for $i \neq j$. We further introduce the set of indices $S = \{ii, ij, jj\}$. A taylor expansion of g at M yields

$$\mathbb{E}_W[g(M_\gamma)] = g(M) + \underbrace{\sum_{pq \in S} \left(\frac{\partial g(M)}{\partial M_{pq}} \right) \mathbb{E}_W[\Delta M_{pq}]}_{T_1} + \underbrace{\frac{1}{2} \sum_{pq, km \in S} \left(\frac{\partial^2 g(M)}{\partial M_{pq} \partial M_{km}} \right) \mathbb{E}_W[\Delta M_{pq} \Delta M_{km}]}_{T_2} + \mathrm{O}(\gamma^3). \tag{66}$$

Note that the choice of the element-wise uniform distribution over $[-\sqrt{3}, \sqrt{3}]$ allows us to deterministically bound the Taylor remainder term by $\mathrm{O}(\gamma^3)$. Now, we compute the derivatives and expectations that appear in the above expansion individually. Let us start with the term T_1. The first-order partial derivative term in T_1 is computed bellow.

$$\frac{\partial g(M)}{\partial M_{pq}} = \begin{cases} -M_{ij}^2 / (M_{ii}^2 M_{jj}) = -g(M) & pq = \{ii, jj\} \\ 2M_{ij} / (M_{ii} M_{jj}) & pq = \{ij\}. \end{cases} \tag{67}$$

The expectation term in T_1 is

$$\mathbb{E}_W[\Delta M_{pq}] = \begin{cases} 0 & pq = \{ij\} \\ \gamma^2 \sum_{k=1}^d M_{kk} = \gamma^2 d & pq = \{ii, jj\}. \end{cases} \tag{68}$$

Given the above formula, we reach the following compact expression for T_1:

$$T_1 = -2\gamma^2 d g(M). \tag{69}$$

The compute T_2 we need to compute second-order partial derivatives of g and also estimate the following expectation:

$$\mathbb{E}_W[\Delta M_{pq} \Delta M_{km}] = \gamma^2 \left(\underbrace{\mathbb{E}_W\left[[WM + MW^\top]_{pq} [WM + MW^\top]_{km} \right]}_{K_{pq,km}} \right) + \mathrm{O}(\gamma^3). \tag{70}$$

We now compute $K_{pq,km}$ in the above formula

$$K_{\alpha,\beta} = \begin{cases} \sum_k M_{kj}^2 + \sum_n M_{in}^2 & \alpha = \{ij\}, \beta = \{ij\} \\ 2 \sum_k M_{kj} M_{ki} & \alpha = \{ij\}, \beta = \{ii\} \\ 4 \sum_k M_{ki}^2 & \alpha = \{ii\}, \beta = \{ii\} \\ 0 & \alpha = \{ii\}, \beta = \{jj\} \end{cases} \tag{71}$$

The second-order partial derivatives of g reads as

$$\frac{\partial^2 g(M)}{\partial M_\alpha \partial M_\beta} = \begin{cases} 2 & \alpha = \{ij\}, \beta = \{ij\} \\ -2M_{ij} & \alpha = \{ij\}, \beta = \{ii\} \\ +2M_{ij}^2 & \alpha = \{ii\}, \beta = \{ii\} \\ M_{ij}^2 & \alpha = \{jj\}, \beta = \{ii\} \end{cases} \tag{72}$$

Now, we replace the computed partial derivatives and the expectations into T_2:

$$T_2 = \sum_k M_{kj}^2 + \sum_n M_{in}^2 - 8\sum_k M_{kj}M_{ij}M_{ki} + 4\sum_k M_{ij}^2 M_{ki}^2 + 4\sum_k M_{ij}^2 M_{kj}^2 \tag{73}$$

Plugging terms T_1 and T_2 into the Taylor expansion yields

$$\begin{aligned}&\mathbb{E}_W\left[g_{ij}(M_+) - g_{ij}(M)\right]/(\gamma^2)\\ &= \sum_k M_{kj}^2 + \sum_n M_{in}^2 - 2dg_{ij}(M) - 8\sum_k M_{kj}M_{ij}M_{ki} + 4\sum_k M_{ij}^2 M_{ki}^2 + 4\sum_k M_{ij}^2 M_{kj}^2 + \mathrm{O}(\gamma)\end{aligned} \tag{74}$$

Summing over $i \neq j$ concludes the proof (note that the diagonal elements are one for the both of matrices M and M_+). ☐

Proof of Proposition 8. Consider the spectral decomposition of matrix M as $M = U\text{diag}(\lambda)U^\top$, then $M^k = U\text{diag}(\lambda^k)U^\top$. Since $\text{Tr}(M^k)$ is equal to the sum of the eigenvalues of M^k, we get

$$\text{Tr}(M^k) = \sum_{i=1}^{d} \lambda_i^k = \|\lambda\|_k^k \tag{75}$$

for $k = 2$ and $k = 3$. The sum of the squared norm of the rows in M is equal to the Frobenius norm of M. Assuming that the rows have equal norm, we get

$$\sum_{k=1}^{d} M_{ik}^2 = \sum_{i=1}^{d}\sum_{k=1}^{d} M_{ik}^2/d = \|M\|_F^2/d = \|\lambda\|_2^2/d. \tag{76}$$

Therefore,

$$\text{Tr}(\text{diag}(M^2)^2) = \sum_{i=1}^{d}\left(\sum_{k=1}^{d} M_{ik}^2\right)^2 = \|\lambda\|_2^4/d \tag{77}$$

holds.

☐

Details of Example 9. Under the assumptions stated in Example 9, we get

$$\|\lambda\|_2^2 \approx d^2 - 2\gamma^2 d, \quad \|\lambda\|_3^3 \approx d^3 - 3\gamma^2 d^2, \quad \|\lambda\|_2^4 \approx d^4 - 4\gamma^2 d^3 \tag{78}$$

where the approximations are obtained by a first-order Taylor approximation of the norms at $\lambda' = (d, 0, \ldots, 0)$, and all small terms $o(\gamma^2)$ are omitted. Using the result of Proposition 8, we get

$$\mathbb{E}\left[\|M_+\|_F^2\right] - \mathbb{E}\left[\|M\|_F^2\right] \approx \gamma^2\delta_F(\lambda) \approx \mathrm{O}(-\gamma^4 d^2). \tag{79}$$

Let λ_+ be the eigenvalues of the matrix M_+, then

$$\sum_{i=1}^{d} \mathbb{E}[\lambda_+^2]_i - \lambda_i^2 = \mathrm{O}(-\gamma^4 d^2) \tag{80}$$

$$\implies \max_i \mathbb{E}[\lambda_+^2]_i - \lambda_1^2 \leq \mathrm{O}(-\gamma^4 d^2) + \sum_{i=2}^{d} \lambda_i^2 \leq \mathrm{O}(-\gamma^4 d^2) + \gamma^4 d = \mathrm{O}(-\gamma^4 d^2). \tag{81}$$

Let $j = \arg\max_i \mathbb{E}\left[[\lambda_+]_i^2\right]$. A straight-forward application of Jensen's inequality yields

$$\mathbb{E}\left[[\lambda_+]_j\right] \leq \sqrt{\mathbb{E}\left[[\lambda_+]_j^2\right]} \leq \lambda_1 - \mathrm{O}(\gamma^4 d). \tag{82}$$

Hence the leading eigenvalue of M_+ is smaller than the one of M. Since the sum of eigenvalues λ_+ and λ are equal, some of the eigenvalues λ_+ are greater than those of λ (in expectation) to compensate $\mathbb{E}[\lambda_+]_j < \lambda_1$. ☐

Details of Example 11. Invoking Prop. 8, we get

$$\mathbb{E}\left[\text{Tr}(M(H)^3)\right] = \|\lambda\|^3, \quad \mathbb{E}\left[\text{diag}(M(H)^2)^2\right] = \|\lambda\|_2^4/d, \tag{83}$$

where $\lambda \in \mathbb{R}^d$ contains the eigenvalues of $M(H)$. Since $H \in \mathcal{H}$, $\|\lambda\|_1 = d$. If the regularity constant is greater than or equal to one, then

$$\|\lambda\|_3^3 \leq \|\lambda\|_2^4/d = \|\lambda\|_2^4/\|\lambda\|_1. \tag{84}$$

A straightforward application of Cauchy-Schwartz yields:

$$\begin{aligned}\|\lambda\|_2^4 = \sum_{i=1}^d \sum_{j=1}^d \lambda_i^2 \lambda_j^2 = \sum_{i=1}^d \sum_{j=1}^d (\lambda_i \lambda_j)^{1/2} (\lambda_i \lambda_j)^{3/2} & \\ \leq \sqrt{\left(\sum_{i,j} \lambda_i \lambda_j\right)\left(\sum_{i,j} \lambda_i^3 \lambda_j^3\right)} &= \|\lambda\|_1 \|\lambda\|_3^3 \end{aligned} \tag{85}$$

The above result together with inequality 84 yields that

$$\|\lambda\|_3^3 = \|\lambda\|_2^4/d = \|\lambda\|_2^4/\|\lambda\|_1. \tag{86}$$

Finally, the above equality is met only when all non-zero eigenvalues are equal.

□

Details of Example 12. Since $\lambda_1 = \Theta(d^\beta)$ and $\lambda_{i>1} = o(d^\beta)$, we get

$$\|\lambda\|_3^3 = \Theta(d^{3\beta}), \quad \|\lambda\|_2^2 = \Theta(d^{2\beta}). \tag{87}$$

Thus, Prop. 8 yields

$$\mathbb{E}\left[\text{Tr}(M^3)\right] = \Theta(d^{3\beta}), \quad \mathbb{E}\left[\text{Tr}(\text{diag}(M^2)^2)\right] = \|\lambda\|_2^4/d = \Theta(d^{4\beta-1}) \tag{88}$$

Therefore,

$$\alpha = \lim_{d\to\infty} \frac{\mathbb{E}\left[\text{Tr}(\text{diag}(M^2)^2)\right]}{\mathbb{E}\left[\text{Tr}(M^3)\right]} = \text{O}(d^{\beta-1}) = 0. \tag{89}$$

As a result, α is less than 0.9 for sufficiently large d. □

Proof of Lemma 16. The proof is based on a contradiction. Suppose that there exits a finite n such that $y_n < \alpha$. Since the sequence is non-increasing, $y_m < \alpha$ for holds for all $m > n$. This yields

$$\begin{aligned}\lim_{N\to\infty} \sum_{k=1}^N y_k/N &= \lim_{N\to\infty} \left(\sum_{k>n}^N y_k/N + \sum_{k\leq n} y_k/N\right) &(90)\\ &< \frac{(N-n)}{N}\alpha + \lim_{N\to\infty} \sum_{k\leq n} y_k/N &(91)\\ &= \frac{(N-n)}{N}\alpha, &(92)\end{aligned}$$

where we used the fact that all y_k are bounded. The above result contradicts the fact that $\lim_{n\to\infty} \sum_{k=1}^N y_k/N > \alpha$. □

F Analysis for Vanilla Linear Networks.

In this section, we prove Lemma 3 that states the rank vanishing problem for vanilla linear networks. Since the proof relies on existing results on products of random matrices (PRM) [9], we first shortly review these results. Let T be the set of $d \times d$ matrices. Then, we review two notions for T: contractiveness and strong irreducibility.

Definition 4 (Contracting set [9]). *T is contracting if there exists a sequence $\{M_n \in T, n \geq 0\}$ such that $M_n/\|M_n\|$ converges to a rank one matrix.*

Definition 5 (Invariant union of proper subspaces [9]). *Consider a family of finite proper linear subspace $V_1, \dots, V_k \subset \mathbb{R}^d$. The union of these subspaces is invariant with respect to T, if $Mv \in V_1$ or V_2 or $\dots$ or V_k holds for $\forall v \in V_1$ or V_2 or $\dots$ or V_k and $\forall M \in T$.*

Example 17. *Consider the following sets*

$$T = \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right), \quad V_1 = \left(span(\underbrace{[0,1]}_{v_1}) \right), \quad V_2 = \left(span(\underbrace{[1,0]}_{v_2}) \right);$$

then, union of V_1 and V_2 is invariant with respect to T because $\alpha T v_1 \in V_2$ and $\alpha T v_2 \in V_1$ hold for $\alpha \neq 0$.

Definition 6 (Strongly irreducible set [9]). *The set T is strongly irreducible if there does not exist a finite family of proper linear subspaces of $\mathbb{R}^d$ such that their union is invariant with respect to T.*

For example, the set T defined in Example 17 is not strongly irreducible.

Lemma 18 (Thm 3.1 of [9]). *Let $W_1, W_2, \dots$ be random $d \times d$ matrices drawn independently from a distribution μ. Let $B_n = \prod_{k=1}^n W_k$. If the support of μ is strongly irreducible and contracting, then any limit point of $\{B_n/\|B_n\|\}_{n=1}^\infty$ is a rank one matrix almost surely.*

This result allows us to prove Lemma 3.

Proof of Lemma 3. Recall the structure of the random weight matrices as $\widehat{W_k} = I + \gamma W_k$ where the coordinates W_k are i.i.d. from (a.) standard Gaussian, (b.) uniform$[-\sqrt{3}, \sqrt{3}]$ (i.e. with variance 1). One can readily check that for the Gaussian weights, the contracting and strong irreducibility hold and one can directly invoke the result of lemma 18 to get part (a.) of Lemma 3. Now, we prove part (b.). Let m be a random integer that obeys the law $p(m = k) = 2^{-k}$. Given the random variable m, we define the random matrix $Y = \prod_{k=1}^m \widehat{W}_k$ and use the notation μ' for its law. Let $\{Y_i = \prod_{j=1}^{m_i} \widehat{W}_k\}_{i=1}^k$ be drawn i.i.d. from μ'. Then, $C_k := Y_k \dots Y_2 Y_1$ is distributed as $B_{\ell_k} := \widehat{W}_{\ell_k} \dots \widehat{W}_2 \widehat{W}_1$ for $\ell_k = \sum_{i=1}^k m_i$. We prove that every limit point of $\{C_k/\|C_k\|\}$ converges to a rank one matrix, which equates the convergence of limit points of $\{B_{\ell_k}/\|B_{\ell_k}\|\}$ to a rank one matrix. To this end, we prove that the support of μ' denoted by $T_{\mu'}$ is contractive and strongly contractive. Then, Lemma 18 implies that the limit points of $\{C_k/\|C_k\|\}$ are rank one.

Contracting. Let $e_1 \in \mathbb{R}^d$ be the first standard basis vector. Since $A_n := (I + \gamma e_1 e_1^\top)^n \in T_{\mu'}$ and its limit point $\{A_n/\|A_n\|\}$ converges to a rank one matrix, $T_{\mu'}$ is contractive.

Strong irreduciblity. Consider an arbitrary family of linear proper subspace of $\mathbb{R}^d$ as $\{V_1, \dots, V_q\}$. Let v be an arbitrary unit norm vector which belongs to one of the subspaces $\{V_i\}_{i=1}^q$. Given v, we define an indexed family of matrices $\{M_\alpha \in T_{\mu'} | \alpha \in \mathbb{R}^d, |\alpha_i| \leq 1\}$ such that

$$M_\alpha = I + \frac{\gamma}{d} \sum_{i=1}^d \alpha_i e_i v^\top \in T_{\mu'}, \tag{93}$$

where e_i is the i-th standard basis[17]. Then, we get

$$M_\alpha v = v + \frac{\gamma}{d} \sum_{i=1}^d \alpha_i e_i. \tag{94}$$

Therefore, $\{M_\alpha v | |\alpha_i| \leq 1\}$ is not contained in any union of finite proper $(m < k)$-dimensional linear subspace of $\mathbb{R}^d$, hence $T_{\mu'}$ is strongly irreducible.

□

---

[17] Notably, the absolute value of each element of $\frac{1}{d} \sum_{i=1}^d \alpha_i e_i v^\top$ is less than 1, hence this matrix belongs to the support of μ.

G Details: Pretraining algorithm

In Section 4, we introduced a pre-training method that effectively obtains a better optimization performance compared ot BN. In this section, we provide more details about the pre-training step. Recall $X \in \mathbb{R}^{d \times N}$ is a minibatch of d-dimensional inputs of size N. Let $H_L(X) \in \mathbb{R}^{d \times N}$ be the hidden representation of input X in the last layer of a MLP. Using gradient descent method, we optimize $r(H_L(X))$ –with respect to the parameters of networks– over different minibatches X. Algorithm 1 presents our pretraining method. As can be seen, the procedure is very simple.

Algorithm 1 Pretraining

```
Input: Training set $S$, a network with parameters $\Theta$ and $L$ layers, and constant $N, M$, and $T$
for $k = 1, 2, \ldots, M$ do
    Draw minibatch $X_k$ of size $N$ i.i.d. from $S$
    for $t = 1, 2, \ldots, T$ do
        Take one GD step on $r(H_L(X_k))$ w.r.t $\Theta$.
    end for
end for
return $\Theta$.
```

H Details: Why the rank matters for gradient based learning.

We now provide an intuitive explanation of why rank one hidden representations prevent randomly initialized networks from learning. Particularly, we argue that these networks essentially map all inputs to a very small subspace[18] such that the final classification layer can no longer disentangle the hidden representations. As a result, the gradients of that layer also align, yielding a learning signal that becomes *independent* of the input.

To make this claim more precise, consider training the linear network from Eq. (6) on a dataset $X \in \mathbb{R}^{d \times N}$, where $x_i \in \mathbb{R}^d$ with d_{out} targets $y_i \in \mathbb{R}^{d_{out}}, i = 1, \ldots, N$. Each column $\widehat{H}_{L,i}^{(\gamma)}$ of the hidden representations in the last hidden layer $\widehat{H}_L^{(\gamma)}$ is the latent representation of datapoint i, which is fed into a final classification layer parametrized by $W_{L+1} \in \mathbb{R}^{d_{out} \times d}$. We optimize $\mathcal{L}(\mathbf{W})$, where $\mathbf{W}$ is a tensor containing all weights $W_1, \ldots, W_{L+1}$ and $\widehat{H}_{L,i}^{(\gamma)}$ is a function of $W_1, \ldots, W_L$ (as detailed in Eq. (6):

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = \sum_{i=1}^{N} \underbrace{\ell\left(y_i, W_{L+1} \widehat{H}_{L,i}^{(\gamma)}(W_1, ..., W_L)\right)}_{:=\mathcal{L}_i(\mathbf{W})}, \tag{95}$$

and $\ell : \mathbb{R}^{d_{out}} \to \mathbb{R}^+$ is a differentiable loss function. Now, if the the hidden representations become rank one (as predicted by Lemma 3 and Fig. 2), one can readily check that the stochastic gradients of any neuron k in the last linear layer, i.e., $\nabla_{W_{L,[k,:]}} \mathcal{L}_i(\mathbf{W}) = (\nabla \ell_i)_k \widehat{H}_{L,i}^{(\gamma)}$, align for both linear and ReLU networks.

Proposition 19. *Consider a network with rank one hidden representations in the last layer $\widehat{H}_L^{(\gamma)}(W_1, ..., W_L)$, then for any neuron k and any two datapoints i, j with non-zero errors $\mathcal{L}_i$ and $\mathcal{L}_j$ we have*

$$\nabla_{W_{L+1,[k,:]}} \mathcal{L}_i(\mathbf{W}) = \underbrace{\frac{c(\nabla \ell_i)_k}{(\nabla \ell_j)_k}}_{\in \mathbb{R}} \nabla_{W_{L+1,[k,:]}} \mathcal{L}_j(\mathbf{W}) \tag{96}$$

$\forall i, j$. That is, all stochastic gradients of neuron k in the final classification layer align along one single direction in $\mathbb{R}^d$.

---

[18] A single line in $\mathbb{R}^d$ in the extreme case of rank one mappings

Proof. The result follows directly from a simple application of the chain rule

$$\begin{aligned}\frac{\partial \mathcal{L}_i(\mathbf{W})}{\partial W_{L+1}} = \frac{\partial \ell(\mathbf{y}_i, W_{L+1}\widehat{H}^{(\gamma)}_{L,i})}{\partial W_{L+1}\widehat{H}^{(\gamma)}_{L,i}} \frac{\partial W_{L+1}\mathbf{h}_{L,i}}{\partial W_{L+1}} &= \nabla_{W_{L+1}\widehat{H}^{(\gamma)}_{L,i}} \ell(\mathbf{y}_i, W_{L+1}\mathbf{h}_{L,i})(\widehat{H}^{(\gamma)}_{L,i})^{\intercal} \\ &= \begin{bmatrix} \nabla\ell_{i,1}\widehat{H}^{(\gamma)}_{L,i,1}, \dots, \nabla\ell_{i,1}\widehat{H}^{(\gamma)}_{L,i,d} \\ \ddots \\ \nabla\ell_{i,d_{out}}\widehat{H}^{(\gamma)}_{L,i,1}, \dots, \nabla\ell_{i,d_{out}}\widehat{H}^{(\gamma)}_{L,i,d} \end{bmatrix} \in \mathbb{R}^{d_{out}\times d}\end{aligned} \tag{97}$$

The same holds for j. Now, if $\widehat{H}^{(\gamma)}_{L,i} = c\widehat{H}^{(\gamma)}_{L,i}, c \in \mathbb{R} \setminus \{0\}$ then

$$\left(\frac{\partial \mathcal{L}_i(\mathbf{W})}{\partial W_{L+1}}\right)_{k,:} = c\underbrace{\frac{\nabla\ell_{i,k}}{\nabla\ell_{j,k}}}_{\in\mathbb{R}}\left(\frac{\partial \mathcal{L}_j(\mathbf{W})}{\partial W_{L+1}}\right)_{k,:}$$

$\square$

To validate this claim, we again train CIFAR-10 on the VGG19 network from Figure 5 (top).

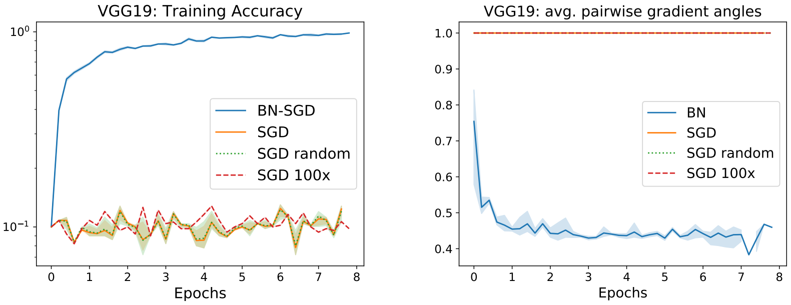


Figure 10: **Directional gradient vanishing** CIFAR-10 on a VGG19 network with BN, SGD, SGD with 100x learning rate and SGD on random data. Average and 95% confidence interval of 5 independent runs.

As expected, the network shows perfectly aligned gradients without BN (right hand side of Fig. 10), which renders it un-trainable. In a next step, we replace the input by images generated randomly from a uniform distribution between 0 and 255 and find that SGD takes almost the exact same path on this data (compare log accuracy on the left hand side). Thus, our results suggest that the commonly accepted vanishing gradient *norm* hypothesis is not descriptive enough since SGD does not take small steps into the *right* direction- but into a *random* one after initialization in deep neural networks. As a result, even a 100x increase in the learning rate does not allow training. We consider our observation as a potential starting point for novel theoretical analysis focusing on understanding the propagation of information through neural networks, whose importance has also been highlighted by [7].

I Additional Experiments

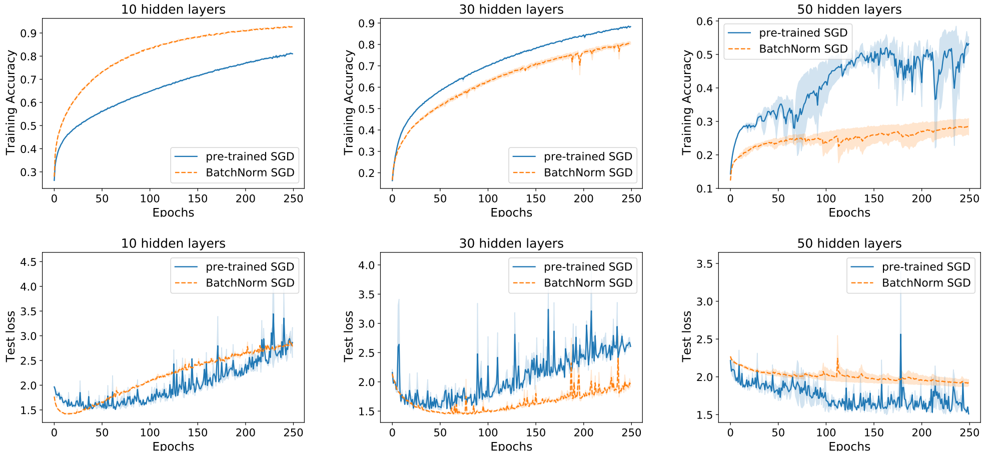


Figure 11: CIFAR-10: Same setting as Fig.4 but now showing accuracy and test loss

Outperforming BN The following Figure shows the result of the experiment of Fig. 4 that is repeated for FashionMNIST dataset. As can be seen, overfitting tends to happen whenever a certain accuracy is achieved on the training set, regardless of the actual method that is used for optimization.

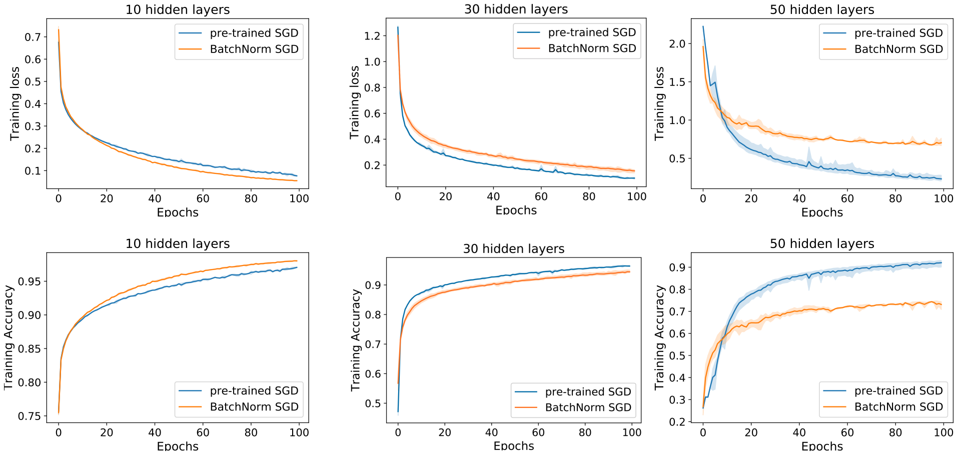


Figure 12: Results of Fig.4 for FashionMNIST

Breaking BN In the following result, we repeated the experiment of Fig. 5 for ResNets.

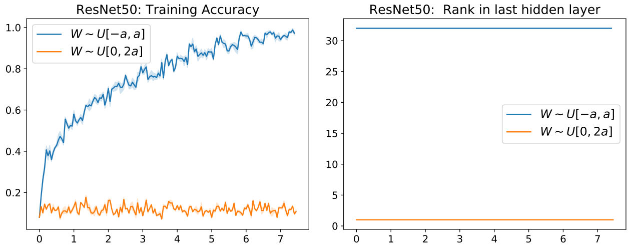


Figure 13: **Breaking Batchnorm:** CIFAR-10 on a ResNet-50 with standard PyTorch initialization as well as a uniform initialization of same variance in $\mathbb{R}^+$. Average and 95% confidence interval of 5 independent runs. This plot also shows results for a BN network without mean deduction/adaption, validating our claim from Section 2.