A Experimental Details

We implemented our methods using PyTorch [37] hooks and an efficient Jacobian approximate algorithm [19].

Figure 1: We generated MNIST-like inputs, where all elements are sampled from the Gaussian distribution $\mathcal{N}(0,1)$. $\mathcal{J}^{0,l}$ data was averaged over 100 different parameter-initializations. Networks were initialized width $N_l = 1000$. For erf plot we initialized at critical point $(\sigma_w, \sigma_b) = (\sqrt{\frac{\pi}{4}}, 0)$, used depth $L = 250$ and the fitting was done with data points collected at depth $l > 100$; for ReLU plot we initialized at critical point $(\sigma_w, \sigma_b) = (\sqrt{2}, 0)$, used depth $L = 100$ and the fitting was done with all data points; for the $\mu = 1$ Pre-LN plot, we initialized both networks at $(\sigma_w, \sigma_b) = (\sqrt{2}, 0)$, used depth $L = 250$ and the fitting was done with $l > 100$ data points.

Figure 2: All the phase diagrams were plotted using $\chi_{\mathcal{J}}^{L-1}$ generated from networks with $L = 50$ and $N_l = 500$. We used hooks to obtain the gradients that go into calculating $\chi_{\mathcal{J}}^{L-1}$. $\chi_{\mathcal{J}}^{L-1}$ data was averaged over 100 different parameter-initializations. Inputs were generated from a normal Gaussian distribution and have dimension 28×28. Generating the data for the figure took approximately 2 days on Google Colab Pro (single Tesla P100 GPU).

Figure 3: In all cases, networks are trained for 10 epochs using stochastic gradient descent with CrossEntropy loss. We used the Fashion MNIST dataset [47]. All networks had depth $L = 50$ and width $N_l = 500$. The learning rates were logarithmically sampled within $(10^{-5}, 1)$. Generating the data for the figure took approximately 12 days on Google Colab Pro (single Tesla P100 GPU).

Figure 4: (1)We made the $\sigma_w^2 - \sigma_b^2$ phase diagram for ResNet110(LayerNorm) by averaging over 100 different parameter-initializations. The $\sigma_w^2 - \mu$ phase diagram was made by averaging over 200 parameters initialization. (3)(4) We used SGD with momentum$= 0.9$ and batch size 128. For selecting the learning rate we ran a grid-search over $0.001, 0.005, 0.01, 0.02, 0.5$ for 10 epochs; with weight decay $\lambda = 10^{-4}$. All models were trained for 50 epochs and averaged over 3 random seeds. It takes 6 GPU days in total on a single NVIDIA RTX 3090 GPU.

Figure 5: (1)(2)We made the phase diagram for MLP-Mixer with 30 blocks and averaged over 100 different parameter-initializations. (3)(4)We used network with $L = 100$, patch size 4×4, hidden size $C = 128$, two MLP dimensions $N_{tm} = N_{cm} = 256$. The $L = 32$ point has doubled widths. All networks have 10 million parameters. Notice that for all Mixer Layers we used NTK initialization. We trained all cases on CIFAR-10 dataset using vanilla SGD paired with CSE. Batch size bs $= 256$, weight decay $\lambda = 10^{-4}$ was selected from $\{10^{-5}, 10^{-4}\}$, mixup rate $\alpha = 0.8$ was selected from $\{0.4, 0.8\}$. We also used RandAgument and horizontal flip with default settings in PyTorch. For all cases we searched learning rates within $\{0.005, 0.01, 0.05, 0.1, 0.2, 0.5\}$. We also tried a linear warm-up schedule for first 3000 iterations, but we did not see any improvement in performance. Generating the data for the figure took approximately 4 days on Google Colab Pro (single Tesla P100 GPU).

B Additional Discussion on ResNet, ResNet with BatchNorm

For Convolution Layers, the NNGP kernel is a 4-index tensor: $\mathcal{K}^l_{\mu\nu;ij}(x, x')$, where the Greek letters(μ, ν) index the channels, whereas the Latin letters (i, j) index the pixels. The infinite width limit in this case is achieved by taking the number of channels to infinity (sequentially). In this limit, most of our equations for MLP can be easily rewritten using the convolutional NNGP kernel. However, in this case, the kernel is only diagonal in channel dimension: $\mathcal{K}^l_{\mu\nu;ij}(x, x') = \mathcal{K}^l_{ij}(x, x')\delta_{\mu\nu}$. This additional structure in the kernel makes it difficult to get a closed-form solution for $\mathcal{J}^{l,l+1}$ in general.

ResNet110 (LayerNorm) In Figure 4 (2), the networks is critical close to $\mu = 1$, as expected from our analysis. One would naively expect the $\mu < 1$ cases also to be critical, since for MLP with ReLU and Pre-LN, $\sigma_b = 0$ is critical regardless of σ_w and μ. However, in Figure 4(2) the region away from $\mu = 1$ is in ordered phase. This is likely a result of the kernel $\mathcal{K}^l_{\mu\nu;ij}(x, x')$ not being diagonal in spatial dimensions. We emphasize that the $\mu = 1$ case stays unaffected by this, since the existence of criticality does not depend on the details of the NNGP kernel in this case. This can be readily seen from (77). We present the Numerical and training results in Figure 4.

ResNet110 (BatchNorm) The operation of BatchNorm on a preactivation (pre-BN) in an MLP can be described as follows:

$$\tilde{h}_i(x) = \frac{h_i(x) - \mu_{i,B}}{\sigma_{i,B}}\,,$$
$$\mu_{i,B} = \frac{1}{|B|}\sum_{x'\in B} h(x') \quad \textit{and} \quad \sigma_{i,B} = \sqrt{\frac{1}{|B|}\sum_{x'\in B}(h(x') - \mu_{i,B})^2}\,, \tag{24}$$

where B is the batch that x belongs to and $|B|$ is batch size.

The works Yang et al. [57], He et al. [18] show that for large batch size, the effect of BatchNorm for NNGP kernel and Jacobian Norm is deterministic. We summarize the results for the pre-BN MLP setup:

$$\mathcal{K}^{l+1}(x,x') = \frac{\sigma_w^2}{N_l}\sum_{j=1}^{N_l}\mathbb{E}_\theta\left[\phi(\tilde{h}_j^l(x))\phi(\tilde{h}_j^l(x'))\right] + \mu^2\mathcal{K}^l(x,x') \tag{25}$$

$$\mathcal{J}^{l,l+1} = \frac{\sigma_w^2}{K^l(x,x) - K^l(x,x')}\mathbb{E}_\theta\left[\phi(\tilde{h}_j^l(x))\phi(\tilde{h}_j^l(x))\right] + \mu^2\,, \tag{26}$$

where x' in the APJN term can be any $x' \neq x$, since for a large batch size all choices are equivalent.

From the above result, we can see that most results we have for LayerNorm can be translated to BatchNorm with an easy replacement $K^l(x,x) \to \left(\mathcal{K}^l(x,x) - \mathcal{K}^l(x,x')\right)$. As a simple example, consider a pre-BN ResNet architecture, but with all the Convolutioal layers replaced with Linear (Fully Connected) layers. For such a network, we have the following result for $\mu < 1$:

$$\mathcal{J}^{l,l+1} = \frac{\pi^2(1-\mu^2)}{(\pi-1)^2} + \mu^2\,. \tag{27}$$

For $\mu = 1$, we have

$$\mathcal{J}^{l,l+1} = 1 + O(l^{-1}) \tag{28}$$

The ResNet results can then be obtained by replacing Fully Connected layers with Convolution layers, in a similar way as discussed in ResNet(LayerNorm) section. We show the numerical results for ResNet with BatchNorm in Figure 6.

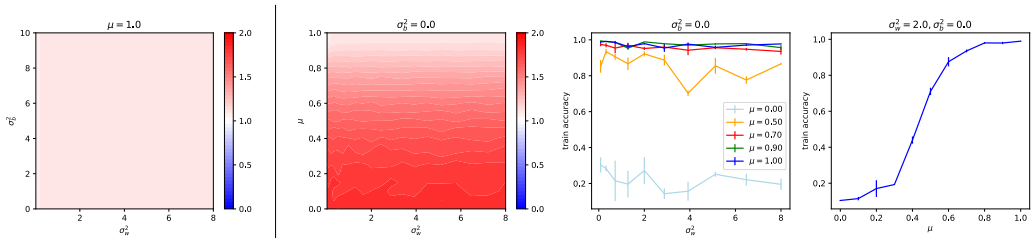

Figure 6: **ResNet110(BatchNorm)**: Left to right: (1)(2) $\mathcal{J}^{L-1,L}$ phase diagrams, with (σ_w^2, σ_b^2) and (σ_w^2, μ). (3)(4) Training curves w.r.t σ_w^2 and μ.

C Technical details for Jacobians and LayerNorm

We will drop the dependence of $h_i^l(x)$ on x throughout the Appendices. It should not cause any confusion since we are *always* considering a single input.

C.1 NNGP Kernel

First, we derive the recurrence relation for the NNGP kernel Eq.(10). As mentioned in the main text, weights and biases are initialized (independently) from standard normal distribution $\mathcal{N}(0, \sigma_w^2/\text{fan\_in})$. We then have

$$\mathbb{E}_\theta[w_{ij}^l w_{mn}^l] = \frac{\sigma_w^2}{N_{l-1}}\delta_{im}\delta_{jn} \text{ and } \mathbb{E}_\theta[b_i^l b_j^l] = \sigma_b^2\delta_{ij} \tag{29}$$

by definition.

We would like to prove theorem 2.4, as a consequence of lemma 2.2. The proof of lemma 2.2 can be found in [40].

Proof of theorem 2.4. One can prove this by definition with lemma 2.2.

$$
\begin{aligned}
\mathcal{K}^{l+1} \equiv & \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta[h_i^{l+1} h_i^{l+1}] \\
= & \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\left(\sum_{j=1}^{N_l} w_{ij}^{l+1} \phi(h_j^l) + b_i^{l+1} \right) \left(\sum_{k=1}^{N_l} w_{ik}^{l+1} \phi(h_k^l) + b_i^{l+1} \right) \right] \\
= & \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\sum_{j=1}^{N_l} \sum_{k=1}^{N_l} w_{ij}^{l+1} w_{ik}^{l+1} \phi(h_j^l) \phi(h_k^l) + b_i^{l+1} b_i^{l+1} \right] \\
= & \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(h_j^l) \phi(h_j^l) + \sigma_b^2 \right] \\
= & \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(h_j^l) \phi(h_j^l) \right] + \sigma_b^2 \,. \qquad (30)
\end{aligned}
$$

□

C.2 Jacobians

Next, we prove theorem 2.5 in the main text.

Proof of theorem 2.5. We start from the definition of the averaged partial Jacobian norm (APJN) ($l > l_0$)

$$
\begin{aligned}
\mathcal{J}^{l_0, l+1} \equiv & \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \right] \\
= & \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \left(\sum_{k=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial h_k^l} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \left(\sum_{m=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial h_m^l} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
= & \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} \left(w_{ik}^{l+1} \phi'(h_k^l) \right) \left(w_{im}^{l+1} \phi'(h_m^l) \right) \left(\frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
= & \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} w_{ik}^{l+1} w_{im}^{l+1} \phi'(h_k^l) \phi'(h_m^l) \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right] \\
= & \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k=1}^{N_l} \frac{\sigma_w^2}{N_l} \mathbb{E}_\theta \left[\phi'(h_k^l) \phi'(h_k^l) \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right] \\
= & \frac{\sigma_w^2}{N_l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(h_k^l) \phi'(h_k^l) \left(\sum_{j=1}^{N_{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \right] . \qquad (31)
\end{aligned}
$$

where we used chain rule and took expectation value over w^{l+1}. Next we take chain rule again:

$$\begin{aligned}
\mathcal{J}^{l_0,l+1} &= \frac{\sigma_w^2}{N_l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(h_k^l)\phi'(h_k^l)\left(\sum_{j=1}^{N_{l_0}}\sum_{m,n=1}^{N_{l-1}} w_{km}^l w_{kn}^l \phi'(h_m^{l-1})\phi'(h_n^{l-1})\frac{\partial h_m^{l-1}}{\partial h_j^{l_0}}\frac{\partial h_n^{l-1}}{\partial h_j^{l_0}}\right)\right] \\
&= \frac{\sigma_w^4}{N_l N_{l-1}}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(h_k^l)\phi'(h_k^l)\left(\sum_{j=1}^{N_{l_0}}\sum_{m=1}^{N_{l-1}} \phi'(h_m^{l-1})\phi'(h_m^{l-1})\frac{\partial h_m^{l-1}}{\partial h_j^{l_0}}\frac{\partial h_m^{l-1}}{\partial h_j^{l_0}}\right) + O(1/N_{l-1})\right] \\
&= \frac{\sigma_w^4}{N_l N_{l-1}}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(h_k^l)\phi'(h_k^l)\right]\mathbb{E}_\theta\left[\left(\sum_{j=1}^{N_{l_0}}\sum_{m=1}^{N_{l-1}} \phi'(h_m^{l-1})\phi'(h_m^{l-1})\frac{\partial h_m^{l-1}}{\partial h_j^{l_0}}\frac{\partial h_m^{l-1}}{\partial h_j^{l_0}}\right)\right] \\
&= \frac{\sigma_w^2}{N_l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(h_k^l)\phi'(h_k^l)\right]\cdot\frac{\sigma_w^2}{N_{l-1}}\mathbb{E}_\theta\left[\left(\sum_{j=1}^{N_{l_0}}\sum_{m=1}^{N_{l-1}} \phi'(h_m^{l-1})\phi'(h_m^{l-1})\frac{\partial h_m^{l-1}}{\partial h_j^{l_0}}\frac{\partial h_m^{l-1}}{\partial h_j^{l_0}}\right)\right] \\
&= \chi_{\mathcal{J}}^l\,\mathcal{J}^{l_0,l}\,,
\end{aligned} \tag{32}$$

where we integrate (by parts) over w^l to get the second line. We take $N_{l-1}\to\infty$ and then $N_l\to\infty$ to get the third line. We rearange terms and use the Eq.(12) to get the fourth and fifth lines. Notice that to get the third line we used the fact in the infinite width limit, the distribution of h_i^l is independent of h_i^{l-1}. Thus we proved

$$\mathcal{J}^{l_0,l+1} = \chi_{\mathcal{J}}^l \mathcal{J}^{l_0,l}\,. \tag{33}$$

□

The critical line is defined by requiring $\chi_{\mathcal{J}}^\star = 1$, where critical points are reached by further requiring $\chi_{\mathcal{K}}^\star = 1$.

As we mentioned in main text, $l_0 = 0$ is subtle since the input dimension is fixed N_0, which can not be assumed to be infinity. Even though for dataset like MNIST, usually N_0 is not significantly smaller than width N_l. We show how to take finite $O(N_0^{-1})$ correction into account by using one example.

Lemma C.1. *Consider a one hidden layer network with a finite input dimension N_0. In the infinite width limit, the APJN is still deterministic and the first step of the recurrence relation is modified to:*

$$\mathcal{J}^{0,2} = \left(\chi_{\mathcal{J}}^1 + \frac{2\sigma_w^2}{N_0}\chi_\Delta^1\sum_k^{N_0}\frac{1}{N_0}h_k^0 h_k^0\right)\mathcal{J}^{0,1}\,, \tag{34}$$

where $\mathcal{J}^{0,1} = \sigma_w^2$.

Proof.

$$
\begin{aligned}
\mathcal{J}^{0,2} =& \frac{1}{N_2}\mathbb{E}_\theta\left[\sum_{i=1}^{N_2}\sum_{j=1}^{N_0}\frac{\partial h_i^2}{\partial h_j^0}\frac{\partial h_i^2}{\partial h_j^0}\right] \\
=& \frac{1}{N_2}\mathbb{E}_\theta\left[\sum_{i=1}^{N_2}\sum_{j=1}^{N_0}\sum_{k,m=1}^{N_1} w_{ik}^2 w_{im}^2 \phi'(h_k^1)\phi'(h_m^1)\frac{\partial h_k^1}{\partial h_j^0}\frac{\partial h_m^1}{\partial h_j^0}\right] \\
=& \frac{1}{N_2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_0}\sum_{k,m=1}^{N_1}\mathbb{E}_\theta[w_{ik}^2 w_{im}^2 w_{kj}^1 w_{mj}^1 \phi'(h_k^1)\phi'(h_m^1)] \\
=& \sum_{j=1}^{N_0}\sum_{k=1}^{N_1}\frac{\sigma_w^2}{N_1}\mathbb{E}_\theta[w_{kj}^1 w_{kj}^1 \phi'(h_k^1)\phi'(h_k^1)] \\
=& \sigma_w^2\left(\chi_{\mathcal{J}}^1 + \frac{2\sigma_w^2}{N_0}\chi_\Delta^1\sum_k^{N_0}\frac{1}{N_0}h_k^0 h_k^0\right) \\
=& \left(\chi_{\mathcal{J}}^1 + \frac{2\sigma_w^2}{N_0}\chi_\Delta^1\sum_k^{N_0}\frac{1}{N_0}h_k^0 h_k^0\right)\mathcal{J}^{0,1}\,,
\end{aligned} \tag{35}
$$

where to get the result we used integrate by parts, then explicitly integrated over w_{ij}^1. We have introduced a coefficient of finite width corrections, χ_Δ^l, defined follows. □

Definition C.2 (Coefficient of Finite Width Corrections)**.**

$$
\chi_\Delta^l = \frac{\sigma_w^2}{N_l}\sum_{i=1}^{N_l}\mathbb{E}_\theta[\phi''(h_i^l)\phi''(h_i^l) + \phi'''(h_i^l)\phi'(h_i^l)]\,. \tag{36}
$$

Remark C.3. Notice that the correction to $\mathcal{J}^{0,2}$ is order $O(N_0^{-1})$. If one calculate the recurrence relation for deeper layers, the correction to $\mathcal{J}^{0,l}$ will be $O(\sum_{l'=0}^{l} N_{l'}^{-1})$, which means the contribution from hidden layers can be ignored in infinite width limit.

The $\mathcal{J}^{0,2}$ example justifies the factorization of the integral when we go from the last line of Eq.(31) to Eq.(33).

Finally, the full Jacobian in infinite width limit can be written as

Lemma C.4 (APJN with $l_0 = 0$)**.** *The APJN (with $l_0 = 0$) of a given network can be written as*

$$
\mathcal{J}^{0,l} = \sigma_w^2\left(\chi_{\mathcal{J}}^1 + \frac{2\sigma_w^2}{N_0}\chi_\Delta^1\sum_k^{N_0}\frac{1}{N_0}h_k^0 h_k^0\right)\prod_{l'=2}^{l-1}\chi_{\mathcal{J}}^{l'}\,. \tag{37}
$$

Note that APJN with $l_0 > 0$ does not receive the $O(N_0^{-1})$ correction.

C.3 APJN and gradients

As mentioned in the main text, APJN is an important tool in studying the exploding and vanishing gradients problem. Its utility stems from the fact that it is a dominant factor in the norm of the gradients. This can be readily by looking at the (squared) L_2 norm of the gradient of any flattened

parameter matrix θ^l, at initialization. In the infinite width limit, one gets

$$\begin{aligned}
\|\nabla_{\theta^l}\mathcal{L}\|_2^2 &= \left(\sum_{all}\frac{\partial\mathcal{L}}{\partial h_i^L}\frac{\partial h_i^L}{\partial h_j^{L-1}}\cdots\frac{\partial h_k^{l+1}}{\partial h_m^l}\frac{\partial h_m^l}{\partial\theta_n^l}\right)^2 \\
&= \sum_{all}\left(\frac{\partial\mathcal{L}}{\partial h_i^L}\frac{\partial\mathcal{L}}{\partial h_{i'}^L}\right)\left(\frac{\partial h_i^L}{\partial h_j^{L-1}}\frac{\partial h_{i'}^L}{\partial h_{j'}^{L-1}}\right)\cdots\left(\frac{\partial h_k^{l+1}}{\partial h_m^l}\frac{\partial h_{k'}^{l+1}}{\partial h_{m'}^l}\right)\left(\frac{\partial h_m^l}{\partial\theta_n^l}\frac{\partial h_{m'}^l}{\partial\theta_n^l}\right) \\
&= \sum_{all}\left(\frac{\partial\mathcal{L}}{\partial h_i^L}\frac{\partial\mathcal{L}}{\partial h_{i'}^L}\right)\left(\frac{\partial h_i^L}{\partial h_j^{L-1}}\frac{\partial h_{i'}^L}{\partial h_{j'}^{L-1}}\right)\cdots\left(\frac{\partial h_k^{l+1}}{\partial h_m^l}\frac{\partial h_{k'}^{l+1}}{\partial h_{m'}^l}\right)\delta_{mm'}\left\|\frac{\partial h^l}{\partial\theta^l}\right\|_F^2 \\
&= \sum_{all}\left(\frac{\partial\mathcal{L}}{\partial h_i^L}\frac{\partial\mathcal{L}}{\partial h_{i'}^L}\right)\left(\frac{\partial h_i^L}{\partial h_j^{L-1}}\frac{\partial h_{i'}^L}{\partial h_{j'}^{L-1}}\right)\cdots\delta_{kk'}\mathcal{J}^{l,l+1}\left\|\frac{\partial h^l}{\partial\theta^l}\right\|_F^2 \\
&= \sum_{i,i',j,j'}\left(\frac{\partial\mathcal{L}}{\partial h_i^L}\frac{\partial\mathcal{L}}{\partial h_{i'}^L}\right)\left(\frac{\partial h_i^L}{\partial h_j^{L-1}}\frac{\partial h_{i'}^L}{\partial h_{j'}^{L-1}}\right)\delta_{jj'}\cdots\mathcal{J}^{l,l+1}\left\|\frac{\partial h^l}{\partial\theta^l}\right\|_F^2 \\
&= \sum_{i,i'}\left(\frac{\partial\mathcal{L}}{\partial h_i^L}\frac{\partial\mathcal{L}}{\partial h_{i'}^L}\right)\delta_{ii'}\mathcal{J}^{L-1,L}\cdots\mathcal{J}^{l,l+1}\left\|\frac{\partial h^l}{\partial\theta^l}\right\|_F^2 \\
&= \left\|\frac{\partial\mathcal{L}}{\partial h^L}\right\|_2^2\mathcal{J}^{L-1,L}\cdots\mathcal{J}^{l,l+1}\left\|\frac{\partial h^l}{\partial\theta^l}\right\|_F^2 \\
&= \left\|\frac{\partial\mathcal{L}}{\partial h^L}\right\|_2^2\mathcal{J}^{l,L}\left\|\frac{\partial h^l}{\partial\theta^l}\right\|_F^2,
\end{aligned} \tag{38}$$

where $\|\cdot\|_2$ denotes the L_2 norm and $\|\cdot\|_F$ denotes the Frobenius norm.

C.4 LayerNorm on Pre-activations

Definition C.5 (Layer Normalization)**.**

$$\tilde{h}_i^l = \frac{h_i^l - \mathbb{E}[h^l]}{\sqrt{\mathbb{E}[(h^l)^2] - \mathbb{E}[h^l]^2}}\gamma_i^l + \beta_i^l\,, \tag{39}$$

where γ_i^l and β_i^l are learnable parameters.

Remark C.6. With only LayerNorm, the (1) is simplified to

$$h_i^{l+1} = \sum_{j=1}^{N_l} w_{ij}^{l+1}\phi(\tilde{h}_j^l) + b_i^{l+1}\,. \tag{40}$$

Remark C.7. In the limit of infinite width, using the law of large numbers, the average over neurons $\mathbb{E}[\cdots]$ can be replaced by the average of parameter-initializations $\mathbb{E}_\theta[\cdots]$. Additionally, in this limit, the preactivations are i.i.d. Gaussian distributed : $h^l \sim \mathcal{N}(0, \mathcal{K}^l)$.

$$\mathbb{E}\left[h^l\right] = \mathbb{E}_\theta\left[h^l\right] = 0\,, \tag{41}$$

$$\mathbb{E}\left[\left(h^l\right)^2\right] = \mathbb{E}_\theta\left[\left(h^l\right)^2\right] = \mathcal{K}^l\,. \tag{42}$$

The normalized preactivation then simplifies to the form of Eq.(20).

Remark C.8. At initialization, the parameters γ_i^l and β_i^l take the values 1 and 0, respectively. This leads to the form in equation (20). In infinite width limit it has the following form

$$\tilde{h}_i^l = \frac{h_i^l - \mathbb{E}_\theta[h^l]}{\sqrt{\mathbb{E}_\theta[(h^l)^2] - \mathbb{E}_\theta[h^l]^2}}\,. \tag{43}$$

Lemma C.9. *With LayerNorm on preactivations, the gaussian average is modified to*

$$\mathbb{E}_\theta\left[O(\tilde{h}_i^l)\right] = \frac{1}{\sqrt{2\pi}}\int d\tilde{h}_i^l\, O(\tilde{h}_i^l)\, e^{-\frac{(\tilde{h}_i^l)^2}{2}}\,. \tag{44}$$

Proof. By definition $\tilde{h}_i^l$ is sampled from a standard normal distribution $\mathcal{N}(0,1)$, then use lemma 2.2 to get the final form. □

Theorem C.10. *In the infinite width limit the recurrence relation for the NNGP kernel with LayerNorm on preactivations is*

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2}{N_l}\sum_{j=1}^{N_l}\mathbb{E}_\theta\left[\phi(\tilde{h}_j^l)\phi(\tilde{h}_j^l)\right] + \sigma_b^2\,. \tag{45}$$

Proof.

$$\begin{aligned}
\mathcal{K}^{l+1} =& \frac{1}{N_{l+1}}\sum_{i=1}^{N_{l+1}}\mathbb{E}_\theta\left[h_i^{l+1}h_i^{l+1}\right] \\
=& \frac{1}{N_{l+1}}\sum_{i=1}^{N_{l+1}}\mathbb{E}_\theta\left[\left(\sum_{j=1}^{N_l} w_{ij}^{l+1}\phi(\tilde{h}_j^l) + b_i^{l+1}\right)\left(\sum_{k=1}^{N_l} w_{ik}^{l+1}\phi(\tilde{h}_k^l) + b_i^{l+1}\right)\right] \\
=& \frac{\sigma_w^2}{N_l}\sum_{j=1}^{N_l}\mathbb{E}_\theta\left[\phi(\tilde{h}_j^l)\phi(\tilde{h}_j^l)\right] + \sigma_b^2\,.
\end{aligned} \tag{46}$$

□

Theorem C.11. *In the infinite width limit the recurrence relation for partial Jacobian with LayerNorm on preactivations is*

$$\mathcal{J}^{l_0,l+1} = \chi_{\mathcal{J}}^l\mathcal{J}^{l_0,l}\,, \tag{47}$$

where $\chi_J^l = \frac{\sigma_w^2}{N_l\mathcal{K}^l}\sum_{i=1}^{N_l}\mathbb{E}_\theta\left[\phi'(\tilde{h}_i^l)^2\right]$.

Proof.

$$\begin{aligned}
\mathcal{J}^{l_0,l+1} &= \frac{1}{N_{l+1}}\mathbb{E}_\theta\left[\sum_{i=1}^{N_{l+1}}\sum_{j=1}^{N_{l_0}}\frac{\partial h_i^{l+1}}{\partial h_j^{l_0}}\frac{\partial h_i^{l+1}}{\partial h_j^{l_0}}\right] \\
&= \frac{1}{N_{l+1}}\mathbb{E}_\theta\left[\sum_{i=1}^{N_{l+1}}\sum_{j=1}^{N_{l_0}}\left(\sum_{k=1}^{N_l}\frac{\partial h_i^{l+1}}{\partial\tilde{h}_k^l}\frac{\partial\tilde{h}_k^l}{\partial h_k^l}\frac{\partial h_k^l}{\partial h_j^{l_0}}\right)\left(\sum_{m=1}^{N_l}\frac{\partial h_i^{l+1}}{\partial\tilde{h}_m^l}\frac{\partial\tilde{h}_m^l}{\partial h_m^l}\frac{\partial h_m^l}{\partial h_j^{l_0}}\right)\right] \\
&= \frac{1}{N_{l+1}}\mathbb{E}_\theta\left[\sum_{i=1}^{N_{l+1}}\sum_{j=1}^{N_{l_0}}\sum_{k,m=1}^{N_l}\left(w_{ik}^{l+1}\phi'(\tilde{h}_k^l)\frac{1}{\sqrt{\mathcal{K}^l}}\right)\left(w_{im}^{l+1}\phi'(\tilde{h}_m^l)\frac{1}{\sqrt{\mathcal{K}^l}}\right)\left(\frac{\partial h_k^l}{\partial h_j^{l_0}}\frac{\partial h_m^l}{\partial h_j^{l_0}}\right)\right] \\
&= \frac{\sigma_w^2}{N_l\mathcal{K}^l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l)\left(\sum_{j=1}^{N_{l_0}}\frac{\partial h_k^l}{\partial h_j^{l_0}}\frac{\partial h_k^l}{\partial h_j^{l_0}}\right)\right] \\
&= \frac{\sigma_w^2}{N_l\mathcal{K}^l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l)\right]\mathcal{J}^{l_0,l} \\
&= \chi_{\mathcal{J}}^l\mathcal{J}^{l_0,l}\,,
\end{aligned} \tag{48}$$

□

C.5 LayerNorm on activations

The general definition of LayerNorm on activations is given as follows.

Definition C.12 (LayerNorm on Activations)**.**

$$\widetilde{\phi(h_i^l)} = \frac{\phi(h_i^l) - \mathbb{E}[\phi(h^l)]}{\sqrt{\mathbb{E}[\phi(h^l)^2] - \mathbb{E}[\phi(h^l)]^2}}\gamma_i^l + \beta_i^l\,. \tag{49}$$

Remark C.13. The recurrence relation for preactivations (Eq.(1)) gets modified to

$$h_i^{l+1} = \sum_{j=1}^{N_l} w_{ij}^{l+1} \widetilde{\phi(h_j^l)} + b_i^{l+1} \,. \tag{50}$$

Remark C.14. At initialization, the parameters γ_i^l and β_i^l take the values 1 and 0, respectively. This leads to the form

$$\begin{aligned} \widetilde{\phi(h_i^l)} &= \frac{\phi(h_i^l) - \mathbb{E}[\phi(h^l)]}{\sqrt{\mathbb{E}[\phi(h^l)^2] - \mathbb{E}[\phi(h^l)]^2}} \\ &= \frac{\phi(h_i^l) - \mathbb{E}_\theta\left[\phi(h^l)\right]}{\sqrt{\mathbb{E}_\theta\left[\phi(h^l)^2\right] - \mathbb{E}_\theta\left[\phi(h^l)\right]^2}} \,, \end{aligned} \tag{51}$$

where the first line follows from the fact that at initialization, the parameters γ_i^l and β_i^l take the values 1 and 0 respectively. In the second line, we have invoked the infinite width limit.

Remark C.15. Evaluating Gaussian average in this case is similar to cases in previous section. The only difference being that the averages are taking over the distribution $h^{l-1} \sim \mathcal{N}(0, \mathcal{K}^{l-1} = \sigma_w^2 + \sigma_b^2)$. Again this can be summarized as

$$\mathbb{E}_\theta\left[O(h_i^l)\right] = \frac{1}{\sqrt{2\pi(\sigma_w^2 + \sigma_b^2)}} \int dh_i^l \, O(h_i^l) \, e^{-\frac{(h_i^l)^2}{2(\sigma_w^2 + \sigma_b^2)}} \,. \tag{52}$$

Next, we calculate the modifications to the recurrence relations for the NNGP kernel and Jacobians.

Theorem C.16. *In the infinite width limit the recurrence relation for the NNGP kernel with LayerNorm on activations is*

$$\mathcal{K}^{l+1} = \sigma_w^2 + \sigma_b^2 \,. \tag{53}$$

Proof.

$$\begin{aligned} \mathcal{K}^{l+1} &= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta\left[h_i^{l+1} h_i^{l+1}\right] \\ &= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta\left[\left(\sum_{j=1}^{N_l} w_{ij}^{l+1} \widetilde{\phi(h_j^l)} + b_i^{l+1}\right)\left(\sum_{k=1}^{N_l} w_{ik}^{l+1} \widetilde{\phi(h_k^l)} + b_i^{l+1}\right)\right] \\ &= \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta\left[\widetilde{\phi(h_j^l)}^2\right] + \sigma_b^2 \\ &= \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta\left[\left(\frac{\phi(h_j^l) - \mathbb{E}_\theta\left[\phi(h^l)\right]}{\sqrt{\mathbb{E}_\theta\left[\phi(h^l)^2\right] - \mathbb{E}_\theta\left[\phi(h^l)\right]^2}}\right)^2\right] + \sigma_b^2 \\ &= \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \frac{\mathbb{E}_\theta\left[\left(\phi(h_j^l) - \mathbb{E}_\theta\left[\phi(h^l)\right]\right)^2\right]}{\mathbb{E}_\theta\left[\phi(h^l)^2\right] - \mathbb{E}_\theta\left[\phi(h^l)\right]^2} + \sigma_b^2 \\ &= \sigma_w^2 + \sigma_b^2 \,. \end{aligned} \tag{54}$$

□

Theorem C.17. *In the infinite width limit the recurrence relation for partial Jacobian with LayerNorm on activations is*

$$\mathcal{J}^{l_0, l+1} = \chi_J^l \mathcal{J}^{l_0, l} \,, \tag{55}$$

where $\chi_J^l \equiv \sigma_w^2 \frac{\mathbb{E}_\theta\left[\phi'(h^l)^2)\right]}{\mathbb{E}_\theta[\phi(h^l)^2] - \mathbb{E}_\theta[\phi(h^l)]^2}$.

Proof.

$$
\begin{aligned}
\mathcal{J}^{l_0,l+1} &= \frac{1}{N_{l+1}}\mathbb{E}_\theta\left[\sum_{i=1}^{N_{l+1}}\sum_{j=1}^{N_{l_0}}\frac{\partial h_i^{l+1}}{\partial h_j^{l_0}}\frac{\partial h_i^{l+1}}{\partial h_j^{l_0}}\right] \\
&= \frac{1}{N_{l+1}}\mathbb{E}_\theta\left[\sum_{i=1}^{N_{l+1}}\sum_{j=1}^{N_{l_0}}\left(\sum_{k=1}^{N_l}\frac{\partial h_i^{l+1}}{\partial h_k^{l}}\frac{\partial h_k^{l}}{\partial h_j^{l_0}}\right)\left(\sum_{m=1}^{N_l}\frac{\partial h_i^{l+1}}{\partial h_m^{l}}\frac{\partial h_m^{l}}{\partial h_j^{l_0}}\right)\right] \\
&= \frac{1}{N_{l+1}}\mathbb{E}_\theta\left[\sum_{i=1}^{N_{l+1}}\sum_{j=1}^{N_{l_0}}\sum_{k,m=1}^{N_l}\left(w_{ik}^{l+1}\widetilde{\phi'(h_k^l)}\right)\left(w_{im}^{l+1}\widetilde{\phi'(h_m^l)}\right)\left(\frac{\partial h_k^{l}}{\partial h_j^{l_0}}\frac{\partial h_m^{l}}{\partial h_j^{l_0}}\right)\right] \\
&= \frac{\sigma_w^2}{N_l}\sum_{k=1}^{N_l}\sum_{j=1}^{N_{l_0}}\mathbb{E}_\theta\left[\widetilde{\phi'(h_k^l)}\widetilde{\phi'(h_k^l)}\left(\sum_{j=1}^{N_{l_0}}\frac{\partial h_k^{l}}{\partial h_j^{l_0}}\frac{\partial h_k^{l}}{\partial h_j^{l_0}}\right)\right] \\
&= \frac{\sigma_w^2}{N_l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\widetilde{\phi'(h_k^l)}^2\right]\mathcal{J}^{l_0,l} \\
&= \sigma_w^2\frac{\mathbb{E}_\theta\left[\phi'(h^l)^2\right]}{\mathbb{E}_\theta\left[\phi(h^l)^2\right]-\mathbb{E}_\theta\left[\phi(h^l)\right]^2}\mathcal{J}^{l_0,l} \\
&= \chi_J^l\mathcal{J}^{l_0,l}\,,
\end{aligned}
\tag{56}
$$

□

D Critical Exponents

To prove theorem 2.7, we first need to find the critical exponent of the NNGP kernel [40].

Lemma D.1. *In the infinite width limit, consider a critically initialized network with a activation function ϕ. The scaling behavior of the fluctuation $\delta\mathcal{K}^l \equiv \mathcal{K}^l - \mathcal{K}^\star$ in non-exponential. If the recurrence relation can be expand to leading order δK^l as $\delta\mathcal{K}^{l+1} \approx \delta\mathcal{K}^l - c_n(\delta\mathcal{K}^l)^n$ for $n \geq 2$. The solution of $\delta\mathcal{K}^l$ is*

$$
\delta\mathcal{K}^l = \frac{1}{c_n(n-1)}\,l^{-\zeta_\mathcal{K}}\,, \tag{57}
$$

where $\zeta_\mathcal{K} = \frac{1}{n-1}$.

Remark D.2. The constant c_n and the order of first non-zero term n is determined by the choice of activation function.

Proof. We can expand the recurrence relation for the NNGP kernel (10) to second order of $\delta\mathcal{K}^l = \mathcal{K}^l - \mathcal{K}^\star$ on both side.

$$
\delta\mathcal{K}^{l+1} \approx \delta\mathcal{K}^l - c_n(\delta\mathcal{K}^l)^n\,. \tag{58}
$$

Use power law ansatz $\delta\mathcal{K}^l = A\,l^{-\zeta_\mathcal{K}}$ then

$$
(l+1)^{-\zeta_\mathcal{K}} = l^{-\zeta_\mathcal{K}} - c_n A\,l^{-n\zeta_\mathcal{K}}\,. \tag{59}
$$

Multiply $l^{\zeta_\mathcal{K}}$ on both side then use Taylor expansion $(\frac{l}{l+1})^{\zeta_\mathcal{K}} \approx 1 - \frac{\zeta_\mathcal{K}}{l}$

$$
\frac{\zeta_\mathcal{K}}{l} = c_n A l^{-(n-1)\zeta_\mathcal{K}}\,. \tag{60}
$$

For arbitrary l, the only non-trivial solution of the equation above is

$$
A = \frac{1}{c_n(n-1)} \text{ and } \zeta_\mathcal{K} = \frac{1}{n-1}\,. \tag{61}
$$

□

Proof of theorem 2.7. We will assume $c_2 \neq 0$. Then use lemma D.1, we can expand $\chi^l_{\mathcal{J}}$ in terms of $\delta\mathcal{K}^l$. To leading order l^{-1}

$$\begin{aligned}\chi^l_{\mathcal{J}} \approx& 1 - d_1 \delta\mathcal{K}^l \\ =& 1 - \frac{d_1}{c_2} l^{-1} \,. \end{aligned} \tag{62}$$

Consider a sufficiently large l. In this case $O(l^{-1})$ approximation is valid. We write recurrence relations of Jacobians as

$$\begin{aligned}\mathcal{J}^{l_0,l} &= \prod_{l'=l_0}^{l-1} \left(1 - \frac{d_1}{c_2} l'^{-1}\right) \mathcal{J}^{l_0,l_0} \\ &\approx c_{l_0} \cdot l^{-\zeta} \,. \end{aligned} \tag{63}$$

When $c_n = 0$ for all $n \geq 2$, from lemma D.1 we have $\delta\mathcal{K}^l = 0$. Thus the Jacobian saturates to some constant. □

We checked the scaling empirically by plotting $\mathcal{J}^{0,l}$ vs. l in a log–log plot and fitting the slope. These results are presented in Fig.1.

E Residual Connections

Definition E.1. We define residual connections by the modified the recurrence relation for preactivations (Eq.(1))

$$h^{l+1}_i = \sum_{j=1}^{N_l} w^{l+1}_{ij} \phi(h^l_j) + b^{l+1}_i + \mu h^l_i \,, \tag{64}$$

where the parameter μ controls the strength of the residual connection.

Remark E.2. Note that this definition requires $N_{l+1} = N_l$. We ensure this by only adding residual connections to the hidden layers, which are of the same width. More generally, one can introduce a tensor parameter μ_{ij}.

Remark E.3. In general, the parameter μ could be layer-dependent (μ^l). But we suppress this dependence here since we are discussing self-similar networks.

Theorem E.4. *In the infinite width limit, the recurrence relation for the NNGP kernel with residual connections is changed by an additional term controlled by μ*

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(h^l_j)\phi(h^l_j)\right] + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \tag{65}$$

Proof.

$$\begin{aligned}\mathcal{K}^{l+1} &= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[h^{l+1}_i h^{l+1}_i\right] \\ &= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\left(\sum_{j=1}^{N_l} w^{l+1}_{ij}\phi(h^l_j) + b^{l+1}_i + \mu h^l_i\right)\left(\sum_{k=1}^{N_l} w^{l+1}_{ik}\phi(h^l_k) + b^{l+1}_i + \mu h^l_i\right)\right] \\ &= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\sum_{j=1}^{N_l}\sum_{k=1}^{N_l} w^{l+1}_{ij} w^{l+1}_{ik} \phi(h^l_j)\phi(h^l_k) + b^{l+1}_i b^{l+1}_i + \mu^2 h^l_i h^l_i\right] \\ &= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(h^l_j)\phi(h^l_j) + \sigma_b^2\right] + \mu^2 \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[h^l_i h^l_i\right] \\ &= \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(h^l_j)\phi(h^l_j)\right] + \sigma_b^2 + \mu^2 \mathcal{K}^l \,, \end{aligned} \tag{66}$$

where we used the fact $N_{l+1} = N_l$ to get the last line. □

Theorem E.5. *In the infinite width limit, the recurrence relation for partial Jacobians with residual connections has a simple multiplicative form*

$$\mathcal{J}^{l_0,l+1} = \chi^l_{\mathcal{J}} \mathcal{J}^{l_0,l}\,, \tag{67}$$

where the recurrence coefficient is shifted to $\chi^l_{\mathcal{J}} = \sigma_w^2 \mathbb{E}_\theta \left[\phi'(h^l_k)\phi'(h^l_k)\right] + \mu^2$.

Proof.

$$\begin{aligned}
\mathcal{J}^{l_0,l+1} &\equiv \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \frac{\partial h^{l+1}_i}{\partial h^{l_0}_j} \frac{\partial h^{l+1}_i}{\partial h^{l_0}_j} \right] \\
&= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \left(\sum_{k=1}^{N_l} \frac{\partial h^{l+1}_i}{\partial h^l_k} \frac{\partial h^l_k}{\partial h^{l_0}_j} \right) \left(\sum_{m=1}^{N_l} \frac{\partial h^{l+1}_i}{\partial h^l_m} \frac{\partial h^l_m}{\partial h^{l_0}_j} \right) \right] \\
&= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} \left(w^{l+1}_{ik} \phi'(h^l_k) + \mu \delta_{ik} \right) \left(w^{l+1}_{im} \phi'(h^l_m) + \mu \delta_{im} \right) \left(\frac{\partial h^l_k}{\partial h^{l_0}_j} \frac{\partial h^l_m}{\partial h^{l_0}_j} \right) \right] \\
&= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} \left(w^{l+1}_{ik} w^{l+1}_{im} \phi'(h^l_k)\phi'(h^l_m) + \mu^2 \delta_{ik}\delta_{im} \right) \frac{\partial h^l_k}{\partial h^{l_0}_j} \frac{\partial h^l_m}{\partial h^{l_0}_j} \right] \\
&= \frac{\sigma_w^2}{N_l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(h^l_k)\phi'(h^l_k) \left(\sum_{j=1}^{N_{l_0}} \frac{\partial h^l_k}{\partial h^{l_0}_j} \frac{\partial h^l_k}{\partial h^{l_0}_j} \right) \right] + \frac{1}{N_l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\mu^2 \left(\sum_{j=1}^{N_{l_0}} \frac{\partial h^l_k}{\partial h^{l_0}_j} \frac{\partial h^l_k}{\partial h^{l_0}_j} \right) \right] \\
&= \left(\sigma_w^2 \mathbb{E}_\theta \left[\phi'(h^l_k)\phi'(h^l_k) \right] + \mu^2 \right) \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{k=1}^{N_l} \sum_{j=1}^{N_{l_0}} \frac{\partial h^l_k}{\partial h^{l_0}_j} \frac{\partial h^l_k}{\partial h^{l_0}_j} \right] \\
&= \left(\sigma_w^2 \mathbb{E}_\theta \left[\phi'(h^l_k)\phi'(h^l_k) \right] + \mu^2 \right) \mathcal{J}^{l_0,l} \\
\mathcal{J}^{l_0,l+1} &= \chi^l_{\mathcal{J}} \mathcal{J}^{l_0,l}\,.
\end{aligned} \tag{68}$$

□

F Residual Connections with LayerNorm on Preactivations (Pre-LN)

We recall the recurrence relation (1):

$$h^{l+1}_i = \sum_{j=1}^{N_l} w^{l+1}_{ij} \phi(\tilde{h}^l_j) + b^{l+1}_i + \mu h^l_i\,. \tag{69}$$

Theorem F.1. *In the infinite width limit, the recurrence relation for the NNGP kernel is then modified to*

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(\tilde{h}^l_j)\phi(\tilde{h}^l_j) \right] + \sigma_b^2 + \mu^2 \mathcal{K}^l\,. \tag{70}$$

Proof.

$$\begin{aligned}
\mathcal{K}^{l+1} =& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[h^{l+1}_i h^{l+1}_i \right] \\
=& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\left(\sum_{j=1}^{N_l} w^{l+1}_{ij} \phi(\tilde{h}^l_j) + b^{l+1}_i + \mu h^l_i \right) \left(\sum_{k=1}^{N_l} w^{l+1}_{ik} \phi(\tilde{h}^l_k) + b^{l+1}_i + \mu h^l_i \right) \right] \\
=& \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(\tilde{h}^l_j)\phi(\tilde{h}^l_j) \right] + \sigma_b^2 + \mu^2 \mathcal{K}^l\,.
\end{aligned} \tag{71}$$

□

Remark F.2. For $\mu < 1$, the recursion relation has a fixed point

$$\mathcal{K}^\star = \frac{\sigma_w^2}{N_{l^\star}(1-\mu^2)} \sum_{j=1}^{N_{l^\star}} \mathbb{E}_\theta \left[\phi(\tilde{h}_j^{l^\star}) \phi(\tilde{h}_j^{l^\star}) \right] + \frac{\sigma_b^2}{1-\mu^2} \,. \tag{72}$$

where the average here is exactly the same as cases for LayerNorm applied to preactivations without residue connections. $l^\star$ labels some very large depth l.

Remark F.3. For $\mu = 1$ case, the solution of (70) is

$$\mathcal{K}^l = \mathcal{K}^0 + \sum_{l'=1}^{l} \left(\frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(\tilde{h}_j^{l'}) \phi(\tilde{h}_j^{l'}) \right] + \sigma_b^2 \right) \,. \tag{73}$$

which is linearly growing since the expectation does not depend on depth. $\mathcal{K}^0$ is the NNGP kernel after the input layer.

Theorem F.4. *In the infinite width limit, the recurrence relation for Jacobians changes by a constant shift in the recursion coefficient.*

$$\mathcal{J}^{l_0,l+1} = \chi_{\mathcal{J}}^l \mathcal{J}^{l_0,l} \,, \tag{74}$$

where for this case

$$\chi_{\mathcal{J}}^l = \frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}_k^l) \phi'(\tilde{h}_k^l) \right] + \mu^2 \,. \tag{75}$$

Proof.

$$\begin{aligned}
\mathcal{J}^{l_0,l+1} &= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \right] \\
&= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \left(\sum_{k=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial \tilde{h}_k^l} \frac{\partial \tilde{h}_k^l}{\partial h_k^l} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \left(\sum_{m=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial \tilde{h}_m^l} \frac{\partial \tilde{h}_m^l}{\partial h_m^l} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
&= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} \left(\frac{w_{ik}^{l+1} \phi'(\tilde{h}_k^l)}{\sqrt{\mathcal{K}^l}} + \mu \delta_{ik} \right) \left(\frac{w_{im}^{l+1} \phi'(\tilde{h}_m^l)}{\sqrt{\mathcal{K}^l}} + \mu \delta_{ik} \right) \left(\frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
&= \mathbb{E}_\theta \left[\left(\frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \phi'(\tilde{h}_k^l) \phi'(\tilde{h}_k^l) + \mu^2 \right) \left(\sum_{j=1}^{N_{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \right] \\
&= \left(\frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}_k^l) \phi'(\tilde{h}_k^l) \right] + \mu^2 \right) \mathcal{J}^{l_0,l} \\
&= \chi_{\mathcal{J}}^l \mathcal{J}^{l_0,l} \,,
\end{aligned} \tag{76}$$

□

Remark F.5. One can directly use results from cases without residue connections. We will momentarily see that the phase boundary does not change with residual connections when $\mu < 1$. However, the correlation length decays way slower when the network is initialized far from criticality.

Remark F.6. As we mentioned above $\mu = 1$ needs extra care. Plug in the result (73) and $\mu = 1$ we find out that

$$\begin{aligned}
\chi_{\mathcal{J}}^l \mid_{\mu=1} &= \frac{\sigma_w^2 \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}_k^l) \phi'(\tilde{h}_k^l) \right]}{N_l \mathcal{K}^0 + \sum_{l'=1}^{l} \left(\sigma_w^2 \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(\tilde{h}_j^l) \phi(\tilde{h}_j^l) \right] + N_l \sigma_b^2 \right)} + 1 \\
&\sim 1 + O\left(\frac{1}{l} \right) \,,
\end{aligned} \tag{77}$$

which leads to power law behaved Jacobians at large depth. Where the exponent ζ is not universal.

Recall that $\xi = |\log \chi_{\mathcal{J}}^\star|^{-1}$, then Theorem 1.3 is a summary of (72) and (75) for $\mu < 1$; and (77) for $\mu = 1$ in $l \to \infty$ limit.

G MLP-Mixer

In this section we would like to analyze an architecture called MLP-Mixer [43], which is based on multi-layer perceptrons (MLPs). A MLP-Mixer (i) chops images into patches, then applies affine transformations per patch, (ii) applies several Mixer Layers, (iii) applies pre-head LayerNorm, Global Average Pooling, an output affine transformation. We will explain the architecture by showing forward pass equations.

Suppose one has a single input with dimension (C_{in}, H_{in}, W_{in}). We label it as $x_{\mu i}$, where the Greek letter labels channels and the Latin letter labels flattened pixels.

First of all the (i) is realized by a special convolutional layer, where kernel size f is equal to the stride s. Then first convolution layer can be written as

$$h^0_{\mu i} = \sum_{j=1}^{f^2} \sum_{\nu=1}^{C_{in}} W^0_{\mu\nu;j} x_{\nu,j+(i-1)s^2} + b^0_{\mu i}\,, \tag{78}$$

where f is the size of filter and s is the stride. In our example $f = s$. Notice in PyTorch both bias and weights are sampled from a uniform distribution $\mathcal{U}(-\sqrt{k}, \sqrt{k})$, where $k = (C_{in} f^2)^{-1}$.

$$\mathbb{E}_\theta[W^0_{\mu\nu;i} W^0_{\rho\sigma;j}] = \frac{1}{3C_{in}f^2}\delta_{\mu\rho}\delta_{\nu\sigma}\delta_{ij}\,, \tag{79}$$

$$\mathbb{E}_\theta[b^0_{\mu i} b^0_{\nu j}] = \frac{1}{3C_{in}f^2}\delta_{\mu\nu}\delta_{ij}\,. \tag{80}$$

Notice that the output of Conv2d: $h^0_{\mu i} \in \mathbb{R}^{C \times N_p}$, where C stands for channels and $N_p = H_{in}W_{in}/f^2$ stands for patches, both of them will be mixed later by Mixer layers.

Next we stack l Mixer Layers. A Mixer Layer contains LayerNorms and two MLPs, where the first one mixed patches i, j (token mixing) with a hidden dimension N_{tm}, the second one mixed channels μ, ν (channel-mixing) with a hidden dimension N_{cm}. Notice that for Mixer Layers we use the standard parameterization.

- First LayerNorm. It acts on channels μ.

$$\tilde{h}^{6l}_{\mu i} = \frac{h^{6l}_{\mu i} - \mathbb{E}_C[h^{6l}_{\rho i}]}{\sqrt{\mathrm{Var}_C[h^{6l}_{\rho i}]}}\,, \tag{81}$$

  where we defined a channel mean $\mathbb{E}_C[h^{6l}_{\rho i}] \equiv \frac{1}{C}\sum_{\rho=1}^{C} h^{6l}_{\rho i}$ and channel variance $\mathrm{Var}_C \equiv \mathbb{E}_C\left[\left(h^{6l}_{\rho i}\right)^2\right] - \left(\mathbb{E}_C[h^{6l}_{\rho i}]\right)^2$.

- First MlpBlock. It mixes patches i, j, preactivations from different channels share the same weight and bias.
  - $6l + 1$: Linear Affine Layer.

$$h^{6l+1}_{\mu j} = \sum_{k=1}^{N_p} w^{6l+1}_{jk} \tilde{h}^{6l}_{\mu k} + b^{6l+1}_{j}\,. \tag{82}$$

  - $6l + 2$: Affine Layer.

$$h^{6l+2}_{\mu i} = \sum_{j=1}^{N_{tm}} w^{6l+2}_{ij} \phi(h^{6l+1}_{\mu j}) + b^{6l+2}_{i}\,, \tag{83}$$

    where N_{tm} stands for hidden dimension of "token mixing".
  - $6l + 3$: Residual Connections.

$$h^{6l+3}_{\mu i} = h^{6l+2}_{\mu i} + \mu h^{6l}_{\mu i}\,. \tag{84}$$

- Second LayerNorm. It again acts on channels μ.

$$\tilde{h}^{6l+3}_{\mu i} = \frac{h^{6l+3}_{\mu i} - \mathbb{E}_C[h^{6l+3}_{\rho i}]}{\sqrt{\mathrm{Var}_C[h^{6l+3}_{\rho i}]}}\,. \tag{85}$$

- Second MlpBlock. It mixes channels μ, ν, preactivations from different patches share the same weight and bias.
  - $6l+4$: Linear Affine Layer.

$$h^{6l+4}_{\nu i} = \sum_{\rho=1}^{C} w^{6l+4}_{\nu\rho}\tilde{h}^{6l+3}_{\rho i} + b^{6l+4}_{\nu}\,. \tag{86}$$

  - $6l+5$. Affine Layer.

$$h^{6l+5}_{\mu i} = \sum_{\nu=1}^{N_{cm}} w^{6l+5}_{\mu\nu}\phi(h^{6l+4}_{\nu i}) + b^{6l+5}_{\mu}\,. \tag{87}$$

  - $6l+6$. Residual Connections.

$$h^{6l+6}_{\mu i} = h^{6l+5}_{\mu i} + \mu h^{6l+3}_{\mu i}. \tag{88}$$

Suppose the network has L Mixer layers. After those layers the network has a pre-head LayerNorm layer, a global average pooling layer and a output layer. The pre-head LayerNorm normalizes over channels μ can be described as the following

$$\tilde{h}^{6L}_{\mu i} = \frac{h^{6L}_{\mu i} - \mathbb{E}_C[h^{6L}_{\rho i}]}{\sqrt{\mathrm{Var}_C[h^{6L}_{\rho i}]}}\,. \tag{89}$$

Global Average Pool over patches i.

$$h^{p}_{\mu} = \frac{1}{N_p}\sum_{i=1}^{N_p}\tilde{h}^{6L}_{\mu i}\,. \tag{90}$$

Output Layer

$$f_{\mu} = \sum_{\nu=1}^{C} w_{\mu\nu}h^{p}_{\nu} + b_{\mu}\,. \tag{91}$$

We plotted phase diagram using the following quantity from repeating Mixer Layers:

$$\chi^{\star}_{\mathcal{J}} = \lim_{L\to\infty}\left(\frac{1}{N_p C}\sum_{i=1}^{N_p}\sum_{\mu=1}^{C}\mathbb{E}_{\theta}\left[\sum_{\rho=1}^{C}\sum_{k=1}^{N_p}\frac{\partial h^{6L}_{\mu i}}{\partial h^{6L-6}_{\rho k}}\frac{\partial h^{6L}_{\mu i}}{\partial h^{6L-6}_{\rho k}}\right]\right)\,. \tag{92}$$

H Results for Scale Invariant Activation Functions

Definition H.1 (Scale invariant activation functions)**.**

$$\phi(x) = a_+\, x\,\Theta(x) + a_-\, x\,\Theta(-x)\,, \tag{93}$$

where $\Theta(x)$ is the Heaviside step function. ReLU is the special case with $a_+ = 1$ and $a_- = 0$.

H.1 NNGP Kernel

First evaluate the average using lemma 2.2

$$\begin{aligned}\mathbb{E}_\theta\left[\phi(h_i^l)\phi(h_i^l)\right] =& \frac{1}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l \left(a_+^2+a_-^2\right)\left(h_i^l\right)^2 e^{-\frac{(h_i^l)^2}{2K^l}} \\ =& \frac{a_+^2+a_-^2}{2}\mathcal{K}^l\,. \end{aligned} \tag{94}$$

Thus we obtain the recurrence relation for the NNGP kernel with scale invariant activation function.

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2(a_+^2+a_-^2)}{2}\mathcal{K}^l + \sigma_b^2\,. \tag{95}$$

Finite fixed point of the recurrence relation above exists only if

$$\chi_{\mathcal{K}}^\star = \frac{\sigma_w^2(a_+^2+a_-^2)}{2} \leq 1\,. \tag{96}$$

As a result

$$\sigma_w^2 \leq \frac{2}{a_+^2+a_-^2}\,. \tag{97}$$

For $\sigma_w^2 = \frac{2}{a_+^2+a_-^2}$ case, finite fixed point exists only if $\sigma_b^2 = 0$.

H.2 Jacobian(s)

The calculation is quite straight forward, by definition

$$\begin{aligned}\chi_{\mathcal{J}}^l =& \sigma_w^2\mathbb{E}_\theta\left[\phi'(h_i^l)\phi'(h_i^l)\right] \\ =& \frac{\sigma_w^2}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l \left[a_+\Theta(h_i^l) - a_-\Theta(h_i^l)\right]^2 e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\ =& \frac{\sigma_w^2(a_+^2+a_-^2)}{2}\,, \end{aligned} \tag{98}$$

where we used the property $x\delta(x) = 0$ for Dirac's delta function to get the first line.

Thus the critical line is defined by

$$\sigma_w = \sqrt{\frac{2}{a_+^2+a_-^2}}\,. \tag{99}$$

For ReLU with $a_+ = 1$ and $a_- = 0$, the network is at critical line when

$$\sigma_w = \sqrt{2}\,, \tag{100}$$

where the critical point is located at

$$(\sigma_w, \sigma_b) = (\sqrt{2}, 0)\,. \tag{101}$$

H.3 Critical Exponents

Since the recurrence relations for the NNGP kernel and Jacobians are linear. Then from lemma D.1 and theorem 2.7

$$\zeta_{\mathcal{K}} = 0 \text{ and } \zeta = 0\,. \tag{102}$$

H.4 LayerNorm on Pre-activations

Use lemma C.9 and combine all known results for scale invariant functions

$$\begin{aligned}\chi^l_{\mathcal{J}} &= \frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}^l_k) \phi'(\tilde{h}^l_k) \right] \Bigg|_{\tilde{\mathcal{K}}^{l-1}=1} \\ &= \frac{\sigma_w^2 (a_+^2 + a_-^2)}{\sigma_w^2 (a_+^2 + a_-^2) + 2\sigma_b^2} \,. \end{aligned} \tag{103}$$

For this case,

$$\chi^l_{\mathcal{J}} \leq 1 \tag{104}$$

is always true. The equality only holds at $\sigma_b = 0$ line.

H.5 LayerNorm on Activations

First we substitute $\mathcal{K}^{l-1} = \sigma_w^2 + \sigma_b^2$ into known results

$$\mathbb{E}_\theta \left[\phi'(h^l_i) \phi'(h^l_i) \right] = \frac{a_+^2 + a_-^2}{2} \,, \tag{105}$$

$$\mathbb{E}_\theta \left[\phi(h^l_i) \phi(h^l_i) \right] = \frac{a_+^2 + a_-^2}{2} (\sigma_w^2 + \sigma_b^2) \,. \tag{106}$$

There is a new expectation value we need to show explicitly

$$\begin{aligned}\mathbb{E}_\theta \left[\phi(h^l_i) \right] &= \frac{1}{\sqrt{2\pi(\sigma_w^2 + \sigma_b^2)}} \int_{-\infty}^{\infty} dh^l_i \phi(h^l_i) e^{-\frac{1}{2} h^l_i (\sigma_w^2 + \sigma_b^2)^{-1} h^l_i} \\ &= \frac{1}{\sqrt{2\pi(\sigma_w^2 + \sigma_b^2)}} \int_{0}^{\infty} dh^l_i (a_+ - a_-) h^l_i e^{-\frac{(h^l_i)^2}{2(\sigma_w^2 + \sigma_b^2)}} \\ &= (a_+ - a_-) \sqrt{\frac{\sigma_w^2 + \sigma_b^2}{2\pi}} \,. \end{aligned} \tag{107}$$

Thus

$$\chi^l_{\mathcal{J}} = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_b^2} \cdot \frac{\pi(a_+^2 + a_-^2)}{\pi(a_+^2 + a_-^2) - (a_+ - a_-)^2} \,. \tag{108}$$

The critical line is defined by $\chi^\star_{\mathcal{J}} = 1$, which can be solved as

$$\sigma_b = \sqrt{\frac{(a_+ - a_-)^2}{\pi(a_+^2 + a_-^2) - (a_+ - a_-)^2}} \sigma_w \,. \tag{109}$$

For ReLU with $a_+ = 1$ and $a_- = 0$

$$\begin{aligned}\sigma_b =& \sqrt{\frac{1}{\pi - 1}} \sigma_w \\ \approx& 0.683 \sigma_w \,. \end{aligned} \tag{110}$$

H.6 Residual Connections

The recurrence relation for the NNGP kernel can be evaluated to be

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2 (a_+^2 + a_-^2)}{2} \mathcal{K}^l + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \tag{111}$$

The condition for the existence of fixed point

$$\chi^\star_{\mathcal{K}} = \frac{\sigma_w^2 (a_+^2 + a_-^2)}{2} + \mu^2 \leq 1 \tag{112}$$

leads us to

$$\sigma_w^2 \leq \frac{2(1-\mu^2)}{a_+^2 + a_-^2}\,. \tag{113}$$

For $\sigma_w^2 = \frac{2(1-\mu^2)}{a_+^2+a_-^2}$, finite fixed point exists only if $\sigma_b^2 = 0$. (Diverges linearly otherwise)

The recurrence coefficient for Jacobian is evaluated to be

$$\chi_{\mathcal{J}}^{\star} = \frac{\sigma_w^2(a_+^2 + a_-^2)}{2} + \mu^2\,. \tag{114}$$

The critical line is defined as

$$\sigma_w = \sqrt{\frac{2(1-\mu^2)}{a_+^2 + a_-^2}}\,. \tag{115}$$

The critical point is located at $\left(\sqrt{\frac{2(1-\mu^2)}{a_+^2+a_-^2}}, 0\right)$.

For ReLU, the critical point is at $\left(\sqrt{2(1-\mu^2)}, 0\right)$.

H.7 Residual Connections with LayerNorm on Preactivations (Pre-LN)

Again use lemma C.9 and combine all known results for scale invariant functions

$$\begin{aligned}
\chi_{\mathcal{J}}^{\star} &= \lim_{l\to\infty}\left(\frac{\sigma_w^2}{N_l\mathcal{K}^l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l)\right]\Big|_{\tilde{\mathcal{K}}^{l-1}=1} + \mu^2\right) \\
&= \frac{\sigma_w^2(a_+^2+a_-^2)(1-\mu^2)}{\sigma_w^2(a_+^2+a_-^2)+2\sigma_b^2} + \mu^2 \\
&= 1 - \frac{2\sigma_b^2(1-\mu^2)}{\sigma_w^2(a_+^2+a_-^2)+2\sigma_b^2}
\end{aligned} \tag{116}$$

Similar to the case without residue connections

$$\chi_{\mathcal{J}}^{l} \leq 1 \tag{117}$$

is always true. The equality only holds at $\sigma_b = 0$ line for $\mu < 1$.

Notice there is a very special case $\mu = 1$, where the whole $\sigma_b - \sigma_w$ plane is critical.

I Results for erf Activation Function

Definition I.1 (erf activation function)**.**

$$\phi(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt\,. \tag{118}$$

I.1 NNGP Kernel

To evaluate lemma 2.2 exactly, we introduce two dummy variables λ_1 and λ_2[45].

$$\begin{aligned}
\mathbb{E}_\theta\left[\phi(\lambda_1 h_i^l)\phi(\lambda_2 h_i^l)\right] &= \int d\lambda_1\int d\lambda_2 \frac{d^2}{d\lambda_1 d\lambda_2}\mathbb{E}_\theta\left[\phi(\lambda_1 h_i^l)\phi(\lambda_2 h_i^l)\right] \\
&= \int d\lambda_1\int d\lambda_2\int dh_i^l \frac{4}{\sqrt{2\pi^3\mathcal{K}^l}}\left(h_i^l\right)^2 e^{-\left(\lambda_1^2+\lambda_2^2+\frac{1}{2\mathcal{K}^l}\right)\left(h_i^l\right)^2} \\
&= \int d\lambda_1\int d\lambda_2 \frac{4\mathcal{K}^l}{\pi\left(1+2\mathcal{K}^l(\lambda_1^2+\lambda_2^2)\right)} \\
&= \frac{2}{\pi}\arcsin\left(\frac{2\mathcal{K}^l\lambda_1\lambda_2}{1+2\mathcal{K}^l(\lambda_1^2+\lambda_2^2)}\right)\,.
\end{aligned} \tag{119}$$

We use the special case where $\lambda_1 = \lambda_2 = 1$.

Thus the recurrence relation for the NNGP kernel with erf activation function is

$$\mathcal{K}^{l+1} = \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{2\mathcal{K}^l}{1+2\mathcal{K}^l}\right) + \sigma_b^2 \,. \tag{120}$$

As in scale invariant case, finite fixed point only exists when

$$\chi_{\mathcal{K}}^\star = \frac{4\sigma_w^2}{\pi} \frac{1}{(1+2\mathcal{K}^\star)\sqrt{1+4\mathcal{K}^\star}} \leq 1 \,. \tag{121}$$

Numerical results show the condition is satisfied everywhere in $\sigma_b - \sigma_w$ plane, where $\chi_{\mathcal{K}}^\star = 1$ is only possible when $\mathcal{K}^\star = 0$.

I.2 Jacobians

Follow the definition

$$\begin{aligned}
\chi_{\mathcal{J}}^l &= \sigma_w^2 \mathbb{E}_\theta \left[\phi'(h_i^l)\phi'(h_i^l)\right] \\
&= \frac{4\sigma_w^2}{\sqrt{2\pi^3\mathcal{K}^l}} \int dh_i^l \, e^{-2(h_i^l)^2} \, e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
&= \frac{4\sigma_w^2}{\pi} \frac{1}{\sqrt{1+4\mathcal{K}^l}} \,.
\end{aligned} \tag{122}$$

To find phase boundary $\chi_{\mathcal{J}}^\star = 1$, we need to combine Eq.(120) and Eq.(122) and evaluate them at $\mathcal{K}^\star$.

$$\mathcal{K}^\star = \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{2\mathcal{K}^\star}{1+2\mathcal{K}^\star}\right) + \sigma_b^2 \,, \tag{123}$$

$$\chi_{\mathcal{J}}^\star = \frac{4\sigma_w^2}{\pi} \frac{1}{\sqrt{1+4\mathcal{K}^\star}} = 1 \,. \tag{124}$$

One can solve equations above and find the critical line

$$\sigma_b = \sqrt{\frac{16\sigma_w^4 - \pi^2}{4\pi^2} - \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{16\sigma_w^4 - \pi^2}{16\sigma_w^4 + \pi^2}\right)} \,. \tag{125}$$

Critical point is reached by further requiring $\chi_{\mathcal{K}}^\star = 1$. Since $\chi_{\mathcal{K}}^\star \leq \chi_{\mathcal{J}}^\star$, the only possible case is $\mathcal{K}^\star = 0$, which is located at

$$(\sigma_w, \sigma_b) = \left(\sqrt{\frac{\pi}{4}}, 0\right) \,. \tag{126}$$

I.3 Critical Exponents

We show how to extract critical exponents of the NNGP kernel and Jacobians of erf activation function.

Critical point for erf is at $(\sigma_b, \sigma_w) = (0, \sqrt{\frac{\pi}{4}})$, with $\mathcal{K}^\star = 0$. Now suppose l is large enough such that the deviation of $\mathcal{K}^l$ from fixed point value $\mathcal{K}^\star$ is small. Define $\delta\mathcal{K}^l \equiv \mathcal{K}^l - \mathcal{K}^\star$. Eq.(120) can be rewritten as

$$\begin{aligned}
\delta\mathcal{K}^{l+1} =& \frac{1}{2} \arcsin\left(\frac{2\delta\mathcal{K}^l}{1+2\delta\mathcal{K}^l}\right) \\
\approx& \delta\mathcal{K}^l - 2(\delta\mathcal{K}^l)^2 \,.
\end{aligned} \tag{127}$$

From lemma D.1

$$A = \frac{1}{2} \text{ and } \zeta_{\mathcal{K}} = 1 \,. \tag{128}$$

Next we analyze critical exponent of Jacobians by expanding (122) around $\mathcal{K}^\star = 0$ critical point $(\sigma_b, \sigma_w) = (0, \sqrt{\frac{\pi}{4}})$.

To leading order l^{-1} we have

$$\begin{aligned}\chi^l_{\mathcal{J}} \approx& 1 - 2\delta K^l \\ \approx& 1 - \frac{1}{l}\,.\end{aligned} \tag{129}$$

Thus the recurrence relation for partial Jacobian, at large l, takes form

$$\mathcal{J}^{l_0,l+1} = \left(1 - \frac{1}{l}\right)\mathcal{J}^{l_0,l}\,. \tag{130}$$

At large l

$$\mathcal{J}^{l_0,l} = c_{l_0}\, l^{-1}\,, \tag{131}$$

with a non-universal constant c_{l_0}.

The critical exponent is

$$\zeta = 1\,, \tag{132}$$

which is the same as $\zeta_{\mathcal{K}}$.

I.4 LayerNorm on Pre-activations

Use lemma C.9, we have

$$\begin{aligned}\chi^l_{\mathcal{J}} =& \frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}^l_k)\phi'(\tilde{h}^l_k)\right]\Bigg|_{\tilde{\mathcal{K}}^{l-1}=1} \\ =& \frac{4\sigma_w^2}{\sqrt{5}\left[2\sigma_w^2 \arcsin\left(\frac{2}{3}\right) + \pi\sigma_b^2\right]}\,.\end{aligned} \tag{133}$$

The critical line is then defined by

$$\begin{aligned}\sigma_b =& \sqrt{\frac{2}{\pi}\left[\frac{2}{\sqrt{5}} - \arcsin\left(\frac{2}{3}\right)\right]}\sigma_w \\ \approx& 0.324\sigma_w\,.\end{aligned} \tag{134}$$

I.5 LayerNorm on Activations

Due to the symmetry of erf activation function $\mathbb{E}_\theta\left[\phi(h^l_i)\right] = 0$, we only need to modify our known results.

$$\mathbb{E}_\theta\left[\phi'(h^l_i)\phi'(h^l_i)\right] = \frac{4}{\pi}\frac{1}{\sqrt{1 + 4(\sigma_w^2 + \sigma_b^2)}}\,, \tag{135}$$

$$\mathbb{E}_\theta\left[\phi(h^l_i)\phi(h^l_i)\right] = \frac{2}{\pi}\arcsin\left(\frac{2(\sigma_w^2 + \sigma_b^2)}{1 + 2(\sigma_w^2 + \sigma_b^2)}\right)\,. \tag{136}$$

Thus

$$\chi^l_{\mathcal{J}} = \frac{2\sigma_w^2}{\sqrt{1 + 4(\sigma_w^2 + \sigma_b^2)}} \cdot \frac{1}{\arcsin\left(\frac{2(\sigma_w^2+\sigma_b^2)}{1+2(\sigma_w^2+\sigma_b^2)}\right)}\,, \tag{137}$$

where the phase boundary is defined by the transcendental equation $\chi^l_{\mathcal{J}} = 1$.

I.6 Residual Connections

The recurrence relation for the NNGP kernel can be evaluated to be

$$\mathcal{K}^{l+1} = \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{2\mathcal{K}^l}{1+2\mathcal{K}^l}\right) + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \tag{138}$$

Finite fixed point only exists when

$$\chi_{\mathcal{K}}^\star = \frac{4\sigma_w^2}{\pi} \frac{1}{(1+2\mathcal{K}^\star)\sqrt{1+4\mathcal{K}^\star}} + \mu^2 \leq 1 \,. \tag{139}$$

Notice that $\chi_{\mathcal{K}}^\star \leq \chi_{\mathcal{J}}^\star$ still holds, where the equality holds only when $\mathcal{K}^\star = 0$.

The recurrence coefficient for Jacobian is evaluated to be

$$\chi_{\mathcal{J}}^\star = \frac{4\sigma_w^2}{\pi} \frac{1}{\sqrt{1+4\mathcal{K}^\star}} + \mu^2 \,. \tag{140}$$

The critical line is defined as

$$\sigma_b = \sqrt{\frac{16\sigma_w^4 - \pi^2(1-\mu^2)^2}{4\pi^2(1-\mu^2)} - \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{16\sigma_w^4 - \pi^2(1-\mu^2)^2}{16\sigma_w^4 + \pi^2(1-\mu^2)^2}\right)} \,. \tag{141}$$

Critical point is reached by further requiring $\chi_{\mathcal{K}}^\star = 1$. Since $\chi_{\mathcal{K}}^\star \leq \chi_{\mathcal{J}}^\star$, the only possible case is $\mathcal{K}^\star = 0$, which is located at

$$(\sigma_w, \sigma_b) = \left(\sqrt{\frac{\pi(1-\mu^2)}{4}}, 0\right) \,. \tag{142}$$

Note that for $\mu = 1$, one needs to put extra efforts into analyzing the scaling behavior. First we notice that $\mathcal{K}^l$ monotonically increases with depth l – the recurrence relation for the NNGP kernel at large l (or large $\mathcal{K}^l$) is

$$\mathcal{K}^{l+1} \approx \sigma_w^2 + \sigma_b^2 + \mathcal{K}^l \,, \tag{143}$$

which regulates the first term in (140).

For $\mu = 1$ at large depth

$$\chi_{\mathcal{J}}^l \sim 1 + \frac{4\sigma_w^2}{\pi\sqrt{C_0 + 4(\sigma_w^2+\sigma_b^2)l}} \,. \tag{144}$$

Here C_0 is a constant that depends on the input.

We can approximate the asymptotic form of $\log \mathcal{J}^{l_0,l}$ as follows

$$\begin{aligned}
\log \mathcal{J}^{l_0,l} &= \log\left(\prod_{l'=l_0}^{l} \chi_{\mathcal{J}}^{l'}\right) \\
&= \sum_{l'=l_0}^{l} \log\left(1 + \frac{4\sigma_w^2}{\pi\sqrt{C_0 + 4(\sigma_w^2+\sigma_b^2)l'}}\right) \\
&\approx \int_{l_0}^{l} dl' \log\left(1 + \frac{4\sigma_w^2}{\pi\sqrt{C_0 + 4(\sigma_w^2+\sigma_b^2)l'}}\right) \\
&\sim 2\tilde{c}\sqrt{l} + O(\log l) \,,
\end{aligned} \tag{145}$$

where $\tilde{c} = \frac{2\sigma_w^2}{\pi\sqrt{\sigma_w^2+\sigma_b^2}}$.

We conclude that at large depth, the APJN for $\mu = 1$, erf networks can be written as

$$\mathcal{J}^{l_0,l} \sim O\left(e^{2\tilde{c}\sqrt{l}+O(\log l)}\right) \,. \tag{146}$$

This result checks out empirically, as shown in Figure 7.[3]

[3] We used NTK parameterization for this experiment. However, we emphasize that it does not affect the final result.

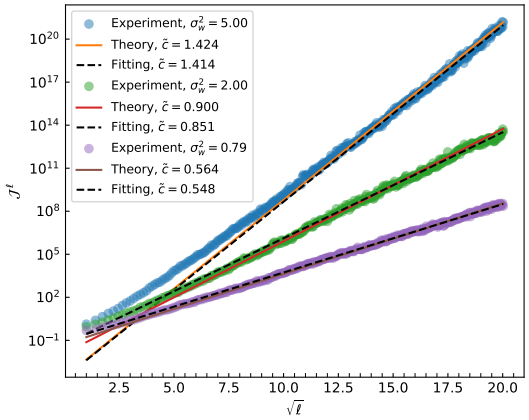

Figure 7: $\log(\mathcal{J}^{l_0,l})$-$\sqrt{l}$ for $\mu = 1$, $\sigma_b^2 = 0$, erf.

I.7 Residual Connections with LayerNorm on Preactivations (Pre-LN)

Use lemma C.9 and results we had without residue connections for erf with LayerNorm on preactivations.

$$\begin{aligned}\chi^*_{\mathcal{J}} =& \lim_{l\to\infty}\left(\frac{\sigma_w^2}{N_l\mathcal{K}^l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l)\right]\Big|_{\tilde{\mathcal{K}}^{l-1}=1} + \mu^2\right)\\ =&\frac{4\sigma_w^2(1-\mu^2)}{\sqrt{5}\left[2\sigma_w^2\arcsin\left(\frac{2}{3}\right)+\pi\sigma_b^2\right]} + \mu^2\,.\end{aligned} \tag{147}$$

The critical line is then defined by

$$\begin{aligned}\sigma_b =& \sqrt{\frac{2}{\pi}\left[\frac{2}{\sqrt{5}} - \arcsin\left(\frac{2}{3}\right)\right]}\sigma_w\\ \approx& 0.324\sigma_w\,.\end{aligned} \tag{148}$$

J Results for GELU Activation Function

Definition J.1 (GELU activation function)**.**

$$\begin{aligned}\phi(x) =& \frac{x}{2}\left[1+\text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]\\ =&\frac{x}{2}\left[1+\frac{2}{\sqrt{\pi}}\int_0^{\frac{x}{\sqrt{2}}}e^{-t^2}dt\right]\,.\end{aligned} \tag{149}$$

J.1 NNGP Kernel

Use lemma 2.2 for GELU

$$\begin{aligned}
\mathbb{E}_\theta\left[\phi(h_i^l)\phi(h_i^l)\right] =& \frac{1}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l \frac{(h_i^l)^2}{4}\left[1+\text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right)\right]^2 e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=& \frac{1}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l \frac{(h_i^l)^2}{4}\left[1+\text{erf}^2\left(\frac{h_i^l}{\sqrt{2}}\right)\right] e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=& \frac{\mathcal{K}^l}{4} + \frac{1}{\sqrt{32\pi\mathcal{K}^l}}\int dh_i^l (h_i^l)^2 \text{erf}^2\left(\frac{h_i^l}{\sqrt{2}}\right) e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=& \frac{\mathcal{K}^l}{4} + \frac{\mathcal{K}^l}{\sqrt{32\pi\mathcal{K}^l}}\int dh_i^l \,\text{erf}^2\left(\frac{h_i^l}{\sqrt{2}}\right) e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
&+ \frac{(\mathcal{K}^l)^2}{\sqrt{32\pi\mathcal{K}^l}}\int dh_i^l \left[\text{erf}'\left(\frac{h_i^l}{\sqrt{2}}\right)\text{erf}'\left(\frac{h_i^l}{\sqrt{2}}\right) + \text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right)\text{erf}''\left(\frac{h_i^l}{\sqrt{2}}\right)\right] e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=& \frac{\mathcal{K}^l}{4} + \frac{\mathcal{K}^l}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^l}{1+\mathcal{K}^l}\right) + \frac{2\mathcal{K}^l}{(1+\mathcal{K}^l)\sqrt{1+2\mathcal{K}^l}}\right], \qquad (150)
\end{aligned}$$

where from the third line to the fourth line we used integrate by parts twice, and to get the last line we used results from erf activations.

Thus the recurrence relation for the NNGP kernel is

$$\mathcal{K}^{l+1} = \left[\frac{\mathcal{K}^l}{4} + \frac{\mathcal{K}^l}{2\pi}\arcsin\left(\frac{\mathcal{K}^l}{1+\mathcal{K}^l}\right) + \frac{(\mathcal{K}^l)^2}{\pi(1+\mathcal{K}^l)\sqrt{1+2\mathcal{K}^l}}\right]\sigma_w^2 + \sigma_b^2\,. \qquad (151)$$

As a result

$$\chi_{\mathcal{K}}^\star = \frac{\sigma_w^2}{4} + \frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right) + \frac{4(\mathcal{K}^\star)^3+11(\mathcal{K}^\star)^2+5\mathcal{K}^\star}{(1+\mathcal{K}^\star)^2(1+2\mathcal{K}^\star)^{\frac{3}{2}}}\right]. \qquad (152)$$

J.2 Jacobians

Follow the definition

$$\begin{aligned}
\chi_{\mathcal{J}}^l =& \sigma_w^2\mathbb{E}_\theta\left[\phi'(h_i^l)\phi'(h_i^l)\right] \\
=& \frac{\sigma_w^2}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l\left[\frac{1}{2}+\frac{1}{2}\text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right) + \frac{e^{-\frac{(h_i^l)^2}{2}}h_i^l}{\sqrt{2\pi}}\right]^2 e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=& \frac{\sigma_w^2}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l\left[\frac{1}{4}+\frac{1}{4}\text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right)^2 + \frac{h_i^l\text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right)e^{-\frac{(h_i^l)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{-(h_i^l)^2}(h_i^l)^2}{2\pi}\right] e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=& \frac{\sigma_w^2}{4} + \frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^l}{1+\mathcal{K}^l}\right) + \frac{\mathcal{K}^l(3+5\mathcal{K}^l)}{(1+\mathcal{K}^l)(1+2\mathcal{K}^l)^{\frac{3}{2}}}\right], \qquad (153)
\end{aligned}$$

where we dropped odd function terms to get the third line, and to get the last line we used known result for erf in the second term, integrate by parts in the third term.

Here to get the critical line is harder. One can use the recurrence relation for the NNGP kernel at fixed point $\mathcal{K}^\star$ and $\chi_{\mathcal{J}}^\star = 1$

$$\mathcal{K}^\star = \frac{\sigma_w^2}{4}\mathcal{K}^\star + \frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right) + \frac{\sigma_w^2\mathcal{K}^\star}{\pi(1+\mathcal{K}^\star)\sqrt{1+2\mathcal{K}^\star}}\right]\mathcal{K}^\star + \sigma_b^2\,, \qquad (154)$$

$$\chi_{\mathcal{J}}^\star = \frac{\sigma_w^2}{4} + \frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right) + \frac{\mathcal{K}^\star(3+5\mathcal{K}^\star)}{(1+\mathcal{K}^\star)(1+2\mathcal{K}^\star)^{\frac{3}{2}}}\right] = 1\,. \qquad (155)$$

Cancel the $\arcsin$ term, σ_w and σ_b then can be written as a function of $\mathcal{K}^\star$

$$\sigma_w = 2\left[1 + \frac{2\mathcal{K}^\star(3+5\mathcal{K}^\star)}{\pi(1+\mathcal{K}^\star)(1+2\mathcal{K}^\star)^{\frac{3}{2}}} + \frac{2}{\pi}\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right)\right]^{-\frac{1}{2}}, \tag{156}$$

$$\sigma_b = \frac{\mathcal{K}^\star}{\sqrt{2\pi}(1+2\mathcal{K}^\star)^{\frac{3}{4}}}\sigma_w\,. \tag{157}$$

One can then scan $\mathcal{K}^\star$ to draw the critical line.

In order to locate critical point, we further require $\chi^\star_{\mathcal{K}} = 1$. To locate the critical point, we solve $\chi^\star_{\mathcal{J}} - \chi^\star_{\mathcal{K}} = 0$ instead. We have

$$\frac{\sigma_w^2[(\mathcal{K}^\star)^3 - 3(\mathcal{K}^\star)^2 - 2\mathcal{K}^\star]}{2\pi(1+\mathcal{K}^\star)^2(1+2\mathcal{K}^\star)^{\frac{3}{2}}} = 0\,, \tag{158}$$

which has two non-negative solutions out of three

$$\mathcal{K}^\star = 0 \text{ and } \mathcal{K}^\star = \frac{3+\sqrt{17}}{2}\,. \tag{159}$$

One can then solve σ_b and σ_w by plugging corresponding $K^\star$ values.

$$(\sigma_w, \sigma_b) = (2, 0)\,,\ \text{for } \mathcal{K}^\star = 0\,, \tag{160}$$

$$(\sigma_w, \sigma_b) \approx (1.408, 0.416)\,,\ \text{for } \mathcal{K}^\star = \frac{3+\sqrt{17}}{2}\,. \tag{161}$$

J.3 Critical Exponents

GELU behaves in a different way compare to erf. First we discuss the $\mathcal{K}^\star = 0$ critical point, which is located at $(\sigma_b, \sigma_w) = (0, 2)$. We expand Eq.(151), and keep next to leading order $\delta\mathcal{K}^l = \mathcal{K}^l - \mathcal{K}^\star$

$$\delta\mathcal{K}^{l+1} \approx \delta\mathcal{K}^l + \frac{6}{\pi}(\delta\mathcal{K}^l)^2\,. \tag{162}$$

From lemma D.1

$$A = -\frac{\pi}{6} \text{ and } \zeta_{\mathcal{K}} = 1\,, \tag{163}$$

which is not possible since $\delta\mathcal{K}^l \geq 0$ for this case. This result means scaling analysis is not working here.

Next, we consider the other fixed point with $\mathcal{K}^\star = \frac{3+\sqrt{17}}{2}$ at $(\sigma_b, \sigma_w) = (0.416, 1.408)$. Expand the NNGP kernel recurrence relation again.

$$\delta\mathcal{K}^{l+1} \approx \delta\mathcal{K}^l + 0.00014(\delta\mathcal{K}^l)^2\,. \tag{164}$$

Following the same analysis, we find

$$\delta\mathcal{K}^l \approx -7142.9\, l^{-1}\,. \tag{165}$$

Looks like scaling analysis works for this case, since $\mathcal{K}^\star > 0$. The solution shows that the critical point is half-stable[40]. If $\mathcal{K}^l < \mathcal{K}^\star$, the fixed point is repealing, while when $\mathcal{K}^l > \mathcal{K}^\star$, the fixed point is attractive. However, the extremely large coefficient in the scaling behavior of $\delta\mathcal{K}^l$ embarrasses the analysis. Since for any network with a reasonable depth, the deviation $\delta\mathcal{K}^l$ is not small.

Now we can expand $\chi^l_{\mathcal{J}}$ at some large depth, up to leading order l^{-1}.

$$\chi^l_{\mathcal{J}} \approx 1 - \frac{66.668}{l}\,. \tag{166}$$

Then

$$\delta\mathcal{J}^{l_0,l} \approx c_{l_0} l^{-66.668}\,, \tag{167}$$

where c_{l_0} is a positive non-universal constant.

Critical exponent

$$\zeta = 66.668\,. \tag{168}$$

Which in practice is not traceable.

J.4 LayerNorm on Pre-activations

Use lemma C.9, we have

$$\begin{aligned}\chi^l_{\mathcal{J}} =& \frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}^l_k) \phi'(\tilde{h}^l_k) \right] \Bigg|_{\tilde{\mathcal{K}}^{l-1}=1} \\ =& \frac{\sigma_w^2 (6\pi + 4\sqrt{3})}{\sigma_w^2 (6\pi + 3\sqrt{3}) + 18\pi\sigma_b^2} \,. \end{aligned} \tag{169}$$

The critical line is then at

$$\begin{aligned}\sigma_b =& \left(6\sqrt{3}\pi\right)^{-\frac{1}{2}} \sigma_w \\ \approx& 0.175 \sigma_w \,. \end{aligned} \tag{170}$$

J.5 LayerNorm on Activations

First we need to evaluate a new expectation value

$$\begin{aligned}\mathbb{E}_\theta \left[\phi(h^l_i) \right] =& \frac{1}{\sqrt{2\pi(\sigma_w^2 + \sigma_b^2)}} \int dh^l_i \frac{h^l_i}{2} \left[1 + \text{erf}\left(\frac{x}{\sqrt{2}} \right) \right] e^{-\frac{(h^l_i)^2}{2(\sigma_w^2 + \sigma_b^2)}} \\ =& \frac{\sigma_w^2 + \sigma_b^2}{\sqrt{2\pi(1 + \sigma_w^2 + \sigma_b^2)}} \,, \end{aligned} \tag{171}$$

where we used integrate by parts to get the result.

The other integrals are modified to

$$\mathbb{E}_\theta \left[\phi'(h^l_i) \phi'(h^l_i) \right] = \frac{1}{4} + \frac{1}{2\pi} \left[\arcsin\left(\frac{\sigma_w^2 + \sigma_b^2}{1 + \sigma_w^2 + \sigma_b^2} \right) + \frac{(\sigma_w^2 + \sigma_b^2)[3 + 5(\sigma_w^2 + \sigma_b^2)]}{(1 + \sigma_w^2 + \sigma_b^2)[1 + 2(\sigma_w^2 + \sigma_b^2)]^{\frac{3}{2}}} \right] , \tag{172}$$

$$\mathbb{E}_\theta \left[\phi(h^l_i) \phi(h^l_i) \right] = \frac{\sigma_w^2 + \sigma_b^2}{4} + \frac{\sigma_w^2 + \sigma_b^2}{2\pi} \arcsin\left(\frac{\sigma_w^2 + \sigma_b^2}{1 + \sigma_w^2 + \sigma_b^2} \right) + \frac{(\sigma_w^2 + \sigma_b^2)^2}{\pi(1 + \sigma_w^2 + \sigma_b^2)\sqrt{1 + 2(\sigma_w^2 + \sigma_b^2)}} \,. \tag{173}$$

One can then combine those results to find $\chi^l_{\mathcal{J}}$

$$\chi^l_{\mathcal{J}} = \frac{\sigma_w^2 \left(1 + \sigma_w^2 + \sigma_b^2\right) \left[\pi + 2 \arcsin\left(\frac{\sigma_w^2 + \sigma_b^2}{1 + \sigma_w^2 + \sigma_b^2} \right) + \frac{2(\sigma_w^2 + \sigma_b^2)(3 + 5(\sigma_w^2 + \sigma_b^2))}{(1 + \sigma_w^2 + \sigma_b^2)(1 + 2(\sigma_w^2 + \sigma_b^2))^{\frac{3}{2}}} \right]}{\pi(\sigma_w^2 + \sigma_b^2)(1 + \sigma_w^2 + \sigma_b^2) - 2(\sigma_w^2 + \sigma_b^2)^2 + \frac{4(\sigma_w^2 + \sigma_b^2)^2}{\sqrt{1 + 2(\sigma_w^2 + \sigma_b^2)}} + 2(\sigma_w^2 + \sigma_b^2)(1 + \sigma_w^2 + \sigma_b^2) \arcsin\left(\frac{\sigma_w^2 + \sigma_b^2}{1 + \sigma_w^2 + \sigma_b^2} \right)} \,. \tag{174}$$

The critical line defined by $\chi^l_{\mathcal{J}} = 1$, one can numerically solve it by scanning over σ_b and σ_w.

J.6 Residual Connections

The recurrence relation for the NNGP kernel is

$$\mathcal{K}^{l+1} = \left[\frac{\mathcal{K}^l}{4} + \frac{\mathcal{K}^l}{2\pi} \arcsin\left(\frac{\mathcal{K}^l}{1 + \mathcal{K}^l} \right) + \frac{(\mathcal{K}^l)^2}{\pi(1 + \mathcal{K}^l)\sqrt{1 + 2\mathcal{K}^l}} \right] \sigma_w^2 + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \tag{175}$$

Fixed point exists if

$$\chi^\star_{\mathcal{K}} = \frac{\sigma_w^2}{4} + \frac{\sigma_w^2}{2\pi} \left[\arcsin\left(\frac{\mathcal{K}^\star}{1 + \mathcal{K}^\star} \right) + \frac{4(\mathcal{K}^\star)^3 + 11(\mathcal{K}^\star)^2 + 5\mathcal{K}^\star}{(1 + \mathcal{K}^\star)^2 (1 + 2\mathcal{K}^\star)^{\frac{3}{2}}} \right] + \mu^2 \leq 1 \,. \tag{176}$$

The recurrence coefficient for Jacobian is

$$\chi_{\mathcal{J}}^{\star} = \frac{\sigma_w^2}{4} + \frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^{\star}}{1+\mathcal{K}^{\star}}\right) + \frac{\mathcal{K}^{\star}(3+5\mathcal{K}^{\star})}{(1+\mathcal{K}^{\star})(1+2\mathcal{K}^{\star})^{\frac{3}{2}}}\right] + \mu^2 \,. \tag{177}$$

Phase boundary is shifted

$$\sigma_w = 2\sqrt{1-\mu^2}\left[1 + \frac{2\mathcal{K}^{\star}(3+5\mathcal{K}^{\star})}{\pi(1+\mathcal{K}^{\star})(1+2\mathcal{K}^{\star})^{\frac{3}{2}}} + \frac{2}{\pi}\arcsin\left(\frac{\mathcal{K}^{\star}}{1+\mathcal{K}^{\star}}\right)\right]^{-\frac{1}{2}} , \tag{178}$$

$$\sigma_b = \frac{\mathcal{K}^{\star}}{\sqrt{2\pi}(1+2\mathcal{K}^{\star})^{\frac{3}{4}}}\sigma_w \,. \tag{179}$$

One can again scan over $\mathcal{K}^{\star}$ to draw the critical line.

In order to locate critical point, we further require $\chi_{\mathcal{K}}^{\star} = 1$. To locate the critical point, we solve $\chi_{\mathcal{J}}^{\star} - \chi_{\mathcal{K}}^{\star} = 0$ instead. We have

$$\frac{\sigma_w^2[(\mathcal{K}^{\star})^3 - 3(\mathcal{K}^{\star})^2 - 2\mathcal{K}^{\star}]}{2\pi(1+\mathcal{K}^{\star})^2(1+2\mathcal{K}^{\star})^{\frac{3}{2}}} = 0 \,, \tag{180}$$

which has two non-negative solutions out of three

$$\mathcal{K}^{\star} = 0 \text{ and } \mathcal{K}^{\star} = \frac{3+\sqrt{17}}{2} \,. \tag{181}$$

One can then solve σ_b and σ_w by plugging corresponding $K^{\star}$ values.

$$(\sigma_w, \sigma_b) = (2\sqrt{1-\mu^2}, 0) \,, \text{ for } \mathcal{K}^{\star} = 0 \,, \tag{182}$$

$$(\sigma_w, \sigma_b) \approx (1.408\sqrt{1-\mu^2}, 0.416\sqrt{1-\mu^2}) \,, \text{ for } \mathcal{K}^{\star} = \frac{3+\sqrt{17}}{2} \,. \tag{183}$$

J.7 Residual Connections with LayerNorm on Preactivations (Pre-LN)

Use lemma C.9 and results we had without residue connections for GELU.

$$\begin{aligned}
\chi_{\mathcal{J}}^{*} =& \lim_{l\to\infty}\left(\frac{\sigma_w^2}{N_l\mathcal{K}^l}\sum_{k=1}^{N_l}\mathbb{E}_{\theta}\left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l)\right]\Big|_{\tilde{\mathcal{K}}^{l-1}=1} + \mu^2\right) \\
=& \frac{\sigma_w^2(6\pi+4\sqrt{3})(1-\mu^2)}{\sigma_w^2(6\pi+3\sqrt{3})+18\pi\sigma_b^2} + \mu^2 \\
=& 1 - \frac{(\sqrt{3}\sigma_w^2 - 18\pi\sigma_b^2)(1-\mu^2)}{\sigma_w^2(6\pi+3\sqrt{3})+18\pi\sigma_b^2} \,.
\end{aligned} \tag{184}$$

The critical line is then at

$$\begin{aligned}
\sigma_b =& \left(6\sqrt{3}\pi\right)^{-\frac{1}{2}}\sigma_w \\
\approx& 0.175\sigma_w \,,
\end{aligned} \tag{185}$$

just like without residue connections.

K Additional Experimental Results

In the following training results, we used NTK parameterization for the linear layers in the MLP. We emphasize that this choice has little effect on the training and convergence in this case, compared to standard initialization.

In figure 8, we compare the performance of deep MLP networks with and without LayerNorm. We note that the case with LayerNorm applied to preactivations continues to train at very large value of σ_w^2. In all cases, networks are trained using stochastic gradient descent with MSE. We used the Fashion MNIST dataset[47]. All networks had depth $L = 50$ and width $N_l = 500$. The learning rates were logarithmically sampled

- within $(10^{-8}, 10^{6})$ for ReLU, $(10^{-5}, 10)$ for LN-ReLU and ReLU-LN;
- within $(10^{-5}, 1)$ for erf, LN-erf and erf-LN;
- within $(10^{-8}, 10)$ for GELU, $(10^{-3}, 10)$ for LN-GELU and GELU-LN, where λ_{max} is the largest eigenvalue of NTK for each σ_w.

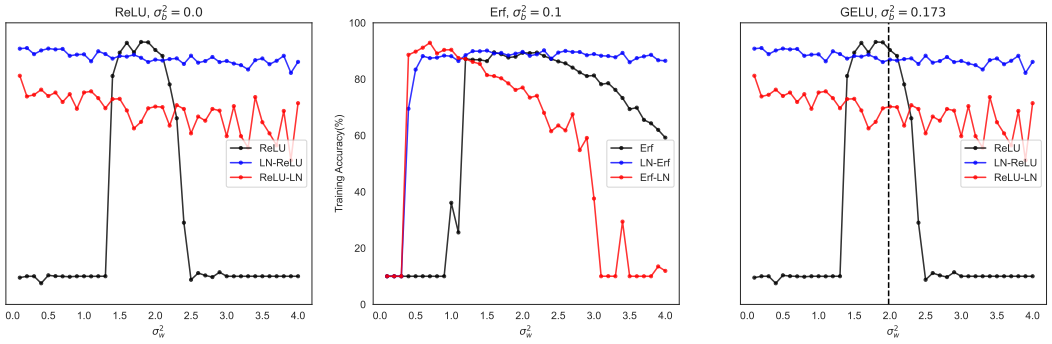

Figure 8: Performance of deep MLP networks at and away from criticality, with and without LayerNorm. The blue plateau, corresponding to LayerNorm applied to preactivations, continues to train at very large values of σ_w^2 without the need to tune the learning rate.

In figure 9, we showed empirically that the critical exponent of partial Jacobians are vanished for erf with LayerNorm.

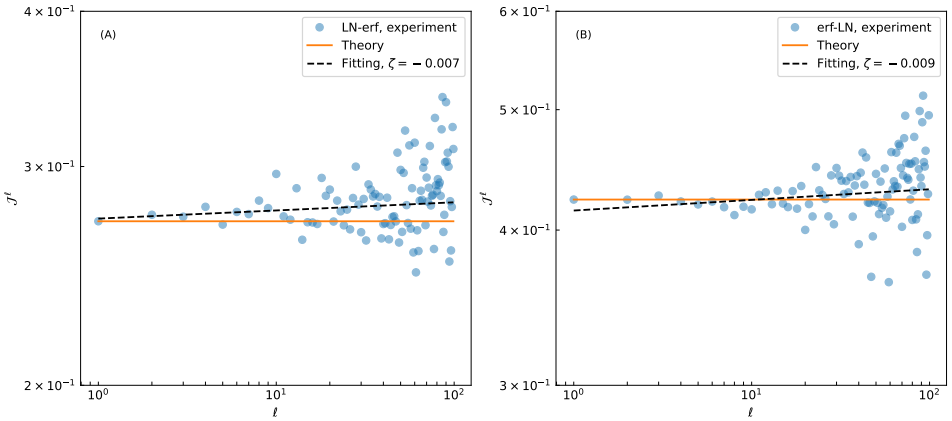

Figure 9: $\log - \log$ plot of partial Jacobian $\mathcal{J}^{0,l}$ vs. l for (A) LN-erf and (B) erf-LN.

In figure 10, we tested $6k$ samples from CIFAR-10 dataset[23] with kernel regression based on neural tangents library [35] [26] [36]. Test accuracy from kernel regression reflects the trainability (training accuracy) with SGD in ordered phase. We found that the trainable depth is be predicted by the correlation length $c\xi$ with LayerNorm applied to preactivations, where the prefactor $c = 28$. The prefactor we had is the same as vanilla cases in [49]. The difference is from the fact that they used $\log_{10}$ and we used $\log_e$.

In figure 11, we explore the broad range in σ_w^2 of the performance of MLP network with erf activation function and LayerNorm on preativations. The network has depth $L = 50$ and width $N_l = 500$; and is trained using SGD on Fashion MNIST. The learning rates are chosen based on a logarithmic scan with a short training time.

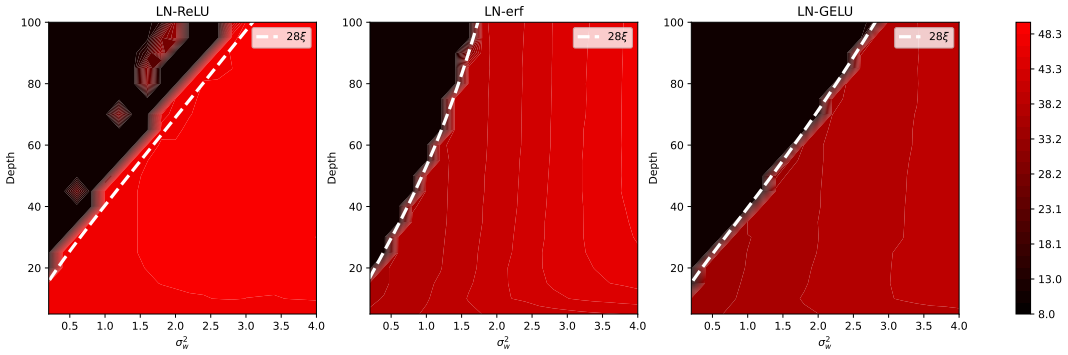

Figure 10: Test accuracy for LayerNorm applied to preactivations. $\sigma_b^2 = 0.5$ for all cases. Correlation lengths calculated using analytical results of $\chi_{\mathcal{J}}^l$.

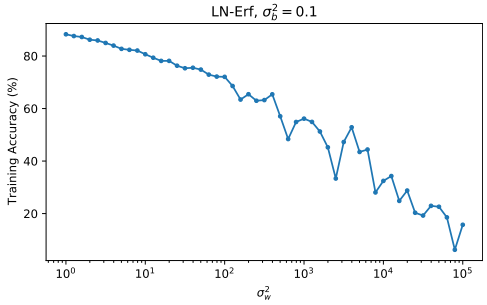

Figure 11: Training performance of MLP networks with erf activation function; and LayerNorm applied to preactivations. It continues to train for several orders of magnitude of σ_w^2 (with learning-rate tuning).