CRITICAL INITIALIZATION OF WIDE AND DEEP NEURAL NETWORKS THROUGH PARTIAL JACOBIANS: GENERAL THEORY AND APPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks are notorious for defying theoretical treatment. However, when the number of parameters in each layer tends to infinity, the network function is a Gaussian process (GP) and quantitatively predictive description is possible. Gaussian approximation allows to formulate criteria for selecting hyperparameters, such as variances of weights and biases, as well as the learning rate. These criteria rely on the notion of criticality defined for deep neural networks. In this work we describe a new practical way to diagnose criticality. We introduce *partial Jacobians* of a network, defined as derivatives of preactivations in layer l with respect to preactivations in layer $l_0 \leq l$. We derive recurrence relations for the norms of partial Jacobians and utilize these relations to analyze criticality of deep fully connected neural networks with LayerNorm and/or residual connections. We derive and implement a simple and cheap numerical test that allows one to select optimal initialization for a broad class of deep neural networks; including fully connected, convolutional and attention layers. Using these tools we show quantitatively that proper stacking of the LayerNorm (applied to preactivations) and residual connections leads to an architecture that is critical for any initialization. Finally, we apply our methods to analyze the MLP-Mixer architecture and show that it is everywhere critical.

1 INTRODUCTION

When the number of parameters in each layer becomes large, the functional space description of deep neural networks simplifies dramatically. The network function, $f(x)$, in this limit, is a Gaussian process (Neal, 1996; Lee et al., 2018) with a kernel – sometimes referred to as neural network Gaussian process (NNGP) kernel (Lee et al., 2018) – determined by the network architecture and hyperparameters (*e.g* depth, precise choices of layers and the activation functions, as well as the distribution of weights and biases). Similar line of reasoning was earlier developed for recurrent neural networks (Molgedey et al., 1992). Furthermore, for special choices of parameterization and MSE loss function, the training dynamics under gradient descent can be solved exactly in terms of the neural tangent kernel (NTK) (Jacot et al., 2018; Lee et al., 2019). A large body of work was devoted to the calculation of the NNGP kernel and NTK for different architectures, calculation of the finite width corrections to these quantities, and empirical investigation of the training dynamics of wide networks (Novak et al., 2018b; Xiao et al., 2018; Hron et al., 2020; Dyer & Gur-Ari, 2019; Andreassen & Dyer, 2020; Lewkowycz & Gur-Ari, 2020; Aitken & Gur-Ari, 2020; Geiger et al., 2020; Hanin, 2021; Roberts et al., 2022; Yaida, 2020; Shankar et al., 2020; Arora et al., 2019b;a; Lee et al., 2020; Yang et al., 2018; Yang & Hu, 2021; Yang, 2019b;a; Matthews et al., 2018; Garriga-Alonso et al., 2018; Allen-Zhu et al., 2019; Tsuchida et al., 2021; Martens et al., 2021).

One important result that arose from these works is that the network architecture determines the most appropriate initialization of the weights and biases (Poole et al., 2016; Schoenholz et al., 2016; Lee et al., 2018). To state this result, we consider networks with/without LayerNorm (Ba et al., 2016) and residual connections (He et al., 2016); the preactivations for which can be defined as follows

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

where $\tilde{h}^l_j = \mathrm{LayerNorm}(h^l_j)$ and the parameter μ controls the strength of residual connections. For the input layer: $h^1_i(x) = \sum_{j=1}^{N_0} w^1_{ij} x_j + b^1_i$. In the $(l+1)$-th layer, weights $w^{l+1}_{ij} \in \mathbb{R}^{N_{l+1} \times N_l}$ and biases $b^{l+1}_i \in \mathbb{R}^{N_{l+1} \times 1}$ are taken from normal distributions $\mathcal{N}(0, \sigma_w^2/N^l)$ and $\mathcal{N}(0, \sigma_b^2)$, respectively. Hyperparameters σ_w and σ_b need to be tuned. $\phi(\cdot)$ is the activation function and $x \in \mathbb{R}^{N_0 \times 1}$ is the input. For results discussed in this work, x can be sampled from either a realistic (*i.e.* highly correlated) dataset or a high entropy distribution.

For a network of depth L, the network function is given by $f(x) = h^L(x)$. Different network architectures and activation functions, ϕ, lead to different "optimal" choices of (σ_w, σ_b). The optimal choice can be understood, using the language of statistical mechanics, as a critical point (or manifold) in the σ_b–σ_w plane. The notion of criticality becomes sharp as the network depth, L, becomes large. Criticality ensures that both NNGP and the norm of gradients remain $O(L^0)$ as the network gets deeper (Roberts et al., 2022). Very deep networks will not train unless initialized critically, since the gradients explode or vanish exponentially. Moreover, high trainability does not imply that the trained model has a great performance (test accuracy) after training.

1.1 Results

Here we focus on two main results of this work: (i) empirical method to check criticality of a neural network and (ii) an architecture based on layer normalization and residual connections that is critical for *any* initialization. First we introduce the notion of a partial Jacobian.

Definition 1.1. Let $h^l_i(x)$ be preactivations of a neural network $f(x)$. The partial Jacobian $J^{l_0,l}_{ij}$ is defined as derivative of preactivations at layer l with respect to preactivations at layer $l_0 \le l$

$$J^{l_0,l}_{ij}(x) = \frac{\partial h^l_j(x)}{\partial h^{l_0}_i(x)} \,. \tag{2}$$

The partial Jacobian is a random matrix with vanishing mean at initialization. We introduce a deterministic measure of the magnitude of $J^{l_0,l}_{ij}$ — its squared Frobenius norm, averaged over parameter-initializations.

Definition 1.2. Let $J^{l_0,l}_{ij}$ be a partial Jacobian of a neural network $f(x)$. Averaged partial Jacobian norm (APJN) is defined as

$$\mathcal{J}^{l_0,l}(x) \equiv \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{i=1}^{N_{l_0}} \left(\frac{\partial h^l_j(x)}{\partial h^{l_0}_i(x)} \right)^2 \right] , \tag{3}$$

where $\mathbb{E}_\theta$ indicates averaging over parameter-initializations.

In what follows, we show that criticality, studied previously in literature, occurs when APJN either remains finite, or varies *algebraically* as l becomes large. To prove this we derive the recurrence relation for $\mathcal{J}^{l_0,l}(x)$ in the limit $N_l \to \infty$ and analyze it at large depth. Algebraic behaviour of APJN with depth is characterized by an architecture-dependent critical exponent, ζ, so that $\mathcal{J}^{l_0,l}(x) \approx l^{-\zeta}$. Such behaviour is familiar from statistical mechanics when a system is tuned to a critical point (Cardy, 1996). Away from criticality, there are two phases: ordered and chaotic. In the ordered phase APJN vanishes exponentially with depth, whereas in the chaotic phase APJN grows exponentially

$$\mathcal{J}^{l_0,l} \approx c_{l_0} e^{\pm \frac{l}{\xi}} \,. \tag{4}$$

Here ξ is the correlation length. It characterizes how fast gradients explode or vanish.

Theorem 1.3 (Main result)**.** *Let $f(x)$ be a deep MLP network with Lipschitz continuous activation $\phi(\cdot)$. Assume that the LayerNorm is applied to preactivations and there are residual connections with strength μ acting according to* (1)*. In the limit $N_l \to \infty$ the correlation length is bounded from below for $\sigma_b^2 < \infty$*

$$\xi \ge \frac{1}{\left| \log \left[(1-\mu^2) \frac{A}{B} + \mu^2 \right] \right|} \,, \tag{5}$$

where the non-negative constants A and B are given by

$$A = E_\theta \left[\frac{1}{N_l} \sum_{k=1}^{N_l} \phi'(\tilde{h}^l_k)^2 \right] , \qquad B = E_\theta \left[\frac{1}{N_l} \sum_{k=1}^{N_l} \phi(\tilde{h}^l_k)^2 \right] , \tag{6}$$

*where $\phi'(\cdot)$ is the derivative of $\phi(\cdot)$. When $\mu = 1$, the correlation length diverges and the network is critical for **any** initialization, with $\zeta = O(1)$.*

In practice Theorem 1.3 means that different choices of initialization bear no effect on trainability of the network provided that LayerNorm and residual connections are arranged as stated.

1.2 Related Work

Some of our results were either surmised or obtained in a different form in the literature. We find that LayerNorm ensures that NNGP kernel remains finite at any depth as suggested in the original work of Ba et al. (2016). LayerNorm also alters the criticality of $\mathcal{J}^{l_0,l}(x)$. It was noted in Xu et al. (2019) that LayerNorm (applied to preactivations) regularizes the backward pass. We formalize this observation by showing that LayerNorm (applied to preactivations) dramatically enhances correlation length (which is not the case for LayerNorm applied to activations). This can be seen from Theorem 1.3, setting $\mu = 0$. When residual connections of strength 1 are combined with erf (or any other erf-like activation function, e.g. tanh), the neural network enters a *subcritical* phase with enhanced correlation length (see Theorem 4.3). A version of this result was discussed in Yang & Schoenholz (2017). When residual connections are introduced on top of LayerNorm, the correlation length ξ is further increased. If residual connections have strength 1 the network enters a critical phase for *any* initialization. Importance of correct ordering of LayerNorm, residual connections and attention layers was discussed in Xiong et al. (2020). Several architectures with the same order of GroupNorm and residual connections were investigated in Yu et al. (2021).

The partial Jacobian has been used to study generalization bounds in Arora et al. (2018). The Jacobian norm (*i.e.* $||J^{0,l}_{ij}||^2$) of trained feed-forward neural networks was studied in Novak et al. (2018a), where it was correlated with generalization. Partial Jacobians with $l_0 = l - 1$ were studied in the context of RNNs (Chen et al., 2018; Can et al., 2020), where they were referred to as *state-to-state* Jacobians.

As the aspect ratio (L/N) of the network approaches 1, the finite width corrections to the Jacobian become more prominent. On the other hand, even with small aspect ratio, the effect of the spectral density of Jacobian becomes important as the depth L becomes *very* large. Pennington et al. (2018) study the spectrum of the input-output Jacobian for MLPs. Xiao et al. (2018) extend the analysis to CNNs, showing that *very* deep vanilla CNNs can be trained by achieving "dynamical isometry".

2 Recurrence Relations

Here we derive the infinite width recurrence relations for the APJN and the norm of preactivations. We use (1) in its simplest form, which has no LayerNorm and $\mu = 0$.

Definition 2.1. We define averaged covariance of preactivations as follows

$$\mathcal{K}^l(x, x') = \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{i=1}^{N_l} h^l_i(x) h^l_i(x') \right] . \tag{7}$$

Lemma 2.2. *When $N_l \to \infty$ for $l = 1, \dots, L-1$, the expectation value over parameter initializations for a general function of preactivations: $\mathcal{O}(h^l(x))$, can be expressed as the averaging over the Gaussian process $h^l(x)$ with covariance $\mathcal{K}^l(x, x')$.*

$$\mathbb{E}_\theta \left[\mathcal{O}(h^l_i(x)) \right] = \frac{1}{\sqrt{2\pi \mathcal{K}^l(x,x)}} \int dh^l \mathcal{O}(h^l_i(x)) e^{-\frac{(h^l_i(x))^2}{2\mathcal{K}^l(x,x)}} . \tag{8}$$

This result has been established in Lee et al. (2018). Note that the density in (8) only depends on the diagonal part of the covariance matrix, $\mathcal{K}^l(x, x)$. We will refer to $\mathcal{K}^l(x, x)$ as NNGP kernel.

Remark 2.3. In the infinite width limit the means appearing in (10)-(12) are self-averaging and, therefore, deterministic. They converge in distribution to their averages over parameterizations.

$$\frac{1}{N_l}\sum_{i=1}^{N_l}\phi(h_i^l)^2 \xrightarrow{N_l\to\infty} \quad \mathbb{E}_\theta\left[\frac{1}{N_l}\sum_{i=1}^{N_l}\phi(h_i^l)^2\right] = \mathbb{E}_\theta\left[\phi(h_i^l)^2\right]\,. \tag{9}$$

When performing analytic calculations we use the infinite width convention; whereas in our finite-width experiments we explicitly average over initializations of θ^l.

Theorem 2.4. *With Lemma 2.2, in the infinite width limit, the NNGP kernel $\mathcal{K}^{l+1}(x,x)$ is deterministic, and can determined recursively via*

$$\mathcal{K}^{l+1}(x,x) = \sigma_w^2\mathbb{E}_\theta\left[\frac{1}{N_l}\sum_{i=1}^{N_l}\phi(h_i^l(x))^2\right] + \sigma_b^2\,. \tag{10}$$

Theorem 2.5. *Let $f(x)$ be an MLP network with a Lipschitz continuous activation function $\phi(x)$. In the infinite width limit, APJN $\mathcal{J}^{l_0,l+1}(x)$ is deterministic and satisfies a recurrence relation*

$$\mathcal{J}^{l_0,l+1}(x) = \chi_{\mathcal{J}}^l\mathcal{J}^{l_0,l}(x)\,, \tag{11}$$

where the factor $\chi_{\mathcal{J}}^l$ is given by

$$\chi_{\mathcal{J}}^l = \mathbb{E}_\theta\left[\frac{\sigma_w^2}{N_l}\sum_{i=1}^{N_l}\left(\phi'(h_i^l(x))\right)^2\right]\,. \tag{12}$$

Theorem 2.4 is due to Lee et al. (2018). Theorem 2.5 is new and is valid only in the limit of infinite width. The proof is in Appendix B. We will drop the explicit dependence on x to improve readability.

The expectation values that appear in (10)-(12) are evaluated using (8). When the integrals can be taken analytically, they lead to explicit equations for the critical lines and/or the critical points. Details of these calculations as well as the derivation of (10)-(12) can be found in the Appendix. A subtlety emerges in (11) when $l_0 = 0$, where a correction of the order $O(N_0^{-1})$ arises for non-scale invariant activation functions. This subtlety is discussed in the Appendix B.

When the depth of the network becomes large, the l-dependence of the expectation values that appear in (7), (12) saturate to a (possibly infinite) constant value; which means that $\mathcal{K}^l$, $\mathcal{J}^{l_0,l}$ and $\chi_{\mathcal{J}}^l$ have reached a fixed point. We denote the corresponding quantities as $\mathcal{K}^\star, \mathcal{J}^{l_0,\star}, \chi_{\mathcal{J}}^\star$. The existence of a fixed point is not obvious and should be checked on a case by case basis. Fixed point analysis for $\mathcal{K}^l$ was done in Poole et al. (2016) for bounded activation functions and in Roberts et al. (2022) for the general case. The stability is formulated in terms of

$$\chi_{\mathcal{K}}^\star = \frac{\partial\mathcal{K}^{l+1}}{\partial\mathcal{K}^l}\Big|_{\mathcal{K}^l=\mathcal{K}^\star}\,. \tag{13}$$

The norm of preactivations remains finite (or behaves algebraically) when $\chi_{\mathcal{K}}^\star = 1$.

Eq. (11) nicely expresses $\mathcal{J}^{l_0,l+1}$ as a linear function of $\mathcal{J}^{l_0,l}$. The behaviour of $\mathcal{J}^{l_0,l+1}$ at large l is determined by $\chi_{\mathcal{J}}^l$. When $\chi_{\mathcal{J}}^l > 1$ partial Jacobians diverge exponentially, while for $\chi_{\mathcal{J}}^l < 1$ partial Jacobians vanish exponentially. Neural networks are trainable only up to a certain depth when initialized $O(1)$ away from criticality, which is determined by the equation

$$\chi_{\mathcal{J}}^\star = 1\,. \tag{14}$$

Eq. (14) is an implicit equation on σ_b, σ_w and generally outputs a critical line in σ_b–σ_w plane. The parameter $\chi_{\mathcal{J}}^\star$ has to be calculated on a case-by-case basis using either (12) or the method presented in the next section. Everywhere on the critical line, $\mathcal{J}^{l_0,l}$ saturates to a constant or behaves algebraically.

When the condition $\chi_{\mathcal{K}}^\star = 1$ is added, we are left with a critical point[1]. This analysis of criticality at infinite width agrees with Roberts et al. (2022), where $\chi_\perp$ is to be identified with $\chi_{\mathcal{J}}^\star$; and Schoenholz et al. (2016); Martens et al. (2021), where their analysis based on the equivalent χ_1 or $C'(1)$ only works for bounded activation functions. In particular, condition (14) together with $\chi_{\mathcal{K}}^\star = 1$ ensures that NTK is $O(1)$ at initialization.

[1] Scale-invariant activation functions are more forgiving: away from the critical point $\mathcal{K}^l$ scales algebraically with l.

2.1 Empirical Diagnostic of Criticality

APJN $\mathcal{J}^{l_0,l}$ provides a clear practical way to diagnose whether the network is critical or not. Proper choice of l_0 and l allows us to minimize the non-universal effects and cleanly extract $\chi^\star_\mathcal{J}$.

Recurrence relation (11), supplemented with the initial condition $\mathcal{J}^{l_0,l_0+1} = \chi^{l_0}_\mathcal{J}$, can be formally solved as

$$\mathcal{J}^{l_0,l} = \prod_{\ell=l_0}^{l-1} \chi^\ell_\mathcal{J} \,. \tag{15}$$

We would like to obtain an estimate of $\chi^\star_\mathcal{J}$ as accurately as possible. To that end, imagine that for some $l' > l_0$ the fixed point has been essentially reached and $\chi^{l'}_\mathcal{J} \approx \chi^\star_\mathcal{J}$. Then the APJN

$$\mathcal{J}^{l_0,l} = (\chi^\star_\mathcal{J})^{l-l'-1} \cdot \prod_{\ell=l_0}^{l'} \chi^\ell_\mathcal{J} \tag{16}$$

depends on the details of how the critical point is approached; which are encoded in the last factor.

Proposition 2.6. *If the network $f(x)$ is homogeneous,* i.e.*, consists of a (possibly complex) block of layers, periodically repeated L times. Then the penultimate APJN provides an accurate estimate of $\chi^\star_\mathcal{J}$:*

$$\mathcal{J}^{L-2,L-1}\Big|_{L\to\infty} = \chi^\star_\mathcal{J} \,. \tag{17}$$

This is a direct consequence of combining (10) *and* (12) *as L goes to infinity. See Figure 4 in Appendix C for numerical justification.*

Proposition 2.6 is the central result of this section and will be heavily used in the remainder of this work.

Note that for deep networks, away from criticality, APJN takes form

$$\mathcal{J}^{l_0,l} \approx c_{l_0} e^{\pm \frac{l}{\xi}} \,, \qquad \xi = |\log \chi^\star_\mathcal{J}|^{-1} \,, \tag{18}$$

where c_{l_0} is a non-universal constant that depends on l_0. If the sign in (18) is positive ($\chi^\star_\mathcal{J} > 1$) the network is in the *chaotic phase*, while when the sign is negative ($\chi^\star_\mathcal{J} < 1$) the network is in the *ordered phase*. ξ has the meaning of correlation length: on the depth scale of approximately $k\xi$ the gradients remain appreciable, and hence the network with the depth of $\approx k\xi$ will train.

We used (17) to map out the σ_b–σ_w phase diagrams of various MLP architectures. The partial Jacobians are calculated numerically with $N_l = 500$, $L = 50$ and averaged over initializations. The details are further elaborated in Appendix A. The location of the critical line agrees remarkably well with our infinite width calculations. Results are presented in Fig. 1. One fortunate outcome of both theory and experiment is that when LayerNorm is applied to *preactivations*, ReLU networks can still be initialized using He initialization (He et al., 2015) which, in our convention, is $(\sqrt{2}, 0)$.

At criticality, $\chi^\star_\mathcal{J} = 1$ and the correlation length diverges; indicating that gradients can propagate arbitrarily far. A more careful analysis of non-linear corrections shows that APJN can exhibit algebraic behaviour with depth and can still vanish in the infinite depth limit, but much slower than the ordered phase.

2.2 Scaling at a Critical Point

At criticality $\chi^l_\mathcal{J}$ saturates to a fixed value $\chi^\star_\mathcal{J} = 1$. If we are interested in $\mathcal{J}^{l_0,l}$ with $l - l_0 = O(L)$ then it is essential to know how exactly $\chi^l_\mathcal{J}$ approaches 1.

Theorem 2.7. *Assume that deep neural network $f(x)$ is initialized critically. Then $l \to \infty$ asymptotics of APJN is given by*

$$\mathcal{J}^{l_0,l}(x) = O(l^{-\zeta}) \,, \tag{19}$$

where ζ is the critical exponent Roberts et al. (2022). Critical exponents can be determined analytically in the limit of infinite width. (For a detailed discussion, see Appendix C.)

3 Layer Normalization

The fact that critical initialization is concentrated on a single point $(\sigma_w^\star, \sigma_b^\star)$ may appear unsettling because great care must be taken to initialize the network critically. The situation can be substantially improved by utilizing the normalization techniques known as LayerNorm (Ba et al., 2016) and GroupNorm (Wu & He, 2018). Our results apply to GroupNorm verbatim in the case when the number of groups is much smaller than the width. LayerNorm can act either on preactivations or on activations (discussed in the Appendix B). Depending on this choice, criticality will occur on different critical *lines* in σ_b–σ_w plane. When LayerNorm is applied to *preactivations* the correlation length is enhanced, allowing to train much deeper networks even far away from criticality.

The LayerNorm applied to preactivations takes the following form

Definition 3.1 (Normalized preactivations)**.**

$$\tilde{h}_i^l = \frac{h_i^l - \mathbb{E}[h^l]}{\sqrt{\mathbb{E}[(h^l)^2] - \mathbb{E}[h^l]^2}} \xrightarrow{N_l \to \infty} \frac{1}{\sqrt{\mathcal{K}^l}} h_i^l \,, \tag{20}$$

where we have introduced $\mathbb{E}[h^l] = \frac{1}{N_l}\sum_{i=1}^{N_l} h_i^l$. In the limit of infinite width $\mathbb{E}[h^l] = 0$ and $\mathbb{E}[(h^l)^2] = \mathcal{K}^l$, defined according to (7).

Normalized preactivations, $\tilde{h}_i^l$, are distributed according to $\mathcal{N}(0,1)$ for all l, σ_w, σ_b. The norms are, therefore, *always* finite and the condition $\chi_{\mathcal{K}}^\star = 1$ is trivially satisfied. This results in a critical line rather than a critical point.

The recurrence relations (10)-12 for the NNGP and partial Jacobians are only slightly modified

$$\mathcal{K}^{l+1} = \sigma_w^2 \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{i=1}^{N_l} \phi(\tilde{h}_i^l)^2 \right] + \sigma_b^2 \,, \qquad \chi_{\mathcal{J}}^l = \frac{\sigma_w^2}{\mathcal{K}^l} \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{i=1}^{N_l} \phi'(\tilde{h}_i^l)^2 \right] . \tag{21}$$

Assuming that the value of $\chi_{\mathcal{J}}^l$ at the fixed point is $\chi_{\mathcal{J}}^\star$, the network is critical when (14) holds.

$\chi_{\mathcal{J}}^l$ (21) changes *very slowly* with l and is also bounded from below, as elaborated in the next section. Thus, $\chi_{\mathcal{J}}^\star$ remains close to 1 for a very wide range of hyperparameters. Consequently, the correlation length is *large* even away from criticality. This leads to much higher trainability of deep networks with LayerNorm on preactivations even away from criticality.

4 Residual (Skip) Connections

Adding residual connections between the network layers is a widely used technique to facilitate the training of deep networks. Originally introduced (He et al., 2016) in the context of convolutional neural networks (LeCun et al., 1998) (CNNs) for image recognition, residual connections have since been used in a variety of networks architectures and tasks.

Consider (1) with non-zero μ and without LayerNorm layers. Then the recurrence relations (10)-(12) for the NNGP kernel and $\chi_{\mathcal{J}}^l$ are modified as follows

$$\mathcal{K}^{l+1} = \sigma_w^2 \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{j=1}^{N_l} \phi(h_j^l)^2 \right] + \sigma_b^2 + \mu^2 \mathcal{K}^l \,, \; \chi_{\mathcal{J}}^l = \sigma_w^2 \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{k=1}^{N_l} \phi'(h_k^l)^2 \right] + \mu^2 \,. \tag{22}$$

Remark 4.1. When $\mu < 1$, the fixed point value of NNGP kernel is scaled by $(1-\mu^2)^{-1}$. For $\mu = 1$, the critical point is formally at $(0, 0)$.

Remark 4.2. For $\mu = 1$, (22) implies that $\chi_{\mathcal{J}}^l \geq 1$, where the equality holds on the $\sigma_w = 0$ axis. Consequently, APJN exponentially diverges as a function of depth l for all $\sigma_w > 0$. In this case, σ_w needs to be taken sufficiently close to 0 to ensure trainability at large depths.

When $\mu < 1$, residual connections amplify the chaotic phase and decrease the correlation length away from criticality for unbounded activation functions.

Solving the recurrence relations (22) for erf activation, we find an effect observed in Yang & Schoenholz (2017) for $\tanh$ activation. They noted that $\tanh$-like MLP networks with skip connections ”hover over the edge of chaos”. We quantify their observation as follows.

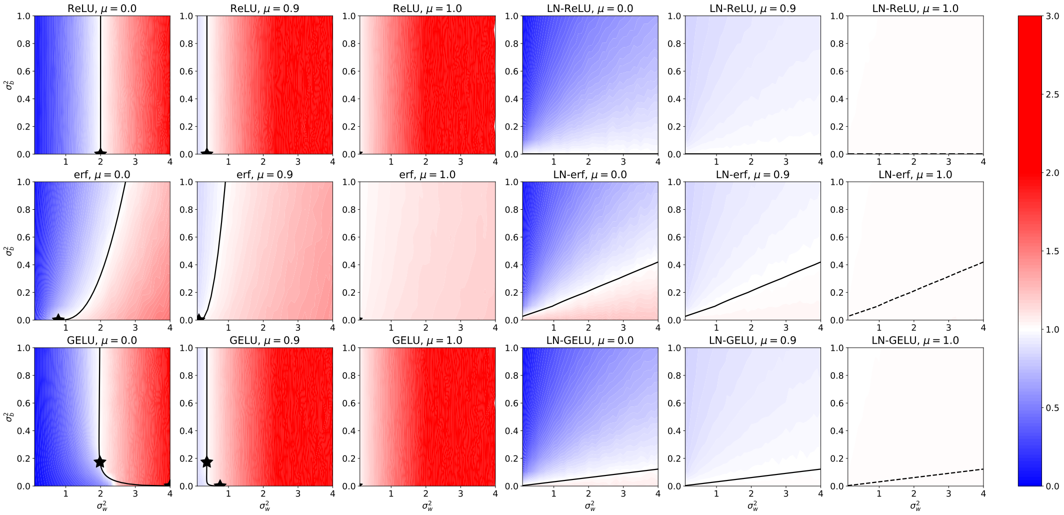

Figure 1: $\chi^\star_\mathcal{J}$ phase diagrams for ReLU (first row), erf (second row) and GELU (third row); with residual connections of variable strengths $\mu = \{0.0, 0.9, 1.0\}$. Both cases: without LayerNorm (first three columns) and with LayerNorm (last three columns) are shown. The solid lines indicate the critical lines obtained through infinite width limit calculations; while the stars indicate the critical points. The dotted lines in the rightmost column correspond to the critical lines for $\mu < 1$ case. For networks with LayerNorm and $\mu = 1$, $\chi^\star_\mathcal{J} = 1$ holds on the entire σ_b–σ_w plane, for all activation functions that we considered. We also note that for erf activation, the case $\mu = 1$ *without* LayerNorm is subcritical and has a large correlation length.

Theorem 4.3. *Let $f(x)$ be a deep MLP network with erf activation function and residual connections of strength $\mu = 1$. Then in the limit $N_l \to \infty$*

- *The NNGP kernel K^l linearly diverges with depth l.*
- *$\chi^l_\mathcal{J}$ approaches 1 from above (as can be seen from Fig. 1) : $\chi^l_\mathcal{J} \approx 1 + \tilde{c}/\sqrt{l}$, where $\tilde{c} = 2\sigma_w^2/(\pi\sqrt{\sigma_w^2 + \sigma_b^2})$ is a non-universal constant.*
- *APJN diverges as a stretched exponential : $\mathcal{J}^{l_0,l} = O(e^{\sqrt{\frac{l}{\lambda}}})$, where $\lambda = 1/(4\tilde{c}^2)$ is the new length scale.*

We will refer to this case as *subcritical*. Although $\chi^\star_\mathcal{J}$ reaches 1, the APJN still diverges with depth faster than any power law. The growth is controlled by the new scale λ. To control the gradient we would like to make λ large, which can be accomplished by decreasing σ_w. In this case the trainability is enhanced (see Fig. 2). Similar results hold for $\tanh$ activation function (Yang & Schoenholz, 2017), however in that case there is no explicit expression for $\tilde{c}$.

5 Residual Connections + LayerNorm

In practice, it is common to use a combination of residual connections and LayerNorm.

Using (1), the recurrence relations (10)-(12) for the NNGP and partial Jacobians are modified as follows

$$\mathcal{K}^{l+1} = \sigma_w^2 \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{j=1}^{N_l} \phi(\tilde{h}^l_j)^2 \right] + \sigma_b^2 + \mu^2 \mathcal{K}^l \,,\ \chi^l_\mathcal{J} = \frac{\sigma_w^2}{\mathcal{K}^l} \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{k=1}^{N_l} \phi'(\tilde{h}^l_k)^2 \right] + \mu^2 \,. \quad (23)$$

Remark 5.1. For $\mu < 1$, (23) implies that the fixed point value of NNGP kernel is scaled by $1 - \mu^2$. Moreover, residual connections do not shift the phase boundary. The interference between residual connections and LayerNorm brings $\chi^l_\mathcal{J}$ closer to 1 on the entire σ_b–σ_w plane (as can be seen from

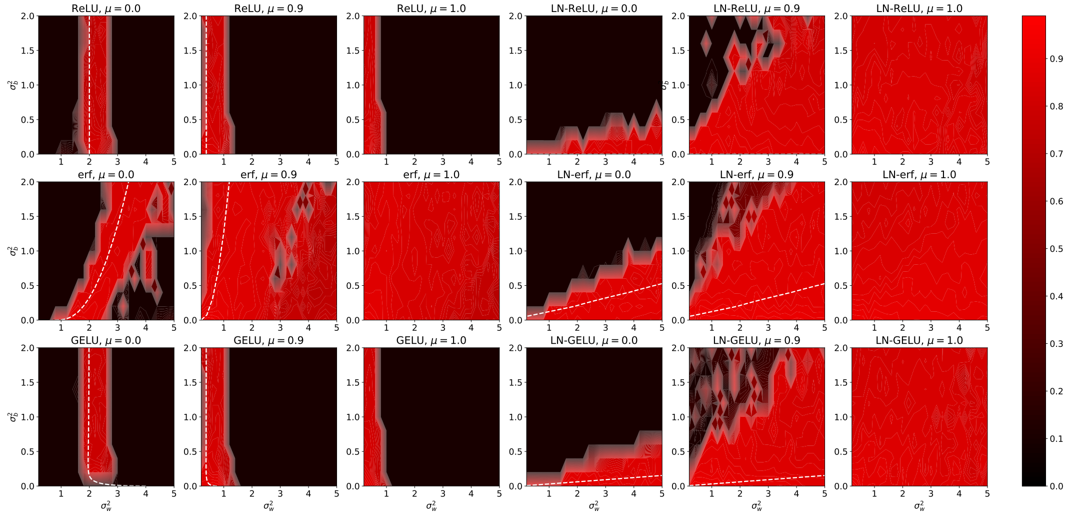

Figure 2: Trainability (Training Accuracy) of deep MLP networks ($N_l = 500$, $L = 50$) featuring ReLU (first row), erf (second row) and GELU (third row); with residual connections of variable strengths $\mu = \{0.0, 0.9, 1.0\}$. Both cases: without LayerNorm (first three columns) and with LayerNorm (last three columns) are shown. Cases with LayerNorm and $\mu = 1$ (last column) train at all values of σ_w^2 and σ_b^2 we considered, in agreement with theory.

Fig. 1). Therefore the correlation length ξ is improved in both the phases, allowing for training of deeper networks. At criticality, Jacobians linearly diverge with depth.

As was mentioned before, the combination of LayerNorm and residual connections dramatically enhances correlation length, leading to a more stable architecture. This observation is formalized by Theorem 1.3. The proof leverages the properties of solutions of (23) close to the fixed point, and is fleshed out in Appendix E.

Remark 5.2. When $\mu = 1$, the correlation length diverges for *any* initialization.

Remark 5.2 provides an alternative perspective on architecture design. On the one hand, given a neural network architecture one can use (17) to initialize it critically. Alternatively, one can add extra layers, such as a combination of residual connections and LayerNorm, to ensure that the network is always critical and will train well no matter which initialization scheme is used.

Remark 5.3. When $\mu = 1$, the condition $\chi_{\mathcal{J}}^{\star} = 1$ holds on the entire $\sigma_b - \sigma_w$ and for any activation function ϕ (see Fig. 1). NNGP kernel diverges linearly, while APJN diverges algebraically with the critical exponent of $\zeta = O(1)$. The exact value of the critical exponent depends on the activation function and the ratio $\frac{\sigma_b}{\sigma_w}$. The trainability is dramatically enhanced as can be seen from Fig. 2.

Remark 5.4. Networks with BatchNorm (Ioffe & Szegedy, 2015), used in conjunction with residual connection of strength $\mu = 1$, also enjoy this *everywhere criticality* and enhanced trainability (Yang et al., 2018; He et al., 2022).

6 MLP-Mixer

MLP-Mixer architecture is a recent example of MLP approach to computer vision (Tolstikhin et al., 2021). Its main ingredients are: patches, MLP layers, LayerNorm and residual connections. As such it can be analyzed using the tools and results presented above. The detailed summary of the MLP-Mixer is presented in the Appendix F, while here we will state the results.

When $\mu = 1$ the MLP-Mixer is everywhere critical due to the interaction between LayerNorm and residual connections. When $\mu < 1$ we can identify a critical line by numerically evaluating $\chi_{\mathcal{J}}^{\star}$ using (17). The phase diagrams are presented in Fig. 3.

To illustrate the importance of critical initialization we trained MLP-Mixer at $\mu = 1$ for various initializations, including a highly unconventional $\sigma_b^2 = 10, \sigma_w^2 = 10$. While the final performance varies by a few percent, the network trains well at a large depth of $L = 100$ mixer blocks. We also trained MLP-Mixer at $\mu = 0.5$. In this case, the model trains well at $L = 100$ when initialized critically; while far away from the critical line ($\sigma_b^2 = 10, \sigma_w^2 = 10$) it starts training after 30 epochs, and the gap between it and critical initialization is larger than $\mu = 1$ cases. The learning curves are presented in Fig. 3. We emphasize that we were interested in trainability and, consequently, did not tune the hyperparameters to achieve the best generalization.

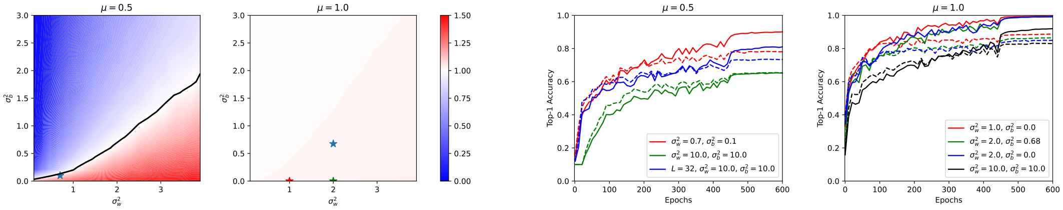

Figure 3: from left to right: (1)(2) $\mu = 0.5$ and $\mu = 1.0$ phase diagrams plotted using $\chi^\star_{\mathcal{J}}$ from repeating Mixer Layers of the GELU MLP-Mixer. Black line indicates the empirical phase boundary. Stars indicate points we selected to train on CIFAR-10. $(\sigma_w^2 = 10, \sigma_b^2 = 10)$ point is outside the phase diagrams. (3)(4) $\mu = 0.5$ and $\mu = 1$ MLP-Mixer training curves. Solid and dashed lines indicate training and validation accuracies, respectively. All the networks are $L = 100$ blocks deep, except for one, which is $L = 32$ blocks deep; all networks have 10 million parameters.

7 Conclusions

We have introduced partial Jacobians and their averaged norms as a tool to analyze the propagation of gradients through deep neural networks at initialization. Using APJN evaluated close to the output, $\mathcal{J}^{L-2,L-1} \approx \chi^\star_{\mathcal{J}}$, we have introduced a very cheap and simple empirical test for criticality. We have also shown that criticality formulated in terms of partial Jacobians is equivalent to criticality studied previously in literature (Poole et al., 2016; Roberts et al., 2022; Martens et al., 2021). APJN will play an important role in quantifying the criticality of inhomogeneous (*i.e.* no periodic stacking of blocks) networks.

We have investigated homogeneous architectures that include fully-connected layers, normalization layers and residual connections. In the limit of infinite width, we showed that (i) in the presence of LayerNorm, the critical point generally becomes a critical line, making the initialization problem much easier, (ii) LayerNorm applied to preactivations enhances correlation length leading to improved trainability, (iii) combination of $\mu = 1$ residual connections and erf activation function enhances correlation length driving the network to a *subcritical* phase with APJN growing according to a stretched exponential law, (iv) combination of residual connections and LayerNorm drastically increases correlation length leading to improved trainability, (v) when $\mu = 1$ and LayerNorm is applied to preactivations *the network is critical on the entire σ_b–σ_w plane.*

We have considered the example of a modern high performance architecture — the MLP-Mixer. We showed that it is critical everywhere and is not sensitive to initialization at $\mu = 1$ due to the interaction between LayerNorm and residual connections. We have also studied MLP-Mixer at $\mu = 0.5$ and showed that it is critical along a line (as expected for an architecture with LayerNorm). We demonstrated empirically that deep (100 blocks) MLP-Mixer trains for a variety of initializations at $\mu = 1$ but only trains well close to the critical line for $\mu = 0.5$.

Our work shows that an architecture can be designed to have a large correlation length leading to a guaranteed trainability with SGD for any initialization scheme.

8 Ethics Statement

We have read the Code of Ethics; have adhered to it while writing this paper; and will adhere to it during the paper submission, review, and discussion process.

9 Reproducibility Statement

We provide the details of all the experiments in Appendix A; including hyperparameter choices, GPU specifications and GPU hours. We also provide the source code in the Supplementary Material.

References

Kyle Aitken and Guy Gur-Ari. On the asymptotics of wide networks with polynomial activations. *arXiv preprint arXiv:2006.06687*, 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Anders Andreassen and Ethan Dyer. Asymptotics of wide convolutional neural networks. *arXiv preprint arXiv:2008.08675*, 2020.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 8141–8150, 2019a.

Sanjeev Arora, Simon S Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. Harnessing the power of infinitely wide deep nets on small-data tasks. In *International Conference on Learning Representations*, 2019b.

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

Tankut Can, Kamesh Krishnamurthy, and David J Schwab. Gating creates slow modes and controls phase-space complexity in grus and lstms. In *Mathematical and Scientific Machine Learning*, pp. 476–511. PMLR, 2020.

John Cardy. *Scaling and renormalization in statistical physics*, volume 5. Cambridge university press, 1996.

Minmin Chen, Jeffrey Pennington, and Samuel Schoenholz. Dynamical isometry and a mean field theory of rnns: Gating enables signal propagation in recurrent neural networks. In *International Conference on Machine Learning*, pp. 873–882. PMLR, 2018.

Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams. In *International Conference on Learning Representations*, 2019.

Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

Boris Hanin. Random neural networks in the infinite width limit as gaussian processes. *arXiv preprint arXiv:2107.01562*, 2021.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Tianyu He, Darshil Doshi, and Andrey Gromov. Autoinit: Automatic initialization via jacobian tuning, 2022.

Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks. In *International Conference on Machine Learning*, pp. 4376–4386. PMLR, 2020.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32:8572–8583, 2019.

Jaehoon Lee, Samuel S Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. 2020.

Aitor Lewkowycz and Guy Gur-Ari. On the training dynamics of deep networks with l_2 regularization. *arXiv preprint arXiv:2006.08643*, 2020.

James Martens, Andy Ballard, Guillaume Desjardins, Grzegorz Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S Schoenholz. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping. *arXiv preprint arXiv:2110.01765*, 2021.

Alexander G de G Matthews, Jiri Hron, Mark Rowland, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

Lutz Molgedey, J Schuchhardt, and Heinz G Schuster. Suppressing chaos in neural networks by noise. *Physical review letters*, 69(26):3717, 1992.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, 1996.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018a.

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2018b.

Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A Alemi, Jascha Sohl-Dickstein, and Samuel S Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*, 2019.

Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1924–1932. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/pennington18a.html.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29:3360–3368, 2016.

Daniel A. Roberts, Sho Yaida, and Boris Hanin. *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press, 2022.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. 2016.

Vaishaal Shankar, Alex Fang, Wenshuo Guo, Sara Fridovich-Keil, Jonathan Ragan-Kelley, Ludwig Schmidt, and Benjamin Recht. Neural kernels without tangents. In *International Conference on Machine Learning*, pp. 8614–8623. PMLR, 2020.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

Russell Tsuchida, Tim Pearce, Chris van der Heide, Fred Roosta, and Marcus Gallagher. Avoiding kernel fixed points: Computing with elu and gelu infinite networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9967–9977, 2021.

Christopher Williams. Computing with infinite networks. In M. C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1997.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pp. 5393–5402. PMLR, 2018.

Lechao Xiao, Jeffrey Pennington, and Sam Schoenholz. Disentangling trainability and generalization in deep learning, 2020.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pp. 10524–10533. PMLR, 2020.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *arXiv preprint arXiv:1911.07013*, 2019.

Sho Yaida. Non-gaussian processes and neural networks at finite widths. In *Mathematical and Scientific Machine Learning*, pp. 165–192. PMLR, 2020.

Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019a.

Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. 2019b.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.

Greg Yang and Samuel S Schoenholz. Mean field residual networks: On the edge of chaos. *arXiv preprint arXiv:1712.08969*, 2017.

Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S Schoenholz. A mean field theory of batch normalization. In *International Conference on Learning Representations*, 2018.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. *arXiv preprint arXiv:2111.11418*, 2021.

A Experimental Details

We implemented our methods using PyTorch (Paszke et al., 2019) hooks and an efficient Jacobian approximate algorithm (Hoffman et al., 2019).

Figure 1: 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 2: In all cases, networks are trained for 10 epochs using stochastic gradient descent with CrossEntropy loss. We used the Fashion MNIST dataset Xiao et al. (2017). 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 3: (a)(b)We made the phase diagram for MLP-Mixer with 30 blocks and averaged over 100 different parameter-initializations. (c)(d)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 $\text{bs} = 256$, 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 performances. Generating the data for the figure took approximately 4 days on Google Colab Pro (single Tesla P100 GPU).

B 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.

B.1 NNGP Kernel

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

$$\mathbb{E}_\theta[w_{ij}^l w_{mn}^l] = \frac{\sigma_w^2}{N_{l-1}} \delta_{im}\delta_{jn} \text{ and } \mathbb{E}_\theta[b_i^l b_j^l] = \sigma_b^2 \delta_{ij} \tag{24}$$

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 Roberts et al. (2022).

Proof of Theorem 2.4. One can prove this by definition with Lemma 2.2.

$$
\begin{aligned}
\mathcal{K}^{l+1} \equiv& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta[h_i^{l+1} h_i^{l+1}] \\
=& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\left(\sum_{j=1}^{N_l} w_{ij}^{l+1} \phi(h_j^l) + b_i^{l+1} \right) \left(\sum_{k=1}^{N_l} w_{ik}^{l+1} \phi(h_k^l) + b_i^{l+1} \right) \right] \\
=& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\sum_{j=1}^{N_l} \sum_{k=1}^{N_l} w_{ij}^{l+1} w_{ik}^{l+1} \phi(h_j^l) \phi(h_k^l) + b_i^{l+1} b_i^{l+1} \right] \\
=& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(h_j^l) \phi(h_j^l) + \sigma_b^2 \right] \\
=& \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(h_j^l) \phi(h_j^l) \right] + \sigma_b^2 \,.
\end{aligned} \tag{25}
$$

□

B.2 Jacobians

Next, we prove Theorem 2.5.

Proof of Theorem 2.5. We start from the definition of the partial Jacobian ($l > l_0$)

$$
\begin{aligned}
\mathcal{J}^{l_0,l+1} \equiv& \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \right] \\
=& \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \left(\sum_{k=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial h_k^l} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \left(\sum_{m=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial h_m^l} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
=& \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} \left(w_{ik}^{l+1} \phi'(h_k^l) \right) \left(w_{im}^{l+1} \phi'(h_m^l) \right) \left(\frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
=& \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} w_{ik}^{l+1} w_{im}^{l+1} \phi'(h_k^l) \phi'(h_m^l) \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right] \\
=& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k=1}^{N_l} \frac{\sigma_w^2}{N_l} \mathbb{E}_\theta \left[\phi'(h_k^l) \phi'(h_k^l) \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right] \\
=& \frac{\sigma_w^2}{N_l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(h_k^l) \phi'(h_k^l) \left(\sum_{j=1}^{N_{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \right] \,.
\end{aligned} \tag{26}
$$

In the infinite width limit, the sum over neurons in a layer self-averages due to the law of large numbers. Also the distribution of h_i^{l+1} is independent of h_i^l. This allows us to represent the expectation value of a product as product of expectation values. (This holds for $l_0 \neq 0$. We will show momentarily that the $l_0 = 0$ case acquires corrections due to finite input width N_0). Thus we have

$$
\begin{aligned}
\mathcal{J}^{l_0,l+1} &= \sigma_w^2 \mathbb{E}_\theta[\phi'(h_k^l)\phi'(h_k^l)] \mathbb{E}_\theta \left[\frac{1}{N_l} \sum_{k=1}^{N_l} \sum_{j=1}^{N_{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right] \\
&= \sigma_w^2 \mathbb{E}_\theta \left[\phi'(h_k^l) \phi'(h_k^l) \right] \mathcal{J}^{l_0,l} \\
\implies \mathcal{J}^{l_0,l+1} &= \chi_{\mathcal{J}}^l \mathcal{J}^{l_0,l} \,,
\end{aligned} \tag{27}
$$

where in the first line we used Lemma 2.2. The first integral is taken over the distribution of h_k^l, and the second integral is taken over the distribution of h_k^{l-1}. $\chi_{\mathcal{J}}^l$ is defined by Eq.(12). □

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

$$\mathcal{J}^{0,2} = \left(\chi_{\mathcal{J}}^1 + \frac{2\sigma_w^2}{N_0} \chi_{\Delta}^1 \sum_k^{N_0} \frac{1}{N_0} h_k^0 h_k^0 \right) \mathcal{J}^{0,1} \,, \tag{28}$$

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

Proof.

$$\begin{aligned}
\mathcal{J}^{0,2} =& \frac{1}{N_2} \mathbb{E}_\theta \left[\sum_{i=1}^{N_2} \sum_{j=1}^{N_0} \frac{\partial h_i^2}{\partial h_j^0} \frac{\partial h_i^2}{\partial h_j^0} \right] \\
=& \frac{1}{N_2} \mathbb{E}_\theta \left[\sum_{i=1}^{N_2} \sum_{j=1}^{N_0} \sum_{k,m=1}^{N_1} w_{ik}^2 w_{im}^2 \phi'(h_k^1) \phi'(h_m^1) \frac{\partial h_k^1}{\partial h_j^0} \frac{\partial h_m^1}{\partial h_j^0} \right] \\
=& \frac{1}{N_2} \sum_{i=1}^{N_2} \sum_{j=1}^{N_0} \sum_{k,m=1}^{N_1} \mathbb{E}_\theta [w_{ik}^2 w_{im}^2 w_{kj}^1 w_{mj}^1 \phi'(h_k^1) \phi'(h_m^1)] \\
=& \sum_{j=1}^{N_0} \sum_{k=1}^{N_1} \frac{\sigma_w^2}{N_1} \mathbb{E}_\theta [w_{kj}^1 w_{kj}^1 \phi'(h_k^1) \phi'(h_k^1)] \\
=& \sigma_w^2 \left(\chi_{\mathcal{J}}^1 + \frac{2\sigma_w^2}{N_0} \chi_{\Delta}^1 \sum_k^{N_0} \frac{1}{N_0} h_k^0 h_k^0 \right) \\
=& \left(\chi_{\mathcal{J}}^1 + \frac{2\sigma_w^2}{N_0} \chi_{\Delta}^1 \sum_k^{N_0} \frac{1}{N_0} h_k^0 h_k^0 \right) \mathcal{J}^{0,1} \,,
\end{aligned} \tag{29}$$

where to get the result we used integrate by parts, then explicitly integrated over w_{ij}^1. □

We defined a new quantity χ_{Δ}^l.

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

$$\chi_{\Delta}^l = \frac{\sigma_w^2}{N_l} \sum_{i=1}^{N_l} \mathbb{E}_\theta [\phi''(h_i^l)\phi''(h_i^l) + \phi'''(h_i^l)\phi'(h_i^l)] \,. \tag{30}$$

Remark B.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 justified factorization of the integral when we go from the last line of Eq.(26) to Eq.(27).

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

Theorem B.4 (Partial Jacobian)**.** *The partial Jacobian of a given network can be written as*

$$\mathcal{J}^{0,l} = \sigma_w^2 \left(\chi_{\mathcal{J}}^1 + \frac{2\sigma_w^2}{N_0} \chi_\Delta^1 \sum_k^{N_0} \frac{1}{N_0} h_k^0 h_k^0 \right) \prod_{l'=2}^{l-1} \chi_{\mathcal{J}}^{l'} \,, \tag{31}$$

where any partial Jacobian with $l_0 > 0$ does not receive an $O(N_0^{-1})$ correction.

B.3 LayerNorm on Pre-activations

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

$$\tilde{h}_i^l = \frac{h_i^l - \mathbb{E}[h^l]}{\sqrt{\mathbb{E}[(h^l)^2] - \mathbb{E}[h^l]^2}} \gamma_i^l + \beta_i^l \,, \tag{32}$$

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

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

$$h_i^{l+1} = \sum_{j=1}^{N_l} w_{ij}^{l+1} \phi(\tilde{h}_j^l) + b_i^{l+1} \,. \tag{33}$$

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

$$\mathbb{E}\left[h^l\right] = \mathbb{E}_\theta\left[h^l\right] = 0 \,, \tag{34}$$

$$\mathbb{E}\left[\left(h^l\right)^2\right] = \mathbb{E}_\theta\left[\left(h^l\right)^2\right] = \mathcal{K}^l \,. \tag{35}$$

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

Remark B.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

$$\tilde{h}_i^l = \frac{h_i^l - \mathbb{E}_\theta[h^l]}{\sqrt{\mathbb{E}_\theta[(h^l)^2] - \mathbb{E}_\theta[h^l]^2}} \,. \tag{36}$$

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

$$\mathbb{E}_\theta\left[O(\tilde{h}_i^l)\right] = \frac{1}{\sqrt{2\pi}} \int d\tilde{h}_i^l \, O(\tilde{h}_i^l) \, e^{-\frac{(\tilde{h}_i^l)^2}{2}} \,. \tag{37}$$

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

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta\left[\phi(\tilde{h}_j^l)\phi(\tilde{h}_j^l)\right] + \sigma_b^2 \,. \tag{38}$$

Proof.

$$\begin{aligned}
\mathcal{K}^{l+1} =& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta\left[h_i^{l+1} h_i^{l+1}\right] \\
=& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta\left[\left(\sum_{j=1}^{N_l} w_{ij}^{l+1}\phi(\tilde{h}_j^l) + b_i^{l+1}\right)\left(\sum_{k=1}^{N_l} w_{ik}^{l+1}\phi(\tilde{h}_k^l) + b_i^{l+1}\right)\right] \\
=& \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta\left[\phi(\tilde{h}_j^l)\phi(\tilde{h}_j^l)\right] + \sigma_b^2 \,.
\end{aligned} \tag{39}$$

□

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

$$\mathcal{J}^{l_0,l+1} = \chi^l_{\mathcal{J}} \mathcal{J}^{l_0,l} \,, \tag{40}$$

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.

$$\begin{aligned}
\mathcal{J}^{l_0,l+1} &= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \right] \\
&= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \left(\sum_{k=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial \tilde{h}_k^l} \frac{\partial \tilde{h}_k^l}{\partial h_k^l} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \left(\sum_{m=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial \tilde{h}_m^l} \frac{\partial \tilde{h}_m^l}{\partial h_m^l} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
&= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} \left(w_{ik}^{l+1} \phi'(\tilde{h}_k^l) \frac{1}{\sqrt{\mathcal{K}^l}} \right) \left(w_{im}^{l+1} \phi'(\tilde{h}_m^l) \frac{1}{\sqrt{\mathcal{K}^l}} \right) \left(\frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
&= \frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}_k^l) \phi'(\tilde{h}_k^l) \left(\sum_{j=1}^{N_{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \right] \\
&= \frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}_k^l) \phi'(\tilde{h}_k^l) \right] \mathcal{J}^{l_0,l} \\
&= \chi^l_{\mathcal{J}} \mathcal{J}^{l_0,l} \,,
\end{aligned} \tag{41}$$

□

B.4 LayerNorm on activations

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

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

$$\widetilde{\phi(h_i^l)} = \frac{\phi(h_i^l) - \mathbb{E}[\phi(h^l)]}{\sqrt{\mathbb{E}[\phi(h^l)^2] - \mathbb{E}[\phi(h^l)]^2}} \gamma_i^l + \beta_i^l \,. \tag{42}$$

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

$$h_i^{l+1} = \sum_{j=1}^{N_l} w_{ij}^{l+1} \widetilde{\phi(h_j^l)} + b_i^{l+1} \,. \tag{43}$$

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

$$\begin{aligned}
\widetilde{\phi(h_i^l)} &= \frac{\phi(h_i^l) - \mathbb{E}[\phi(h^l)]}{\sqrt{\mathbb{E}[\phi(h^l)^2] - \mathbb{E}[\phi(h^l)]^2}} \\
&= \frac{\phi(h_i^l) - \mathbb{E}_\theta\left[\phi(h^l)\right]}{\sqrt{\mathbb{E}_\theta\left[\phi(h^l)^2\right] - \mathbb{E}_\theta\left[\phi(h^l)\right]^2}} \,,
\end{aligned} \tag{44}$$

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 B.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

$$\mathbb{E}_\theta\left[O(h_i^l)\right] = \frac{1}{\sqrt{2\pi(\sigma_w^2 + \sigma_b^2)}} \int dh_i^l \, O(h_i^l) \, e^{-\frac{(h_i^l)^2}{2(\sigma_w^2 + \sigma_b^2)}} \,. \tag{45}$$

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

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

$$\mathcal{K}^{l+1} = \sigma_w^2 + \sigma_b^2 \,. \tag{46}$$

Proof.

$$\begin{aligned}
\mathcal{K}^{l+1} &= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[h_i^{l+1} h_i^{l+1} \right] \\
&= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\left(\sum_{j=1}^{N_l} w_{ij}^{l+1} \widetilde{\phi(h_j^l)} + b_i^{l+1} \right) \left(\sum_{k=1}^{N_l} w_{ik}^{l+1} \widetilde{\phi(h_k^l)} + b_i^{l+1} \right) \right] \\
&= \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\widetilde{\phi(h_j^l)}^2 \right] + \sigma_b^2 \\
&= \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\left(\frac{\phi(h_j^l) - \mathbb{E}_\theta \left[\phi(h^l) \right]}{\sqrt{\mathbb{E}_\theta \left[\phi(h^l)^2 \right] - \mathbb{E}_\theta \left[\phi(h^l) \right]^2}} \right)^2 \right] + \sigma_b^2 \\
&= \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \frac{\mathbb{E}_\theta \left[\left(\phi(h_j^l) - \mathbb{E}_\theta \left[\phi(h^l) \right] \right)^2 \right]}{\mathbb{E}_\theta \left[\phi(h^l)^2 \right] - \mathbb{E}_\theta \left[\phi(h^l) \right]^2} + \sigma_b^2 \\
&= \sigma_w^2 + \sigma_b^2 \,.
\end{aligned} \tag{47}$$

□

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

$$\mathcal{J}^{l_0,l+1} = \chi_J^l \mathcal{J}^{l_0,l} \,, \tag{48}$$

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.

$$\begin{aligned}
\mathcal{J}^{l_0,l+1} &= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \right] \\
&= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \left(\sum_{k=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial h_k^l} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \left(\sum_{m=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial h_m^l} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
&= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} \left(w_{ik}^{l+1} \widetilde{\phi'(h_k^l)} \right) \left(w_{im}^{l+1} \widetilde{\phi'(h_m^l)} \right) \left(\frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\
&= \frac{\sigma_w^2}{N_l} \sum_{k=1}^{N_l} \sum_{j=1}^{N_{l_0}} \mathbb{E}_\theta \left[\widetilde{\phi'(h_k^l)} \widetilde{\phi'(h_k^l)} \left(\sum_{j=1}^{N_{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \right] \\
&= \frac{\sigma_w^2}{N_l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\widetilde{\phi'(h_k^l)}^2 \right] \mathcal{J}^{l_0,l} \\
&= \sigma_w^2 \frac{\mathbb{E}_\theta \left[\phi'(h^l)^2 \right]}{\mathbb{E}_\theta \left[\phi(h^l)^2 \right] - \mathbb{E}_\theta \left[\phi(h^l) \right]^2} \mathcal{J}^{l_0,l} \\
&= \chi_J^l \mathcal{J}^{l_0,l} \,,
\end{aligned} \tag{49}$$

□

C Critical Exponents

To prove Theorem 2.7, we first need to find the critical exponent of the NNGP kernel Roberts et al. (2022).

Lemma C.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 $\delta\mathcal{K}^l$ 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*

$$\delta\mathcal{K}^l = \frac{1}{c_n(n-1)}\, l^{-\zeta_\mathcal{K}}\,, \tag{50}$$

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

Remark C.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.

$$\delta\mathcal{K}^{l+1} \approx \delta\mathcal{K}^l - c_n(\delta\mathcal{K}^l)^n\,. \tag{51}$$

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

$$(l+1)^{-\zeta_\mathcal{K}} = l^{-\zeta_\mathcal{K}} - c_n A\, l^{-n\zeta_\mathcal{K}}\,. \tag{52}$$

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

$$\frac{\zeta_\mathcal{K}}{l} = c_n A l^{-(n-1)\zeta_\mathcal{K}}\,. \tag{53}$$

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

$$A = \frac{1}{c_n(n-1)} \text{ and } \zeta_\mathcal{K} = \frac{1}{n-1}\,. \tag{54}$$

□

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

$$\begin{aligned}\chi^l_\mathcal{J} \approx& 1 - d_1\delta\mathcal{K}^l \\ =& 1 - \frac{d_1}{c_2} l^{-1}\,.\end{aligned} \tag{55}$$

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

$$\begin{aligned}\mathcal{J}^{l_0,l} =& \prod_{l'=l_0}^{l-1}\left(1 - \frac{d_1}{c_2} l'^{-1}\right)\mathcal{J}^{l_0,l_0} \\ \approx& c_{l_0}\cdot l^{-\zeta}\,.\end{aligned} \tag{56}$$

When $c_n = 0$ for all $n \geq 2$, from Lemma C.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.4. The agreement with infinite width calculation (following sections) is excellent. [2]

[2] We note that for this particular experiment, we used NTK parameterization for MLP. However, we emphasize that this does not affect the results.

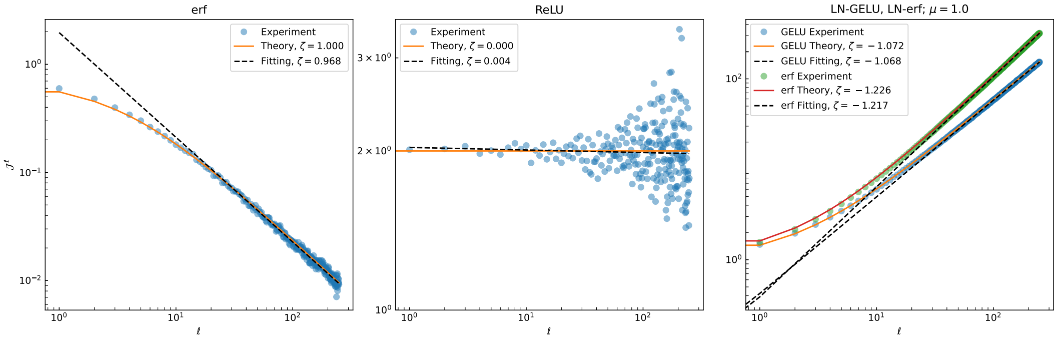

Figure 4: log–log plot of the partial Jacobian $\mathcal{J}^{0,l}$ vs. l for erf, ReLU and erf together with GELU (with LayerNorm applied to preactivations and residual connections of strength 1) activation functions. The critical exponents predicted from the infinite width analysis are in agreement with the data. The fluctuations get larger towards the output because the aspect ratio (*i.e.* L/N_l) approaches $1/4$.

D Residual Connections

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

$$h_i^{l+1} = \sum_{j=1}^{N_l} w_{ij}^{l+1} \phi(h_j^l) + b_i^{l+1} + \mu h_i^l \,, \tag{57}$$

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

Remark D.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 D.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 D.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*

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(h_j^l) \phi(h_j^l) \right] + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \tag{58}$$

Proof.

$$\begin{aligned}
\mathcal{K}^{l+1} &= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[h_i^{l+1} h_i^{l+1} \right] \\
&= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\left(\sum_{j=1}^{N_l} w_{ij}^{l+1} \phi(h_j^l) + b_i^{l+1} + \mu h_i^l \right) \left(\sum_{k=1}^{N_l} w_{ik}^{l+1} \phi(h_k^l) + b_i^{l+1} + \mu h_i^l \right) \right] \\
&= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\sum_{j=1}^{N_l} \sum_{k=1}^{N_l} w_{ij}^{l+1} w_{ik}^{l+1} \phi(h_j^l) \phi(h_k^l) + b_i^{l+1} b_i^{l+1} + \mu^2 h_i^l h_i^l \right] \\
&= \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(h_j^l) \phi(h_j^l) + \sigma_b^2 \right] + \mu^2 \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[h_i^l h_i^l \right] \\
&= \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(h_j^l) \phi(h_j^l) \right] + \sigma_b^2 + \mu^2 \mathcal{K}^l \,,
\end{aligned} \tag{59}$$

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

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

$$\mathcal{J}^{l_0,l+1} = \chi^l_{\mathcal{J}} \mathcal{J}^{l_0,l} \,, \tag{60}$$

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_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) + \mu \delta_{ik} \right) \left(w_{im}^{l+1} \phi'(h_m^l) + \mu \delta_{im} \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} \left(w_{ik}^{l+1} w_{im}^{l+1} \phi'(h_k^l) \phi'(h_m^l) + \mu^2 \delta_{ik} \delta_{im} \right) \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_m^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] + \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_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \right] \\
&= \left(\sigma_w^2 \mathbb{E}_\theta \left[\phi'(h_k^l) \phi'(h_k^l) \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_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right] \\
&= \left(\sigma_w^2 \mathbb{E}_\theta \left[\phi'(h_k^l) \phi'(h_k^l) \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{61}$$

□

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

We recall the recurrence relation (1):

$$h_i^{l+1} = \sum_{j=1}^{N_l} w_{ij}^{l+1} \phi(\tilde{h}_j^l) + b_i^{l+1} + \mu h_i^l \,. \tag{62}$$

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

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(\tilde{h}_j^l) \phi(\tilde{h}_j^l) \right] + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \tag{63}$$

Proof.

$$\begin{aligned}\mathcal{K}^{l+1} =& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[h_i^{l+1} h_i^{l+1} \right] \\ =& \frac{1}{N_{l+1}} \sum_{i=1}^{N_{l+1}} \mathbb{E}_\theta \left[\left(\sum_{j=1}^{N_l} w_{ij}^{l+1} \phi(\tilde{h}_j^l) + b_i^{l+1} + \mu h_i^l \right) \left(\sum_{k=1}^{N_l} w_{ik}^{l+1} \phi(\tilde{h}_k^l) + b_i^{l+1} + \mu h_i^l \right) \right] \\ =& \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(\tilde{h}_j^l) \phi(\tilde{h}_j^l) \right] + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \end{aligned} \tag{64}$$

□

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

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

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 E.3. For $\mu = 1$ case, the solution of (63) is

$$\mathcal{K}^l = \mathcal{K}^0 + \sum_{l'=1}^{l} \left(\frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \mathbb{E}_\theta \left[\phi(\tilde{h}_j^{l'}) \phi(\tilde{h}_j^{l'}) \right] + \sigma_b^2 \right) \,. \tag{66}$$

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

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

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

where for this case

$$\chi_{\mathcal{J}}^l = \frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}_k^l) \phi'(\tilde{h}_k^l) \right] + \mu^2 \,. \tag{68}$$

Proof.

$$\begin{aligned}\mathcal{J}^{l_0,l+1} &= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \frac{\partial h_i^{l+1}}{\partial h_j^{l_0}} \right] \\ &= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \left(\sum_{k=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial \tilde{h}_k^l} \frac{\partial \tilde{h}_k^l}{\partial h_k^l} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \left(\sum_{m=1}^{N_l} \frac{\partial h_i^{l+1}}{\partial \tilde{h}_m^l} \frac{\partial \tilde{h}_m^l}{\partial h_m^l} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\ &= \frac{1}{N_{l+1}} \mathbb{E}_\theta \left[\sum_{i=1}^{N_{l+1}} \sum_{j=1}^{N_{l_0}} \sum_{k,m=1}^{N_l} \left(\frac{w_{ik}^{l+1} \phi'(\tilde{h}_k^l)}{\sqrt{\mathcal{K}^l}} + \mu \delta_{ik} \right) \left(\frac{w_{im}^{l+1} \phi'(\tilde{h}_m^l)}{\sqrt{\mathcal{K}^l}} + \mu \delta_{ik} \right) \left(\frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_m^l}{\partial h_j^{l_0}} \right) \right] \\ &= \mathbb{E}_\theta \left[\left(\frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \phi'(\tilde{h}_k^l) \phi'(\tilde{h}_k^l) + \mu^2 \right) \left(\sum_{j=1}^{N_{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \frac{\partial h_k^l}{\partial h_j^{l_0}} \right) \right] \\ &= \left(\frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}_k^l) \phi'(\tilde{h}_k^l) \right] + \mu^2 \right) \mathcal{J}^{l_0,l} \\ &= \chi_{\mathcal{J}}^l \mathcal{J}^{l_0,l} \,, \end{aligned} \tag{69}$$

□

Remark E.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 E.6. As we mentioned above $\mu = 1$ needs extra care. Plug in the result (66) and $\mu = 1$ we find out that

$$\begin{aligned}\chi^l_{\mathcal{J}}\,|_{\mu=1} &= \frac{\sigma_w^2 \sum_{k=1}^{N_l} \mathbb{E}_\theta\left[\phi'(\tilde{h}^l_k)\phi'(\tilde{h}^l_k)\right]}{N_l\mathcal{K}^0 + \sum_{l'=1}^{l}\left(\sigma_w^2\sum_{j=1}^{N_l}\mathbb{E}_\theta\left[\phi(\tilde{h}^l_j)\phi(\tilde{h}^l_j)\right] + N_l\sigma_b^2\right)} + 1 \\ &\sim 1 + O\left(\frac{1}{l}\right)\,, \end{aligned} \tag{70}$$

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

Recall that $\xi = |\log\chi^\star_{\mathcal{J}}|^{-1}$, then Theorem 1.3 is a summary of (68) and (70) in $l \to \infty$ limit.

F MLP-Mixer

In this section we would like to analyze an architecture called MLP-Mixer Tolstikhin et al. (2021), 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

$$h^0_{\mu i} = \sum_{j=1}^{f^2}\sum_{\nu=1}^{C_{in}} W^0_{\mu\nu;j}\, x_{\nu, j+(i-1)s^2} + b^0_{\mu i}\,, \tag{71}$$

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 $k = (C_{in}f^2)^{-1}$.

$$\mathbb{E}_\theta[W^0_{\mu\nu;i}W^0_{\rho\sigma;j}] = \frac{1}{3C_{in}f^2}\delta_{\mu\rho}\delta_{\nu\sigma}\delta_{ij}\,, \tag{72}$$

$$\mathbb{E}_\theta[b^0_{\mu i}b^0_{\nu j}] = \frac{1}{3C_{in}f^2}\delta_{\mu\nu}\delta_{ij}\,. \tag{73}$$

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.

$$\tilde{h}^{6l}_{\mu i} = \frac{h^{6l}_{\mu i} - \mathbb{E}_C[h^{6l}_{\rho i}]}{\sqrt{\mathrm{Var}_C[h^{6l}_{\rho i}]}}\,, \tag{74}$$

  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.

$$h_{\mu j}^{6l+1} = \sum_{k=1}^{N_p} w_{jk}^{6l+1} \tilde{h}_{\mu k}^{6l} + b_j^{6l+1} \,. \tag{75}$$

- $6l + 2$: Affine Layer.

$$h_{\mu i}^{6l+2} = \sum_{j=1}^{N_{tm}} w_{ij}^{6l+2} \phi(h_{\mu j}^{6l+1}) + b_i^{6l+2} \,, \tag{76}$$

where N_{tm} stands for hidden dimension of ”token mixing”.

- $6l + 3$: Residual Connections.

$$h_{\mu i}^{6l+3} = h_{\mu i}^{6l+2} + \mu h_{\mu i}^{6l} \,. \tag{77}$$

- Second LayerNorm. It again acts on channels μ.

$$\tilde{h}_{\mu i}^{6l+3} = \frac{h_{\mu i}^{6l+3} - \mathbb{E}_C[h_{\rho i}^{6l+3}]}{\sqrt{\mathrm{Var}_C[h_{\rho i}^{6l+3}]}} \,. \tag{78}$$

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

$$h_{\nu i}^{6l+4} = \sum_{\rho=1}^{C} w_{\nu\rho}^{6l+4} \tilde{h}_{\rho i}^{6l+3} + b_\nu^{6l+4} \,. \tag{79}$$

  - $6l + 5$. Affine Layer.

$$h_{\mu i}^{6l+5} = \sum_{\nu=1}^{N_{cm}} w_{\mu\nu}^{6l+5} \phi(h_{\nu i}^{6l+4}) + b_\mu^{6l+5} \,. \tag{80}$$

  - $6l + 6$. Residual Connections.

$$h_{\mu i}^{6l+6} = h_{\mu i}^{6l+5} + \mu h_{\mu i}^{6l+3}. \tag{81}$$

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

$$\tilde{h}_{\mu i}^{6L} = \frac{h_{\mu i}^{6L} - \mathbb{E}_C[h_{\rho i}^{6L}]}{\sqrt{\mathrm{Var}_C[h_{\rho i}^{6L}]}} \,. \tag{82}$$

Global Average Pool over patches i.

$$h_\mu^p = \frac{1}{N_p} \sum_{i=1}^{N_p} \tilde{h}_{\mu i}^{6L} \,. \tag{83}$$

Output Layer

$$f_\mu = \sum_{\nu=1}^{C} w_{\mu\nu} h_\nu^p + b_\mu \,. \tag{84}$$

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

$$\chi_{\mathcal{J}}^\star = \lim_{L\to\infty} \left(\frac{1}{N_p C} \sum_{i=1}^{N_p} \sum_{\mu=1}^{C} \mathbb{E}_\theta \left[\sum_{\rho=1}^{C} \sum_{k=1}^{N_p} \frac{\partial h_{\mu i}^{6L}}{\partial h_{\rho k}^{6L-6}} \frac{\partial h_{\mu i}^{6L}}{\partial h_{\rho k}^{6L-6}} \right] \right) \,. \tag{85}$$

G Results for Scale Invariant Activation Functions

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

$$\phi(x) = a_+ \, x \, \Theta(x) + a_- \, x \, \Theta(-x) \,, \tag{86}$$

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

G.1 NNGP Kernel

First evaluate the average using Lemma 2.2

$$\begin{aligned}\mathbb{E}_\theta \left[\phi(h_i^l)\phi(h_i^l)\right] =& \frac{1}{\sqrt{2\pi\mathcal{K}^l}} \int dh_i^l \left(a_+^2 + a_-^2\right) \left(h_i^l\right)^2 e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\ =& \frac{a_+^2 + a_-^2}{2} \mathcal{K}^l \,.\end{aligned} \tag{87}$$

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

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2 (a_+^2 + a_-^2)}{2} \mathcal{K}^l + \sigma_b^2 \,. \tag{88}$$

Finite fixed point of the recurrence relation above exists only if

$$\chi_{\mathcal{K}}^\star = \frac{\sigma_w^2 (a_+^2 + a_-^2)}{2} \le 1 \,. \tag{89}$$

As a result

$$\sigma_w^2 \le \frac{2}{a_+^2 + a_-^2} \,. \tag{90}$$

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

G.2 Jacobian(s)

The calculation is quite straight forward, by definition

$$\begin{aligned}\chi_{\mathcal{J}}^l =& \sigma_w^2 \mathbb{E}_\theta \left[\phi'(h_i^l)\phi'(h_i^l)\right] \\ =& \frac{\sigma_w^2}{\sqrt{2\pi\mathcal{K}^l}} \int dh_i^l \left[a_+ \Theta(h_i^l) - a_- \Theta(h_i^l)\right]^2 e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\ =& \frac{\sigma_w^2 (a_+^2 + a_-^2)}{2} \,,\end{aligned} \tag{91}$$

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

$$\sigma_w = \sqrt{\frac{2}{a_+^2 + a_-^2}} \,. \tag{92}$$

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

$$\sigma_w = \sqrt{2} \,, \tag{93}$$

where the critical point is located at

$$(\sigma_w, \sigma_b) = (\sqrt{2}, 0) \,. \tag{94}$$

G.3 Critical Exponents

Since the recurrence relations for the NNGP kernel and Jacobians are linear. Then from Lemma C.1 and Theorem 2.7

$$\zeta_{\mathcal{K}} = 0 \text{ and } \zeta = 0 \,. \tag{95}$$

G.4 LayerNorm on Pre-activations

Use Lemma B.9 and combine all known results for scale invariant functions

$$\begin{aligned}\chi_{\mathcal{J}}^{l} &= \frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l) \right] \Bigg|_{\tilde{\mathcal{K}}^{l-1}=1} \\ &= \frac{\sigma_w^2 (a_+^2 + a_-^2)}{\sigma_w^2 (a_+^2 + a_-^2) + 2\sigma_b^2} \,. \end{aligned} \tag{96}$$

For this case,

$$\chi_{\mathcal{J}}^{l} \le 1 \tag{97}$$

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

G.5 LayerNorm on Activations

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

$$\mathbb{E}_\theta \left[\phi'(h_i^l)\phi'(h_i^l) \right] = \frac{a_+^2 + a_-^2}{2} \,, \tag{98}$$

$$\mathbb{E}_\theta \left[\phi(h_i^l)\phi(h_i^l) \right] = \frac{a_+^2 + a_-^2}{2} (\sigma_w^2 + \sigma_b^2) \,. \tag{99}$$

There is a new expectation value we need to show explicitly

$$\begin{aligned}\mathbb{E}_\theta \left[\phi(h_i^l) \right] &= \frac{1}{\sqrt{2\pi(\sigma_w^2 + \sigma_b^2)}} \int_{-\infty}^{\infty} dh_i^l \phi(h_i^l) e^{-\frac{1}{2} h_i^l (\sigma_w^2 + \sigma_b^2)^{-1} h_i^l} \\ &= \frac{1}{\sqrt{2\pi(\sigma_w^2 + \sigma_b^2)}} \int_{0}^{\infty} dh_i^l (a_+ - a_-) h_i^l e^{-\frac{(h_i^l)^2}{2(\sigma_w^2 + \sigma_b^2)}} \\ &= (a_+ - a_-) \sqrt{\frac{\sigma_w^2 + \sigma_b^2}{2\pi}} \,. \end{aligned} \tag{100}$$

Thus

$$\chi_{\mathcal{J}}^{l} = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_b^2} \cdot \frac{\pi(a_+^2 + a_-^2)}{\pi(a_+^2 + a_-^2) - (a_+ - a_-)^2} \,. \tag{101}$$

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

$$\sigma_b = \sqrt{\frac{(a_+ - a_-)^2}{\pi(a_+^2 + a_-^2) - (a_+ - a_-)^2}} \sigma_w \,. \tag{102}$$

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

$$\begin{aligned}\sigma_b =& \sqrt{\frac{1}{\pi - 1}} \sigma_w \\ \approx& 0.683 \sigma_w \,. \end{aligned} \tag{103}$$

G.6 Residual Connections

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

$$\mathcal{K}^{l+1} = \frac{\sigma_w^2 (a_+^2 + a_-^2)}{2} \mathcal{K}^l + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \tag{104}$$

The condition for the existence of fixed point

$$\chi_{\mathcal{K}}^{\star} = \frac{\sigma_w^2 (a_+^2 + a_-^2)}{2} + \mu^2 \le 1 \tag{105}$$

leads us to

$$\sigma_w^2 \leq \frac{2(1-\mu^2)}{a_+^2 + a_-^2}\,. \tag{106}$$

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

$$\chi_{\mathcal{J}}^{\star} = \frac{\sigma_w^2(a_+^2 + a_-^2)}{2} + \mu^2\,. \tag{107}$$

The critical line is defined as

$$\sigma_w = \sqrt{\frac{2(1-\mu^2)}{a_+^2 + a_-^2}}\,. \tag{108}$$

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 $\left(\sqrt{2(1-\mu^2)}, 0\right)$.

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

Again use Lemma B.9 and combine all known results for scale invariant functions

$$\begin{aligned}
\chi_{\mathcal{J}}^{\star} &= \lim_{l\to\infty}\left(\frac{\sigma_w^2}{N_l \mathcal{K}^l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l)\right]\Big|_{\tilde{\mathcal{K}}^{l-1}=1} + \mu^2\right) \\
&= \frac{\sigma_w^2(a_+^2+a_-^2)(1-\mu^2)}{\sigma_w^2(a_+^2+a_-^2) + 2\sigma_b^2} + \mu^2 \\
&= 1 - \frac{2\sigma_b^2(1-\mu^2)}{\sigma_w^2(a_+^2+a_-^2) + 2\sigma_b^2}
\end{aligned} \tag{109}$$

Similar to the case without residue connections

$$\chi_{\mathcal{J}}^l \leq 1 \tag{110}$$

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.

H Results for erf Activation Function

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

$$\phi(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt\,. \tag{111}$$

H.1 NNGP Kernel

To evaluate Lemma 2.2 exactly, we introduce two dummy variables λ_1 and λ_2Williams (1997).

$$\begin{aligned}
\mathbb{E}_\theta\left[\phi(\lambda_1 h_i^l)\phi(\lambda_2 h_i^l)\right] &= \int d\lambda_1 \int d\lambda_2 \frac{d^2}{d\lambda_1 d\lambda_2}\mathbb{E}_\theta\left[\phi(\lambda_1 h_i^l)\phi(\lambda_2 h_i^l)\right] \\
&= \int d\lambda_1 \int d\lambda_2 \int dh_i^l \frac{4}{\sqrt{2\pi^3 \mathcal{K}^l}}\left(h_i^l\right)^2 e^{-\left(\lambda_1^2+\lambda_2^2+\frac{1}{2\mathcal{K}^l}\right)\left(h_i^l\right)^2} \\
&= \int d\lambda_1 \int d\lambda_2 \frac{4\mathcal{K}^l}{\pi\left(1+2\mathcal{K}^l(\lambda_1^2+\lambda_2^2)\right)} \\
&= \frac{2}{\pi}\arcsin\left(\frac{2\mathcal{K}^l\lambda_1\lambda_2}{1+2\mathcal{K}^l(\lambda_1^2+\lambda_2^2)}\right)\,.
\end{aligned} \tag{112}$$

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

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

$$\mathcal{K}^{l+1} = \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{2\mathcal{K}^l}{1+2\mathcal{K}^l}\right) + \sigma_b^2 \,. \tag{113}$$

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

$$\chi_{\mathcal{K}}^\star = \frac{4\sigma_w^2}{\pi} \frac{1}{(1+2\mathcal{K}^\star)\sqrt{1+4\mathcal{K}^\star}} \leq 1 \,. \tag{114}$$

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

H.2 Jacobians

Follow the definition

$$\begin{aligned}
\chi_{\mathcal{J}}^l &= \sigma_w^2 \mathbb{E}_\theta \left[\phi'(h_i^l)\phi'(h_i^l)\right] \\
&= \frac{4\sigma_w^2}{\sqrt{2\pi^3\mathcal{K}^l}} \int dh_i^l \; e^{-2(h_i^l)^2} \; e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
&= \frac{4\sigma_w^2}{\pi} \frac{1}{\sqrt{1+4\mathcal{K}^l}} \,.
\end{aligned} \tag{115}$$

To find phase boundary $\chi_{\mathcal{J}}^\star = 1$, we need to combine Eq.(113) and Eq.(115) and evaluate them at $\mathcal{K}^\star$.

$$\mathcal{K}^\star = \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{2\mathcal{K}^\star}{1+2\mathcal{K}^\star}\right) + \sigma_b^2 \,, \tag{116}$$

$$\chi_{\mathcal{J}}^\star = \frac{4\sigma_w^2}{\pi} \frac{1}{\sqrt{1+4\mathcal{K}^\star}} = 1 \,. \tag{117}$$

One can solve equations above and find the critical line

$$\sigma_b = \sqrt{\frac{16\sigma_w^4 - \pi^2}{4\pi^2} - \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{16\sigma_w^4 - \pi^2}{16\sigma_w^4 + \pi^2}\right)} \,. \tag{118}$$

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

$$(\sigma_w, \sigma_b) = \left(\sqrt{\frac{\pi}{4}}, 0\right) \,. \tag{119}$$

H.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.(113) can be rewritten as

$$\begin{aligned}
\delta\mathcal{K}^{l+1} =& \frac{1}{2} \arcsin\left(\frac{2\delta\mathcal{K}^l}{1+2\delta\mathcal{K}^l}\right) \\
\approx& \delta\mathcal{K}^l - 2(\delta\mathcal{K}^l)^2 \,.
\end{aligned} \tag{120}$$

From Lemma C.1

$$A = \frac{1}{2} \text{ and } \zeta_{\mathcal{K}} = 1 \,. \tag{121}$$

Next we analyze critical exponent of Jacobians by expanding (115) around $\mathcal{K}^\star = 0$ critical point $(\sigma_b, \sigma_w) = (0, \sqrt{\frac{\pi}{4}})$.

To leading order l^{-1} we have

$$\begin{aligned}\chi_{\mathcal{J}}^l \approx & 1 - 2\delta K^l \\ \approx & 1 - \frac{1}{l}\,.\end{aligned} \tag{122}$$

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

$$\mathcal{J}^{l_0, l+1} = \left(1 - \frac{1}{l}\right)\mathcal{J}^{l_0, l}\,. \tag{123}$$

At large l

$$\mathcal{J}^{l_0, l} = c_{l_0}\, l^{-1}\,, \tag{124}$$

with a non-universal constant c_{l_0}.

The critical exponent is

$$\zeta = 1\,, \tag{125}$$

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

H.4 LayerNorm on Pre-activations

Use Lemma B.9, we have

$$\begin{aligned}\chi_{\mathcal{J}}^l = & \frac{\sigma_w^2}{N_l \mathcal{K}^l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l)\right]\Bigg|_{\tilde{\mathcal{K}}^{l-1}=1} \\ = & \frac{4\sigma_w^2}{\sqrt{5}\left[2\sigma_w^2 \arcsin\left(\frac{2}{3}\right) + \pi\sigma_b^2\right]}\,.\end{aligned} \tag{126}$$

The critical line is then defined by

$$\begin{aligned}\sigma_b = & \sqrt{\frac{2}{\pi}\left[\frac{2}{\sqrt{5}} - \arcsin\left(\frac{2}{3}\right)\right]}\sigma_w \\ \approx & 0.324\sigma_w\,.\end{aligned} \tag{127}$$

H.5 LayerNorm on Activations

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

$$\mathbb{E}_\theta\left[\phi'(h_i^l)\phi'(h_i^l)\right] = \frac{4}{\pi}\frac{1}{\sqrt{1 + 4(\sigma_w^2 + \sigma_b^2)}}\,, \tag{128}$$

$$\mathbb{E}_\theta\left[\phi(h_i^l)\phi(h_i^l)\right] = \frac{2}{\pi}\arcsin\left(\frac{2(\sigma_w^2 + \sigma_b^2)}{1 + 2(\sigma_w^2 + \sigma_b^2)}\right)\,. \tag{129}$$

Thus

$$\chi_{\mathcal{J}}^l = \frac{2\sigma_w^2}{\sqrt{1 + 4(\sigma_w^2 + \sigma_b^2)}} \cdot \frac{1}{\arcsin\left(\frac{2(\sigma_w^2 + \sigma_b^2)}{1 + 2(\sigma_w^2 + \sigma_b^2)}\right)}\,, \tag{130}$$

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

H.6 Residual Connections

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

$$\mathcal{K}^{l+1} = \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{2\mathcal{K}^l}{1+2\mathcal{K}^l}\right) + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \tag{131}$$

Finite fixed point only exists when

$$\chi_{\mathcal{K}}^\star = \frac{4\sigma_w^2}{\pi} \frac{1}{(1+2\mathcal{K}^\star)\sqrt{1+4\mathcal{K}^\star}} + \mu^2 \leq 1 \,. \tag{132}$$

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

$$\chi_{\mathcal{J}}^\star = \frac{4\sigma_w^2}{\pi} \frac{1}{\sqrt{1+4\mathcal{K}^\star}} + \mu^2 \,. \tag{133}$$

The critical line is defined as

$$\sigma_b = \sqrt{\frac{16\sigma_w^4 - \pi^2(1-\mu^2)^2}{4\pi^2(1-\mu^2)} - \frac{2\sigma_w^2}{\pi} \arcsin\left(\frac{16\sigma_w^4 - \pi^2(1-\mu^2)^2}{16\sigma_w^4 + \pi^2(1-\mu^2)^2}\right)} \,. \tag{134}$$

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

$$(\sigma_w, \sigma_b) = \left(\sqrt{\frac{\pi(1-\mu^2)}{4}}, 0\right) \,. \tag{135}$$

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

$$\mathcal{K}^{l+1} \approx \sigma_w^2 + \sigma_b^2 + \mathcal{K}^l \,, \tag{136}$$

which regulates the first term in (133).

For $\mu = 1$ at large depth

$$\chi_{\mathcal{J}}^l \sim 1 + \frac{4\sigma_w^2}{\pi\sqrt{C_0 + 4(\sigma_w^2 + \sigma_b^2)l}} \,. \tag{137}$$

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

$$\begin{aligned}
\log \mathcal{J}^{l_0,l} &= \log\left(\prod_{l'=l_0}^{l} \chi_{\mathcal{J}}^{l'}\right) \\
&= \sum_{l'=l_0}^{l} \log\left(1 + \frac{4\sigma_w^2}{\pi\sqrt{C_0 + 4(\sigma_w^2 + \sigma_b^2)l'}}\right) \\
&\approx \int_{l_0}^{l} dl' \log\left(1 + \frac{4\sigma_w^2}{\pi\sqrt{C_0 + 4(\sigma_w^2 + \sigma_b^2)l'}}\right) \\
&\sim 2\tilde{c}\sqrt{l} + O(\log l) \,,
\end{aligned} \tag{138}$$

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

$$\mathcal{J}^{l_0,l} \sim O\left(e^{2\tilde{c}\sqrt{l} + O(\log l)}\right) \,. \tag{139}$$

This result checks out empirically, as shown in Figure 5.[3]

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

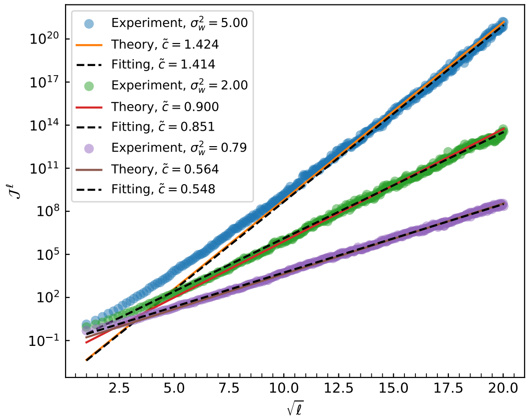

Figure 5: $\log(\mathcal{J}^{l_0,l})$-$\sqrt{l}$ for $\mu = 1$, $\sigma_b^2 = 0$, erf.

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

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

$$\begin{aligned}\chi_{\mathcal{J}}^* &= \lim_{l\to\infty}\left(\frac{\sigma_w^2}{N_l\mathcal{K}^l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l)\right]\Big|_{\tilde{\mathcal{K}}^{l-1}=1}+\mu^2\right)\\ &=\frac{4\sigma_w^2(1-\mu^2)}{\sqrt{5}\left[2\sigma_w^2\arcsin\left(\frac{2}{3}\right)+\pi\sigma_b^2\right]}+\mu^2\,. \end{aligned}\tag{140}$$

The critical line is then defined by

$$\begin{aligned}\sigma_b &= \sqrt{\frac{2}{\pi}\left[\frac{2}{\sqrt{5}}-\arcsin\left(\frac{2}{3}\right)\right]}\sigma_w\\ &\approx 0.324\sigma_w\,. \end{aligned}\tag{141}$$

I Results for GELU Activation Function

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

$$\begin{aligned}\phi(x) &= \frac{x}{2}\left[1+\text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]\\ &= \frac{x}{2}\left[1+\frac{2}{\sqrt{\pi}}\int_0^{\frac{x}{\sqrt{2}}}e^{-t^2}dt\right]\,. \end{aligned}\tag{142}$$

I.1 NNGP Kernel

Use Lemma 2.2 for GELU

$$\begin{aligned}
\mathbb{E}_\theta\left[\phi(h_i^l)\phi(h_i^l)\right] =&\frac{1}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l \frac{(h_i^l)^2}{4}\left[1+\text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right)\right]^2 e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=&\frac{1}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l\, \frac{(h_i^l)^2}{4}\left[1+\text{erf}^2\left(\frac{h_i^l}{\sqrt{2}}\right)\right] e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=&\frac{\mathcal{K}^l}{4}+\frac{1}{\sqrt{32\pi\mathcal{K}^l}}\int dh_i^l\,(h_i^l)^2\text{erf}^2\left(\frac{h_i^l}{\sqrt{2}}\right) e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=&\frac{\mathcal{K}^l}{4}+\frac{\mathcal{K}^l}{\sqrt{32\pi\mathcal{K}^l}}\int dh_i^l\,\text{erf}^2\left(\frac{h_i^l}{\sqrt{2}}\right) e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
&+\frac{(\mathcal{K}^l)^2}{\sqrt{32\pi\mathcal{K}^l}}\int dh_i^l\left[\text{erf}'\left(\frac{h_i^l}{\sqrt{2}}\right)\text{erf}'\left(\frac{h_i^l}{\sqrt{2}}\right)+\text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right)\text{erf}''\left(\frac{h_i^l}{\sqrt{2}}\right)\right] e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=&\frac{\mathcal{K}^l}{4}+\frac{\mathcal{K}^l}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^l}{1+\mathcal{K}^l}\right)+\frac{2\mathcal{K}^l}{(1+\mathcal{K}^l)\sqrt{1+2\mathcal{K}^l}}\right], \qquad (143)
\end{aligned}$$

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

$$\mathcal{K}^{l+1}=\left[\frac{\mathcal{K}^l}{4}+\frac{\mathcal{K}^l}{2\pi}\arcsin\left(\frac{\mathcal{K}^l}{1+\mathcal{K}^l}\right)+\frac{(\mathcal{K}^l)^2}{\pi(1+\mathcal{K}^l)\sqrt{1+2\mathcal{K}^l}}\right]\sigma_w^2+\sigma_b^2\,. \qquad (144)$$

As a result

$$\chi_{\mathcal{K}}^\star=\frac{\sigma_w^2}{4}+\frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right)+\frac{4(\mathcal{K}^\star)^3+11(\mathcal{K}^\star)^2+5\mathcal{K}^\star}{(1+\mathcal{K}^\star)^2(1+2\mathcal{K}^\star)^{\frac{3}{2}}}\right]. \qquad (145)$$

I.2 Jacobians

Follow the definition

$$\begin{aligned}
\chi_{\mathcal{J}}^l =&\sigma_w^2\mathbb{E}_\theta\left[\phi'(h_i^l)\phi'(h_i^l)\right] \\
=&\frac{\sigma_w^2}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l\left[\frac{1}{2}+\frac{1}{2}\text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right)+\frac{e^{-\frac{(h_i^l)^2}{2}}h_i^l}{\sqrt{2\pi}}\right]^2 e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=&\frac{\sigma_w^2}{\sqrt{2\pi\mathcal{K}^l}}\int dh_i^l\left[\frac{1}{4}+\frac{1}{4}\text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right)^2+\frac{h_i^l\text{erf}\left(\frac{h_i^l}{\sqrt{2}}\right)e^{-\frac{(h_i^l)^2}{2}}}{\sqrt{2\pi}}+\frac{e^{-(h_i^l)^2}(h_i^l)^2}{2\pi}\right] e^{-\frac{(h_i^l)^2}{2\mathcal{K}^l}} \\
=&\frac{\sigma_w^2}{4}+\frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^l}{1+\mathcal{K}^l}\right)+\frac{\mathcal{K}^l(3+5\mathcal{K}^l)}{(1+\mathcal{K}^l)(1+2\mathcal{K}^l)^{\frac{3}{2}}}\right], \qquad (146)
\end{aligned}$$

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 $\chi_{\mathcal{J}}^\star=1$

$$\mathcal{K}^\star=\frac{\sigma_w^2}{4}\mathcal{K}^\star+\frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right)+\frac{\sigma_w^2\mathcal{K}^\star}{\pi(1+\mathcal{K}^\star)\sqrt{1+2\mathcal{K}^\star}}\right]\mathcal{K}^\star+\sigma_b^2\,, \qquad (147)$$

$$\chi_{\mathcal{J}}^\star=\frac{\sigma_w^2}{4}+\frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right)+\frac{\mathcal{K}^\star(3+5\mathcal{K}^\star)}{(1+\mathcal{K}^\star)(1+2\mathcal{K}^\star)^{\frac{3}{2}}}\right]=1\,. \qquad (148)$$

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

$$\sigma_w = 2\left[1 + \frac{2\mathcal{K}^\star(3+5\mathcal{K}^\star)}{\pi(1+\mathcal{K}^\star)(1+2\mathcal{K}^\star)^{\frac{3}{2}}} + \frac{2}{\pi}\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right)\right]^{-\frac{1}{2}}\,, \tag{149}$$

$$\sigma_b = \frac{\mathcal{K}^\star}{\sqrt{2\pi}(1+2\mathcal{K}^\star)^{\frac{3}{4}}}\sigma_w\,. \tag{150}$$

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

$$\frac{\sigma_w^2[(\mathcal{K}^\star)^3 - 3(\mathcal{K}^\star)^2 - 2\mathcal{K}^\star]}{2\pi(1+\mathcal{K}^\star)^2(1+2\mathcal{K}^\star)^{\frac{3}{2}}} = 0\,, \tag{151}$$

which has two non-negative solutions out of three

$$\mathcal{K}^\star = 0 \text{ and } \mathcal{K}^\star = \frac{3+\sqrt{17}}{2}\,. \tag{152}$$

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

$$(\sigma_w, \sigma_b) = (2, 0)\,, \text{ for } \mathcal{K}^\star = 0\,, \tag{153}$$

$$(\sigma_w, \sigma_b) \approx (1.408, 0.416)\,, \text{ for } \mathcal{K}^\star = \frac{3+\sqrt{17}}{2}\,. \tag{154}$$

I.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.(144), and keep next to leading order $\delta\mathcal{K}^l = \mathcal{K}^l - \mathcal{K}^\star$

$$\delta\mathcal{K}^{l+1} \approx \delta\mathcal{K}^l + \frac{6}{\pi}(\delta\mathcal{K}^l)^2\,. \tag{155}$$

From Lemma C.1

$$A = -\frac{\pi}{6} \text{ and } \zeta_{\mathcal{K}} = 1\,, \tag{156}$$

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.

$$\delta\mathcal{K}^{l+1} \approx \delta\mathcal{K}^l + 0.00014(\delta\mathcal{K}^l)^2\,. \tag{157}$$

Following the same analysis, we find

$$\delta\mathcal{K}^l \approx -7142.9\, l^{-1}\,. \tag{158}$$

Looks like scaling analysis works for this case, since $\mathcal{K}^\star > 0$. The solution shows that the critical point is half-stableRoberts et al. (2022). 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}.

$$\chi^l_{\mathcal{J}} \approx 1 - \frac{66.668}{l}\,. \tag{159}$$

Then

$$\delta\mathcal{J}^{l_0,l} \approx c_{l_0} l^{-66.668}\,, \tag{160}$$

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

Critical exponent

$$\zeta = 66.668\,. \tag{161}$$

Which in practice is not traceable.

I.4 LayerNorm on Pre-activations

Use Lemma B.9, we have

$$\begin{aligned}\chi^l_{\mathcal{J}} = & \frac{\sigma_w^2}{N_l \mathcal{K}^l} \sum_{k=1}^{N_l} \mathbb{E}_\theta \left[\phi'(\tilde{h}^l_k) \phi'(\tilde{h}^l_k) \right] \Bigg|_{\tilde{\mathcal{K}}^{l-1}=1} \\ = & \frac{\sigma_w^2 (6\pi + 4\sqrt{3})}{\sigma_w^2 (6\pi + 3\sqrt{3}) + 18\pi \sigma_b^2} \,. \end{aligned} \tag{162}$$

The critical line is then at

$$\begin{aligned}\sigma_b = & \left(6\sqrt{3}\pi\right)^{-\frac{1}{2}} \sigma_w \\ \approx & 0.175 \sigma_w \,. \end{aligned} \tag{163}$$

I.5 LayerNorm on Activations

First we need to evaluate a new expectation value

$$\begin{aligned}\mathbb{E}_\theta \left[\phi(h^l_i) \right] = & \frac{1}{\sqrt{2\pi(\sigma_w^2 + \sigma_b^2)}} \int dh^l_i \frac{h^l_i}{2} \left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}} \right) \right] e^{-\frac{(h^l_i)^2}{2(\sigma_w^2 + \sigma_b^2)}} \\ = & \frac{\sigma_w^2 + \sigma_b^2}{\sqrt{2\pi(1 + \sigma_w^2 + \sigma_b^2)}} \,, \end{aligned} \tag{164}$$

where we used integrate by parts to get the result.

The other integrals are modified to

$$\mathbb{E}_\theta \left[\phi'(h^l_i) \phi'(h^l_i) \right] = \frac{1}{4} + \frac{1}{2\pi} \left[\arcsin\left(\frac{\sigma_w^2 + \sigma_b^2}{1 + \sigma_w^2 + \sigma_b^2} \right) + \frac{(\sigma_w^2 + \sigma_b^2)[3 + 5(\sigma_w^2 + \sigma_b^2)]}{(1 + \sigma_w^2 + \sigma_b^2)[1 + 2(\sigma_w^2 + \sigma_b^2)]^{\frac{3}{2}}} \right] \,, \tag{165}$$

$$\mathbb{E}_\theta \left[\phi(h^l_i) \phi(h^l_i) \right] = \frac{\sigma_w^2 + \sigma_b^2}{4} + \frac{\sigma_w^2 + \sigma_b^2}{2\pi} \arcsin\left(\frac{\sigma_w^2 + \sigma_b^2}{1 + \sigma_w^2 + \sigma_b^2} \right) + \frac{(\sigma_w^2 + \sigma_b^2)^2}{\pi(1 + \sigma_w^2 + \sigma_b^2)\sqrt{1 + 2(\sigma_w^2 + \sigma_b^2)}} \,. \tag{166}$$

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

$$\chi^l_{\mathcal{J}} = \frac{\sigma_w^2 \left(1 + \sigma_w^2 + \sigma_b^2\right) \left[\pi + 2\arcsin\left(\frac{\sigma_w^2 + \sigma_b^2}{1 + \sigma_w^2 + \sigma_b^2} \right) + \frac{2(\sigma_w^2 + \sigma_b^2)(3 + 5(\sigma_w^2 + \sigma_b^2))}{(1 + \sigma_w^2 + \sigma_b^2)(1 + 2(\sigma_w^2 + \sigma_b^2))^{\frac{3}{2}}} \right]}{\pi(\sigma_w^2 + \sigma_b^2)(1 + \sigma_w^2 + \sigma_b^2) - 2(\sigma_w^2 + \sigma_b^2)^2 + \frac{4(\sigma_w^2 + \sigma_b^2)^2}{\sqrt{1 + 2(\sigma_w^2 + \sigma_b^2)}} + 2(\sigma_w^2 + \sigma_b^2)(1 + \sigma_w^2 + \sigma_b^2)\arcsin\left(\frac{\sigma_w^2 + \sigma_b^2}{1 + \sigma_w^2 + \sigma_b^2} \right)} \,. \tag{167}$$

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

I.6 Residual Connections

The recurrence relation for the NNGP kernel is

$$\mathcal{K}^{l+1} = \left[\frac{\mathcal{K}^l}{4} + \frac{\mathcal{K}^l}{2\pi} \arcsin\left(\frac{\mathcal{K}^l}{1 + \mathcal{K}^l} \right) + \frac{(\mathcal{K}^l)^2}{\pi(1 + \mathcal{K}^l)\sqrt{1 + 2\mathcal{K}^l}} \right] \sigma_w^2 + \sigma_b^2 + \mu^2 \mathcal{K}^l \,. \tag{168}$$

Fixed point exists if

$$\chi^\star_{\mathcal{K}} = \frac{\sigma_w^2}{4} + \frac{\sigma_w^2}{2\pi} \left[\arcsin\left(\frac{\mathcal{K}^\star}{1 + \mathcal{K}^\star} \right) + \frac{4(\mathcal{K}^\star)^3 + 11(\mathcal{K}^\star)^2 + 5\mathcal{K}^\star}{(1 + \mathcal{K}^\star)^2 (1 + 2\mathcal{K}^\star)^{\frac{3}{2}}} \right] + \mu^2 \leq 1 \,. \tag{169}$$

The recurrence coefficient for Jacobian is

$$\chi_{\mathcal{J}}^{\star} = \frac{\sigma_w^2}{4} + \frac{\sigma_w^2}{2\pi}\left[\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right) + \frac{\mathcal{K}^\star(3+5\mathcal{K}^\star)}{(1+\mathcal{K}^\star)(1+2\mathcal{K}^\star)^{\frac{3}{2}}}\right] + \mu^2 \,. \tag{170}$$

Phase boundary is shifted

$$\sigma_w = 2\sqrt{1-\mu^2}\left[1 + \frac{2\mathcal{K}^\star(3+5\mathcal{K}^\star)}{\pi(1+\mathcal{K}^\star)(1+2\mathcal{K}^\star)^{\frac{3}{2}}} + \frac{2}{\pi}\arcsin\left(\frac{\mathcal{K}^\star}{1+\mathcal{K}^\star}\right)\right]^{-\frac{1}{2}} \,, \tag{171}$$

$$\sigma_b = \frac{\mathcal{K}^\star}{\sqrt{2\pi}(1+2\mathcal{K}^\star)^{\frac{3}{4}}}\sigma_w \,. \tag{172}$$

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

$$\frac{\sigma_w^2[(\mathcal{K}^\star)^3 - 3(\mathcal{K}^\star)^2 - 2\mathcal{K}^\star]}{2\pi(1+\mathcal{K}^\star)^2(1+2\mathcal{K}^\star)^{\frac{3}{2}}} = 0 \,, \tag{173}$$

which has two non-negative solutions out of three

$$\mathcal{K}^\star = 0 \text{ and } \mathcal{K}^\star = \frac{3+\sqrt{17}}{2} \,. \tag{174}$$

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

$$(\sigma_w, \sigma_b) = (2\sqrt{1-\mu^2}, 0) \,, \text{ for } \mathcal{K}^\star = 0 \,, \tag{175}$$

$$(\sigma_w, \sigma_b) \approx (1.408\sqrt{1-\mu^2}, 0.416\sqrt{1-\mu^2}) \,, \text{ for } \mathcal{K}^\star = \frac{3+\sqrt{17}}{2} \,. \tag{176}$$

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

Use Lemma B.9 and results we had without residue connections for GELU.

$$\begin{aligned}\chi_{\mathcal{J}}^* =& \lim_{l\to\infty}\left(\frac{\sigma_w^2}{N_l\mathcal{K}^l}\sum_{k=1}^{N_l}\mathbb{E}_\theta\left[\phi'(\tilde{h}_k^l)\phi'(\tilde{h}_k^l)\right]\Bigg|_{\tilde{\mathcal{K}}^{l-1}=1} + \mu^2\right)\\ =& \frac{\sigma_w^2(6\pi+4\sqrt{3})(1-\mu^2)}{\sigma_w^2(6\pi+3\sqrt{3})+18\pi\sigma_b^2} + \mu^2\\ =& 1 - \frac{(\sqrt{3}\sigma_w^2 - 18\pi\sigma_b^2)(1-\mu^2)}{\sigma_w^2(6\pi+3\sqrt{3})+18\pi\sigma_b^2} \,.\end{aligned} \tag{177}$$

The critical line is then at

$$\begin{aligned}\sigma_b =& \left(6\sqrt{3}\pi\right)^{-\frac{1}{2}}\sigma_w\\ \approx& 0.175\sigma_w \,,\end{aligned} \tag{178}$$

just like without residue connections.

J 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 6, 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 datasetXiao et al. (2017). 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.

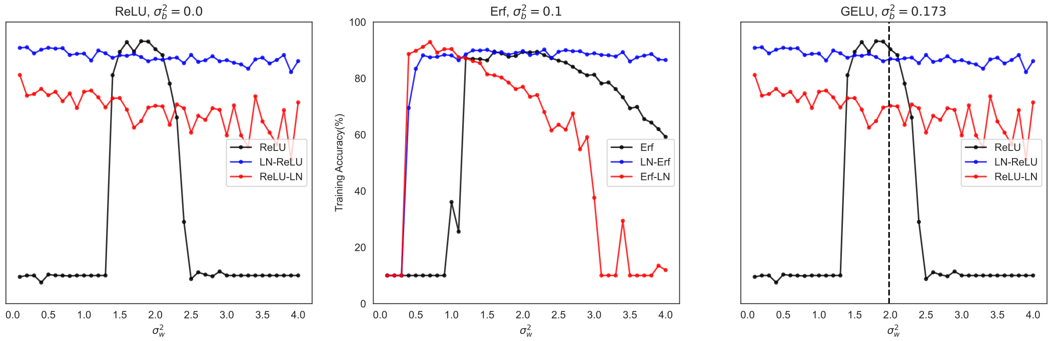

Figure 6: Performance of deep MLP networks at and away from criticality, with and without LayerNorm. The blue plateau, corresponding to LayerNorm applied to preactivations, continues to train at very large values of σ_w^2 without the need to tune the learning rate.

In figure 7, we showed empirically that the critical exponent of partial Jacobians are vanished for erf with LayerNorm.

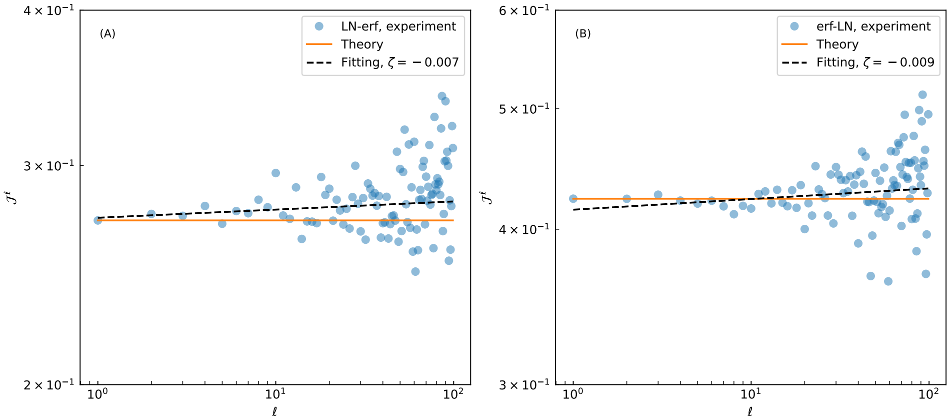

Figure 7: $\log - \log$ plot of partial Jacobian $\mathcal{J}^{0,l}$ vs. l for (A) LN-erf and (B) erf-LN.

In figure 8, we tested $6k$ samples from CIFAR-10 datasetKrizhevsky et al. (2009) with kernel regression based on neural tangents library Novak et al. (2019) Lee et al. (2019) Novak et al. (2020). 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 Xiao et al. (2020). The difference is from the fact that they used $\log_{10}$ and we used $\log_e$.

In figure 9, 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.

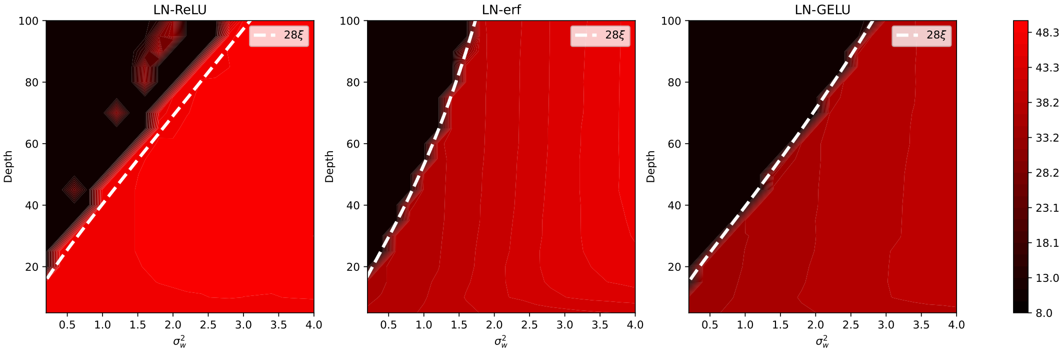

Figure 8: Test accuracy for LayerNorm applied to preactivations. $\sigma_b^2 = 0.5$ for all cases. Correlation lengths calculated using analytical results of $\chi_{\mathcal{J}}^l$.

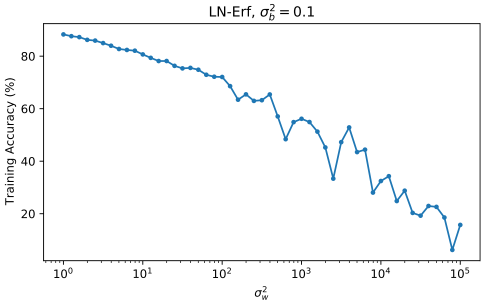

Figure 9: Training performance of MLP networks with erf activation function; and LayerNorm applied to preactivations. It continues to train for several orders of magnitude of σ_w^2 (with learning-rate tuning).