When are smooth-ReLUs ReLU-like?

Anonymous authors
Paper under double-blind review

Abstract

ReLU is one of the most popular activations in deep learning, especially thanks to its stabilizing effect on training. However, because it is non-differentiable at the origin, it complicates the use of analysis methods that examine derivatives, such as the Neural Tangent Kernel (NTK). Many smooth relaxations try to retain the practical benefits of ReLU while increasing network regularity. Although their success has ranged widely, some notable architectures (e.g., the BERT family) do utilize them. We present a theoretical characterization of smooth-ReLUs within fully-connected feed-forward neural networks. In addition to the well-known SWISH and GeLU, we introduce GumbelLU, AlgebraicLU, and GudermanLU, as new relaxations. All these activations can be characterized by a positive temperature parameter which we can lower to continuously improve the approximation. By studying the interplay of initialization schemes with temperature, we confirm that when these relaxations converge uniformly to ReLU, the statistical properties of the corresponding neural networks at initialization also converge to those of ReLU networks. Moreover, we derive temperature-dependent critical initialization schemes with which networks based on these activations exhibit stable ReLU-like behavior at any temperature. Finally, we empirically study both classes of networks on MNIST and CIFAR-10 in the full-batch training regime. We show that, while all networks exhibit very similar train loss trajectories at criticality, smooth-ReLU networks feature differentiable NTKs throughout training, whereas ReLU networks exhibit stochastic NTK fluctuations. Our results clarify how smooth-ReLU relaxations reproduce the practical benefits of ReLU in everywhere-smooth neural networks.

1 Introduction

In recent decades, deep learning has shown tremendous success in e.g. computer vision, natural language processing, and drug discovery (LeCun et al., 2015). For instance, EfficientNet (Tan and Le, 2019) achieved state-of-the-art performance in image classification on CIFAR100 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009). In natural language processing, GPT (Radford et al., 2018) and its newer versions (Radford et al., 2019; Brown et al., 2020) were capable of producing human-like responses to various reading comprehension tasks. As the name suggests, deep learning neural architectures are composed of many layers sequentially applied. The most basic architecture, the fully connected feed-forward network (FFN), consists of an alternating sequence of linear layers and non-linear layers called activations. One of the most popular activations is ReLU (Jarrett et al., 2009; Nair and Hinton, 2010), given its computational simplicity, gradient stability, and expressive power (Nair and Hinton, 2010; Raghu et al., 2017). Beyond ReLU, EfficientNet and others use SWISH (Ramachandran et al., 2017; Elfwing et al., 2018; Alcaide, 2018; Chieng et al., 2018; Howard et al., 2019). GPT-2 and ALBERT (Lan et al., 2020) employ the GeLU activation (Hendrycks and Gimpel, 2016). Both are smooth approximations to ReLU, with the advantage of the existence of higher-order derivatives to study their properties (Hanin and Nica, 2020b; Li et al., 2021). Activation smoothness is required to ensure smoothness of the network's input-output mapping; this is in turn necessary for certain applications (e.g., physics-informed neural networks and neural network methods to solving differential equations require the network output to be a smooth function of its input, (Raissi et al.,

2019)), certain architectures (e.g., the existence of neural ODEs is typically proven assuming smooth activations (Chen et al., 2018)), and for certain theoretical analysis techniques (e.g., differential Neural Tangent Kernel (NTK) approaches Roberts et al. (2022)).

Our focus is to understand the properties of smooth ReLU relaxations, that we call smooth-ReLUs, of the type $\sigma_T(z) = za(z/T)$, with T a positive *temperature* parameter and $a(\cdot)$ any sigmoid function ranging from 0 to 1. As the temperature is lowered, the smooth function $a(z/T)$ converges to the Heaviside function $H(z)$ pointwise. Different choices of $a(\cdot)$ correspond to different smooth approximations to ReLU. The well known SWISH and GeLU can be represented this way, and we introduce new activations with this property. Inspired by recent works based on renormalization group and field theory (Roberts et al., 2022), we study the stability of information propagation through the layers of FFNs with smooth-ReLU variants. In particular, we study statistical properties over random network initializations, compare initialization and training dynamics of various smooth-ReLUs, and analyze to what extent they echo ReLU.

Initializing weights and biases by sampling a standard normal distribution results in the exponential explosion of the variance of the representations. Weight initializations whose variance is proportional to the inverse of the width can train deeper linear networks, more reliably and rapidly, by keeping the variance of representations and gradients constant with depth. However, for non-linear activations, an additional proportionality factor, that is activation dependent, is necessary to obtain depth stability (He et al., 2015; Roberts et al., 2022). Based on this observation, we quantify how similar different smooth-ReLU functions are to ReLU by comparing the variance of network representations at initialization.

We also compare the behaviour of training loss during full-batch training. To gain further insight into training dynamics, we also use the NTK theory (Jacot et al., 2018). In recent years, kernel methods have found wide use to theoretically understand neural networks' performance (Allen-Zhu et al., 2019; Zou et al., 2020; Liu et al., 2022; 2020; Yang, 2020). A flurry of works have extended the NTK analysis to all kinds of neural networks, with Arora et al. (2019) extending NTK to convolutional networks, and more recently Feng and Kolter (2020) investigating the NTK analysis for large-depth limits in Deep Equilibrium Models Bai et al. (2019).

Our work aims to improve our understanding of the expressivity and training of fully-connected deep smooth-ReLU networks and devise simple prescriptions to improve them. Specifically, our contributions are as follows:

- Derive conditions of existence of stable initialization schemes for smooth-ReLUs at any temperature, and provide code for computing the necessary hyper-parameters.
- Introduce GumbelLU, AlgebraicLU and GudermanLU. We note that, in certain aspects of the analysis, GudermanLU is closer to ReLU than the other smooth-ReLUs analysed, including SWISH and GeLU.
- Demonstrate experimentally that representation variance and NTK of smooth-ReLUs resemble those of ReLU under the stable initialization scheme we provide, for any temperature. Also, the same applies for very low temperatures with standard initialization.
- Show that smooth-ReLUs typically have stable and continuous NTK updates during training, while ReLU shows stochasticity instead. We observe that the smooth-ReLU NTK updates are similar under our stable initialization for different temperatures.

2 Preliminaries

2.1 Initialization Covariance

To better understand the trainability of a FFN, we pay attention to how the magnitude of the input propagates through the layers. Let's denote the layer preactivation as,

$$z^{(\ell)}_{i;\alpha} = \sum_{j}^{n} W^{(\ell)}_{ij} \sigma(z^{\ell-1}_{j;\alpha}) + b^{(\ell)}_{i}, \tag{1}$$

where i, j denote neuron indices, α_k data indices, and l and n denote depth and width respectively. The activation function is denoted by σ. We assume the network weights and biases are drawn from normal distributions with means zero and variances $\mathbb{E}\left[b_i^{(l)} b_j^{(l)}\right] = \delta_{ij} C_b^{(l)}$ and $\mathbb{E}\left[W_{i_1 j_1}^{(l)} W_{i_2 j_2}^{(l)}\right] = \delta_{i_1 i_2} \delta_{j_1 j_2} C_W^{(l)} / n_{l-1}$. For simplicity, we assume all layers to have equal width $n_\ell = n$, possibly different from the input size n_0, and the same initialization hyper-parameters, i.e. $C_b^{(l)} = C_b$ and $C_W^{(l)} = C_W$.

We are interested in the magnitude of the second order correlation metric defined as

$$K_{\alpha_1 \alpha_2}^{(\ell)} = \left\langle \frac{1}{n} \sum_{j=1}^{n} z_{j;\alpha_1}^{(\ell)} z_{j;\alpha_2}^{(\ell)} \right\rangle, \tag{2}$$

where $\langle \cdot \rangle$ denotes ensemble averaging with respect to the initializations of the weights of the architecture, and α_i denote different input data samples for different i. Intuitively, this metric measures magnitude change per layer of the layer output, or representation. Given the progressive transformations through layers, the following recursion is satisfied

$$K_{\alpha_1 \alpha_2}^{(\ell+1)} = C_W \left\langle \sigma(z_{\alpha_1}) \sigma(z_{\alpha_2}) \right\rangle_{K = K_{\alpha_1 \alpha_2}^{(\ell)}} + C_b \tag{3}$$

as shown in Theorem 2.2 in (Hanin, 2022). For clarity, we omit input subscript, unless required. When this metric remains constant in magnitude with depth, we call the architecture stable, and the initialization as critical.

As an illustration, we take into consideration linear networks, where the activation is the identity function. In these networks the metric behaves like a scalar dynamic system. The resulting representation variance is $K_{\alpha_1 \alpha_1}^{(\ell+1)} = C_W K_{\alpha_1 \alpha_1}^{(\ell)} + C_b$. If $C_W > 1$, the metric grows exponentially with depth. Instead, if $C_W < 1$, the metric vanishes exponentially to zero with depth, eliminating any data-dependence in the outputs. Finally, if $C_W = 1$ (Glorot and Bengio, 2010), it prevents both, the exponential explosion and vanishing of the variance. In other words, it is a critical initialization. For the same reason, in a non-linear neural network, if the metric increases or decreases exponentially, the variance either diverges to infinity or converges to zero. Both outcomes are undesirable as they lead to exploding or vanishing representations and gradients, and make training very difficult.

In addition, one may be concerned with higher order correlations of the activation and how they grow with depth. For instance, in the infinite-depth-and-width limit for fully connected networks with ReLU activation, the statistics is characterized by the log-Gaussian distribution (Hanin and Nica, 2020a). Here we focus on the $l/n \ll 1$ regime with $l \gg 1$, where we can safely assume that the quadratic moment dominates the distribution (Roberts et al., 2022). In App. A we compute higher moments and confirm this assumption experimentally.

2.2 Neural Tangent Kernels

The Neural Tangent Kernel (NTK) theory is widely used to study the dynamics of infinitely-wide deep neural networks under gradient descent (Jacot et al., 2018; Seleznova and Kutyniok, 2020). Here we focus on the finite width version defined as,

$$\Theta_{i,\alpha_1;j,\alpha_2} = \sum_{\mu,\nu} \partial_{\phi_\mu} z_{i;\alpha_1}^{(L)} \partial_{\phi_\nu} z_{j;\alpha_2}^{(L)} \tag{4}$$

where ϕ represents the parameters of the network, and we sum over all the intermediate layer parameters. For the samples collected we compute the ensemble average,

$$\tilde{\Theta} = \left\langle \frac{1}{n} \sum_{i,j,\alpha_1,\alpha_2} |\Theta_{i,\alpha_1;j,\alpha_2}| \right\rangle \tag{5}$$

Introducing a new index t to indicate training iteration, we define the NTK update as

$$\Delta \tilde{\Theta}_t = \tilde{\Theta}_t - \tilde{\Theta}_{t-1} \tag{6}$$

In general, during training, there is an initial phase of rapid decrease in the loss function followed by a longer phase of very small updates. Based on this observation, and since $\Delta\tilde{\Theta}_t$ is related to the parameters update, we expect it to first increase in the initial epochs and proceed to decrease in magnitude as learning improves. As such, analysing this 'observable' gives another useful indication of whether training converges, in addition to the loss function curve. For a very small learning rate, we have $d\tilde{\Theta}/dt \approx \Delta\tilde{\Theta}_t$, and we expect it to be continuous for smooth functions, but not so for ReLU which is discontinuous at zero.

3 Theory for smooth-ReLU critical initialization

As mentioned, we consider the initialization to be critical when it allows the network to keep the variance constant with depth, i.e. $K^{(\ell+1)}_{\alpha_1\alpha_1} = K^{(\ell)}_{\alpha_1\alpha_1} = K^*$, K^* being the fixed-point value (Glorot and Bengio, 2010; He et al., 2015; Roberts et al., 2022). Here, we are interested in keeping the diagonal components of $K^{(\ell)}_{\alpha_1\alpha_2}$ constant. In App. F and App. D, we prove the following Lemma on the conditions that have to be satisfied for a network with this type of activations to admit critical Gaussian initializations:

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*

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

Notice that even if the criticality exists, it might have an undesirable $K^* = 0$, as it is the case for $a(z) = \arctan(z)$ function (see App. F). Therefore:

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 $K^* \neq 0$.*

We prove that the critical covariance and hyper-parameters can be uniquely determined for all temperatures, if they are known for a specific temperature, which we chose to be $T = 1$ for reference:

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 $(C^*_{W,T}, C^*_{B,T}) = (C^*_W, T^2 C^*_B)$.*

A simple consequence is that both the critical representation covariance K^* and bias variance C^*_B tend to zero as we decrease the temperature:

Corollary 1. $\lim_{T\to 0} K^*_T = \lim_{T\to 0} C^*_{B,T} = 0$.

This is an interesting outcome if we keep in mind that in ReLU networks any value of K serves as a fixed point when critically initialized with $C_W = 2$ and $C_b = 0$ (He et al., 2015), known as He initialization. For smooth-ReLU networks the critical variance tends to zero as the temperature approaches zero and the activation becomes closer in shape to ReLU. However, as we show in the next section in Fig. 1, the scaled layer update, $\Delta K/K^*$, shows the same dependence at both $T = 0.1$ and $T = 1$. In other words, the scaled quantities in this figure have no temperature dependence. Since K^* is proportional to T^2, so is the layer update proportional to T^2. As such, the ΔK tends to zero at low temperatures, effectively making every K behave like a fixed point at $T = 0$.

In Table 1 we show several differentiable relaxations of ReLU, in addition to the well known SWISH and GeLU, together with their critical variances and hyper-parameter initializations. These values were obtained for infinitesimal variations as shown in App. B. In App. G we compute analytically the critical point for GeLU. We introduce GumbelLU, AlgebraicLU, and GudermanLU, by simply observing that there are many functions that can serve as smooth relaxations of the Heaviside function with a temperature parameter. Strictly speaking, these activation functions are different from ReLU, judged for example by C^*_W, which for ReLU has an exact value of $C^*_W = 2$, while it is very close to 2 but never exactly so

for its relaxations. However, the formulas we use to calculate C_W^* and C_b^* are based on first-order perturbation theory, and the inclusion of higher orders could bring $C_W^* = 2$. As we show in this work, the difference in C_W^* is rather small for practical purposes. The last column comments on the stability of the critical point, and semi-stable refers to the fixed point being stable only for perturbations to the left or right, but not both, see App. C.

Activation	$a(u)$	K^*	C_b^*	C_W^*	Stability
SWISH	$(1+\exp(-u))^{-1}$	14.320	0.555	1.988	Semi-stable
GeLU	$\frac{1}{2}\left(1+\text{Erf}(\frac{u}{\sqrt{2}})\right)$	$\frac{1}{2}\left(3+\sqrt{17}\right)$	0.173	1.983	Semi-stable
GumbelLU	$\exp\left(-\exp\left(-u\right)\right)$	21.123	0.606	1.988	Semi-stable
AlgebraicLU	$\frac{1}{2}\left(\frac{u}{\sqrt{(u)^2+1}}+1\right)$	20.210	0.334	2.005	Unstable
GudermanLU	$\frac{1}{2}+\frac{2}{\pi}\tan^{-1}\left(\tanh\left(u\right)\right)$	3.154	0.103	1.990	Semi-stable
ReLU (He)	Heaviside(u)	—	0	2	Stable

Table 1: Critical hyper-parameters for different smooth-ReLU activations. The more similar C_b^* and C_W^* are to their ReLU counterpart, the more ReLU-like the activation can be considered to be in the low temperature limit or when critically initialized (see App. I).

4 Numerical Results

4.1 Metric Layer Update

To better understand how the input magnitude changes with depth, we define the layer update of the metric as $\Delta K = K^{\ell+1} - K^{\ell}$ where C_W^* and C_b^* are tuned to criticality for a given temperature. The experimental results in this and the following subsection were computed from 20 ensembles of 1000 samples each. The input data was sampled from a standard normal distribution and kept the same across all architecture samples, that resulted in K spanning values below K^* up to $100K^*$. In Fig. 1 we compare the experimental layer update, denoted by points with one standard deviation uncertainty bands, and theoretical prediction in dashed lines (see App. B), for SWISH and GeLU, $T = 0.1, 1$, and depth $l = 10$. The horizontal axis denotes the magnitudes of the output of the last layer. As the figure confirms, the theoretical and experimental re-scaled update match, for all temperatures in the entire domain. As shown in the App. C, very near K^* there is another numerically critical point (except for AlgebraicLU), and both points scale together with temperature. There is no qualitative difference between various temperatures and only at $T = 0$ these two fixed points are both equal to zero. As a reminder, for ReLU, once we set $C_W = 2, C_B = 0$, any value of K serves as a fixed point since there is no variance magnitude change between layers. For all the activations in Table 1, $\Delta K \propto K$ for $K \gg K^*$ or $K \ll K^*$, so the update remains exponential in nature, but the exponent is a rather small value. In practice, a deep, but not terribly deep, neural architecture, if properly initialized, should be both expressive and trainable. As an additional confirmation, we also compute the layer update for ReLU initialized at the critical parameters of the two other activations. It is remarkable how well it matches them throughout the domain, with the biggest discrepancies at small values of K^*, in the inset.

To compare critical and He initializations in practice, in Fig. 2 we depict the layer update for He, at depth $l = 10$ and width $n = 320$, at the two temperatures from the previous figure. As a reminder, at $T = 0.1$ the activations look similar to ReLU, while at $T = 1$ they have their natural shape. At $T = 0.1$ we recover the stable critical behavior of ReLU for all three activations. The layer update is essentially zero within the uncertainty band, which can be further reduced by increasing sampling size. At $T = 1$ the two activations are very different from ReLU. For both temperatures, the layer update is negative, which indicates that the magnitude decreases with depth.

In summary, ReLU networks initialized with smooth-ReLU critical hyper-parameters behave like them except at very low values of K. Also, smooth-ReLUs initialized with ReLU

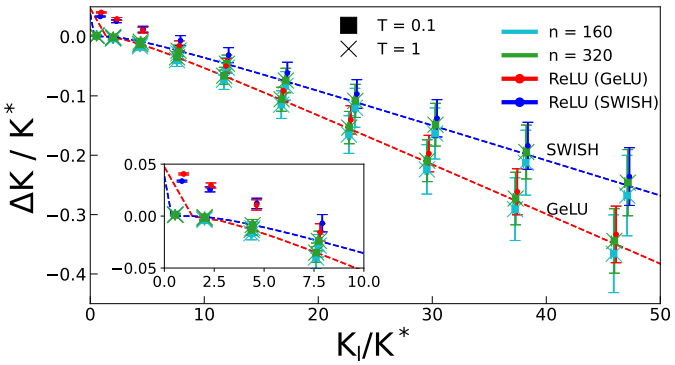
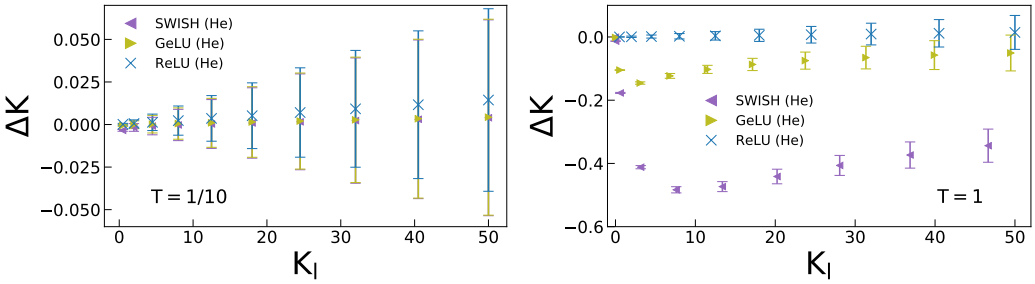


Figure 1: Layer update experimental results, for $T = (0.1, 1)$, critical initializations, for the SWISH and GeLU activation functions. The theoretical results in dashed lines match surprisingly well the experimental values.



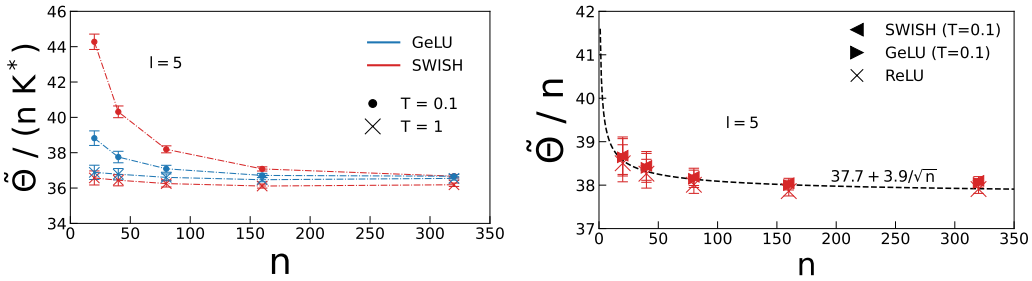
(a) $T = 0.1$ (b) $T = 1$

Figure 2: Experimental results on the layer update for $T = 0.1$, He and critical initializations, at depth $L = 10$ and width $N = 320$. Figure (a) shows how the two relaxations converge to ReLU behavior at low temperatures for the He initialization. In figure (b) we plot them for $T = 1$, and notice sharp differences between each other and ReLU. Both cases confirm that at low temperatures ($T \leq 1$) architectures with continuous relaxations are both differentiable throughout the domain and behave like ReLU.

critical hyper-parameters behave like ReLU only at low temperatures. smooth-ReLUs initialized with their respective hyper-parameters are markedly different from ReLU with He initialization, but the magnitude of the layer update, ΔK, is very small with respect to K. We have confirmed this empirically for GeLU and SWISH in Fig. 1 and Fig. 2.

4.2 NTK at Initialization

In sub-figure 3 (a) we plot ensemble averaged NTK for critical initializations of SWISH and GeLU, as a function of width, for a fixed depth $l = 5$. As it can be seen, $\tilde{\Theta}/(nK^*)$ is approximately a constant with some deviation for smaller width values, irrespective of temperature. Then, in sub-figure (b) we compare them at $T = 0.1$, to ReLU under He initialization. We fit $\tilde{\Theta}/n$ to confirm the constant behavior in the previous plot. We do notice a sub-leading $\sim n^{-1/2}$ which quickly disappears as width increases. It is remarkable that in both cases we observe a constant $\tilde{\Theta}/n$. Not only the magnitude of the output, K, but also the gradients with respect to the parameters, $\tilde{\Theta}$, are very similar between FNN with ReLU and the smooth approximations we study, for low temperatures. We have numerically confirmed the same behaviour to be true for all the functions listed in table 1.



(a) SWISH and GeLU, critical initialization

(b) ReLU, SWISH and GeLU at $T = 0.1$, He initialization

Figure 3: Experimental results for the NTK at $T = 0.1$ and $l = 5$ as function of width. Figure (a) shows the SWISH and GeLU activations initialized at criticality, and figure (b) compares ReLU with SWISH and GELU with He initialization. Both cases confirm that at low temperatures architectures with continuous relaxations are differentiable throughout the domain and behave like ReLU when initialized in the same way.

4.3 Training dynamics

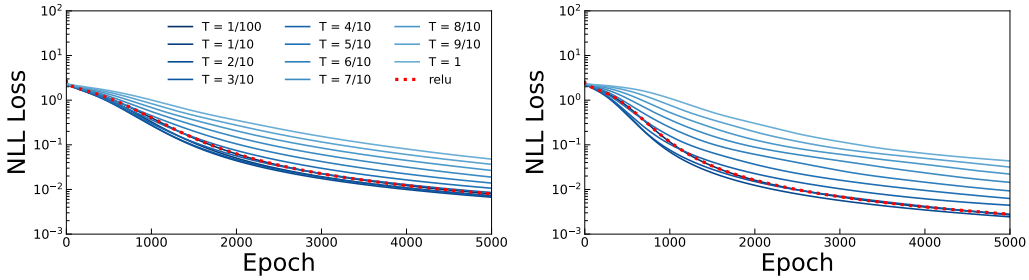
The previous section confirmed that smooth-ReLUs are ReLU-like at initialization, if the network is tuned to criticality. Here we study their training properties. We focus on the CIFAR10 (Krizhevsky et al.) and MNIST (LeCun and Cortes, 2010) datasets. To remove mini-batch stochasticity and better resolve the dynamics, we train with full-batch gradient descent at fixed learning rate of 10^{-3}. Due to computational memory limitations, we select the first 100 images for each category for training, and 10 samples per category for testing. To understand the impact of initialization, we compare the learning curves for He (critical for ReLU), and the SWISH critical initialization, for a wide range of temperatures. The trends we observe are the same across various seeds for the random number generator. For clarity of presentation, all the plots displayed in this section were obtained with seed equal to 1. We tested with various depths from 4 to 16 and for the presentation that follows we display results for $l = 8, 12$. The widths were set equal to the number of pixels of the input, 784 for MNIST and 1024 for CIFAR10. The cost function is the negative log-likelihood. Here we show the training curves for SWISH and ReLU on CIFAR10, and other activations can be found in App. I, where we also show results for the MNIST dataset.

4.3.1 Training Loss function

Fig. 4 shows the training curves on the CIFAR10 dataset, for SWISH and ReLU activations. When using He initialization, training worsens as the temperature is increased. In other words, as the activation becomes increasingly different from ReLU, training takes longer to converge. As the depth is increased, learning becomes harder for these high-temperature, He-initialized smooth-ReLU networks. This finding agrees qualitatively with what we observed in Fig. 2: at low temperatures SWISH behaves like ReLU (and training is stable) while at higher temperatures it becomes less stable (slower learning made worse by greater depths). When SWISH is critically initialized, all loss curves are closer to each other, for all temperatures and for ReLU. In both initialization schemes we notice that the deeper networks have a steeper decrease of loss in the first 1000 epochs region.

4.3.2 Neural Tangent Kernel

The one step update of the empirical NTK, $\Delta\tilde{\Theta}_t$ during training is plotted in Fig. 5. It is smooth for SWISH but not so for ReLU. As the ReLU curve shows no clear trend, in App. I Fig. 9, 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



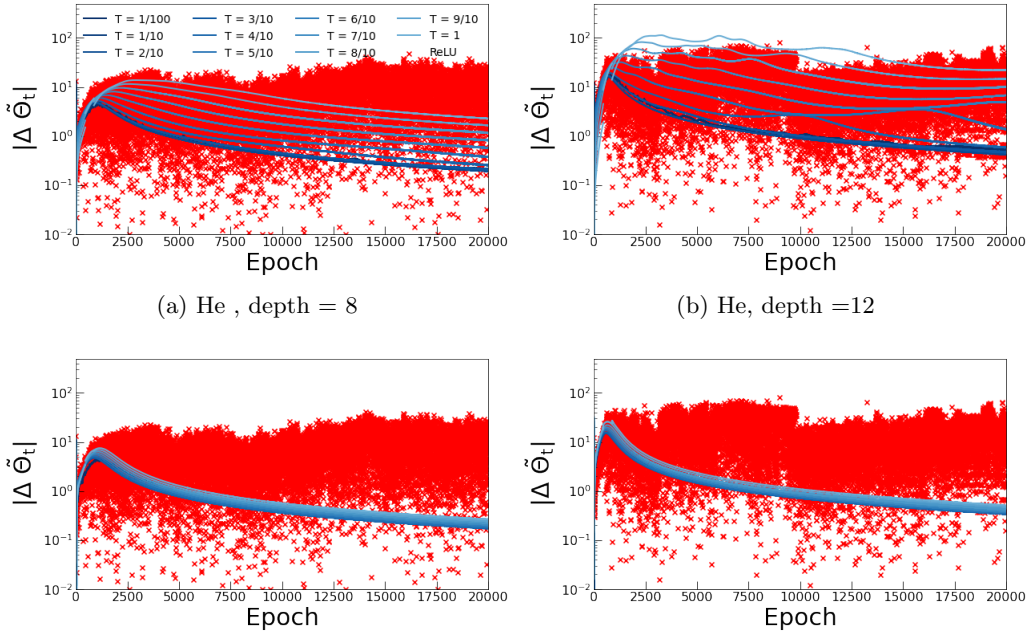
(a) He, depth = 8

(b) He, depth =12

(c) Critical, depth = 8

(d) Critical, depth =12

Figure 4: Loss function 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 5: NTK update for Cifar10 training with SWISH activation, He and critical initialization.

results. The $\Delta\tilde{\Theta}_t$ for smooth-ReLUs first increases and then proceeds to decrease, as the task is learned and convergence achieved. For He there is a clear temperature dependence and the NTK update gets larger with increasing T and l, while for the critical initialization the various curves are closer to each other and less sensitive to depth. This result is a further confirmation of our observations throughout this work. However, ReLU NTK update does not show a clear trend. 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. Remarkably, very low temperature SWISH, $T = 0.01$, starts showing some level of stochasticity in the $\Delta\tilde{\Theta}_t$ as well.

5 Summary and Conclusion

In this work we study differentiable approximations to ReLU, in fully-connected feedforward neural networks and restrict ourselves to a temperature interpolation of the form $\sigma_T(z) = za(z/T)$. We study approximations where $a(z/T)$ approaches the Heaviside function at low temperatures. Inspired by recent field theory treatments, we study the signal magnitude propagation with depth subject to small perturbations. We develop several theorems to explain their behavior and discuss how to make information propagation stable through critical initialization. We find that for small temperatures these activations remain close to ReLU in both initialization statistics and training dynamics. For temperatures of order one or higher, typical initializations (e.g., He) hinder information propagation through exponential decay of signal magnitude. On the other hand, the critical initialization stabilizes the network statistics for all temperatures and results in training curves similar to ReLU. Despite the activations never strictly becoming ReLU, they show ReLU-like properties in practice, while retaining differentiability. This is further confirmed by our neural tangent kernel analysis. At initialization, these activations display NTK values very close to ReLU, with $\tilde{\Theta}/n$ approximately constant. During training they retain a well-defined (continuous) NTK update which helps explain learning and convergence. In other words, smooth-ReLU networks have continuous updates during training, while ReLU networks have noisy updates. In future work, we would like to extend our analysis to convolutional and residual networks. It may be of interest to apply this analysis to neural ODE's, where differentiable activations play a major role. In addition, it would be interesting to analyse data normalization in conjunction with critical initialization.

References

E. Alcaide. E-swish: Adjusting activations to different network depths. *arXiv preprint arXiv:1801.07145*, 2018.

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

S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

S. Bai, J. Z. Kolter, and V. Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

H. H. Chieng, N. Wahid, P. Ong, and S. R. K. Perla. Flatten-t swish: a thresholded relu-swish-like activation function for deep learning. *arXiv preprint arXiv:1812.06247*, 2018.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

S. Elfwing, E. Uchibe, and K. Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Z. Feng and J. Z. Kolter. On the neural tangent kernel of equilibrium models. 2020.

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

B. Hanin. Correlation functions in random fully connected neural networks at finite width. *arXiv preprint arXiv:2204.01058*, 2022.

B. Hanin and M. Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, May 2020a. ISSN 1432-0916. doi: 10.1007/s00220-019-03624-z. URL https://doi.org/10.1007/s00220-019-03624-z.

B. Hanin and M. Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020b.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th international conference on computer vision*, pages 2146–2153. IEEE, 2009.

A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

M. Li, M. Nica, and D. Roy. The future is log-gaussian: Resnets and their infinite-depth-and-width limit at initialization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7852–7864. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/412758d043dd247bddea07c7ec558c31-Paper.pdf.

C. Liu, L. Zhu, and M. Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33: 15954–15964, 2020.

C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pages 2847–2854. PMLR, 2017.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

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

A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

M. Seleznova and G. Kutyniok. Analyzing finite neural networks: Can we trust neural tangent kernel theory? *CoRR*, abs/2012.04477, 2020. URL https://arxiv.org/abs/2012.04477.

S. B. Shlosman. Signs of the Ising model Ursell functions. *Communications in Mathematical Physics*, 102(4):679 – 686, 1986. doi: cmp/1104114545. URL https://doi.org/.

M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

G. Yang. Tensor programs ii: Neural tangent kernel for any architecture, 2020. URL https://arxiv.org/abs/2006.14548.

D. Zou, Y. Cao, D. Zhou, and Q. Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

Appendices

A 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

$$Z^{(c)}_{i,\alpha_1,j,\alpha_2} = \langle z_{i,\alpha_1} z_{j,\alpha_2} \rangle - \langle z_{i,\alpha_1} \rangle \langle z_{j,\alpha_2} \rangle = \langle z_{i,\alpha_1} z_{j,\alpha_2} \rangle. \tag{7}$$

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

$$\begin{aligned} Z^{(c)}_{i,\alpha_1,j,\alpha_2,k,\alpha_3} &= \langle z_{i,\alpha_1} z_{j,\alpha_2} z_{k,\alpha_3} \rangle - \langle z_{i,\alpha_1} \rangle \langle z_{j,\alpha_2} z_{k,\alpha_3} \rangle \\ &- \langle z_{j,\alpha_2} \rangle \langle z_{i,\alpha_1} z_{k,\alpha_3} \rangle - \langle z_{k,\alpha_3} \rangle \langle z_{i,\alpha_1} z_{j,\alpha_2} \rangle + 2 \langle z_{i,\alpha_1} \rangle \langle z_{j,\alpha_2} \rangle \langle z_{k,\alpha_3} \rangle. \end{aligned} \tag{8}$$

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,

$$\begin{aligned} \hat{K}^{(\ell)}_{\alpha_1\alpha_2} &= \frac{1}{n} \sum_{j=1}^{n} z^{(\ell)}_{j;\alpha_1} z^{(\ell)}_{j;\alpha_2}, \\ \Delta \hat{K}^{(\ell)}_{\alpha_1\alpha_2} &= \hat{K}^{(\ell)}_{\alpha_1\alpha_2} - K^{(\ell)}_{\alpha_1\alpha_2} \\ V^{(\ell)}_{\alpha_1\alpha_2;\alpha_3\alpha_4} &= \langle \Delta \hat{K}^{(\ell)}_{\alpha_1\alpha_2} \Delta \hat{K}^{(\ell)}_{\alpha_3\alpha_4} \rangle \end{aligned} \tag{9}$$

If the first and third cumulants can be ignored,

$$Z^{(c)\ell}_{i,\alpha_1,j,\alpha_2,k,\alpha_3,r,\alpha_4} = \frac{1}{n} \left(\delta_{i,j}\delta_{k,r} V^{\ell}_{\alpha_1\alpha_2;\alpha_3\alpha_4} + \delta_{i,k}\delta_{j,r} V^{\ell}_{\alpha_1\alpha_3;\alpha_2\alpha_4} + \delta_{i,r}\delta_{j,k} V^{\ell}_{\alpha_1\alpha_4;\alpha_2\alpha_3} \right) \tag{10}$$

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,

$$Z^{(4)}_c = 3nV, \tag{11}$$

where we have suppressed the the depth superscript for simplicity of notation. In fig. 6 (a) we verify this relation and also compute the ratio of the third over the second cumulant in fig. 6 (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.

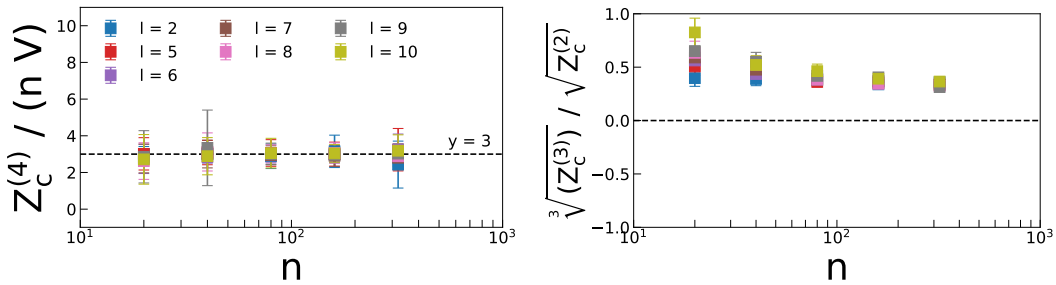
B Fluctuations through layers

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

$$\chi_{\|}(K) = C_W \frac{dg}{dK} = \frac{C_W}{2K^2} \left\langle \sigma(z)\sigma(z) \left(z^2 - K\right) \right\rangle_K . \tag{12}$$

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,

$$\chi_{\perp}(K) = C_W \left\langle \frac{d\sigma(z)}{dz} \frac{d\sigma(z)}{dz} \right\rangle_K \tag{13}$$



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

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

Figure 6: The experimental results the correlations for the SWISH activation with critial initialization at $T = 1$.

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,

$$\chi_{||}(K^*) = \chi_\perp(K^*) = 1. \tag{14}$$

We will call such a case 'critical'.

As an illustration, let's analyze the ReLU activation

$$\sigma(z) = \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases} \tag{15}$$

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.

B.1 Susceptibilities for ReLu function

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

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

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

Criticality implies $C_W = 2$.

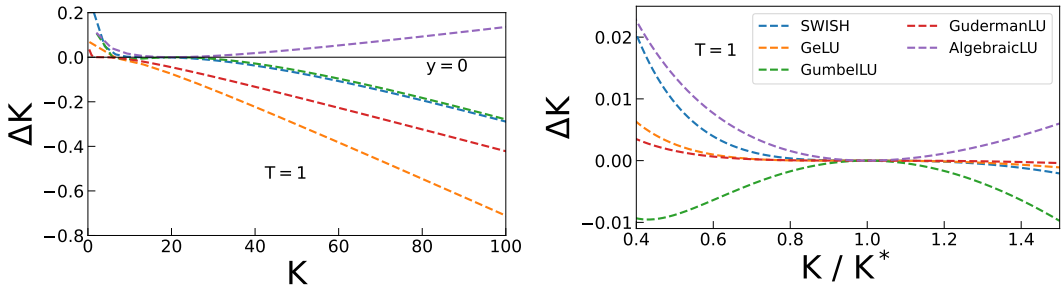
These results do not depend on K^* and the recursion equation is

$$\begin{aligned} K^* =& C_b + C_W \frac{K^* 2}{2} \\ =& C_b + K^* \end{aligned} \tag{16}$$

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

C Critical point

In Fig. 7 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. 8 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 7: Theoretical layer update for all activation functions.

D 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*

- $2\langle za(z)a'(z)\rangle_{K^*} = -\langle z^2 a(z)a''(z)\rangle_{K^*}$
- $z^3 a(z)^2$ *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

$$\sigma(z) = za(z) \tag{17}$$

$$\sigma'(z) = a(z) + za'(z) \tag{18}$$

$$\sigma''(z) = 2a'(z) + za''(z) \tag{19}$$

$$\tag{20}$$

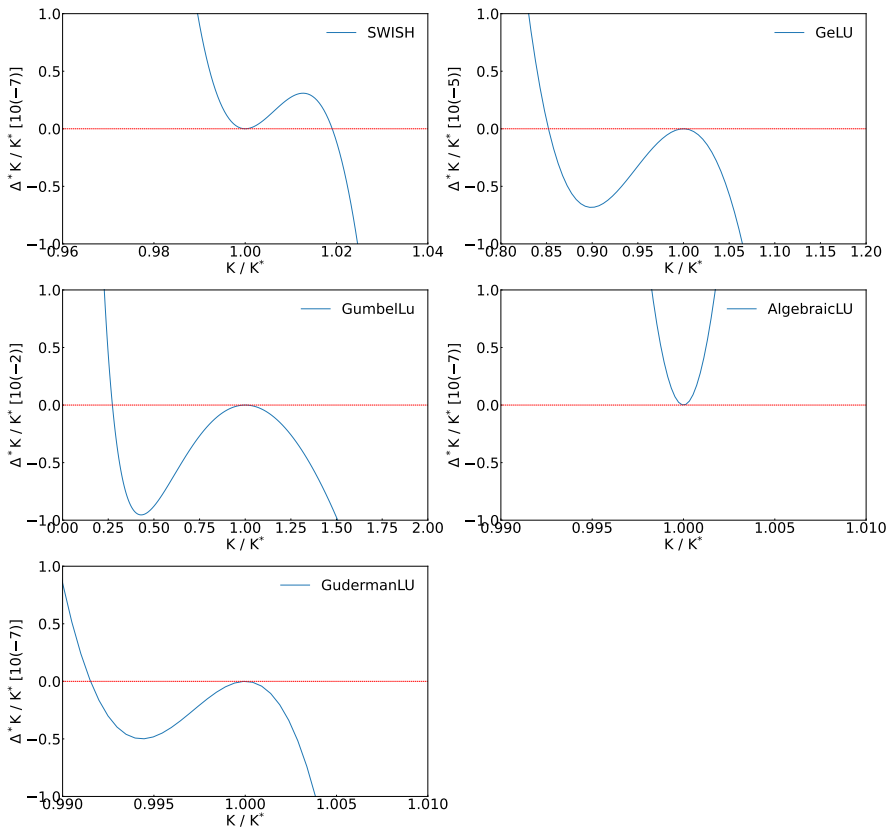


Figure 8: Zoomed in view of the theoretical layer update for all activation functions. Apart from AlgebraicLu, they all display 2 fixed points.

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

$$\chi_\perp = \chi_\parallel + C_W \langle \sigma\sigma'' \rangle_K \tag{21}$$

Then, the following equation has to be true at criticality

$$\langle \sigma\sigma'' \rangle_K = 0 \tag{22}$$

Therefore, substituting the activation definition we have that

$$\langle za(z)\Big(2a'(z) + za''(z)\Big)\rangle_K = 0 \tag{23}$$

$$2\langle za(z)a'(z)\rangle = -\langle z^2 a(z)a''(z)\rangle_K \tag{24}$$

$$\tag{25}$$

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

$$\frac{da_T(z)}{dz} = \frac{d}{dz}a(z/T) \tag{26}$$
$$= \frac{1}{T}\frac{d}{d(z/T)}a(z/T) \tag{27}$$
$$= \frac{1}{T}\frac{da(u)}{du}, \tag{28}$$
$$\tag{29}$$

and that

$$\frac{d^2a_T(z)}{dz^2} = \frac{1}{T^2}\frac{d^2a(u)}{du^2}, \tag{30}$$
$$\tag{31}$$

Then we substitute in the criticality condition found in Lemma 1

$$2\langle za(z)a'(z)\rangle = -\langle z^2a(z)a''(z)\rangle_K \tag{32}$$
$$\tag{33}$$

$$\langle za_T(z)a_T'(z)\rangle_{K_T} = \frac{1}{\sqrt{2\pi K}}\int dz\ e^{-\frac{z^2}{2K}}\ za_T(z)a_T'(z) \tag{34}$$
$$= \frac{1}{\sqrt{2\pi K_T}}\int Tdu\ e^{-\frac{(uT)^2}{2K_T}}(uT)a(u)\frac{1}{T}\frac{da(u)}{du} \qquad u = z/T \tag{35}$$
$$= \frac{1}{T}\frac{1}{\sqrt{2\pi(K_T/T^2)}}\int du\ e^{-\frac{u^2}{2(K_T/T^2)}}ua(u)\frac{da(u)}{du} \tag{36}$$
$$= \frac{1}{T}\langle za(z)a'(z)\rangle_{K_T/T^2} \tag{37}$$

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

$$\left\langle z^2a_T(z)a_T''(z)\right\rangle_{K_T} = \frac{1}{\sqrt{2\pi K}}\int dz\ e^{-\frac{z^2}{2K}}\ z^2a_T(z)a_T''(z) \tag{38}$$
$$= \frac{1}{\sqrt{2\pi K_T}}\int Tdu\ e^{-\frac{(uT)^2}{2K_T}}(uT)^2a(u)\frac{1}{T^2}\frac{d^2a(u)}{du^2} \qquad u = z/T \tag{39}$$
$$= \frac{1}{T}\frac{1}{\sqrt{2\pi(K_T/T^2)}}\int du\ e^{-\frac{u^2}{2(K_T/T^2)}}u^2a(u)\frac{d^2a(u)}{du^2} \tag{40}$$
$$= \frac{1}{T}\left\langle z^2a(z)a''(z)\right\rangle_{K_T/T^2} \tag{41}$$

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. □

E 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 $(C_{W,T}^*, C_{B,T}^*) = (C_W^*, T^2C_B^*)$.*

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,

$$g_T(K) = \langle \sigma_T(z)\sigma_T(z)\rangle_K = \frac{1}{\sqrt{2\pi K}} \int dz\ e^{-\frac{z^2}{2K}}\ z^2 a(z/T)^2 \tag{42}$$

$$= \frac{1}{\sqrt{2\pi K}} \int T du\ e^{-\frac{(uT)^2}{2K}} (uT)^2 a(u)^2 \qquad u = z/T \tag{43}$$

$$= T^2 \frac{1}{\sqrt{2\pi (K/T^2)}} \int du\ e^{-\frac{u^2}{2(K/T^2)}} u^2 a(u)^2 \tag{44}$$

$$= T^2 g(K/T^2) \tag{45}$$

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

$$\chi_{\parallel,T}(K_T^*) = C_{W,T} \frac{dg_T(K_T)}{dK_T}\Big|_{K_T^*} \tag{46}$$

$$= \frac{1}{T^2} C_{W,T} \frac{dg_T(K_T)}{d(K_T/T^2)}\Big|_{K_T^*} \tag{47}$$

$$= \frac{T^2}{T^2} C_{W,T} \frac{dg(K_T/T^2)}{d(K_T/T^2)}\Big|_{K_T^*} \tag{48}$$

$$= C_{W,T} \frac{dg(K_T/T^2)}{d(K_T/T^2)}\Big|_{K_T^*} \tag{49}$$

$$= C_{W,T} \frac{dg(K)}{d(K)}\Big|_{K^*} \tag{50}$$

$$\chi_{\parallel}(K) = C_W \frac{dg(K)}{dK}\Big|_{K^*} \tag{51}$$

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:

$$K_T = C_{b,T} + C_{W,T} g_T(K_T) \tag{52}$$

$$\tag{53}$$

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

$$K_T^* = C_{b,T} + C_{W,T} g_T(K_T^*) \tag{54}$$

$$= C_{b,T} + C_{W,T} T^2 g(K_T^*/T^2) \tag{55}$$

$$K_T^*/T^2 = C_{b,T}/T^2 + C_{W,T} g(K_T^*/T^2) \tag{56}$$

$$K^* = C_{b,T}/T^2 + C_W g(K^*) \tag{57}$$

$$\tag{58}$$

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

$$K^* = C_b + C_W g(K^*) \tag{59}$$

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

□

F 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

$$\langle \sigma\sigma'' \rangle = \left\langle \frac{T^3 z \left(2\tan^{-1}\left(\frac{z}{T}\right) + \pi\right)}{\pi^2 \left(T^2 + z^2\right)^2} \right\rangle \tag{60}$$

$$= \frac{2T^3}{\pi^2} \left\langle \frac{z \tan^{-1}\left(\frac{z}{T}\right)}{\left(T^2 + z^2\right)^2} \right\rangle \geq 0 \tag{61}$$

$$\tag{62}$$

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

□

G Analytical K^* for GeLU

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

$$\langle \sigma\sigma'' \rangle = -\frac{K_T T \left(K_T^2 - 3K_T T^2 - 2T^4\right)}{2\pi \left(K_T + T^2\right)^2 \left(2K_T + T^2\right)^{3/2}} \tag{63}$$

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

$$K_T^* T \left(K_T^{*\,2} - 3K_T^* T^2 - 2T^4\right) = 0 \tag{64}$$

$$K_T^* = 0, \;\; K_T^* = \frac{T^2}{2}\left(3 \pm \sqrt{17}\right) \tag{65}$$

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

H 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

$$\frac{d}{dK}\langle \sigma\sigma \rangle_K = \frac{d}{dK}\frac{1}{\sqrt{2\pi K}} \int e^{\frac{-x^2}{2K}} \sigma\sigma \tag{66}$$

$$= -\frac{1}{2K\sqrt{2\pi K}} \int e^{\frac{-x^2}{2K}} \sigma\sigma + x^2 \frac{1}{2K^2} \frac{1}{\sqrt{2\pi K}} \int e^{\frac{-x^2}{2K}} \sigma\sigma \tag{67}$$

$$= \frac{1}{2K^2} \langle (x^2 - K)\sigma\sigma \rangle_K \tag{68}$$

Then, integrating by parts

$$u = x\sigma^2 \qquad du = \sigma^2 dx + 2x\sigma\sigma' dx \tag{69}$$

$$dv = xe^{\frac{-x^2}{2K}} dx \qquad v = -Ke^{\frac{-x^2}{2K}} \tag{70}$$

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

$$\int udv = uv - \int vdu \tag{71}$$
$$= \int Ke^{\frac{-x^2}{2K}}\sigma^2 dx + \int Ke^{\frac{-x^2}{2K}} 2x\sigma\sigma' dx \tag{72}$$
$$= Kg(K) + 2K\langle x\sigma\sigma'\rangle_K \tag{73}$$

Integrating again by parts the second integral we get

$$u = \sigma\sigma' \qquad du = \sigma\sigma'' dx + \sigma'\sigma' dx \tag{74}$$
$$dv = xe^{\frac{-x^2}{2K}} dx \qquad v = -Ke^{\frac{-x^2}{2K}} \tag{75}$$

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

$$\int udv = uv - \int vdu \tag{76}$$
$$= \int Ke^{\frac{-x^2}{2K}}\sigma\sigma'' dx + \int Ke^{\frac{-x^2}{2K}}\sigma'\sigma' dx \tag{77}$$
$$= K\langle\sigma\sigma''\rangle + K\langle\sigma'\sigma'\rangle \tag{78}$$

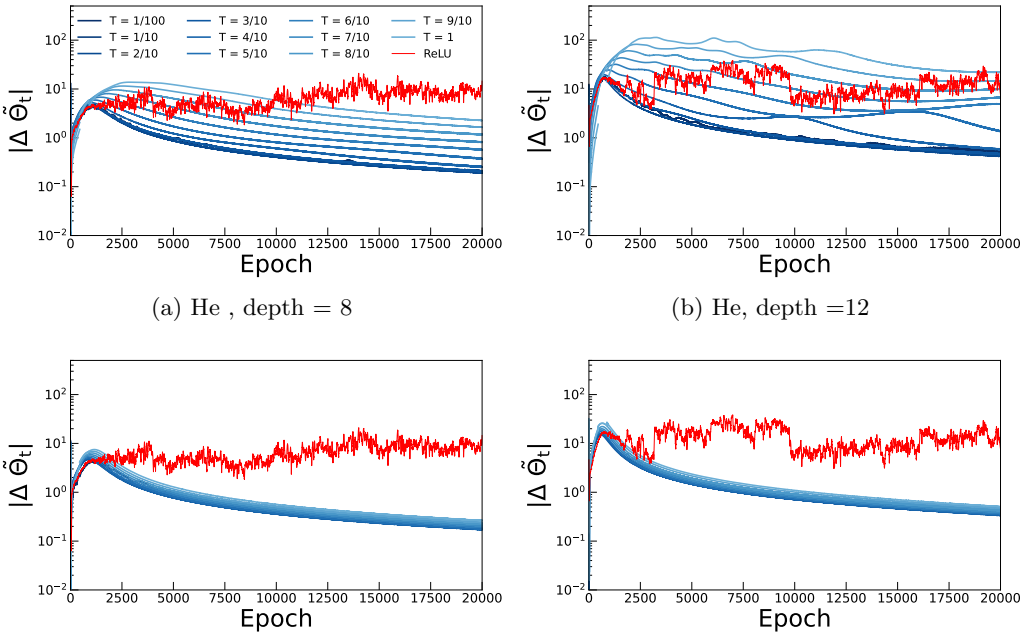
Substituting in the initial integral we get

$$\frac{d}{dK}\langle\sigma\sigma\rangle_K = \frac{1}{2K^2}\langle(x^2 - K)\sigma\sigma\rangle_K \tag{79}$$
$$= \frac{1}{2K^2}\langle x^2\sigma\sigma\rangle_K - \frac{1}{2K}\langle\sigma\sigma\rangle_K \tag{80}$$
$$= \frac{1}{2K^2}(K\langle\sigma\sigma\rangle_K + 2K\langle x\sigma\sigma'\rangle_K) - \frac{1}{2K}\langle\sigma\sigma\rangle_K \tag{81}$$
$$= \frac{1}{K}\langle x\sigma\sigma'\rangle_K \tag{82}$$
$$= \langle\sigma\sigma''\rangle + \langle\sigma'\sigma'\rangle \tag{83}$$

□

I Additional loss and NTK results through training

In Fig. 9 and what follows we display window filtered NTK update for ReLU. 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. 10 and Fig. 11, with the CIFAR10 dataset, and GeLU, Fig. 12 and Fig. 13, 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.



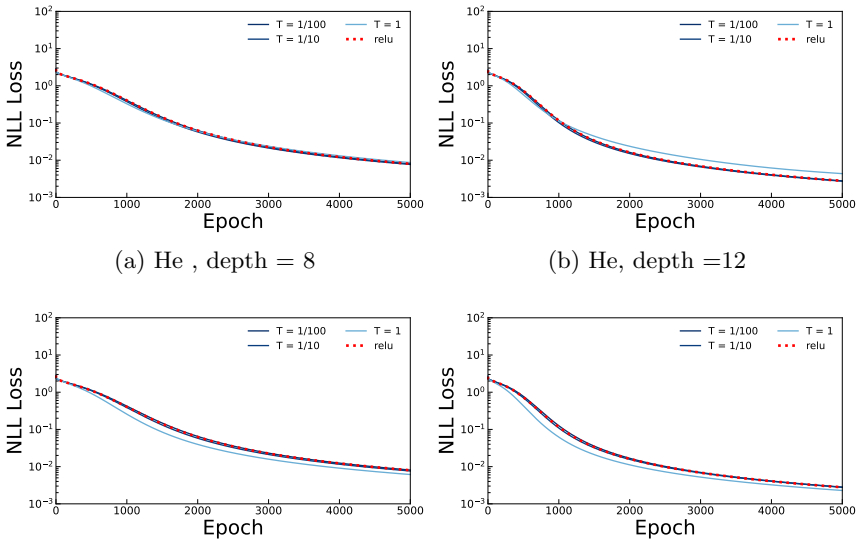
(a) He , depth = 8

(b) He, depth =12

(c) Critical, depth = 8

(d) Critical, depth =12

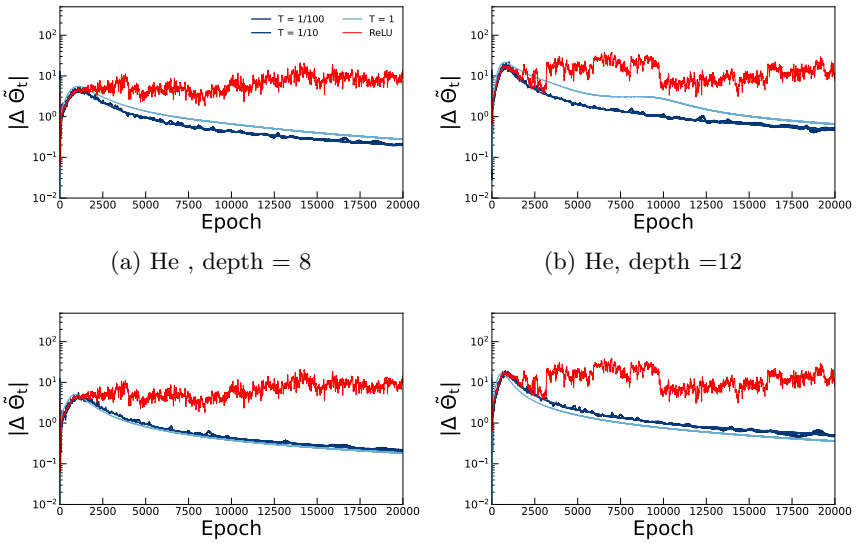
Figure 9: NTK update for Cifar10 training with SWISH activation, He and critical initialization.



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

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

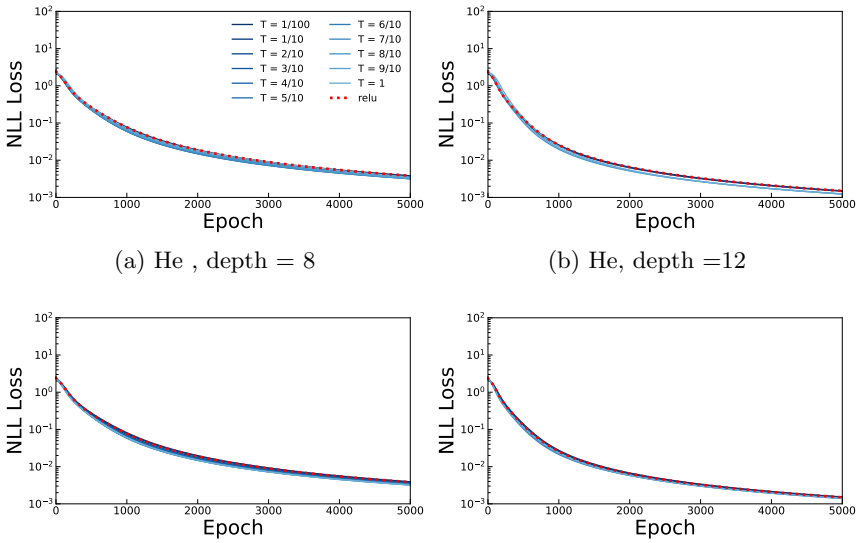
Figure 10: Loss function for Cifar10 training with GudermanLU activation, He and critical initialization.



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

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

Figure 11: NTK update for Cifar10 training with GudermanLU activation, He and critical initialization.



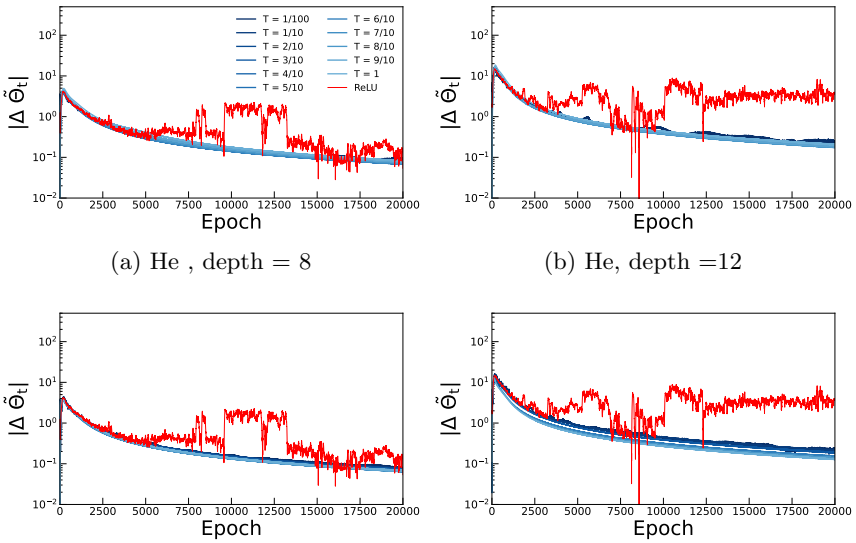
(a) He , depth = 8

(b) He, depth =12

(c) Critical, depth = 8

(d) Critical, depth =12

Figure 12: 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 13: NTK update for MNIST training with GeLU activation, He and critical initialization.