Appendices

A Fluctuations through layers

We are interested in analyzing the fixed point solutions in Eq. 3, which may be stable (attractors), unstable (repulsors), or meta-stable (attractive for certain values of K and repulsive for others K). Let's define the Gaussian average as

$$\langle f(z)\rangle_K = \frac{1}{\sqrt{2\pi K}}\int_{-\infty}^{\infty} dz e^{-\frac{z^2}{2K}} f(z). \tag{7}$$

For a given input ($\alpha_1 = \alpha_2$), the infinitesimal magnitude variation is characterized by the parellel susceptibility (Roberts et al., 2022)

$$\chi_{\|}(K) = C_W \frac{dg}{dK} = \frac{C_W}{2K^2}\left\langle \sigma(z)\sigma(z)\left(z^2 - K\right)\right\rangle_K. \tag{8}$$

The susceptibility $\chi_{||}(K)$ characterizes how 'susceptible' the kernel is to perturbations around the fixed point. The kernel value exponentially expands away from or contracts towards the fixed-point value, according to whether $\chi_{||}(K^*) > 1$ or $\chi_{||}(K^*) < 1$. For two different inputs ($\alpha_1 \neq \alpha_2$), the infinitesimal variation of magnitude differences is characterized by the perpendicular susceptibility,

$$\chi_{\perp}(K) = C_W \left\langle \frac{d\sigma(z)}{dz}\frac{d\sigma(z)}{dz}\right\rangle_K. \tag{9}$$

The input difference either explodes ($\chi_{\perp}(K^*) > 1$) or shrinks ($\chi_{\perp}(K^*) < 1$) (Roberts et al., 2022). From this discussion it becomes clear that, in an ideal scenario,

$$\chi_{||}(K^*) = \chi_{\perp}(K^*) = 1. \tag{10}$$

We will call such a case 'critical'.

As an illustration, let's analyze the ReLU activation

$$\sigma(z) = \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases} \tag{11}$$

The susceptibility integrals can be evaluated analytically, $\chi_{\|}(K) = \chi_{\perp}(K) = \frac{C_W}{2}$, independent of K. This suggests a parametric family of fixed points, and ($C_W = 2$, $C_b = 0$) leads to the critical initialization for any value of K.

A.1 Susceptibilities for ReLu function

ReLU is defined as $\sigma(z) = \max(z, 0)$.

$$\begin{aligned}
\chi_{\|}(K) =& C_W \frac{d}{dK}\langle\sigma(z)\sigma(z)\rangle_K \\
=& C_W \frac{d}{dK}\Big[\frac{1}{\sqrt{2\pi K}}\int_{-\infty}^{\infty} dz e^{-\frac{z^2}{2K}}\sigma(z)\sigma(z)\Big] \\
=& C_W \frac{d}{dK}\Big[\frac{1}{\sqrt{2\pi K}}\int_{0}^{\infty} dz e^{-\frac{z^2}{2K}} z^2\Big] \\
=& C_W \frac{d}{dK}\left(\frac{K}{2}\right) \\
=& \frac{C_W}{2}
\end{aligned}$$

$$
\begin{aligned}
\chi_{\perp}(K) =& C_W \langle \sigma'(z)\sigma'(z) \rangle_K \\
=& C_W \Big[\frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz e^{-\frac{z^2}{2K}} \sigma'(z)\sigma'(z) \Big] \\
=& C_W \Big[\frac{1}{\sqrt{2\pi K}} \int_{0}^{\infty} dz e^{-\frac{z^2}{2K}} \Big] \\
=& \frac{C_W}{2}
\end{aligned}
$$

Criticality implies $C_W = 2$.

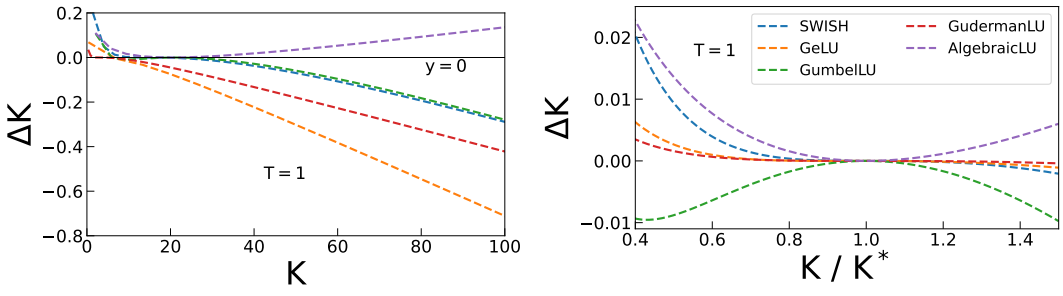
These results do not depend on K^* and the recursion equation is

$$
\begin{aligned}
K^* =& C_b + C_W \frac{K^* 2}{2} \\
=& C_b + K^*
\end{aligned}
\tag{12}
$$

Thus, for $(C_b, C_W) = (0, 2)$ there is family of K^* that serve as critical kernels.

B Critical point

In Fig. 6 we plot the dependence of the layer update on the metric, the expectation value of the two point correlator, for all smooth-ReLU activations from Table 1 for a wide range in K in subplot (a), and closer to the fixed point in subplot (b). In Fig. 7 we rescale the axes and we zoom in to reveal the fixed point obtained by solving Eq. 3, and the adjacent fixed point. AlgebraicLU is the only activation without a secondary fixed point. While in the neighborhood of the critical value there is a power law layer update, away from the point it is exponential. For low temperatures, the region between the fixed point from Eq. 3 and the second fixed point is very small, and most typically, one would observe the exponential update from either side towards the fixed point region, apart from AlgebaricLU which has only positive updates. This is why in Table 1, we refer to this point as unstable, and the others as semi-stable. From these figures one may be tempted to conclude that as $T \to 0$ the update becomes almost 0 and hence, the ReLU-like behavior is obtained. As this transition seems rather smooth, nothing analogous to a phase transition should happen as the temperature is lowered.



(a) Large span in K

(b) Span in the neighborhood of K^*

Figure 6: Theoretical layer update for all activation functions.

C Condition for criticality in smooth-ReLU activations

Lemma 1 (Criticality Condition)**.** *A FFN network with activation $\sigma(z) = za(z)$ admits a critical initialization scheme if and only if the following conditions are satisfied*

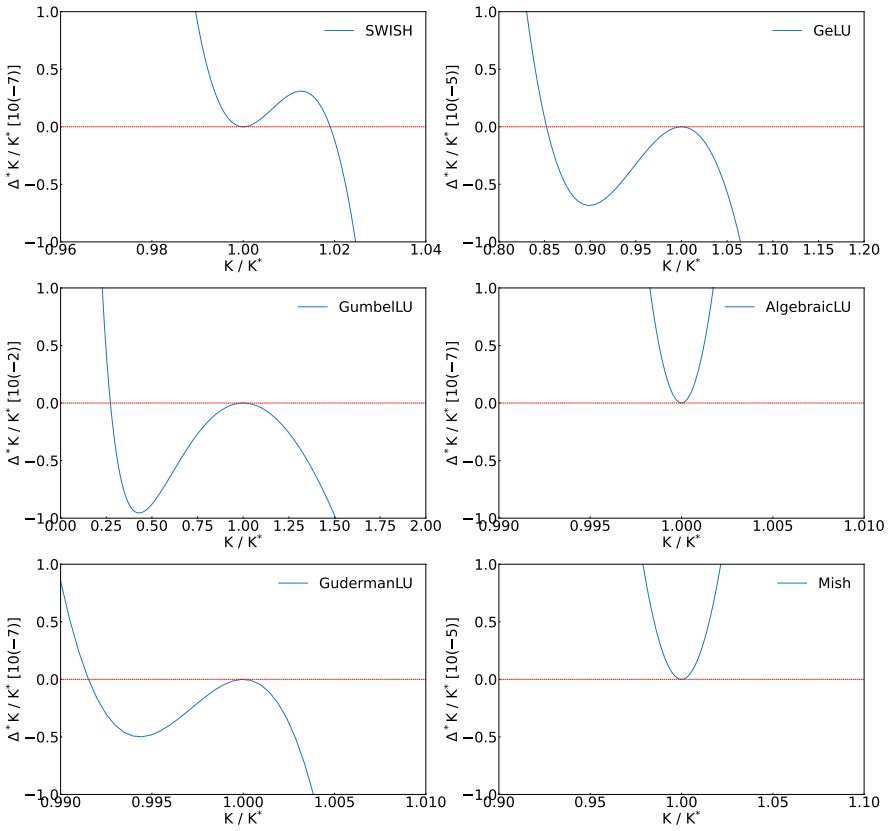


Figure 7: Zoomed in view of the theoretical layer update for all activation functions. Apart from AlgebraicLU and Mish, they all display 2 fixed points.

- $2\langle za(z)a'(z)\rangle_{K^*} = -\langle z^2 a(z)a''(z)\rangle_{K^*}$
- $z^3a(z)^2$ *and* $a(z)(a(z)+za'(z))$ *are sub-Gaussian* ($\lim_{z\to\pm\infty} f(z)e^{-z^2/2K} = 0$)

Proof. Substituting the activation definition

$$\sigma(z) = za(z) \tag{13}$$
$$\sigma'(z) = a(z) + za'(z) \tag{14}$$
$$\sigma''(z) = 2a'(z) + za''(z) \tag{15}$$
$$\tag{16}$$

Using Lemma 4, if $x\sigma^2$ and $\sigma\sigma'$ are sub-Gaussian, the susceptibilities satisfy

$$\chi_\perp = \chi_\parallel + C_W\langle\sigma\sigma''\rangle_K \tag{17}$$

Then, the following equation has to be true at criticality

$$\langle\sigma\sigma''\rangle_K = 0 \tag{18}$$

Therefore, substituting the activation definition we have that

$$\langle za(z)\Big(2a'(z) + za''(z)\Big)\rangle_K = 0 \tag{19}$$

$$2\langle za(z)a'(z)\rangle = -\langle z^2 a(z)a''(z)\rangle_K \tag{20}$$

$$\tag{21}$$

as we wanted to prove.

□

Corollary 2 (T Criticality)**.** *Given the activation $\sigma_T(z) = za(z/T)$, if K^* is critical for $T = 1$, then $K_T^* = K^*T^2$ is critical for $T \neq 1$.*

Proof. Noticing that

$$\frac{da_T(z)}{dz} = \frac{d}{dz}a(z/T) \tag{22}$$

$$= \frac{1}{T}\frac{d}{d(z/T)}a(z/T) \tag{23}$$

$$= \frac{1}{T}\frac{da(u)}{du}, \tag{24}$$

$$\tag{25}$$

and that

$$\frac{d^2a_T(z)}{dz^2} = \frac{1}{T^2}\frac{d^2a(u)}{du^2}, \tag{26}$$

$$\tag{27}$$

Then we substitute in the criticality condition found in Lemma 1

$$2\langle za(z)a'(z)\rangle = -\langle z^2 a(z)a''(z)\rangle_K \tag{28}$$

$$\tag{29}$$

$$\langle za_T(z)a_T'(z)\rangle_{K_T} = \frac{1}{\sqrt{2\pi K}}\int dz\; e^{-\frac{z^2}{2K}}\, za_T(z)a_T'(z) \tag{30}$$

$$= \frac{1}{\sqrt{2\pi K_T}}\int Tdu\; e^{-\frac{(uT)^2}{2K_T}}(uT)a(u)\frac{1}{T}\frac{da(u)}{du} \qquad u = z/T \tag{31}$$

$$= \frac{1}{T}\frac{1}{\sqrt{2\pi(K_T/T^2)}}\int du\; e^{-\frac{u^2}{2(K_T/T^2)}}ua(u)\frac{da(u)}{du} \tag{32}$$

$$= \frac{1}{T}\langle za(z)a'(z)\rangle_{K_T/T^2} \tag{33}$$

and similarly for $\langle z^2a(z)a''(z)\rangle_K$:

$$\left\langle z^2a_T(z)a_T''(z)\right\rangle_{K_T} = \frac{1}{\sqrt{2\pi K}}\int dz\; e^{-\frac{z^2}{2K}}\, z^2a_T(z)a_T''(z) \tag{34}$$

$$= \frac{1}{\sqrt{2\pi K_T}}\int Tdu\; e^{-\frac{(uT)^2}{2K_T}}(uT)^2a(u)\frac{1}{T^2}\frac{d^2a(u)}{du^2} \qquad u = z/T \tag{35}$$

$$= \frac{1}{T}\frac{1}{\sqrt{2\pi(K_T/T^2)}}\int du\; e^{-\frac{u^2}{2(K_T/T^2)}}u^2a(u)\frac{d^2a(u)}{du^2} \tag{36}$$

$$= \frac{1}{T}\left\langle z^2a(z)a''(z)\right\rangle_{K_T/T^2} \tag{37}$$

Therefore if K^* is critical for $T = 1$, then $K_T^* = K^*T^2$ is critical for $T \neq 1$, as we wanted to prove. □

D Susceptibility scaling with T

Lemma 3 (Temperature Criticality Smooth ReLU)**.** *If the activation $\sigma(z) = za(z)$ is critical at K^* for (C_W^*, C_B^*), then the activation $\sigma_T(z) = za(z/T)$ is critical at $K_T^* = K^* T^2$ for $(C_{W,T}^*, C_{B,T}^*) = (C_W^*, T^2 C_B^*)$.*

Proof. The activation considered in this work is of the form $\sigma_T(z) = za(z/T)$, and we want to know how the statistics relates to $\sigma(z) = za(z)$ There are many ways to search for the critical point for a given activation function. For cases that can not be solved analytically, the following steps can be implemented,

To study the T dependence, at first, we compute the T dependence of the auxiliary function,

$$
\begin{aligned}
g_T(K) = \langle \sigma_T(z)\sigma_T(z) \rangle_K &= \frac{1}{\sqrt{2\pi K}} \int dz\ e^{-\frac{z^2}{2K}}\, z^2 a(z/T)^2 && (38)\\
&= \frac{1}{\sqrt{2\pi K}} \int T du\ e^{-\frac{(uT)^2}{2K}} (uT)^2 a(u)^2 && u = z/T \quad (39)\\
&= T^2 \frac{1}{\sqrt{2\pi (K/T^2)}} \int du\ e^{-\frac{u^2}{2(K/T^2)}} u^2 a(u)^2 && (40)\\
&= T^2 g(K/T^2) && (41)
\end{aligned}
$$

Then, the parallel susceptibilities, given that according to Corollary 2 at criticality $K_T^* = T^2 K^*$, have to satisfy

$$
\begin{aligned}
\chi_{\parallel,T}(K_T^*) &= C_{W,T} \frac{dg_T(K_T)}{dK_T}\Big|_{K_T^*} && (42)\\
&= \frac{1}{T^2} C_{W,T} \frac{dg_T(K_T)}{d(K_T/T^2)}\Big|_{K_T^*} && (43)\\
&= \frac{T^2}{T^2} C_{W,T} \frac{dg(K_T/T^2)}{d(K_T/T^2)}\Big|_{K_T^*} && (44)\\
&= C_{W,T} \frac{dg(K_T/T^2)}{d(K_T/T^2)}\Big|_{K_T^*} && (45)\\
&= C_{W,T} \frac{dg(K)}{d(K)}\Big|_{K^*} && (46)\\
\chi_{\parallel}(K) &= C_W \frac{dg(K)}{dK}\Big|_{K^*} && (47)
\end{aligned}
$$

Since at criticality both susceptibilities have to be equal to one, $\chi_{\parallel,T}(K_T^*) = \chi_{\parallel}(K^*)$ gives that $C_{W,T} = C_W$.

To conclude the proof, we need to retrieve the T dependence of C_b. A necessary condition to be at criticality is that there is a critical variance K^*, that can satisfy a layer-wise fixed point equation:

$$
\begin{aligned}
K_T &= C_{b,T} + C_{W,T} g_T(K_T) && (48)\\
& && (49)
\end{aligned}
$$

Using the results above $g_T(K_T) = T^2 g(K_T/T^2)$, and $C_{W,T} = C_W$, and Corollary 2, dividing both sides by T^2 we have that at criticality

$$
\begin{aligned}
K_T^* &= C_{b,T} + C_{W,T} g_T(K_T^*) && (50)\\
&= C_{b,T} + C_{W,T} T^2 g(K_T^*/T^2) && (51)\\
K_T^*/T^2 &= C_{b,T}/T^2 + C_{W,T} g(K_T^*/T^2) && (52)\\
K^* &= C_{b,T}/T^2 + C_W g(K^*) && (53)\\
& && (54)
\end{aligned}
$$

Comparing with the layer-wise fixed point equation for $T = 1$

$$K^* = C_b + C_W g(K^*) \tag{55}$$

We conclude that $C_{b,T} = T^2 C_b$, concluding the proof.

□

E arctanLU only has $K^* = 0$

Lemma 2 (Zero Criticality)**.** *For neural networks with activations of the form $za(z/T)$, the condition that $a(z/T) \to H(z)$ as $T \to 0$ is not sufficient to ensure the existence of a critical initialization scheme with a non-zero fixed point $K^* \neq 0$.*

Proof. The activation $a(z/T) = \left(\frac{\tan^{-1}\left(\frac{z}{T}\right)}{\pi} + \frac{1}{2}\right)$ tends to a Heaviside as $T \to 0$, however using Lemma 4

$$\langle \sigma\sigma'' \rangle = \left\langle \frac{T^3 z \left(2\tan^{-1}\left(\frac{z}{T}\right) + \pi\right)}{\pi^2 \left(T^2 + z^2\right)^2} \right\rangle \tag{56}$$

$$= \frac{2T^3}{\pi^2} \left\langle \frac{z \tan^{-1}\left(\frac{z}{T}\right)}{\left(T^2 + z^2\right)^2} \right\rangle \geq 0 \tag{57}$$

$$\tag{58}$$

The expression above is 0 only for $K = 0$ since the integrand is even.

□

F Analytical K^* for GeLU

For the activation $\sigma(z, T) = \frac{z}{2}(1 + \text{erf}(\frac{z}{T\sqrt{2}})$,

$$\langle \sigma\sigma'' \rangle = -\frac{K_T T \left(K_T^2 - 3K_T T^2 - 2T^4\right)}{2\pi \left(K_T + T^2\right)^2 \left(2K_T + T^2\right)^{3/2}} \tag{59}$$

The activation function is critical when $\langle \sigma\sigma'' \rangle = 0$:

$$K_T^* T \left(K_T^{*\,2} - 3K_T^* T^2 - 2T^4\right) = 0 \tag{60}$$

$$K_T^* = 0, \; K_T^* = \frac{T^2}{2}\left(3 \pm \sqrt{17}\right) \tag{61}$$

We can recover K^* by setting $T = 1$, $K^* = \frac{1}{2}\left(3 + \sqrt{17}\right) \approx 3.56$.

G Equivalent statements for criticality

The following result appeared already in Hanin (2022) without proof, we prove it here for completeness, even if the proof is simple. We say that $f(x)$ is sub-Gaussian if $f(x)e^{-x^2/2K} \xrightarrow{x \to \pm\infty} 0$ for the K of interest.

Lemma 4. *If $x\sigma^2$ and $\sigma\sigma'$ are sub-gaussian, then $\frac{d}{dK}\langle \sigma\sigma \rangle_K = \langle \sigma\sigma'' \rangle + \langle \sigma'\sigma' \rangle$*

Proof. Let's start by noticing that

$$\frac{d}{dK}\langle\sigma\sigma\rangle_K = \frac{d}{dK}\frac{1}{\sqrt{2\pi K}}\int e^{\frac{-x^2}{2K}}\sigma\sigma \tag{62}$$

$$= -\frac{1}{2K\sqrt{2\pi K}}\int e^{\frac{-x^2}{2K}}\sigma\sigma + x^2\frac{1}{2K^2}\frac{1}{\sqrt{2\pi K}}\int e^{\frac{-x^2}{2K}}\sigma\sigma \tag{63}$$

$$= \frac{1}{2K^2}\langle(x^2-K)\sigma\sigma\rangle_K \tag{64}$$

Then, integrating by parts

$$u = x\sigma^2 \qquad du = \sigma^2 dx + 2x\sigma\sigma' dx \tag{65}$$

$$dv = xe^{\frac{-x^2}{2K}}dx \qquad v = -Ke^{\frac{-x^2}{2K}} \tag{66}$$

and given that, if $x\sigma^2$ is sub-gaussian, $uv = 0$ at infinity then

$$\int udv = uv - \int vdu \tag{67}$$

$$= \int Ke^{\frac{-x^2}{2K}}\sigma^2 dx + \int Ke^{\frac{-x^2}{2K}}2x\sigma\sigma' dx \tag{68}$$

$$= Kg(K) + 2K\langle x\sigma\sigma'\rangle_K \tag{69}$$

Integrating again by parts the second integral we get

$$u = \sigma\sigma' \qquad du = \sigma\sigma'' dx + \sigma'\sigma' dx \tag{70}$$

$$dv = xe^{\frac{-x^2}{2K}}dx \qquad v = -Ke^{\frac{-x^2}{2K}} \tag{71}$$

Performing the integration by parts and noticing that if $\sigma\sigma'$ is sub-gaussian, $uv = 0$ at infinity,

$$\int udv = uv - \int vdu \tag{72}$$

$$= \int Ke^{\frac{-x^2}{2K}}\sigma\sigma'' dx + \int Ke^{\frac{-x^2}{2K}}\sigma'\sigma' dx \tag{73}$$

$$= K\langle\sigma\sigma''\rangle + K\langle\sigma'\sigma'\rangle \tag{74}$$

Substituting in the initial integral we get

$$\frac{d}{dK}\langle\sigma\sigma\rangle_K = \frac{1}{2K^2}\langle(x^2-K)\sigma\sigma\rangle_K \tag{75}$$

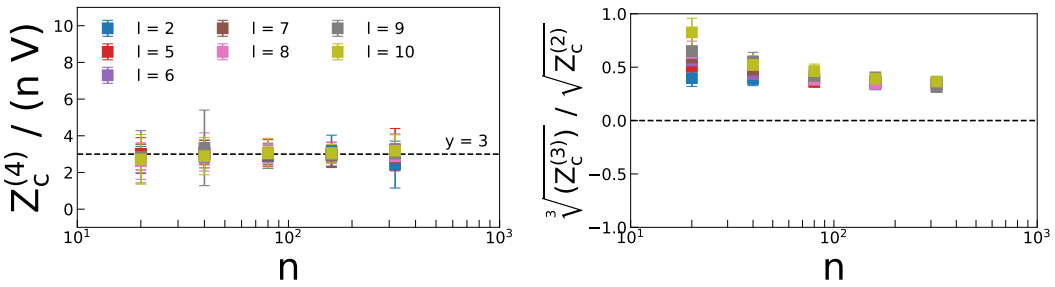
$$= \frac{1}{2K^2}\langle x^2\sigma\sigma\rangle_K - \frac{1}{2K}\langle\sigma\sigma\rangle_K \tag{76}$$

$$= \frac{1}{2K^2}(K\langle\sigma\sigma\rangle_K + 2K\langle x\sigma\sigma'\rangle_K) - \frac{1}{2K}\langle\sigma\sigma\rangle_K \tag{77}$$

$$= \frac{1}{K}\langle x\sigma\sigma'\rangle_K \tag{78}$$

$$= \langle\sigma\sigma''\rangle + \langle\sigma'\sigma'\rangle \tag{79}$$

□



(a) 4-point correlator vs 2-point correlator fluctuations

(b) 3-point correlator vs 2-point correlator

Figure 8: The experimental results the correlations for the SWISH activation with critial initialization at $T = 1$.

H Verification of near Gaussian behaviour

Due to the Gaussian initialization of the weights of the networks, the first cumulant of the final layer pre-activation is $\langle z_{i,\alpha_1} \rangle = 0$. The next order, the 2-point correlator is

$$Z^{(c)}_{i,\alpha_1,j,\alpha_2} = \langle z_{i,\alpha_1} z_{j,\alpha_2} \rangle - \langle z_{i,\alpha_1} \rangle \langle z_{j,\alpha_2} \rangle = \langle z_{i,\alpha_1} z_{j,\alpha_2} \rangle. \tag{80}$$

Similarly, one can compute higher order cumulants (Shlosman, 1986), for instance,

$$\begin{aligned} Z^{(c)}_{i,\alpha_1,j,\alpha_2,k,\alpha_3} = &\langle z_{i,\alpha_1} z_{j,\alpha_2} z_{k,\alpha_3} \rangle - \langle z_{i,\alpha_1} \rangle \langle z_{j,\alpha_2} z_{k,\alpha_3} \rangle \\ &- \langle z_{j,\alpha_2} \rangle \langle z_{i,\alpha_1} z_{k,\alpha_3} \rangle - \langle z_{k,\alpha_3} \rangle \langle z_{i,\alpha_1} z_{j,\alpha_2} \rangle + 2 \langle z_{i,\alpha_1} \rangle \langle z_{j,\alpha_2} \rangle \langle z_{k,\alpha_3} \rangle. \end{aligned} \tag{81}$$

If the pre-activation through the layers remained exactly Gaussian, the only cumulant present would be the 2-nd order. However, higher order correlations do develop and their strength depends on both width and depth. For very wide networks ($l/n \ll 1$) the higher moments are sub-leading with the most important being the 4-point correlator. We can define the fluctuation on the 2-point correlator based on the metric previously introduced,

$$\begin{aligned} \hat{K}^{(\ell)}_{\alpha_1\alpha_2} =& \frac{1}{n} \sum_{j=1}^{n} z^{(\ell)}_{j;\alpha_1} z^{(\ell)}_{j;\alpha_2}, \\ \Delta \hat{K}^{(\ell)}_{\alpha_1\alpha_2} =& \hat{K}^{(\ell)}_{\alpha_1\alpha_2} - K^{(\ell)}_{\alpha_1\alpha_2} \\ V^{(\ell)}_{\alpha_1\alpha_2;\alpha_3\alpha_4} =& \langle \Delta \hat{K}^{(\ell)}_{\alpha_1\alpha_2} \Delta \hat{K}^{(\ell)}_{\alpha_3\alpha_4} \rangle \end{aligned} \tag{82}$$

If the first and third cumulants can be ignored,

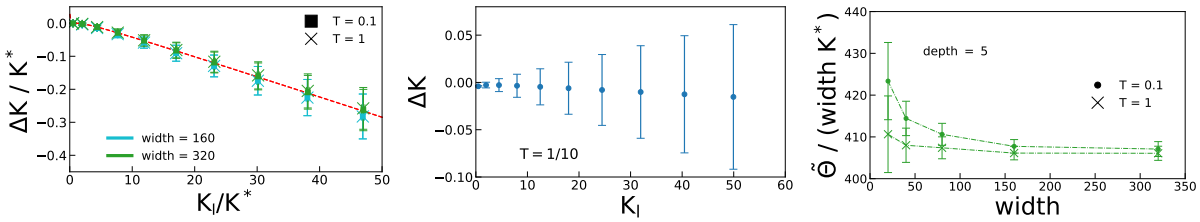
$$Z^{(c)\ell}_{i,\alpha_1,j,\alpha_2,k,\alpha_3,r,\alpha_4} = \frac{1}{n} \left(\delta_{i,j}\delta_{k,r} V^{\ell}_{\alpha_1\alpha_2;\alpha_3\alpha_4} + \delta_{i,k}\delta_{j,r} V^{\ell}_{\alpha_1\alpha_3;\alpha_2\alpha_4} + \delta_{i,r}\delta_{j,k} V^{\ell}_{\alpha_1\alpha_4;\alpha_2\alpha_3} \right) \tag{83}$$

For more details, we refer the reader to (Roberts et al., 2022). By marginalizing over all the indices and taking the absolute value we obtain the following relationship,

$$Z^{(4)}_c = 3nV, \tag{84}$$

where we have suppressed the the depth superscript for simplicity of notation. In fig. 8 (a) we verify this relation and also compute the ratio of the third over the second cumulant in fig. 8 (b). We have taken respective roots to account for the powers of the output in each. The plots were generated for the SWISH activation with the samples collected for section 4.1, and similar results were obtained for other activations. As predicted theoretically, we confirm empirically in the figure that the third moment goes to zero with increasing width for a fixed depth and there is a factor of 3 between $Z^{(4)}_c$ and nV.

I Initialization statistics for GumbelLU

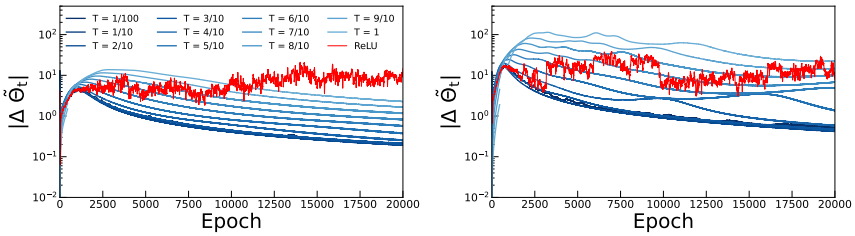


(a) Critical, depth $n = 10$ (b) He et al, depth $n = 10$ (c) NTK for critical initialization

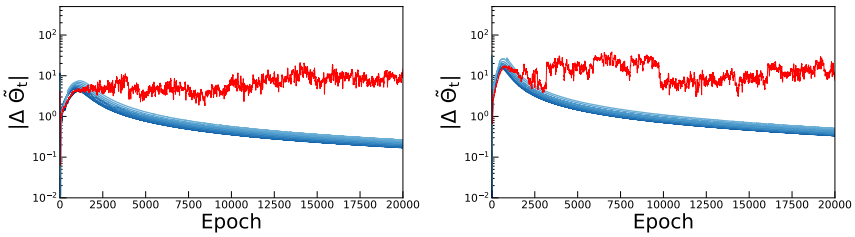
Figure 9: Initialization plots for GumbelLu; (a) shows the layer update for critical initialization at various temperatures and widths and in dashed red the respective theoretical prediction, (b) depicts the layer update for He et al initialization and $T = 0.1$, and (c) portrays the NTK for critical initialization at the same temperatures as (a). These results are in complete agreement with the results in the main text and our theoretical expectations.

J Additional loss and NTK results through training

As the ReLU curve shows no clear trend, Fig. 10, we apply a Savitzky-Golay filter (Savitzky and Golay, 1964; Virtanen et al., 2020) with a window of 80 epochs and a polynomial order of 3, to improve the clarity of the plot. We tested several epochs and polynomial orders for the window filtering with similar results. This allows us to discern some trends in the ReLU curves, particularly in the early stages of training where the window filtered ReLU seems to agree with smooth-ReLU. We plot additional loss and NTK training curves, for critical and He initializations, for two additional smooth-ReLUs: GudermanLU, Fig. 11 and Fig. 12, with the CIFAR10 dataset, and GeLU, Fig. 13 and Fig. 14, with the MNIST dataset. As the plots show these two activations are less sensitive to temperature, even for He initialization. From Table 1 we can see that their hyper-parameters are closer to ReLU than SWISH or GumbelLU. It is interesting to notice that, after using the window filter, in the initial faster training phase, the ReLU $\Delta\tilde{\Theta}_t$ follows closely the low temperature SWISH $\Delta\tilde{\Theta}_t$, while on the later stages of training, the ReLU NTK update seems to resemble stochastic noise without a specific pattern.

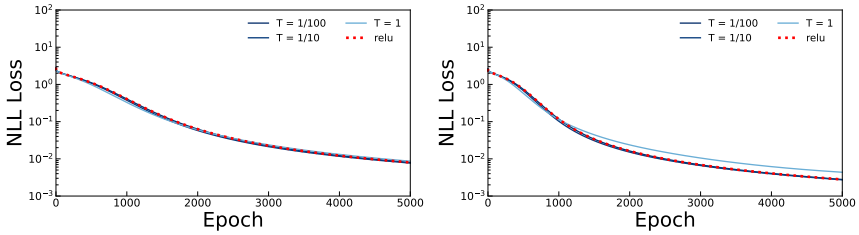


(a) He , depth = 8 (b) He, depth =12

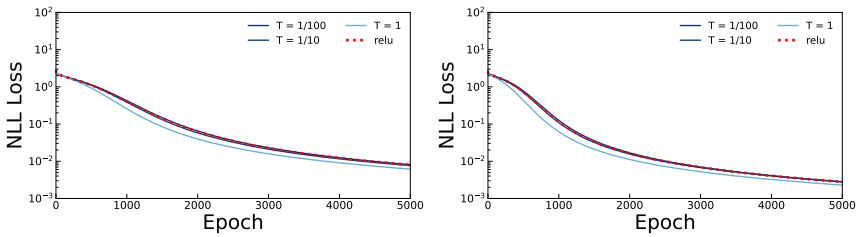


(c) Critical, depth = 8 (d) Critical, depth =12

Figure 10: NTK update for Cifar10 training with SWISH activation, He and critical initialization.

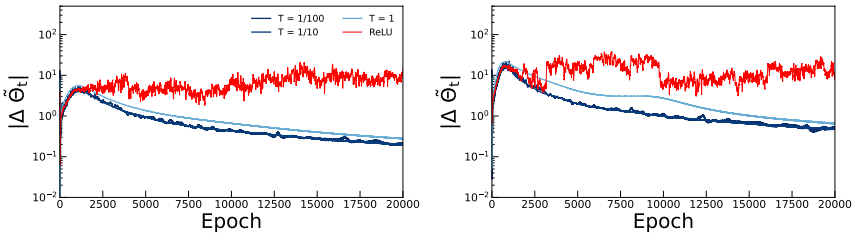


(a) He , depth = 8 (b) He, depth =12

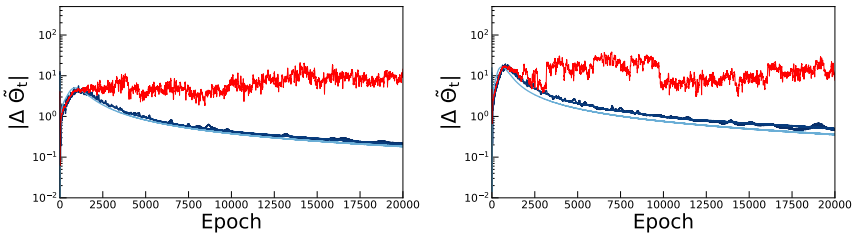


(c) Critical, depth = 8 (d) Critical, depth =12

Figure 11: Loss function for Cifar10 training with GudermanLU activation, He and critical initialization.

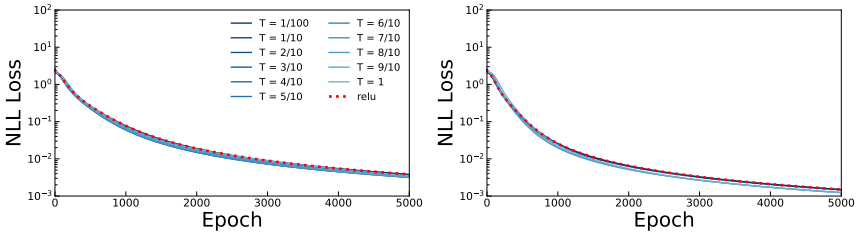


(a) He , depth = 8 (b) He, depth =12

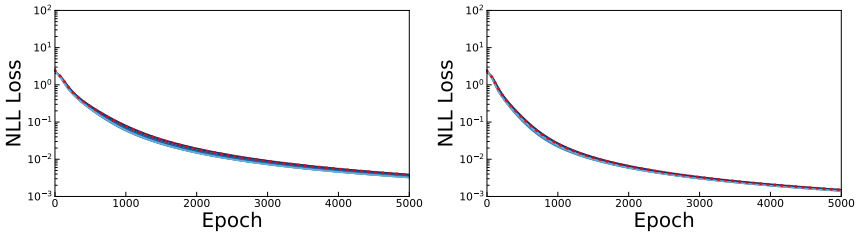


(c) Critical, depth = 8 (d) Critical, depth =12

Figure 12: NTK update for Cifar10 training with GudermanLU activation, He and critical initialization.

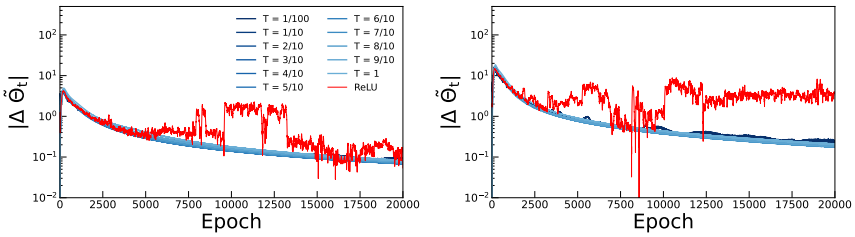


(a) He , depth = 8 (b) He, depth =12



(c) Critical, depth = 8 (d) Critical, depth =12

Figure 13: Loss function for MNIST training with GeLU activation, He and critical initialization.



(a) He , depth = 8

(b) He, depth =12

(c) Critical, depth = 8

(d) Critical, depth =12

Figure 14: NTK update for MNIST training with GeLU activation, He and critical initialization.

K EfficientNet and Transformer

We check whether the closer C_W^* and C_b^* are to ReLU critical values, the more the activation behaves like ReLU, for more complex tasks and architectures, see Fig. 15. We see that B0 EfficientNet trained on CIFAR100 shows GudermanLU being closer to ReLU than SWISH, and both get closer to ReLU when the temperature is lowered, confirming our assumption. Same holds true when we train the small Transformer on the English-German translation task, especially at the beginning of training and less so when the small differences accumulate towards the end. We report only one random seed for clarity, but the same behavior was observed for several seeds.

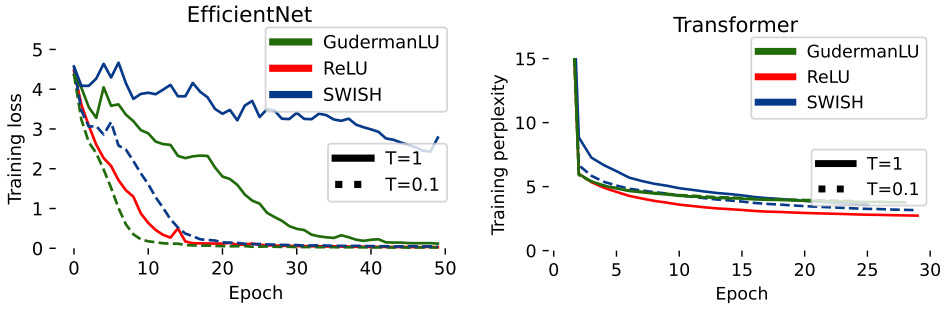


Figure 15: GudermanLU is more ReLU-like than Swish in more complex tasks. Left panel, we train the EfficientNet on CIFAR100, and right panel, we train the Transformer on the WMT'14 English-German Translation task.