On the Initialisation of Wide Low-Rank Feedforward Neural Networks

Thiziri Nait Saada [1] **Jared Tanner** [1]

Abstract

The edge-of-chaos dynamics of wide randomly initialized low-rank feedforward networks are analyzed. Formulae for the optimal weight and bias variances are extended from the full-rank to low-rank setting and are shown to follow from multiplicative scaling. The principle second order effect, the variance of the input-output Jacobian, is derived and shown to increase as the rank to width ratio decreases. These results inform practitioners how to randomly initialize feedforward networks with a reduced number of learnable parameters while in the same ambient dimension, allowing reductions in the computational cost and memory constraints of the associated network.

1. Introduction

Neural networks being applied to new settings, limiting transfer learning, are typically initialized with i.i.d. random entries. The edge-of-chaos theory of (Poole et al., 2016) determine the appropriate scaling of the weight matrices and biases so that intermediate layer representations (1) and the median of the input-output Jacobian's spectra (10) are to first order independent of the layer. Without this normalization there is typically an *exponential* growth in the magnitude of these intermediate representations and gradients as they progress between layers of the network; such a disparity of scale inhibits the early training of the network (Glorot & Bengio, 2010).

For instance, consider an untrained fully connected neural network whose weights and biases are set to be respectively identically and independently distributed with respect to a Gaussian distributions: $W_{ij}^{(l)} \sim \mathcal{N}(0, \frac{\sigma_W^2}{N_{l-1}})$, $b_i^{(l)} \sim \mathcal{N}(0, \sigma_b^2)$ with N_l the width at layer l. Starting such a network, with nonlinear activation $\phi : \mathbb{R} \to \mathbb{R}$, from an input vector $z^0 := x^0 \in \mathbb{R}^{N_0}$, the data propagation is then given by the following equations,

$$h_j^{(l)} = \sum_{k=1}^{N_{l-1}} W_{jk}^{(l)} z_k^{(l)} + b_j^{(l)}, \qquad z_k^{(l)} = \phi(h_k^{(l-1)}) \quad (1)$$

where we call $h^{(l)}$ the preactivation vector at layer l.

It has been shown by (Poole et al., 2016) that the pre-activation vectors h^l have geometric properties of length $q^l := N_l^{-1} \left(h^l\right)^T h^l$ and the pairwise covariance $q_{12}^l := N_l^{-1} \left(h_1^l\right)^T h_2^l$ of two inputs $x^{0,1}$ and $x^{0,2}$ which propagate through the network according to functions of the network entries' variances and nonlinear activation $(\sigma_b, \sigma_W, \phi)$. These propagation maps were computed by (Poole et al., 2016) in the limiting setting of infinitely wide networks and either i.i.d. Gaussian entries or scaled randomly drawn orthnormal matrices. Here we extend this setting to their low-rank analogous.

Consider rank $r_l := \gamma_l N_l$ weight matrices, $W^{(l)} \in \mathbb{R}^{N_l \times N_{l-1}}$, formed as

$$W_{ij}^{(l)} = \sum_{k=1}^{r_l} \alpha_{k,j}^{(l)} (C_k^l)_i, \quad (2)$$

where the scalars $\left(\alpha_{k,i}^{(l)}\right)_{1 \le i \le N_{l-1}} \in \mathbb{R} \overset{\text{iid}}{\sim} \mathcal{N}(0, \frac{\sigma_\alpha^2}{N_{l-1}})$ and the columns $C_1^{(l)}, \dots, C_{r_l}^{(l)}$ are drawn jointly as the matrix $C^{(l)} := [C_1^{(l)}, \dots, C_{r_l}^{(l)}] \in \mathbb{R}^{N_l \times r_l}$ from the Grassmannian of rank r matrices with orthonormal columns having zero mean and variance $1/N_l$. Similarly, consider bias vectors within the same column span as $W^{(l)}$, given by $b^{(l)}(C_1^{(l)} + \cdots + C_{r_l}^{(l)})$, where $b^{(l)} \in \mathbb{R} \sim \mathcal{N}(0, \sigma_b^2)$. It is shown in Appendix A.2 that, in the large width limit, the preactivation vector $h^{(l)}$ follows a Gaussian distribution over the $r-$dimensional column span of $W^{(l)}$ with a non-diagonal covariance; this differs from the full rank setting in (Poole et al., 2016) where the entries in (1) are independent.

We extend the pre-activation length and correlation maps to this low-rank setting:

$$q^l = \gamma_l \left(\sigma_\alpha^2 \int_{\mathbb{R}} \phi^2(\sqrt{q^{l-1}} z) Dz + \sigma_b^2\right) \quad (3)$$

$$:= \mathcal{V}(q^{(l-1)} | \sigma_\alpha, \sigma_b, \gamma_l) \quad (4)$$

[1]Mathematical Institute, University of Oxford, UK. Correspondence to: Thiziri Nait Saada <thiziri.naitsaada@maths.ox.ac.uk>.

Preprint. Under review by the International Conference on Machine Learning (ICML).

where $Dz := \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$ is the Gaussian probability measure, and

$$q_{12}^l = \gamma_l \left(\sigma_\alpha^2 \int_{\mathbb{R}^2} \phi(u_1)\phi(u_2) Dz_1 Dz_2 + \sigma_b^2 \right), \quad (5)$$

$$:= \mathcal{C}(q_{ab}^{l-1}, q_{aa}^{l-1}, q_{bb}^{l-1} | \sigma_\alpha, \sigma_b, \gamma_l) \quad (6)$$

with $u_1 = \sqrt{q_{11}^{l-1}} z_1, u_2 = \sqrt{q_{22}^{l-1}} (c_{12}^{l-1} z_1 + \sqrt{1-(c_{12}^{l-1})^2} z_2$ and

$$c_{12}^l = q_{12}^l (q_{11}^l q_{22}^l)^{-\frac{1}{2}}. \quad (7)$$

Equations (3) and (5) are derived in Appendix A.3 and Appendix A.4 respectively. These equations exactly recover the equations by (Poole et al., 2016) when $\gamma_l = 1$, and show that by appropriately rescaling σ_2^2 and σ_b^2 by γ_r the low-rank maps remain consistent with the full rank setting.

These two mappings (3) and (5) are functions of the network entries variances, the rank at each layer γ_l and the nonlinear activation $(\sigma_b, \sigma_W, \gamma_l, \phi)$ which determine the existence of eventual stable fixed points of q^l and q_{ab}^l as well as the dynamics they follow through the network.

The dominant quantity determining the dynamics of the network is

$$\chi_\gamma := \gamma \sigma_\alpha^2 \int_{\mathbb{R}} \left(\phi'(\sqrt{q^*} z) \right)^2 Dz \quad (8)$$

which is equal to two fundamental quantities. First, χ_γ is equal to the gradient of the correlation function (7) evaluated at correlation $c_{12}^l = 1$,

$$\chi_\gamma = \frac{\partial c_{12}^l}{\partial c_{12}^{l-1}} |_{c_{12}^{l-1}=1} \quad (9)$$

A detailed derivation of the equivalence of (8) and (9) is given in Appendix A.5. When there exists a fixed point q^* such that $\mathcal{V}(q^*) = q^*$, and $\chi < 1$, then inputs with small initial correlation converge to correlation 1 at an exponential rate; this phase is referred to as *ordered*. Alternatively, when $\chi > 1$ the fixed point $c^* = 1$ becomes unstable, meaning that an input and its arbitrarily small perturbation have correlation q_{ab}^l decreasing with layers; this is referred to as the *chaotic* phase due to all nearby points on a data manifold diverging as they progress through the network. In the ordered phase, the output function of the network is constant whereas in the chaotic phase it is non-smooth everywhere.

In both cases ($\chi > 1$ or $\chi < 1$), in (Schoenholz et al., 2016), the mappings $\mathcal{V}$ and $\frac{1}{q^*}\mathcal{C}$ are shown to converge exponentially fast to their fixed point, when they exist. Therefore, the data geometry is quickly lost as it is propagated through layers. The boundary between these phases, where $\chi = 1$, is referred to as the edge-of-chaos and determines the scaling of $(\sigma_w, \sigma_b, \gamma_l)$, as functions of nonlinear activation $\phi(\cdot)$, which ensures a sub-exponential asymptotic behaviour of these maps towards their fixed point and thus a deeper data propagation along layers which facilitates early training of the network.

Second, the quantity χ_γ in (8) is equal to the median singular value of the the matrix $D^{(l)} W^{(l)}$ where $D^{(l)}$ is the diagonal matrix with at layer l with entries $D_{ii}^{(l)} = \phi'(h_i^l)$; for details see Appendix A.9. Defining the Jacobian matrix $J \in \mathbb{R}^{N_L \times N_0}$ of the input-output map as

$$J := \frac{\partial z^L}{\partial z^0} = \prod_{l=1}^{L} D^{(l)} W^{(l)}, \quad (10)$$

we see that the average singular value of J is equal to χ_γ^L. If $\chi_\gamma = 1$ the average singular value of J is fixed at 1 throughout the network, while if χ_γ is greater than or less than 1 the average singular value deviates from 1 at an exponential rate. Further note that the growth of a perturbation from a layer to the following one is given by the average squared singular value of $D^{(l)} W^{(l)}$.

1.1. Main contributions

This manuscript extends the edge-of-chaos analysis of random feed-forward networks to the setting of low-rank matrices, following the work of (Poole et al., 2016). This work is motivated by the recent challenges faced to store in memory the constantly growing number of parameters used to train large Deep Learning models, see (Price & Tanner, 2022) and references therein.

As shown in equations (3), (6), and (8), despite the dependence between entries in the low-rank weight matrices (2), that the edge-of-chaos curve defined by $\chi_\gamma = 1$ can be retained by scaling the weight and bias variances σ_w^2 and σ_b^2 respectively by the ratio of the weight matrix rank r_l to layer width $\gamma_l := r_l / N_l$, see Figure 1 and contrast with Figure 10. That is, a simple re-scaling retains the dominant first order dynamics of a feedforward network when the weight matrices are initialized to be low-rank.

In Section 2 we show that additional first order dynamics are similarly modified through a multiplicative scaling by the rank to width factor $\gamma_l = r_l / N_l$. In particular, we demonstrate the role of γ_l on the length and correlation depth scale as well as the training gradient vectors.

However, in Section 3 we show that important second order properties of the dynamics, specifically the variance of the singular values of the input-output Jacobian given in (10), is modified by the reduced rank in a way that cannot be overcome with simple re-scaling. This result alerts practitioners

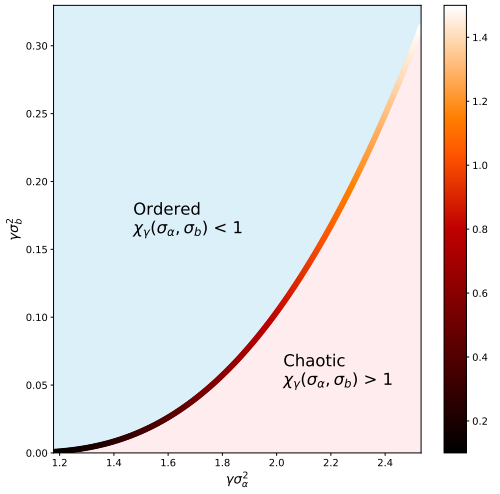

Figure 1. Edge of Chaos curve of a low-rank neural network where the rank is proportional to the width by a factor γ and nonlinear activation $\phi(x) = tanh(x)$. The plot is generated with $\gamma = \frac{1}{4}$, where the axis are rescaled by γ.

to anticipated greater variability in training low-rank weight matrices and suggests that methods to reduce the variance of the spectrum may be increasingly important in this setting, see (Murray et al., 2021).

The manuscript then concludes with numerical experiments in Section 4 which demonstrate that empirical measurements on the Jacobian are consistent with the established formula and a brief summary and future work in Section 5.

2. Network dynamics and data propagation

The parameter χ_γ further controls the length and correlation depth scaling as well as the relative magnitude of training gradients computed via back-propagration for the sum-of-squares loss function.

2.1. Depth scales a functions of χ_γ

The role of χ_γ on the achievable depth scale was pioneered by (Schoenholz et al., 2016) for full-rank feedforward networks. In this subsection we extend their results to the low-rank setting with the suitably adapted spectral mean χ_γ given in (8).

2.1.1. Length depth scale

Assuming there exists a fixed point q^* such that $\mathcal{V}(q^*) = q^*$, then the dynamics of $\mathcal{V}(q)$ can be linearized around q^* to obtain stability conditions and a rate of decay which determine how deeply data can propagate through the network before converging towards the fixed point. Following the computations done in (Schoenholz et al., 2016), setting a perturbation around the fixed point $q^* + \epsilon_l$, then around the fixed point, ϵ_l evolves as $e^{-\frac{l}{\xi_{q,\gamma}}}$, when $\gamma_l =' gamma$ is fixed along layers and we define the following quantities,

$$\xi_{q,\gamma}^{-1} := -\log\left(\chi_\gamma + \gamma\sigma_\alpha^2 \int Dz\phi''(\sqrt{q^*}z)\phi(\sqrt{q^*}z)\right).$$

Details are given in Appendix A.6. Given that $\gamma \in (0, 1]$, we can see the convergence gets faster towards the fixed point when increasing γ. Note that when $\gamma\sigma_\alpha^2 = \sigma_W^2$, we recover the results of a full-rank feedforward neural network in (Schoenholz et al., 2016).

2.1.2. Correlation depth scale

Similarly, we compute the dynamical evolution of the correlation map around its fixed point by considering a perturbation ϵ_l and we obtain that (see Appendix A.7), when all the ranks are set to be proportional to the width with the same coefficient of proportionality $\gamma_l = \gamma$ at any layer l, then the perturbation vanishes exponentially fast $\epsilon_l = \mathcal{O}(e^{-\frac{l}{\xi_{q,\gamma}}})$ where

$$\xi_{c,\gamma}^{-1} := -\log\left(\chi_\gamma\right).$$

We recover that the correlation depth scale diverges to $+\infty$ when $\chi_\gamma \to 1$, yielding again the key role of this quantity, even in the low-rank case. As $\gamma \in (0, 1]$, we can see the convergence gets faster towards the fixed point when increasing γ, which highlights the tension between low-rank and the depth to which data can propagate along layers. Note again we recover previous results from (Schoenholz et al., 2016) after appropriate scaling of the variance.

2.2. Layerwise scaling of the training gradient for the sum-of-squares loss function

As already shown in previous works ((Schoenholz et al., 2016), and (Poole et al., 2016)), there exists an direct link between the capacity for a network to propagate data through layers of a network in the forward pass and to backpropagate gradients of any given error function E. In this section, we extend the results known for full-rank feedforward neural networks with infinite width to the low-rank case, with rank $r_l = \gamma_l N_l$ evolving proportionally to the width.

The derivative of the training error follows by the chain rule,

$$\frac{\partial E}{\partial h_i^{(l)}} := \delta_i^l = \Big(\sum_{k=1}^{N_{l+1}} \delta_k^{l+1} W_{ki}^{(l+1)}\Big)\phi'(h_i^{(l)}),$$
$$\frac{\partial E}{\partial W_{ij}^{(l)}} = \delta_i^l \phi(h_j^{(l-1)}),$$
$$\frac{\partial E}{\partial \alpha_{ij}^{(l)}} = \Big(\sum_{m=1}^{r_l} \delta_m^l (C_i^l)_m\Big)\phi(h_j^{(l-1)}).$$

Consider the propagation of the gradients $\frac{\partial E}{\partial \alpha_{ij}^{(l)}}$ of the error with respect to our trainable parameters $\alpha^{(l)} := \big(\alpha_{i,j}^{(l)}\big)_{i,j}$, which are initalized $\Big(\alpha_{k,i}^{(l)}\Big)_{1\leq i\leq N_{l-1}} \in \mathbb{R} \overset{\text{iid}}{\sim} \mathcal{N}(0, \frac{\sigma_\alpha^2}{N_{l-1}})$. The length of this gradient along layers $||\nabla_{\alpha^{(l)}} E||_2^2$ is proportional to $\tilde{q}^l := \mathbb{E}((\delta_1^l)^2)$ (see Appendix A.8 for proofs). In our analysis of the variance of the training error we treat the backpropagated weights as independent from the forwarded weights, which wile not strictly true is commonly done due to its efficacy in aiding computations which reflect the observed backward dynamics of the network, see (Pennington & Bahri, 2017). Considering an input vector $x^{0,a}$, and $\tilde{q}_{aa}^l := \tilde{q}^l(x^{0,a})$,

$$\tilde{q}_{aa}^l = \tilde{q}_{aa}^{l+1} \frac{N_{l+1}}{N_l} \chi_{l+1},$$

see Appendix A.8.

With constant width along layers $\frac{N_{l+1}}{N_l} \approx 1$, then the sequence is asymptotically exponential and $\tilde{q}_{aa}^l = \tilde{q}_{aa}^L \prod_{k=l+1}^L \chi_{\gamma_k}$, or, if the proportional coefficient of the rank $\gamma_l = \gamma$ is constant along layers, $\tilde{q}_{aa}^l = \mathcal{O}(e^{\frac{l}{\xi_{\Delta,\gamma}}})$, where

$$\xi_{\Delta,\gamma}^{-1} := -\log(\chi_\gamma)$$

The same critical point is observed in the low-rank setting $\gamma < 1$ as in previous works (Schoenholz et al., 2016) given by $\chi_\gamma = 1$:

- When $\chi_\gamma > 1$, then $||\nabla_{\alpha^{(l)}} E||_2^2$ grows exponentially after $|\xi_{\nabla,\gamma}|$ layers. This is the chaotic phase with the network is being exponentially-sensitive to perturbations.
- When $\chi_\gamma < 1$, then $||\nabla_{\alpha^{(l)}} E||_2^2$ vanishes at an exponential rate after $\xi_{\nabla,\gamma}$ layers. This is the ordered phase with the network is being insensitive to perturbations.
- When $\chi_\gamma = 1$, then $||\nabla_{\alpha^{(l)}} E||_2^2$ remains of the same scale across even after an infinite number of layers which is referred to as the edge-of-chaos.

3. Dynamical isometry

Using tools from Random Matrix Theory, (Pennington et al., 2018) provides a method to compute the moments of the spectral distribution of the Jacobian, revealing secondary information beyond the mean of the spectra. We review the most essential equations to derive the variance of the Jacobian'spectrum here but we refer the reader to (Tao, 2012) for more details on the random matrix transforms.

3.1. Review of the computation of the variance of the Jacobian

In this section, we review a set of definitions of random matrix transforms that allow the calculation of the spectra of the product of matrices in terms of their individual spectra. Let X be a random matrix with spectral density ρ_X

$$\rho_X(\lambda) := \langle \frac{1}{N} \sum_{i=1}^N \delta(\lambda - \lambda_i) \rangle_X,$$

where $\langle . \rangle$ is the average with respect to the distribution of the random matrix X, and δ is the usual dirac distribution.

For a probability density ρ_X and $z \in \mathbb{C} \setminus \mathbb{R}$, the Stieltjes Transform G_X and its inverse are given by

$$G_X(z) := \int \frac{\rho_X(t)}{t - z} dt,$$
$$\rho_X(\lambda) = -\pi^{-1} \lim_{\epsilon \to 0^+} \text{Im}\big(G_X(\lambda + \epsilon i)\big).$$

The moment generating function is $M_X(z) := zG_x(z) - 1 = \sum_{k=1}^{\infty} m_k z^{-k}$ and the $\mathcal{S}_X$ Transform is defined as $\mathcal{S}_X(z) := \frac{1+z}{xM_X^{-1}(z)}$. The interest of using the $\mathcal{S}$ Transform here is that it has the following multiplicative property, which in our case is desirable as the Jacobian is a product of random matrices: if X and Y are freely independent, then $\mathcal{S}_{XY} = \mathcal{S}_X \mathcal{S}_Y$.

In (Pennington et al., 2018), the authors start with establishing $\mathcal{S}_{JJ^T} = S_{D^2}^L S_{W^TW}^L$, assuming the input vector is chosen such that $q^l \approx q^*$ so that distribution of D^2 is independent of l and we already had the weights identically distributed along layers. The strategy here to compute the spectral density of ρ_{JJ^T} (and thus the density of the singular values of the Jacobian J) starts with computing the $\mathcal{S}$ Transforms of W^TW and D^2 from their spectral density, determined by respectively, the way of sampling the weights at initialisation and the choice of the activation function in the network. Note that in this study we focus only on two possible distributions for the low-rank weights matrix - either scaled Gaussian weights or scaled orthogonal matrices, that are defined more precisely in the next sections. Once that $\mathcal{S}_{JJ^T}$ is obtained by multiplying S_{W^TW} and S_{D^2}, rather than inverting it back to find ρ_{JJ^T}, the authors show

there is a way to shortcut these steps and obtain directly the moments of ρ_{JJ^T} based on the following set of equations. Defining

$$m_k := \int \lambda^k \rho_{JJ^T}(\lambda) d\lambda$$
$$S_{W^TW}(z) := \gamma^{-1}\sigma_\alpha^{-2}\big(1 + \sum_{k=1}^{\infty} s_k z^k\big)$$
$$\mu_k = \int Dz\big(\phi'(\sqrt{q^*}z)\big)^{2k}$$

then as derived in (Pennington et al., 2018), the first two moments of the spectrum of the Jacobian are

$$m_1 = (\gamma\sigma_\alpha^2\mu_1)^L$$
$$m_2 = (\gamma\sigma_\alpha^2\mu_1)^{2L} L\left(\frac{\mu_2}{\mu_1^2} + \frac{1}{L} - 1 - s_1\right).$$

The first moment m_1 recovers the previous statement that the average squared singular value is equal to $m_1 = \chi_\gamma^L$ and the edge-of-chaos given by $\chi_\gamma = \gamma\sigma_\alpha^2\mu_1 = 1$ is consistent with previous results as the gradient either vanishes or grows exponentially along with the median of the Jacobian's spectra. Moreover, the variance of the spectrum of JJ^T about its mean $\chi_\gamma = 1$ can now be computed

$$\sigma^2_{JJ^T} := m_2 - m_1^2 = L\left(\frac{\mu_2}{\mu_1^2} - 1 - s_1\right). \quad (11)$$

The variance $\sigma^2_{JJ^T}$ grows linearly with depth as in the full-rank setting, recovering the full-rank result when $\gamma = 1$. As in the edge-of-chaos axes scaling in Figure 1, $\gamma\sigma_\alpha^2$ plays the same role as σ_W^2. Note that $\frac{\mu_2}{\mu_1^2} \geq 1$ and consequently $\sigma^2_{JJ^T}$ as given in (11) is only independent of depth L if $s_1 = 0$ which is only achieved here in the case of full-rank, i.e. $\gamma = 1$ orthogonal matrices.

3.2. Low-Rank Orthogonal weights

Consider a weight matrix whose r first columns are orthonormal columns sampled from a normal distribution, and the rest is 0, such that $W^TW = \left(\begin{array}{c|c} \sigma_\alpha^2\mathbb{I}_r & 0 \\ \hline 0 & \mathbb{O}_{N-r} \end{array}\right)$. Therefore the spectral distribution of $\sigma_\alpha^{-2}W^TW$ is trivially given by

$$\rho_{\sigma_\alpha^{-2}W^TW}(z) = \gamma\delta(z-1) + (1-\gamma)\delta(z),$$

from which the $\mathcal{S}$ Transform is computed, see Appendix A.11, to obtain $s_1 = -(\gamma^{-1} - 1)$. When $\gamma = 1$, the known result in the full-rank orthogonal case is retrieved.

3.3. Low-Rank Gaussian weights

With weights at any layer l given by 2, the matrix can be rewritten as the product $W^l = C^lA^l$, where $C^l \in \mathbb{R}^{N_l \times r_l}$ with $C^l_{ij} = (C^l_j)_i$, and $A^l \in \mathbb{R}^{r_l \times N_{l-1}}$ with $A^l_{ij} \overset{iid}{\sim} \mathcal{N}(0, \frac{\sigma_\alpha^2}{N_{l-1}})$. As ${C^l}^TC^l = \mathbb{I}_{r_l}$ by construction, then ${W^{(l)}}^TW^{(l)} = {A^l}^T{C^l}^TC^lA^l = {A^l}^TA^l$ which is a Wishart matrix, whose spectral density is known and given by the Marčenko Pastur distribution (Marčenko & Pastur, 1967) where some mass is added at 0 since the matrix A^l is not full-rank and contains some 0 eigenvalues. Recall that $r_l = \gamma N_l$.

Table 1. Transforms of weights. LR stands for Low-Rank.

RANDOM MATRIX W	$S_{W^TW}(z)$	s_1
LR SCALED ORTHOGONAL	$\gamma^{-1}\sigma_\alpha^{-2}\frac{1+z}{1+\gamma^{-1}z}$	$1 - \frac{1}{\gamma}$
LR SCALED GAUSSIAN	$\gamma^{-1}\sigma_\alpha^{-2}\frac{1+z}{1+z(1+\gamma^{-1})+\gamma^{-1}z^2}$	$-\frac{1}{\gamma}$

$$\rho_{A^TA}(\lambda) = (1-\gamma)_+\delta(\lambda) + \gamma\frac{\sqrt{(\lambda^+ - \lambda)(\lambda - \lambda^-)}}{2\pi\lambda\sigma_\alpha^2}\mathbb{1}_{[\lambda^-,\lambda^+]}(\lambda),$$

where $x_+ = \max(0, x), \lambda^- = (1 - \frac{1}{\gamma})^2$ and $\lambda^+ = (1+\frac{1}{\gamma})^2$. The $\mathcal{S}$ Transform S_{W^TW} can be computed (see Appendix A.10) and expanded around 0, which gives $s_1 = -\frac{1}{\gamma}$. Note that when $\gamma = 1$, one recovers the result given in (Pennington et al., 2018).

The $\mathcal{S}$ Transforms and first moments in both orthogonal and Gaussian cases are summarized in Table 1.

4. Numerical experiments

In this section, we give empirical evidence in agreement with the theoretical results established above. Its interest is two-fold:

- The variance of the spectrum of the Jacobian does indeed still grow with depth even in the low-rank setting as emphasized in Figure 2. Moreover, at a fixed depth, the rank to width ratio plays a key role in how the spectrum of the Jacobian spreads out around its mean value, which is 1 when the network is initialised on the edge-of-chaos.
- Figure 3 shows that the advantage that Scaled Orthogonal Weights have over Scaled Gaussian Weights in Feedforward networks presented in (Pennington et al., 2018) is lost for low-rank matrices. Indeed, from (11), one can see that in both situations, it is not possible to adjust either the activation function nor q^* through a careful choice of variances for the weights and the biases, unless $\gamma = 1$ and $W^{(l)}$ is a scaled orthonormal matrix.

In Figure 2 and Figure 3, the variance of the spectrum of the Jacobian is computed in the low-rank Gaussian and Or-

thogonal cases when the activation function is chosen to be the identity. Although such a choice of activation function completely destroys the network's expressivity power, it is a simple example of situations in the full-rank case where Gaussian distributed weight matrices lead to ill conditioned Jacobians as depth increases. This still holds in the low-rank setting as shown in the plot since the variance $\sigma^2_{JJ^T} > 0$. Simulations are performed on a 1000- layer wide feedforward network, initialised and fed with a random input, whose length is set to be equal to q^* so that the network would already be at its equilibrium state without passing by a transient phase.

The source code can be found at `shorturl.at/syLP9`.

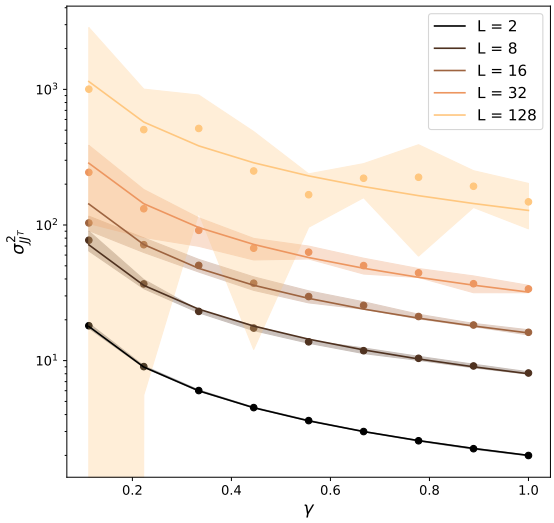

Figure 2. Evolution of the variance of the spectrum of JJ^T with respect to γ where γ is the proportionality coefficient giving the rank of the weights matrices at layer l, whose width is N_l, $r_l = \gamma N_l$. Points are obtained empirically and averaged over 5 simulations when the lines are derived from the theory, see (11). Confidence intervals of 1 standard deviation around each mean point are shown. The weights are chosen to be low-rank Scaled Gaussian and the activation function is linear $\phi : x \mapsto x$. The same seed is used to initialise the weight matrices for each simulation and q^* is set to 0.5. The $y-$axis is shown in log scale.

5. Summary and further work

Herein the edge-of-chaos theory of (Poole et al., 2016) and (Schoenholz et al., 2016) has been extended from the setting of full-rank weight matrices to the low-rank setting. Suitable scaling by the ratio of the rank to width factor $\gamma_l := r_l/N_l$ recovers the phenomenon driven by the mean

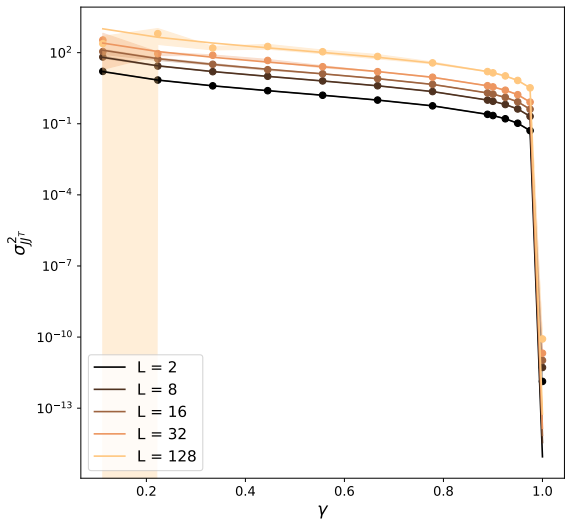

Figure 3. Evolution of the variance of the spectrum of JJ^T with respect to γ where γ is the proportionality coefficient giving the rank of the weights matrices at layer l, whose width is N_l, $r_l = \gamma N_l$. Points are obtained empirically and averaged over 3 simulations when the lines are derived from the theory, see (11). Confidence intervals of 1 standard deviation around each mean point are shown. The weights are chosen to be low-rank Scaled Orthogonal and the activation function is linear $\phi : x \mapsto x$. The same seed is used to initialise the weight matrices for each simulation and q^* is set to 0.5. The $y-$axis is shown in log scale.

of the Jacobian's spectra which defines the edge-of-chaos. Moreover, the variance of the Jacobian's spectra is shown to be strictly increasing with decreasing γ_l which suggests greater variability in the initial training of low-rank feedforward networks.

The edge-of-chaos initialisation scheme has been successfully generalised to a large set of different settings, including changes of architectures as CNNs (Xiao et al., 2018), LSTMs and GRUs (Gilboa et al., 2019), RNNs (Chen et al., 2018), ResNets (Yang & Schoenholz, 2017) and to extra features like dropout (Schoenholz et al., 2016), (Huang et al., 2019) or batch normalisation (Yang et al., 2019) and pruning (Hayou et al., 2020). It has been improved with changes of activation functions (Hayou et al., 2019), (Murray et al., 2021) to enable the data to propagate even deeper through the network. As a future work, each of these settings could be extended to the setting of low-rank weight matrices.

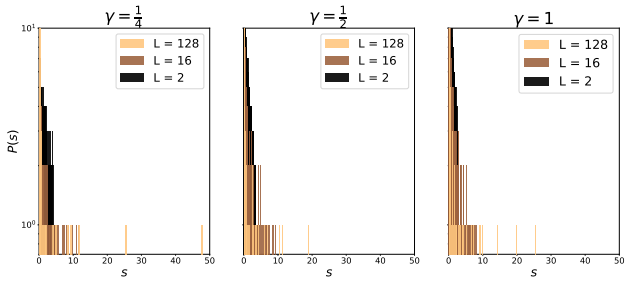

Figure 4. Singular values of the Jacobian J with respect to the depth of the network, whose weight matrices are low-rank Scaled Gaussian. The rank to width ratio γ increases on each plot from left to right when the width is kept constant to 1000. The activation function is erf. The same seed is used to initialise the weight matrices for each simulation and q^* is set to 0.5. The $y-$axis is shown in log scale.

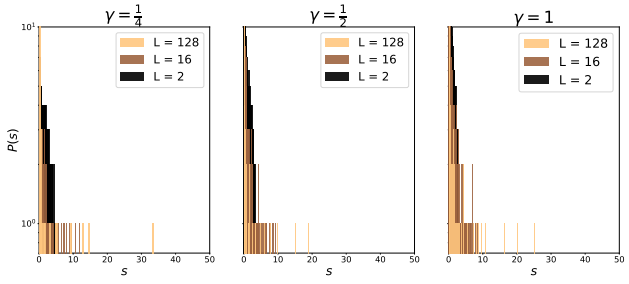

Figure 5. Singular values of the Jacobian J with respect to the depth of the network, whose weight matrices are low-rank Scaled Gaussian. The rank to width ratio γ increases on each plot from left to right when the width is kept constant to 1000. The activation function is tanh. The same seed is used to initialise the weight matrices for each simulation and q^* is set to 0.5. The $y-$axis is shown in log scale.

Acknowledgments

TNS is financially supported by the Engineering and Physical Sciences Research Council (EPSRC). JT is supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA) and thanks UCLA Department of Mathematics for kindly hosting him during the completion of this manuscript.

References

Chen, M., Pennington, J., and Schoenholz, S. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 873–882. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/chen18i.html.

Gilboa, D., Chang, B., Chen, M., Yang, G., Schoenholz, S. S., Chi, E. H., and Pennington, J. Dynamical isometry and a mean field theory of lstms and grus, 2019. URL https://arxiv.org/abs/1901.08987.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

Hayou, S., Doucet, A., and Rousseau, J. On the impact of the activation function on deep neural networks training, 2019. URL https://arxiv.org/abs/1902.06853.

Hayou, S., Ton, J.-F., Doucet, A., and Teh, Y. W. Robust pruning at initialization, 2020. URL https://arxiv.org/abs/2002.08797.

Huang, W., Da Xu, R. Y., Du, W., Zeng, Y., and Zhao, Y. Mean field theory for deep dropout networks: digging up gradient backpropagation deeply. 2019. doi: 10.48550/ARXIV.1912.09132. URL https://arxiv.org/abs/1912.09132.

Marčenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, apr 1967. doi: 10.1070/SM1967v001n04ABEH001994. URL https://dx.doi.org/10.1070/SM1967v001n04ABEH001994.

Murray, M., Abrol, V., and Tanner, J. Activation function design for deep networks: linearity and effective initialisation, 2021. URL https://arxiv.org/abs/2105.07741.

Pennington, J. and Bahri, Y. Geometry of neural network loss surfaces via random matrix theory. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2798–2806. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/pennington17a.html.

Pennington, J., Schoenholz, S. S., and Ganguli, S. The emergence of spectral universality in deep networks, 2018. URL https://arxiv.org/abs/1802.09979.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos, 2016. URL https://arxiv.org/abs/1606.05340.

Price, I. and Tanner, J. Improved projection learning for lower dimensional feature maps, 2022. URL https://arxiv.org/abs/2210.15170.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation, 2016. URL https://arxiv.org/abs/1611.01232.

Tao, T. *Topics in Random Matrix Theory*. Graduate studies in mathematics. American Mathematical Soc., 2012. ISBN 9780821885079. URL https://books.google.co.uk/books?id=Hjq_JHLNPT0C.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/xiao18a.html.

Yang, G. and Schoenholz, S. S. Mean field residual networks: On the edge of chaos, 2017. URL https://arxiv.org/abs/1712.08969.

Yang, G., Pennington, J., Rao, V., Sohl-Dickstein, J., and Schoenholz, S. S. A mean field theory of batch normalization, 2019. URL https://arxiv.org/abs/1902.08129.

A. Supplementary Material

A.1. Preliminary lemma

The following lemma is used later in the proofs contained in Appendix A.2.

Lemma A.1. *Let $\gamma \in \mathbb{R}^*$. If $C_1, \dots, C_{\gamma n} \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{n})$, then $C_1^2 + \cdots + C_{\gamma n}^2 \to \gamma$ in probability when $n \to \infty$.*

Proof. If X is a random variable, let us denote by F_X its cumulative distribution function. Let $x \in \mathbb{R}$.

$$\begin{aligned} F_{C_1^2+\cdots+C_{\gamma n}^2}(x) &= \mathbb{P}(C_1^2 + \cdots + C_{\gamma n}^2 \leq x) \\ &= \mathbb{P}(\sqrt{n}\frac{C_1^2 + \cdots + C_{\gamma n}^2 - \gamma}{\sqrt{2\gamma}} \leq \sqrt{n}\frac{x-\gamma}{\sqrt{2\gamma}}) \\ &= \mathbb{P}(\frac{C_1^2 + \cdots + C_{\gamma n}^2 - (n\gamma)\frac{1}{n}}{\frac{\sqrt{2}}{n}\sqrt{\gamma n}} \leq \sqrt{n}\frac{x-\gamma}{\sqrt{2\gamma}}). \end{aligned}$$

where $\mathbb{E}(C_1^2) = \frac{1}{n}$ and $\mathbb{V}(C_1^2) = \frac{\sqrt{2}}{n}$. Thus the Central Limit theorem holds and gives that the left hand side converges in distribution to a standard normal Gaussian when $n \to \infty$. The right hand side tends to $\text{sign}(x-\gamma)\infty$. Thus

$$F_{C_1^2+\cdots+C_{\gamma n}^2}(x) \to \mathbb{1}_{x \geq \gamma}(x).$$

As we have a convergence in distribution towards a constant, the convergence in probability follows.

A.2. Distribution of hidden layers

Let us now consider at layer l the weight matrix, $W^{(l)} \in \mathbb{R}^{N_l \times N_{l-1}}$, being of rank r_l:

$$W^{(l)} = \left[\begin{array}{c|c|c|c} \alpha_{1,1}^{(l)} C_1^{(l)} + \cdots + \alpha_{r_l,1}^{(l)} C_{r_l}^{(l)} & \alpha_{1,2}^{(l)} C_1^{(l)} + \cdots + \alpha_{r_l,2}^{(l)} C_{r_l}^{(l)} & \dots & \alpha_{1,N_{l-1}}^{(l)} C_1^{(l)} + \alpha_{r_l,N_{l-1}}^{(l)} C_{r_l}^{(l)} \end{array} \right],$$

where, at any layer l, for any $k \in [\![1, r_l]\!]$, the scalars $\left(\alpha_{k,i}^{(l)}\right)_{1 \leq i \leq N_{l-1}} \in \mathbb{R}$ are identically and independently drawn from a Gaussian distribution $\mathcal{N}(0, \frac{\sigma_\alpha^2}{N_{l-1}})$ and the columns $C_1^{(l)}, \dots, C_{r_l}^{(l)}$ are drawn jointly as the matrix $C^{(l)} := [C_1^{(l)}, \dots, C_{r_l}^{(l)}] \in \mathbb{R}^{N_l \times r_l}$ from the Grassmannian of rank r matrices with orthonormal columns having zero mean and variance $1/N_l$. Similarly, in this section, we consider a random bias at layer l along the directions given by $C_1^{(l)}, \dots, C_{r_l}^{(l)}$, i.e. a bias of the form $b^{(l)}(C_1^{(l)} + \cdots + C_{r_l}^{(l)})$, where $b^{(l)} \in \mathbb{R} \sim \mathcal{N}(0, \sigma_b^2)$.

Thus, at layer l, whose width is N_l, the preactivation vector $h^{(l)} \in \mathbb{R}^{N_l}$ is given by

$$\begin{aligned} h^{(l)} &= W^{(l)}\phi(h^{(l-1)}) + b^{(l)}(C_1^{(l)} + \cdots + C_{r_l}^{(l)}) \\ &= C_1^{(l)} \underbrace{\left(\sum_{j=1}^{N_{l-1}} \alpha_{1,j}^{(l)} \phi(h_j^{(l-1)}) + b^{(l)} \right)}_{:= z_1^{(l)}} + C_2^{(l)} \underbrace{\left(\sum_{j=1}^{N_{l-1}} \alpha_{2,j}^{(l)} \phi(h_j^{(l-1)}) + b^{(l)} \right)}_{:= z_2^{(l)}} + \cdots + C_{r_l}^{l} \underbrace{\left(\sum_{j=1}^{N_{l-1}} \alpha_{r_l,j}^{(l)} \phi(h_j^{(l-1)}) + b^{(l)} \right)}_{:= z_{r_l}^{(l)}} \\ &= \sum_{k=1}^{r_l} \underbrace{z_k^{(l)}}_{\in \mathbb{R}} C_k^{(l)}, \end{aligned}$$

where the scalars $z_k^{(l)}$ follow a Gaussian distribution, given the preactivation vector at the previous layer $z_k^{(l)}|_{h^{(l-1)}} \sim \mathcal{N}\left(0, \frac{\sigma_\alpha^2}{N_{l-1}} \sum_{j=1}^{N_{l-1}} \phi(h_j^{(l-1)})^2 + \sigma_b^2\right)$, which is given using the Central Limit Theorem in the large width N_{l-1} regime.

A.3. Length recursion formula

This being said, one can compute the length of the (random) preactivation vector, at layer l.

$$\begin{aligned} q^l &:= \frac{1}{N_l}||h^{(l)}||_2^2 = \frac{1}{N_l}\sum_{j=1}^{N_l}(h_j^{(l)})^2 \\ &= \frac{1}{N_l}\left((z_1^{(l)})^2||C_1^{(l)}||_2^2 + \cdots + (z_{r_l}^{(l)})^2||C_{r_l}^{(l)}||_2^2\right) \quad \text{using Pythagore's theorem} \\ &= \frac{1}{N_l}\left((z_1^{(l)})^2 + \ldots (z_{r_l}^{(l)})^2\right) \quad \text{since for any } k,\ ||C_k^{(l)}||_2 = 1 \\ &= \frac{1}{N_l}\sum_{k=1}^{r_l}(z_k^{(l)})^2 \end{aligned}$$

Therefore, given $h^{(l-1)}$, q^l is a sum of r_l χ^2 distributions.

Let us now consider at any layer, a rank that is proportional to the width: $r_l = \gamma_l N_l$, where $\gamma_l \in (0, 1]$. Thus,

$$q^l = \gamma_l \frac{1}{\gamma_l N_l}\sum_{k=1}^{\gamma_l N_l}(z_k^{(l)})^2. \tag{12}$$

Then all $z_k^{(l)}$ are independent and identically distributed Gaussian variables. The Law of Large Numbers holds and $q^l \to \gamma_l \mathbb{V}\big(z_1^{(l)}\big)$ when $N_l \to \infty$, with

$$\mathbb{V}\big(z_1^{(l)}\big) = \frac{\sigma_\alpha^2}{N_{l-1}}\sum_{j=1}^{N_{l-1}}\phi(h_j^{(l-1)})^2 + \sigma_b^2.$$

On the other hand,

$$h_j^{(l-1)} = \sum_{k=1}^{r_{N_{l-1}}} z_k^{(l-1)}\big(C_k^{(l-1)}\big)_j = \big(z^{(l-1)}\big)^T\big(C_\cdot^{(l-1)}\big)_j$$

denoting

$$\big(C_\cdot^{(l)}\big)_j := \begin{pmatrix} \big(C_1^{(l)}\big)_j \\ \vdots \\ \big(C_{r_l}^{(l)}\big)_j \end{pmatrix}$$

We know the distribution of $z^{(l-1)} \sim \mathcal{N}\left(\mathbb{O}_{\mathbb{R}^{r_{l-1}}}, \frac{q^{l-1}}{\gamma_{l-1}}\mathbb{I}_{r_{l-1}}\right)$ and so, given $C^{(l-1)}$, $h_j^{(l-1)} \sim \mathcal{N}\left(0, \big(C_\cdot^{(l-1)}\big)_j^T \frac{q^{l-1}}{\gamma_{l-1}}\mathbb{I}_{r_{l-1}}\big(C_\cdot^{(l-1)}\big)_j\right)$.

In the asymptotic limit approximation, when $N_{l-1} \to \infty$, then $\big(C_\cdot^{(l-1)}\big)_j^T\big(C_\cdot^{(l-1)}\big)_j \to \gamma_{l-1}$ in probability as shown in Lemma A.1. Therefore, $h_j^{(l-1)} \sim \mathcal{N}\big(0, q^{l-1}\big)$.

In the limit when $N_{l-1} \to \infty$, the Law of Large Numbers enables us to conclude,

$$\begin{aligned} q^l &= \gamma_l\big(\frac{\sigma_\alpha^2}{N_{l-1}}\sum_{j=1}^{N_{l-1}}\phi(h_j^{(l-1)})^2 + \sigma_b^2\big) \\ &\to \gamma_l\left(\sigma_\alpha^2\int_{\mathbb{R}}\phi^2(\sqrt{q^{l-1}}z)Dz + \sigma_b^2\right) := \mathcal{V}(q^{(l-1)}|\sigma_\alpha^2, \sigma_b^2, \gamma_l). \end{aligned}$$

Note that when $\forall l, \gamma_l = 1$ and $\sigma_\alpha = \sigma_W$, one recovers the formula from (Poole et al., 2016). Alternatively, by rescaling the variances by γ_l at every layer, e.g. $\sigma_\alpha^2 \to \frac{\sigma_W^2}{\gamma_l}$ and $\sigma_b^2 \to \frac{\sigma_b^2}{\gamma_l}$, we recover the formulae of (Poole et al., 2016).

A.4. Correlation recursion formula

Let us denote by $x^{0,1}$ and $x^{0,2}$ two input data. Then one can define the following 2 by 2 matrix

$$(q_{ab}^l)_{1\leq a,b\leq 2} = \frac{1}{N_l}\sum_{i=1}^{N_l}\begin{pmatrix} h_i^{(l)}(x^{0,1})^2 & h_i^{(l)}(x^{0,1})h_i^{(l)}(x^{0,2}) \\ h_i^{(l)}(x^{0,1})h_i^{(l)}(x^{0,2}) & h_i^{(l)}(x^{0,2})^2 \end{pmatrix}$$

where, for $i \in [\![1, N_l]\!]$, $h_i^{(l)}(x^{0,a}) = \sum_{r=1}^{r_l} z_k^{(l)}(x^{0,a})\big(C_k^{(l)}\big)_i$.

So,

$$\begin{aligned}\frac{1}{N_l}\sum_{i=1}^{N_l} h_i^{(l)}(x^{0,1})h_i^{(l)}(x^{0,2}) &= \frac{1}{N_l}\sum_{i=1}^{N_l}\left(\sum_{k=1}^{r_l} z_k^{(l)}(x^{0,1})\big(C_k^{(l)}\big)_i\right)\left(\sum_{p=1}^{r_l} z_p^{(l)}(x^{0,2})\big(C_p^{(l)}\big)_i\right) \\ &= \frac{1}{N_l}\sum_{k=1}^{r_l} z_k^{(l)}(x^{0,1})z_k^{(l)}(x^{0,2})\Bigg(\underbrace{\sum_{i=1}^{N_l}\big(C_k^{(l)}\big)_i^2}_{||C_k^{(l)}||_2^2=1}\Bigg) + \frac{1}{N_l}\sum_{\substack{1\leq k,p\leq r_l \\ k\neq p}} z_k^{(l)}(x^{0,1})z_p^{(l)}(x^{0,2})\Bigg(\underbrace{\sum_{i=1}^{N_l}\big(C_k^{(l)}\big)_i\big(C_p^{(l)}\big)_i}_{\langle C_k^{(l)}, C_p^{(l)}\rangle = 0}\Bigg) \\ &= \frac{1}{N_l}\sum_{k=1}^{r_l} z_k^{(l)}(x^{0,1})z_k^{(l)}(x^{0,2})\end{aligned}$$

Therefore, when the rank is proportional to the width and the width $N_l \to \infty$ as previously, the Law of Large Numbers gives

$$= \gamma_l \frac{1}{\gamma_l N_l}\sum_{k=1}^{r_l} z_k^{(l)}(x^{0,1})z_k^{(l)}(x^{0,2}) \to \gamma_l Cov\left(z_1^{(l)}(x^{0,2}), z_1^{(l)}(x^{0,2})\right)$$

On the other hand,

$$\begin{aligned}Cov\left(z_1^{(l)}(x^{0,1}), z_1^{(l)}(x^{0,2})\right) &= \sum_{1\leq i,j\leq N_{l-1}} \phi(h_i^{(l-1)}(x^{0,1}))\phi(h_j^{(l-1)}(x^{0,2}))\underbrace{Cov(\alpha_{1,i}^{(l)}, \alpha_{1,j}^{(l)})}_{\frac{\sigma_\alpha}{N_{l-1}}\delta_{i,j}} \\ &\quad + \sum_{1\leq i\leq N_{l-1}} \phi(h_i^{(l-1)}(x^{0,1}))\underbrace{Cov(\alpha_{1,i}^{(l)}, b_1^{(l)})}_{=0} \\ &\quad + \sum_{1\leq i\leq N_{l-1}} \phi(h_i^{(l-1)}(x^{0,2}))\underbrace{Cov(\alpha_{1,i}^{(l)}, b_1^{(l)})}_{=0} + \underbrace{Cov(b_1^{(l)}, b_1^{(l)})}_{\sigma_b^2}\end{aligned}$$

Thus, when $N_{l-1} \to \infty$, the Law of Large Numbers enables us to conclude

$$q_{12}^l = \gamma_l\left(\sigma_\alpha^2\int_{\mathbb{R}^2}\phi(\sqrt{q_{11}^{l-1}}z_1)\phi(\sqrt{q_{22}^{l-1}}(c_{12}^{l-1}z_1 + \sqrt{1-(c_{12}^{l-1})^2}z_2)Dz_1Dz_2 + \sigma_b^2\right)$$

with $c_{12}^l = q_{12}^l(q_{11}^l q_{22}^l)^{-\frac{1}{2}}$.

Note that if we consider the short convergence of the variance, compared to the covariance one, as observed in (Poole et al., 2016), then we can assume $q_{11}^l \approx q_{22}^l \approx q^*$. We can then rescale the previous covariance map to get the correlation map as follows:

$$c_{12}^l = \frac{\gamma_l}{q^*}\left(\sigma_\alpha^2 \int_{\mathbb{R}^2} \phi(\sqrt{q^*}z_1)\phi(\sqrt{q^*}(c_{12}^{l-1}z_1 + \sqrt{1-(c_{12}^{l-1})^2}z_2)Dz_1Dz_2 + \sigma_b^2\right)$$

We observe that 1 is always a fixed point as $1 = \frac{\gamma_l}{q^*}\left(\sigma_\alpha^2 \int_{\mathbb{R}} Dz\phi^2(\sqrt{q^*}z) + \sigma_b^2\right) = \frac{1}{q^*}\mathcal{V}(q^*|\sigma_\alpha^2, \sigma_b^2, \gamma) = \frac{q^*}{q^*}$.

For completeness we replicate correlation maps and dynamics of correlations through layers using the same parameters as in (Poole et al., 2016) suitably modified for the low-rank setting, see Figure 6 - 9. The dynamics of the low-rank networks are observed to be consistent, under appropriate scaling by γ_l,with those of the full-rank networks in (Poole et al., 2016).

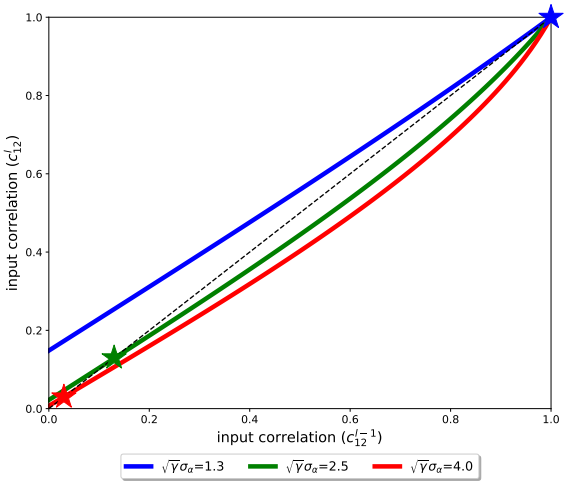

Figure 6. Correlation map of a low-rank neural network where the rank is proportional to the width by a factor γ. The nonlinear activation function is $\phi = \tanh$. The map is given by (6) and the integral is computed numerically. The dashed line is the identity function and stars represent fixed points of the correlation map.

A.5. Derivative of the correlation map

In this section, we extend the computations of the derivative of the correlation map.

$$\frac{\partial c_{12}^l}{\partial c_{12}^{l-1}} = \frac{\gamma_l}{q^*}\left(\sigma_\alpha^2 \int_{\mathbb{R}^2} \phi(u_1)\phi'(u_2)\big(\sqrt{q^*}z_1 - \sqrt{q^*}\frac{c_{12}^{l-1}}{\sqrt{1-(c_{12}^{l-1})^2}}z_2\big)Dz_1Dz_2\right)$$

Using, for F smooth enough, the identity $\int_{\mathbb{R}} F(z)zDz = \int_{\mathbb{R}} F'(z)Dz$ to the functions $G : z_1 \mapsto \phi(\sqrt{q^*}z_1)\int_{z_2}\phi'\big(\sqrt{q^*}(c_{12}^{l-1}z_1 + \sqrt{1-(c_{12}^{l-1})^2}z_2)\big)Dz_2$ and $H : z_2 \mapsto \int_{z_1}\phi(\sqrt{q^*}z_1)\phi'\big(\sqrt{q^*}(c_{12}^{l-1}z_1 + \sqrt{1-(c_{12}^{l-1})^2}z_2)\big)Dz_1$, we obtain

$$\frac{\partial c_{12}^l}{\partial c_{12}^{l-1}} = \gamma_l\sigma_\alpha^2 \int_{\mathbb{R}^2} \phi'(\sqrt{q^*}z_1)\phi'\big(\sqrt{q^*}(c_{12}^{l-1}z_1 + \sqrt{1-(c_{12}^{l-1})^2}z_2)\big)Dz_1Dz_2,$$

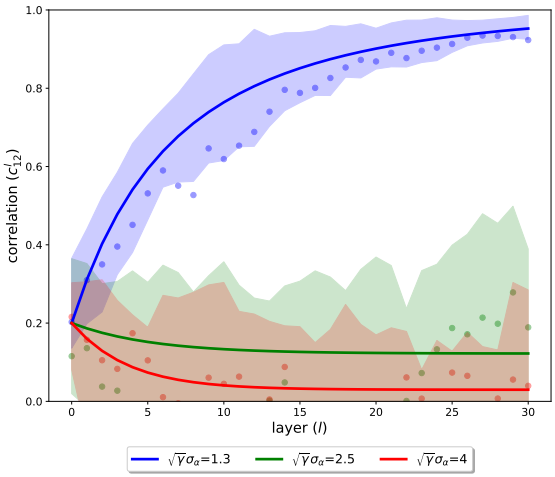

Figure 7. Dynamic of the correlation through layers, starting from two input vectors with correlation $c_{12}^0 = 0.2$. Points are obtained empirically and averaged over 5 simulations when the lines are derived from the theory, see (6). Confidence intervals of 2 standard deviations around each point are shown. For each point on the plot, we generated a pair of points with correlation c_{12}^0, passed them through a Wide low-rank network initialised and computed the correlation between both preactivation vectors across layers. The network has constant width $N = 1000$. $\sigma_b = 0.3, \phi = \tanh, \gamma = \frac{1}{4}$.

which, evaluated at its fixed point $c_{12}^{l-1} = 1$ gives

$$\chi_\gamma := \frac{\partial c_{12}^l}{\partial c_{12}^{l-1}}|_{c_{12}^{l-1}=1} = \gamma_l \sigma_\alpha^2 \int_{\mathbb{R}} \left(\phi'(\sqrt{q^*}z)\right)^2 Dz.$$

The edge-of-chaos level set defined by $\chi_\gamma = 1$ for nonlinear activation $\phi(x) = tanh(x)$ is shown in Figure 1 with axes $\gamma\sigma_w^2$ and $\gamma\sigma_b^2$. Figure 10 show the analogous edge-of-chaos plot for a full-rank matrix as given, which is identical to that of 1 but with axes σ_w^2 and σ_b^2.

A.6. Length depth scale

Recall that $q^l = \gamma_l\left(\sigma_\alpha^2 \int Dz\phi^2(\sqrt{q^{l-1}}z) + \sigma_b^2\right)$ and q^* is a fixed point assumed to exist when $\gamma_l = \gamma$ at any layer l. We then define the perturbation $\epsilon_l \to 0$ such that $q^l = q^* + \epsilon_l$. We can then expand the relation around its fixed point, as done

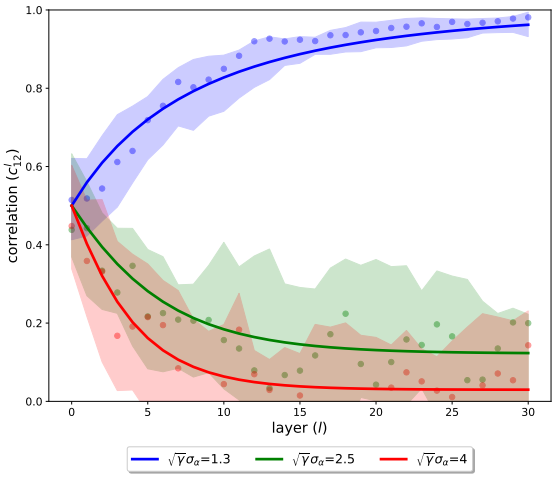

Figure 8. Dynamic of the correlation through layers, starting from two input vectors with correlation $c_{12}^0 = 0.5$.. Points are obtained empirically and averaged over 5 simulations when the lines are derived from the theory, see (6). Confidence intervals of 2 standard deviations around each point are shown. For each point on the plot, we generated a pair of points with correlation c_{12}^0, passed them through a Wide low-rank network initialised and computed the correlation between both preactivation vectors across layers. The network has constant width $N = 1000$. $\sigma_b = 0.3, \phi = \tanh, \gamma = \frac{1}{4}$.

in the case of feedforward neural network in (Schoenholz et al., 2016).

$$
\begin{aligned}
q^{l+1} &= q^* + \epsilon_{l+1} = \gamma\big(\sigma_\alpha^2 \int Dz\phi^2((\epsilon_l + q^*)^{\frac{1}{2}}z) + \sigma_b^2\big) \\
&= \gamma\left(\sigma_\alpha^2 \int Dz\phi\big(\sqrt{q^*}z + \frac{1}{2}\frac{\epsilon_l}{\sqrt{q^*}}z + \mathcal{O}(\epsilon_l^2)\big)^2 + \sigma_b^2\right) \text{ expanding the square root} \\
&= \gamma\left(\sigma_\alpha^2 \int Dz\big((\phi(\sqrt{q^*}z) + \phi'(\sqrt{q^*}z)\frac{\epsilon_l}{2\sqrt{q^*}}z + \mathcal{O}(\epsilon_l^2)\big)^2 + \sigma_b^2\right) \text{ expanding } \phi \text{ around } \sqrt{q^*}z \\
&= \gamma\left(\sigma_\alpha^2 \int Dz\phi^2(\sqrt{q^*}z) + \int Dz\phi'(\sqrt{q^*}z)\phi(\sqrt{q^*}z)\frac{\epsilon_l}{\sqrt{q^*}}z + \mathcal{O}(\epsilon_l^2) + \sigma_b^2\right) \\
&= q^* + \gamma \int Dz\phi'(\sqrt{q^*}z)\phi(\sqrt{q^*}z)\frac{\epsilon_l}{\sqrt{q^*}}z + \mathcal{O}(\epsilon_l^2) \text{ by definition of } q^* \\
&= q^* + \epsilon_l\gamma\sigma_\alpha^2\left(\int Dz\big(\phi'(\sqrt{q^*}z)\big)^2 + \int Dz\phi''(\sqrt{q^*}z)\phi(\sqrt{q^*}z)\right) + \mathcal{O}(\epsilon_l^2) \text{ using } \int DzF(z)z = \int DzF'(z)
\end{aligned}
$$

Note that in the proof above we assumed the activation function ϕ to be smooth enough to use its Taylor expansion around the point $\sqrt{q^*}z$ for any z.

Therefore by identification, $\epsilon_{l+1} = \epsilon_l\big(\chi_\gamma + \gamma\sigma_\alpha^2 \int Dz\phi''(\sqrt{q^*}z)\phi(\sqrt{q^*}z)\big) + \mathcal{O}(\epsilon_l^2)$, which concludes the proof.

A.7. Correlation depth scale

Let consider the computation is done at a layer l deep enough so that the variance map has already converged towards its fixed point $q_{11}^l = q_{22}^l = q^*$. We generate a perturbation $\epsilon_l \underset{l\to\infty}{\to} 0$ around the fixed point c^* and analyse how

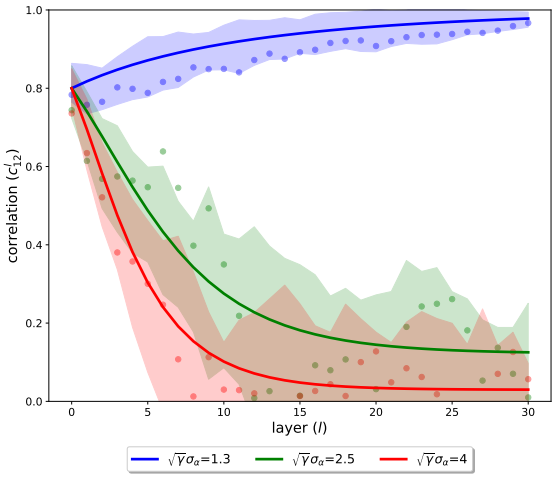

Figure 9. Dynamic of the correlation through layers, starting from two input vectors with correlation $c_{12}^0 = 0.8$.. Points are obtained empirically and averaged over 5 simulations when the lines are derived from the theory, see (6). Confidence intervals of 2 standard deviations around each point are shown. For each point on the plot, we generated a pair of points with correlation c_{12}^0, passed them through a Wide low-rank network initialised and computed the correlation between both preactivation vectors across layers. The network has constant width $N = 1000$. $\sigma_b = 0.3, \phi = \tanh, \gamma = \frac{1}{4}$.

it propagates: $c_{12}^l = c^* + \epsilon_l$. Additionally, we introduce $u_1^l = \sqrt{q^*}z = u_1^*, u_2^* = \sqrt{q^*}(c^* z_1 + \sqrt{1-(c^*)^2}z_2)$ and $u_2^l = \sqrt{q^*}(c_{12}^l z_1 + \sqrt{1-(c_{12}^l)^2}z_2)$. Following the same strategy as in the previous section (using expansions), it is shown in (Schoenholz et al., 2016) that

$$u_2^l = \begin{cases} u_2^* + \sqrt{q^*}\epsilon_l\big(z_1 - \frac{c^*}{\sqrt{1-c^{*2}}}z_2\big) + \mathcal{O}(\epsilon_l^2) & \text{when } c^* < 1, \\ u_2^* + \sqrt{2q^*\epsilon_l}z_2 - \epsilon_l\sqrt{q^*}z_1 + \mathcal{O}(\epsilon_l^{\frac{3}{2}}) & \text{when } c^* = 1. \end{cases}$$

Therefore, in the first case $c^* < 1$,

$$\begin{aligned}
c_{12}^{l+1} = c^* + \epsilon_l &= \frac{\gamma_l}{q^*}\left(\sigma_\alpha^2 \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\phi(u_2^l) + \sigma_b^2\right) \\
&= \frac{\gamma_l}{q^*}\left(\sigma_\alpha^2 \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\phi(u_2^* + \sqrt{q^*}\epsilon_l\big(z_1 - \frac{c^*}{\sqrt{1-c^{*2}}}z_2\big) + \mathcal{O}(\epsilon_l^2)) + \sigma_b^2\right) \\
&= \frac{\gamma_l}{q^*}\left(\sigma_\alpha^2 \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)[\phi(u_2^*) + \phi'(u_2^*)\sqrt{q^*}\epsilon_l\big(z_1 - \frac{c^*}{\sqrt{1-c^{*2}}}z_2\big)] + \mathcal{O}(\epsilon_l^2) + \sigma_b^2\right) \text{ expanding } \phi \text{ around } u_2^* \\
&= c^* + \frac{\gamma_l}{\sqrt{q^*}}\sigma_\alpha^2\epsilon_l \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\phi'(u_2^*)z_1 - \frac{c^*}{\sqrt{1-c^{*2}}}\frac{\gamma_l}{\sqrt{q^*}}\int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\phi'(u_2^*)z_2 + \mathcal{O}(\epsilon_l^2) \\
&= c^* + \gamma_l\sigma_\alpha^2\epsilon_l \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\big(\phi'(u_2^*) + c^*\phi''(u_2^*)\big) - c^*\gamma_l \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\phi\prime\prime(u_2^*) + \mathcal{O}(\epsilon_l^2) \\
&= c^* + \gamma_l\sigma_\alpha^2\epsilon_l \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\phi'(u_2^*) + \mathcal{O}(\epsilon_l^2)
\end{aligned}$$

where the second to last line is obtained using $\int DzF(z)z = \int DzF'(z)$, for ϕ smooth enough. We can then identify $\epsilon_{l+1} = \epsilon_l\gamma_l\sigma_\alpha^2 \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\phi'(u_2^*) + \mathcal{O}(\epsilon_l^2)$.

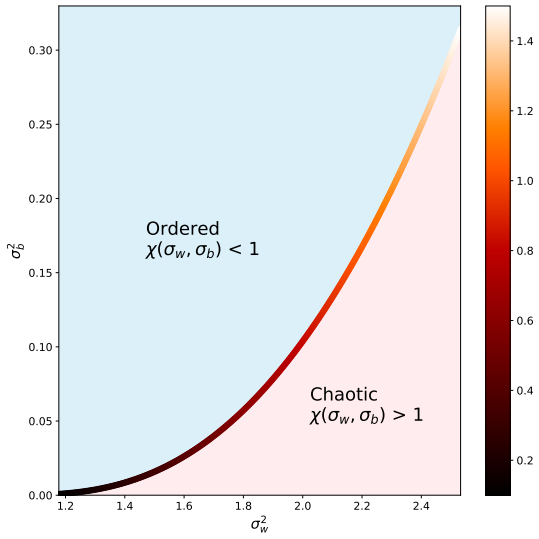

Figure 10. Original edge-of-chaos curve of the full-rank feedforward neural network with nonlinear activation $\phi(x) = tanh(x)$.

In the second case $c^* = 1$, $u_1^* = u_2^*$, $c_{12}^l = 1 - \epsilon_l$ and we expand ϕ around u_2^* until to the second order.

$$
\begin{aligned}
c_{12}^{l+1} = 1 - \epsilon_{l+1} &= \frac{\gamma_l}{q^*}\left(\sigma_\alpha^2 \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\phi(u_2^l) + \sigma_b^2\right) \\
&= \frac{\gamma_l}{q^*}\left(\sigma_\alpha^2 \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\phi\big(u_2^* + \sqrt{q^*}\epsilon_l\big(z_1 - \frac{c^*}{\sqrt{1-c^{*2}}}z_2\big) + \mathcal{O}(\epsilon_l^2)\big) + \sigma_b^2\right) \\
&= \frac{\gamma_l}{q^*}\left(\sigma_\alpha^2 \int_{\mathbb{R}^2} Dz_1 Dz_2 \phi(u_1^*)\left(\phi(u_2^*) + \phi'(u_2^*)\big(\sqrt{2q^*\epsilon_l}z_2 - \sqrt{q^*}\epsilon_l z_1\big) + \phi''(u_2^*)\frac{1}{2}\big(\sqrt{2q^*\epsilon_l}z_2 - \sqrt{q^*}\epsilon_l z_1\big)^2 + \mathcal{O}(\epsilon_l^{\frac{3}{2}})\right) + \sigma_b^2\right) \\
&= c^* + \frac{\gamma_l}{\sqrt{q^*}}\sqrt{2\epsilon_l}\sigma_\alpha^2 \int_{\mathbb{R}} Dz_1 \phi(u_1^*)\phi'(u_2^*) \underbrace{\int_{\mathbb{R}} z_2 Dz_2}_{=0} - \frac{\gamma_l}{\sqrt{q^*}}\epsilon_l \sigma_\alpha^2 \int_{\mathbb{R}} Dz_1 \phi(u_1^*)\phi'(u_2^*) z_1 \underbrace{\int_{\mathbb{R}} Dz_2}_{=1} \\
&\qquad + \gamma_l \epsilon_l \sigma_\alpha^2 \int_{\mathbb{R}} Dz_1 \phi(u_1^*)\phi''(u_2^*) \underbrace{\int_{\mathbb{R}} z_2^2 Dz_2}_{=1} + \mathcal{O}(\epsilon_l^{\frac{3}{2}}) \\
&= c^* - \frac{\gamma_l}{\sqrt{q^*}}\epsilon_l \sigma_\alpha^2 \int_{\mathbb{R}} Dz_1 \phi(u_1^*)\phi'(u_2^*) z_1 + \gamma_l \epsilon_l \sigma_\alpha^2 \int_{\mathbb{R}} Dz_1 \phi(u_1^*)\phi''(u_2^*) + \mathcal{O}(\epsilon_l^{\frac{3}{2}}) \\
&= c^* - \gamma_l \epsilon_l \sigma_\alpha^2 \int_{\mathbb{R}} Dz_1 (\phi'(u_1^*))^2 - \gamma_l \epsilon_l \sigma_\alpha^2 \int_{\mathbb{R}} Dz_1 \phi(u_1^*)\phi''(u_2^*) + \gamma_l \epsilon_l \sigma_\alpha^2 \int_{\mathbb{R}} Dz_1 \phi(u_1^*)\phi''(u_2^*) + \mathcal{O}(\epsilon_l^{\frac{3}{2}}) \\
&= c^* - \epsilon_l \gamma_l \sigma_\alpha^2 \int_{\mathbb{R}} Dz (\phi'(\sqrt{q^*}z))^2 + \mathcal{O}(\epsilon_l^{\frac{3}{2}})
\end{aligned}
$$

Therefore, $\epsilon_{l+1} = \epsilon_l \gamma_l \sigma_\alpha^2 \int_{\mathbb{R}} Dz (\phi'(\sqrt{q^*}z))^2 + \mathcal{O}(\epsilon_l^{\frac{3}{2}})$, which concludes the proof.

Figure 11 shows the analytically calculated correlation depth scales as a function of $\gamma\sigma_\alpha^2$ as well as simulations with networks of width $N = 1000$ and nonlinear activation $\phi(x) = tanh(x)$. The networks depth scale are observed to be consistent with the analytic calculations; in particular showing the depth scale asymptotes at $\chi_\gamma = 1$ for the different choices of σ_b^2.

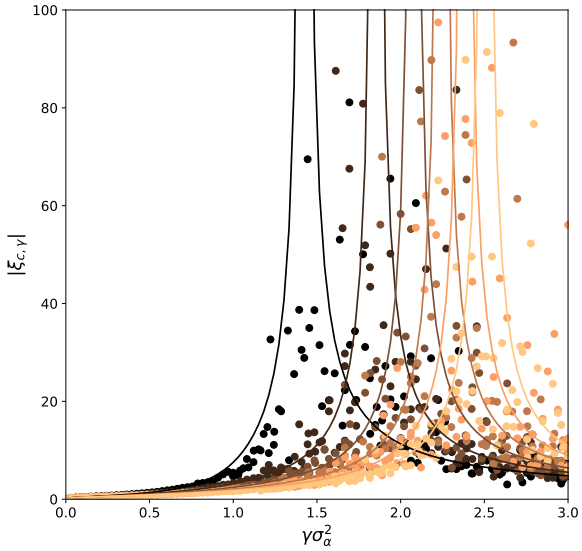

Figure 11. Correlation depth scale with respect to $\gamma\sigma_\alpha^2$ diverging when $\chi_\gamma = 1$. Points are obtained empirically when the lines are derived from the theory. The variance of the bias varies from $\sigma_b^2 = 0.01\gamma^{-1}$ (black) to $\sigma_b^2 = 0.3\gamma^{-1}$ (yellow). The network has constant width $N = 1000$. $\phi = \tanh, \gamma = \frac{1}{4}$.

A.8. Backpropagation

Recall that $h_i^{(l)} = \sum_{k=1}^{r_l} \left(\sum_{j=1}^{N_{l-1}} \alpha_{k,j}^{(l)} \phi(h_j^{(l-1)}) + b^{(l)} \right) (C_k^{(l)})_i$. Then the chain rule immediately gives,

$$\delta_j^l := \frac{\partial E}{\partial h_j^{(l)}} = \Big(\sum_{k=1}^{N_{l+1}} \delta_k^{l+1} W_{kj}^{(l+1)} \Big) \phi'(h_j^{(l)})$$

Because the trainable parameters of our network are the coefficients $\alpha_{ij}^{(l)}$, we compute the gradient of the error loss E with respect to them. So we need to adapt the proof from (Schoenholz et al., 2016), derived in the standard feedforward case as follows.

$$\begin{aligned} ||\nabla_{\alpha_{ij}^{(l)}} E||_2^2 &= \sum_{i,j} \Big(\frac{\partial E}{\partial \alpha_{ij}^{(l)}} \Big)^2 \\ &\underset{N_l, N_{l+1} \to \infty}{\approx} N_l N_{l+1} \mathbb{E}\left(\Big(\frac{\partial E}{\partial \alpha_{ij}^{(l)}} \Big)^2 \right). \end{aligned}$$

Since, assuming all these partial derivatives to be identically and independently distributed, the Law of Large numbers holds.

On the one hand,

$$\frac{\partial E}{\partial \alpha^{(l)}} = \sum_{m=1}^{N_l} \frac{\partial E}{\partial h_m^{(l)}} \frac{\partial h_m^{(l)}}{\partial \alpha_{ij}^{(l)}} = \sum_{m=1}^{N_l} \delta_m^l \phi(h_j^{(l-1)}) (C_i^{(l)})_m = \left(\sum_{m=1}^{N_l} \delta_m^l (C_i^{(l)})_m \right) \phi(h_j^{(l-1)}).$$

Therefore, assuming independence between the weights used for the forward pass and the weights backpropagated,

$$\begin{aligned}\mathbb{E}\Big(\big(\frac{\partial E}{\partial \alpha_{ij}^{(l)}}\big)^2\Big) &= \mathbb{E}\Big(\big(\sum_{m=1}^{N_l}\delta_m^l (C_i^{(l)})_m\big)^2\Big)\mathbb{E}\big(\phi^2(h_j^{(l-1)})\big)\\ &= \mathbb{E}\Big(\sum_{m=1}^{N_l}(\delta_m^l)^2(C_i^{(l)})_m^2 + \sum_{p,m}\delta_m^l (C_i^{(l)})_m\Big)\mathbb{E}\big(\phi^2(h_j^{(l-1)})\big)\\ &= \frac{1}{N_l}\sum_{m=1}^{N_l}\mathbb{E}\big((\delta_m^l)^2\big)\mathbb{E}\big(\phi^2(h_j^{(l-1)})\big) \text{ since } (C_1^{(l)})_m \overset{\text{iid}}{\sim} \mathcal{N}(0,\frac{1}{N_l})\\ &= \mathbb{E}\big((\delta_1^l)^2\big)\mathbb{E}\big(\phi^2(h_j^{(l-1)})\big).\end{aligned}$$

Therefore, the length of the gradient loss is proportional to the variance of δ_1^l.

On the other hand, denoting with a subscript the function f_a when fed with input data $x^{0,a}$,

$$\begin{aligned}\tilde{q}_{aa}^l &:= \mathbb{E}\big((\delta_{1,a}^l)^2\big)\\ &= \mathbb{E}\Big(\big(\sum_{k=1}^{N_{l+1}}\delta_{k,a}^{l+1}W_{kj}^{(l+1)}\big)^2\Big)\mathbb{E}\Big(\big(\phi'(h_{j,a}^{(l)})\big)^2\Big)\\ &= \Big(\sum_{k=1}^{N_{l+1}}\mathbb{E}\big((\delta_{k,a}^{l+1})^2\big)\mathbb{E}\big((W_{kj}^{(l+1)})^2\big)\Big)\mathbb{E}\Big(\big(\phi'(h_{j,a}^{(l)})\big)^2\Big)\end{aligned}$$

where we used again the assumed independence. The first and second order moments of $W_{ij}^{(l)}$ are given by $\mathbb{E}(W_{ij}^{(l)}) = 0$ and $\mathbb{E}\big((W_{ij}^{(l)})^2\big) = \mathbb{E}\big((\sum_{p=1}^{r_l}\alpha_{p,j}^{(l)}(C_p^{(l)})_i)^2\big) = \sum_{p=1}^{r_l}\mathbb{V}(\alpha_{p,j}^{(l)})\mathbb{V}((C_p^{(l)})_i) = r_l\frac{\sigma_\alpha^2}{N_{l-1}}\frac{1}{N_l}$. Thus,

$$\begin{aligned}\tilde{q}_{aa}^l &= r_{l+1}\frac{\sigma_\alpha^2}{N_l}\frac{1}{N_{l+1}}\Big(\sum_{k=1}^{N_{l+1}}\mathbb{E}\big((\delta_{k,a}^{l+1})^2\big)\Big)\mathbb{E}\Big(\big(\phi'(h_{j,a}^{(l)})\big)^2\Big)\\ &= r_{l+1}\frac{\sigma_\alpha^2}{N_l}\frac{1}{N_{l+1}}N_{l+1}\tilde{q}_{aa}^{l+1}\mathbb{E}\Big(\big(\phi'(h_{j,a}^{(l)})\big)^2\Big)\\ &= r_{l+1}\frac{\sigma_\alpha^2}{N_l}\frac{1}{N_{l+1}}N_{l+1}\tilde{q}_{aa}^{l+1}\int Dz\big(\phi'(\sqrt{q_{aa}^{l-1}}z)\big)^2 \text{ as } h_{j,a}^{(l-1)} \sim \mathcal{N}(0, q_{aa}^{l-1}).\end{aligned}$$

Considering that the computation is done at a layer deep enough, since q^{l-1} converges to q^* shortly, then $q_{aa}^{l-1} \approx q^*$, and, as $r_l = \gamma_l N_l$,

$$\tilde{q}_{aa}^l = \tilde{q}_{aa}^{l+1}\gamma_{l+1}\sigma_\alpha^2\int Dz\big(\phi'(\sqrt{q^*}z)\big)^2 = \tilde{q}_{aa}^{l+1}\chi_{l+1}.$$

Figure 12 demonstrates the exponential evolution of $||\nabla_{\alpha^{(l)}}E||_2^2$ from the final layer, $L = 250$, to the earlier layers. The analytic expressions are shown to be consistent with simulation from random low-rank networks with nonlinear activation $\phi(x) = tanh(x)$, rank to width scale $\gamma = 1/4$, the bias variance σ_b^2 held fixed and σ_α^2 varying.

A.9. Average singular value of $D^l W^{(l)}$

As a preliminary, let us first note that $\frac{1}{N_l}\text{Tr}\Big(D^l W^{(l)}(D^l W^{(l)})^T\Big) = \frac{1}{N_l}\sum_{k=1}^{N_l}\lambda_k\Big(D^l W^{(l)}(D^l W^{(l)})^T\Big) = \frac{1}{N_l}\sum_{k=1}^{N_l}\sigma_k^2\Big(D^l W^{(l)}\Big)$, where $\lambda_k(M), \sigma_k(M)$ represents the k-th eigenvalue and singular value, respectively, of the matrix

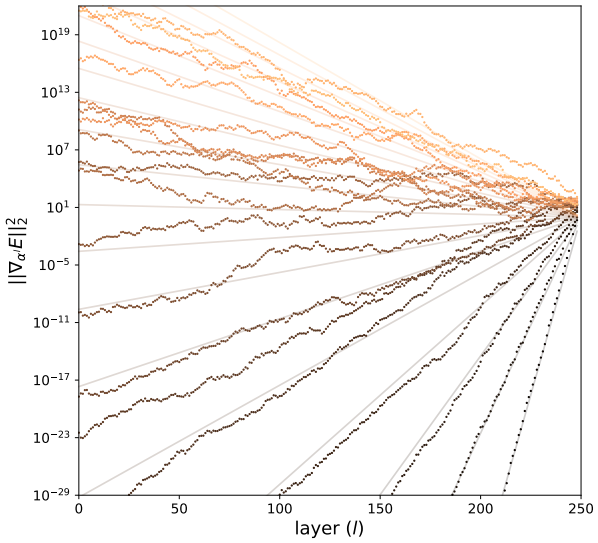

Figure 12. Exponential evolution of the propagation of the L_2-norm of the gradient with respect to the depth for a 250 layer deep random neural network with a cross entropy loss on MNIST dataset. Points are obtained empirically when the lines are derived from the theory. The variance of the weights $\gamma\sigma_\alpha^2$ varies from $0.01\gamma^{-1}$ (black) to $0.3\gamma^{-1}$ (yellow) when the variance of the bias $\gamma\sigma_b^2$ is kept fixed to 0.05. $\phi = \tanh, \gamma = \frac{1}{4}$.

M. Therefore, it appears clearly now that $\frac{1}{N_l}\text{Tr}\left(D^l W^{(l)})(D^l W^{(l)})^T\right)$ gives the empirical mean squared singular value of $D^l W^{(l)}$.

Let us now show that $\lim_{N_l\to\infty}\frac{1}{N_l}\mathbb{E}_{W^{(l)}}\text{Tr}\left(D^l W^{(l)}(D^l W^{(l)})^T\right) = \chi$ in the infinite width limit and when q^l is at its fixed point q^*.

$$
\begin{aligned}
\frac{1}{N_l}\mathbb{E}_{W^{(l)}}\text{Tr}\left(D^l W^{(l)}(D^l W^{(l)})^T\right) &= \frac{1}{N_l}\mathbb{E}_{W^{(l)}}\left(\sum_{j=1}^{N_l}\sum_{i=1}^{N_{l-1}}\phi'(h_i^{(l)})^2(W_{ij}^{(l)})^2\right)\\
&= \frac{1}{N_l}\sum_{j=1}^{N_l}\sum_{i=1}^{N_{l-1}}\mathbb{E}_{W^{(l)}}\left(\phi'(h_i^{(l)})^2(W_{ij}^{(l)})^2\right)\\
&= \frac{1}{N_l}\sum_{j=1}^{N_l}\sum_{i=1}^{N_{l-1}}\phi'(h_i^{(l)})^2\mathbb{E}_{W^{(l)}}\left((W_{ij}^{(l)})^2\right) \text{ considering } l \text{ big enough so that } q^l \approx q^*\\
&= \frac{1}{N_l}\sum_{j=1}^{N_l}\sum_{i=1}^{N_{l-1}}\phi'(h_i^{(l)})^2\left(\gamma_l\frac{\sigma_\alpha^2}{N_{l-1}}\right)\\
&= \gamma_l\sigma_\alpha^2\frac{1}{N_{l-1}}\sum_{i=1}^{N_{l-1}}\phi'(h_i^{(l)})^2\\
&\to \gamma_l\sigma_\alpha^2\int Dz\phi'(\sqrt{q^*}z)^2 \text{ using the Law of Large Numbers with } N_{l-1}\to\infty,\\
&= \chi,
\end{aligned}
$$

where we used, from the previous section, $\mathbb{E}_{W^{(l)}}\big((W_{ij}^{(l)})^2\big) = r_l \frac{\sigma_\alpha^2}{N_{l-1}} \frac{1}{N_l} = \gamma_l \frac{\sigma_\alpha^2}{N_{l-1}}$.

A.10. Computation of S_{W^TW} for low-rank Gaussian weights

Recall that $A_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \frac{\sigma_\alpha^2}{N})$ and $\text{rank}(A) = \gamma N$. Thus, its spectral density is given by the Marčenko Pastur distribution, where we first consider the matrix $\sigma_\alpha^{-2} A^T A$ as the variance of each coefficient is $\frac{1}{N}$ to make the computation simpler before appropriately rescaling the $\mathcal{S}$ Transform using the fact that if one rescales B by σ, then $S_{\sigma B} = \sigma^{-1} S_B$.

$$\rho_{\sigma_\alpha^{-2} A^T A}(\lambda) = (1-\gamma)_+ \delta(\lambda) + \gamma \frac{\sqrt{(\lambda^+ - \lambda)(\lambda - \lambda^-)}}{2\pi\lambda} \mathbb{1}_{[\lambda^-, \lambda^+]}(\lambda),$$

where $x_+ = \max(0, x)$, $\lambda^- = (1 - \frac{1}{\gamma})^2$ and $\lambda^+ = (1 + \frac{1}{\gamma})^2$. The Stieltjes Transform is known to be

$$G_{\sigma_\alpha^{-2} A^T A}(z) = \gamma \frac{z + \gamma^{-1} - 1 - \sqrt{(\lambda^+ - z)(z - \lambda^-)}}{2z},$$

from which we can easily compute the moment generating function

$$M_{\sigma_\alpha^{-2} A^T A}(z) = z G_{\sigma_\alpha^{-2} A^T A}(z) - 1 = \frac{1}{2}\big(-1 - \gamma + \gamma z - \gamma \sqrt{(\lambda^+ - z)(z - \lambda^-)}\big),$$

whose invert is

$$M^{-1}_{\sigma_\alpha^{-2} A^T A}(z) = \frac{\gamma + z(1+\gamma) + z^2}{\gamma z}.$$

And therefore

$$S_{\sigma_\alpha^{-2} A^T A}(z) = \frac{1+z}{z M^{-1}_{\sigma_\alpha^{-2} A^T A}(z)} = \gamma \frac{1+z}{\gamma + z(1+\gamma) + z^2} = \frac{1+z}{1 + z(1+\gamma^{-1}) + \gamma^{-1} z^2}.$$

Note that when $\gamma = 1$, the weight matrix is full rank and we get the same result as in (Pennington et al., 2018). Rescaling the matrix by σ_α^2 to match our original distribution gives

$$S_{A^T A}(z) = \sigma_\alpha^{-2} \frac{1+z}{1 + z(1+\gamma^{-1}) + \gamma^{-1} z^2}.$$

Now note that as we have $W_{ij} \sim \mathcal{N}(0, \gamma \frac{\sigma_\alpha^2}{N})$, the scaling property of the $\mathcal{S}$ transform gives

$$S_{W^TW} = S_{(\sqrt{\gamma}A)^T \sqrt{\gamma} A} = \sqrt{\gamma}^{-2} S_{A^T A} = \gamma^{-1} S_{A^T A}.$$

We can now expand S_{W^TW} around 0 and identify from $S_{W^TW}(z) := \gamma^{-1}\sigma_\alpha^{-2}\big(1 + \sum_{k=1}^{\infty} s_k z^k\big)$

$$\begin{aligned} S_{W^TW}(z) &= \gamma^{-1}\sigma_\alpha^{-2} \frac{1+z}{1 + z(1+\gamma^{-1}) + \gamma^{-1} z^2} = \gamma^{-1}\sigma_\alpha^{-2}\big(1 - \frac{1}{\gamma} z + \frac{1 - 4\gamma}{\gamma^2} z^2 + \dots\big) \\ &\implies s_1 = -\frac{1}{\gamma}. \end{aligned}$$

A.11. Computation of S_{W^TW} for low-rank Orthogonal weights

Recall the spectral density of W^TW,

$$\rho_{\sigma_\alpha^{-2}W^TW}(z) = \gamma\delta(z-1) + (1-\gamma)\delta(z).$$

Then, the following computations are straightforward

$$\begin{aligned}
G_{\sigma_\alpha^{-2}W^TW}(z) &= \gamma(z-1)^{-1} + (1-\gamma)z^{-1}, \\
M_{\sigma_\alpha^{-2}W^TW}(z) &= zG_{\sigma_\alpha^{-2}W^TW}(z) - 1 = \gamma(z-1)^{-1} \\
M^{-1}_{\sigma_\alpha^{-2}W^TW}(z) &= \frac{\gamma+z}{z} \\
S_{\sigma_\alpha^{-2}W^TW}(z) &= \frac{1+z}{zM^{-1}_{\sigma_\alpha^{-2}W^TW}(z)} = \frac{1+z}{\gamma+z} = \gamma^{-1}(1+z)(1+\gamma^{-1}z)^{-1}
\end{aligned}$$

Rescaling by σ_α^2, expanding around 0 and then identifying from $S_{W^TW}(z) := \gamma^{-1}\sigma_\alpha^{-2}\big(1 + \sum_{k=1}^{\infty} s_k z^k\big)$ gives

$$\begin{aligned}
S_{W^TW}(z) &= \sigma_\alpha^{-2}S_{\sigma_\alpha^{-2}W^TW}(z) = \sigma_\alpha^{-2}\gamma^{-1}(1+z)(1+\gamma^{-1}z)^{-1} \\
&= \sigma_\alpha^{-2}\gamma^{-1}(1+z)\sum_{k=0}^{\infty}(-\frac{z}{\gamma})^k = \gamma^{-1}\sigma_\alpha^{-2}\big(1 - (\gamma^{-1}-1)z + (\gamma^{-2}-\gamma^{-1})z^2 + \dots\big) \\
\implies s_1 &= -(\gamma^{-1}-1).
\end{aligned}$$