A Notation

Let $\|\cdot\|$ denote the Euclidean (L_2) norm. We use x_i to denote the i-th component of a vector, and A_{ij} to denote the entries of a matrix A. For a $d\times d$ symmetric positive semi-definite matrix Σ, the notation $\mathcal{N}(m,\Sigma)$ denotes the d-dimensional multi-variate normal distribution with mean m and covariance matrix Σ. I_d denotes the $d\times d$ identity matrix. For a real number $r\in\mathbb{R}$ and a non-negative integer k, we introduce the binomial coefficient

$$\binom{r}{k} := \frac{r(r-1)(r-2)\cdots(r-k+1)}{k!}.$$

For $z>0$, (Euler's) Gamma function is defined as the integral

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx.$$

For $z>0$, the digamma function is defined as

$$\psi_0(z) := \left(\frac{d}{dz}\Gamma(z)\right)/\Gamma(z)$$

and the trigamma function

$$\psi_1(z) := \frac{d}{dz}\psi_0(z).$$

For $x,y>0$, the Beta function is defined as the integral

$$\mathrm{B}(x,y) := \int_0^1 t^{x-1}(1-t)^{y-1}\,dt.$$

If a sequence of random variables X_k converges to a random variable X in distribution as $k\to\infty$, we denote this by $X_k \Rightarrow X$. The set of positive integers will be denoted by $\mathbb{Z}_+$. Let f,g be real valued functions, defined on some unbounded subset of $\mathbb{R}$, and let $g(x)$ be strictly positive for all large enough values of x. We denote $f(x)=\mathcal{O}(g(x))$ as $x\to\infty$ if there exists a constant $M>0$ and $x_0\in\mathbb{R}$ such that $|f(x)|\le Mg(x)$ for all $x\ge x_0$. We denote $f(x)=o(g(x))$ as $x\to\infty$ if for all $\varepsilon>0$ there exists a constant x_0 such that $|f(x)|\le\varepsilon g(x)$ for all $x\ge x_0$.

B Proof of Theorem 1

Proof. When the activation function is linear, k-th layer output $x^{(k)}$ obeys a linear recursion where the proof technique of Cohen and Newman (1984) about the moments of random Gaussian matrix products are directly applicable. Our main proof idea is to extend this proof technique to non-linear recursions obeyed by $x^{(k)}$ when ReLU activation is used, where we exploit piecewise-linearity properties of the ReLU function. We first note that for $x^{(k)}\neq 0$,

$$\begin{aligned}\frac{\|x^{(k+1)}\|}{\|x^{(k)}\|} &= \frac{\|\phi_0(W^{(k+1)}x^{(k)})\|}{\|x^{(k)}\|}\\ &= \left\|\phi_0\left(W^{(k+1)}\frac{x^{(k)}}{\|x^{(k)}\|}\right)\right\|,\end{aligned} \tag{B.1}$$

which we used the equality $\phi_0(cy)=c\phi_0(y)$ for any given $c>0$ and arbitrary vector y (with the choice of $y=W^{(k)}x^{(k)}$ and $c=1/\|x^{(k)}\|$). On the other hand, the entries of the $W^{(k)}$ matrix are i.i.d. Gaussians, where each row is a spherically symmetric random vector (in the sense of (Fourdrinier et al., 2018, Ch. 4)) with i.i.d. entries. From this symmetry property it follows that the distribution of $W^{(k)}z$ is independent of the choice of z on the unit sphere in $\mathbb{R}^d$. Therefore, if we choose $z=e_1$, we have

$$W^{(k)}\frac{x^{(k)}}{\|x^{(k)}\|} \sim W^{(k)}e_1,$$

where $e_1=[1,0,\ldots,0]^T$ is the first basis vector. Therefore from (B.1), we obtain

$$\frac{\|x^{(k+1)}\|}{\|x^{(k)}\|} \sim \left\|\phi_0\left(W^{(k+1)}e_1\right)\right\|,$$

which says that the distribution of the ratio $\frac{\|x^{(k+1)}\|}{\|x^{(k)}\|}$ is independent of $x^{(k)}$ and the history $x^{(j)}$ for $j<k$. Then, by the independence of the random variables $\frac{\|x^{(j+1)}\|}{\|x^{(j)}\|}$, we can write

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\|x^{(k)}\|}{\|x^{(0)}\|}\right)^s\right] &= \mathbb{E}\left[\Pi_{j=1}^k \frac{\|x^{(j)}\|^s}{\|x^{(j-1)}\|^s}\right]\\ &= \Pi_{j=1}^k \mathbb{E}\left[\frac{\|x^{(j)}\|^s}{\|x^{(j-1)}\|^s}\right]\\ &= \Pi_{j=1}^k \mathbb{E}\left\|\phi_0\left(W^{(j)}e_1\right)\right\|^s\\ &= \left(\sigma^s\,\mathbb{E}\left\|\phi_0\left(z\right)\right\|^s\right)^k,\end{aligned} \tag{B.2}$$

where z is a d-dimensional random vector with standard normal distribution $\mathcal{N}(0,I_d)$. The rest of the proof is about explicit computation of the term $\mathbb{E}\left\|\phi_0\left(z\right)\right\|^s$ which appear in the product (B.2) and showing that it is equal to $I_0(s,d)$ where $I_0(s,d)$ is defined by (3.1). Note that

$$\mathbb{E}\left\|\phi_0\left(z\right)\right\|^s = \mathbb{E}\left[\phi_0^2(z_1)+\phi_0^2(z_2)+\cdots+\phi_0^2(z_d)\right]^{s/2},$$

where $z=(z_1,z_2,\ldots,z_d)$ and z_i are i.i.d. standard normal random variables. We first note that the function $\phi_0(x):\mathbb{R}^d\to\mathbb{R}$ has a piecewise linear structure on $\mathbb{R}^d$ depending on the sign of the components x_i of a vector

x. In particular, we observe that by the definition of the ϕ_0 function,

$$\begin{aligned}\|\phi_0(z)\|^2 =&\phi_0^2(z_1)+\phi_0^2(z_2)+\cdots+\phi_0^2(z_d)\\ =&\sum_{i:z_i>0}(z_i)^2,\end{aligned} \tag{B.3}$$

which depends on the orthant that the vector z resides in $\mathbb{R}^d$. In particular, there are 2^d (open) orthants in dimension d, where each orthant is defined by a system of inequalities:

$$\varepsilon_1 x_1>0,\quad \varepsilon_2 x_2>0,\quad \varepsilon_3 x_3>0,\quad \ldots \varepsilon_n x_n>0,$$

where each ε_i is 1 or -1. Therefore, we can identify each orthant from an element of the set $\{+,-\}^d$. For example, the non-negative (open) orthant corresponds to $\{+,+,\ldots,+\}$ whereas the non-positive (open) orthant corresponds to $\{-,-,\ldots,-\}$. On every quadrant that corresponds to n plus signs and $d-n$ minus signs (with an arbitrary order of the signs), the distribution of (B.3) is the same as the distribution of

$$Y_n := \chi^2(n), \tag{B.4}$$

where $\chi^2(n)$ denotes a chi-squared distribution with n degrees of freedom as long as $n\geq 1$. If we choose a random quadrant; with probability

$$p_d(n)=\binom{d}{n}\frac{1}{2^d}, \tag{B.5}$$

we will be in such a quadrant. [1] Therefore, we can interpret $\|\phi_0(z)\|^2$ as a mixture of chi-square distributions with weights from the Binomial distribution. It follows from (B.3)–(B.5) that we can write

$$\mathbb{E}\left\|\phi_0\left(z\right)\right\|^s=\sum_{n=1}^d p_d(n)\mathbb{E}(Y_n^{s/2}).$$

The moments of Y_n are explicitly known, and we have

$$\mathbb{E}(Y_n^s)=2^s\frac{\Gamma(n/2+s)}{\Gamma(n/2)}\quad \text{for}\quad s\geq 0,$$

(see (Walck, 1996, Sec. 8)) for any $s\geq 0$ where $\Gamma(\cdot)$ denotes the Gamma function. Therefore, we obtain

$$\mathbb{E}\left\|\phi_0\left(z\right)\right\|^s=I_0(s,d)=\sum_{n=0}^d p_d(n)2^{s/2}\frac{\Gamma(n/2+s/2)}{\Gamma(n/2)}.$$

[1] Note that $p_d(n)=\mathbb{P}(B_d=n)$ where $B_d\sim \mathrm{Bi}(d,\frac{1}{2})$ is a random variable with a Binomial distribution, where the parameter d represents the total number of Bernouilli trials and the parameter $\frac{1}{2}$ is the success probability for each trial.

We conclude from (B.2) that (3.1) holds. This also implies directly that part (ii) and (iii) are true. Finally, for any $p>s$, we have $\bar{\sigma}_0(s,d)>\bar{\sigma}_0(p,d)$ by Corollary 9. Therefore, if $\sigma=\bar{\sigma}_0(s,d)$, then $\sigma>\bar{\sigma}_0(p,d)$ and by part (iii), we obtain $\mathbb{E}\|x^{(k)}\|^p\to\infty$ exponentially fast in k. This completes the proof. □

Remark 16. *In the setting of Theorem 1, in the special case when $s=2$, we obtain $I_0(2,d)=d/2$ and we obtain $I_0(2,d)=2\sum_{n=1}^d\binom{d}{n}\frac{1}{2^d}\frac{\Gamma(n/2+1)}{\Gamma(n/2)}=\sum_{n=0}^d\binom{d}{n}\frac{1}{2^d}n=\frac{d}{2}$ where we used the identity $\Gamma(x+1)=x\Gamma(x)$ for $x>0$ and the last equality can be obtained from the properties of the Binomial distributions, see e.g. (Walck, 1996, Section 5.2). Therefore, from part (i) of Theorem 1, $\bar{\sigma}_0(2,d)=1/\sqrt{I_0(2,d)}=\sqrt{2}/\sqrt{d}$. In particular, the choice of $\bar{\sigma}_0(2,d)$ corresponds to Kaiming initialization. Theorem 1 is more general in the sense that it is applicable to any moment $s>0$.*

C Proof of Corollary 4

Proof. This result follows from analyzing the asymptotics of $I_0(s,d)$ for large d. It is known that for any real $\alpha>0$ and $z>0$, we can write the series expansion

$$\frac{\Gamma(z+\alpha)}{\Gamma(z)}=z^\alpha S(z,\alpha),$$

with

$$\begin{aligned}S(z,\alpha):=&\sum_{m=0}^\infty A_m(\alpha)(\frac{1}{z})^m\\ =&1+\frac{\alpha(\alpha-1)}{2z}+\mathcal{O}(\frac{1}{z^2}),\end{aligned} \tag{C.1}$$

where $A_m(\alpha)$ are coefficients of the expansion that admits an explicit representation (see (Tricomi et al., 1951)). Therefore, choosing $z=n/2$ and $\alpha=s/2$,

$$\frac{\Gamma(n/2+s/2)}{\Gamma(n/2)}=(\frac{n}{2})^{s/2}S(n/2,s/2), \tag{C.2}$$

so that

$$I_0(s,d)=\sum_{n=1}^d p_d(n)n^{s/2}S(n/2,s/2).$$

Since the Γ function is log-convex (Merkle, 1996), we also have

$$\begin{aligned}\Gamma(\frac{n}{2}+\frac{s}{2}) &= \Gamma\left((1-\frac{s}{2})\frac{n}{2}+\frac{s}{2}(\frac{n}{2}+1)\right)\\ &\leq \left(\Gamma(\frac{n}{2})\right)^{1-\frac{s}{2}}\left(\Gamma(\frac{n}{2}+1)\right)^{\frac{s}{2}}\\ &= \Gamma(\frac{n}{2})(\frac{n}{2})^{s/2},\end{aligned}$$

where we used the identity $\Gamma(z+1) = z\Gamma(z)$ for $z > 0$. Therefore, we see from (C.2) that

$$0 \le S(n/2, s/2) \le 1, \tag{C.3}$$

for every $s > 0$ and $n > 0$. Note that

$$\frac{(\frac{d}{2})^{s/2}}{I_0(s,d)} = \frac{1}{\mathbb{E}(F_d(B_d))}, \tag{C.4}$$

where B_d is a Binomial random variable, i.e.

$$\mathbb{P}(B_d = n) = \binom{d}{n}\frac{1}{2^d} \quad \text{for} \quad n = 0, 1, \dots, d, \tag{C.5}$$

and

$$F_d(X) := \begin{cases} 2^{s/2}\frac{X^{s/2}}{d^{s/2}} S(X/2, s/2) & \text{if} \quad X > 0, \\ 0 & \text{if} \quad X = 0, \end{cases}$$

satisfying for all $X > 0$

$$F_d(X) = 2^{s/2}\frac{X^{s/2}}{d^{s/2}}\left(1 + \frac{\frac{s}{2}(\frac{s}{2}-1)}{X} + \mathcal{O}(\frac{1}{X^2})\right) \tag{C.6}$$

where we used (C.1). By the normal approximation of the binomial distribution, we also have

$$Z_d := \frac{B_d - \mathbb{E}(B_d)}{\sqrt{\text{var}B_d}} = \frac{B_d - \frac{d}{2}}{\sqrt{d}/2} \Rightarrow \mathcal{N}(0,1) \tag{C.7}$$

in distribution. We also have

$$\begin{aligned} & \mathbb{E}(F_d(B_d)) \\ = \; & \mathbb{E}\left(F_d(\frac{d}{2} + \frac{\sqrt{d}}{2}Z_d)\right) \\ = \; & 2^{s/2}\mathbb{E}\left[\frac{(\frac{d}{2} + \frac{\sqrt{d}}{2}Z_d)^{s/2}}{d^{s/2}} S(\frac{d}{4} + \frac{\sqrt{d}}{4}Z_d, s/2)\right] \\ = \; & \mathbb{E}\left[(1 + \frac{1}{\sqrt{d}}Z_d)^{s/2} S(\frac{d}{4} + \frac{\sqrt{d}}{4}Z_d, s/2)\right]. \end{aligned}$$

Using the Binomial expansion formula,

$$(1+x)^{s/2} = \sum_{k=0}^{\infty}\binom{s/2}{k}x^k \quad \text{for} \quad |x| < 1,$$

for $Z_d < \sqrt{d}$, we can write

$$\begin{aligned} & (1 + \frac{1}{\sqrt{d}}Z_d)^{s/2} S(\frac{d}{4} + \frac{\sqrt{d}}{4}Z_d, s/2) \\ = & \left[\sum_{k=0}^{\infty}\binom{s/2}{k}\frac{1}{(\sqrt{d})^k}Z_d^k\right]\left[\sum_{m=0}^{M} A_m(s/2)\left(\frac{2}{\frac{d}{2} + \frac{\sqrt{d}}{2}Z_d}\right)^m\right] \\ = & \left[\sum_{k=0}^{\infty}\binom{s/2}{k}\frac{1}{(\sqrt{d})^k}Z_d^k\right] \\ & \left[\sum_{m=0}^{M} A_m(s/2)\frac{4^m}{d^m}\left(\sum_{\ell=0}^{\infty}\frac{(-1)^l}{\sqrt{d}^\ell}Z_d^\ell\right)^m\right] \\ = & \left(1 + \binom{s/2}{1}\frac{1}{\sqrt{d}}Z_d + \binom{s/2}{2}\frac{1}{d}Z_d^2 + \dots\right) \\ & \cdot\left(1 + \binom{s/2}{2}\frac{4}{d}\left(\sum_{\ell=0}^{\infty}\frac{(-1)^l}{\sqrt{d}^\ell}Z_d^\ell\right) + \dots\right) \\ = & 1 + \binom{s/2}{2}\frac{4}{d} + \left[\binom{s/2}{1}\frac{1}{\sqrt{d}} + \binom{s/2}{2}\frac{4}{d\sqrt{d}}\right]Z_d \\ & + \left[\binom{s/2}{2}\frac{1}{d} + \binom{s/2}{2}\frac{4}{d^2}(\frac{s^2}{8} - \frac{3s}{4} + 1)\right]Z_d^2 \\ & + \dots \end{aligned} \tag{C.8}$$

where we used the identity $A_1(s/2) = \frac{\frac{s}{2}(\frac{s}{2}-1)}{2}$. Since $\mathbb{P}(Z_d \ge \sqrt{d}) = \mathcal{O}(e^{-d/2})$ and the function S is non-negative and bounded by 1 according to (C.3), we have

$$\begin{aligned} & \mathbb{E}\left[(1 + \frac{1}{\sqrt{d}}Z_d)^{s/2} S_M(\frac{d}{2} + \frac{\sqrt{d}}{2}Z_d, s)\right] \\ = & \mathcal{O}(e^{-\frac{d}{2}}) + \mathbb{E}\left[1 + \binom{s/2}{1}\frac{1}{\sqrt{d}}Z_d + \binom{s/2}{2}\frac{5}{d}Z_d^2 + \dots\right] \\ = & \mathcal{O}(e^{-d/2}) + 1 + \binom{s/2}{2}\frac{4}{d} + \binom{s/2}{2}\frac{1}{d} \\ & + \binom{s/2}{2}\frac{4}{d^2}(\frac{s^2}{8} - \frac{3s}{4} + 1) + o(\frac{1}{d^2}) \\ = & 1 + \binom{s/2}{2}\frac{5}{d} + o(\frac{1}{d}), \end{aligned} \tag{C.9}$$

where we used the fact that $\mathbb{E}(Z_d^k) \to \mathbb{E}(Z^k)$ as $d \to \infty$ for any fixed k implied by (C.7) where Z is a standard-normal variable in $\mathbb{R}$ which satisfies $\mathbb{E}(Z) = 0$ and $\mathbb{E}(Z^2) = 1$. Then, it follows from (C.4) that

$$\begin{aligned} \frac{(\frac{d}{2})^{s/2}}{I_0(s,d)} = & 1 - \binom{s/2}{2}\frac{5}{d} + o(\frac{1}{d}) \\ = & 1 + \frac{5s(2-s)}{8d} + o(\frac{1}{d}), \end{aligned} \tag{C.10}$$

which implies

$$\begin{aligned}\bar{\sigma}_0(s,d) =& \frac{1}{\sqrt[s]{I_0(s,d)}} \\ =& \frac{\sqrt{2}}{\sqrt{d}}\left(1+\frac{5s(2-s)}{8d}+o(\frac{1}{d})\right)^{1/s} \\ =& \frac{\sqrt{2}}{\sqrt{d}}\left(1+\frac{5(2-s)}{8d}+o(\frac{1}{d})\right).\end{aligned}$$

Similarly, taking square of both sides,

$$\bar{\sigma}_0^2(s,d) = \frac{1}{\sqrt[s/2]{I_0(s,d)}} = \frac{2}{d}\left(1+\frac{5(2-s)}{4d}+o(\frac{1}{d})\right)$$

which completes the proof.

□

D Probability of zero network output for ReLU activations

When X is a Gaussian random variable with distribution $\mathcal{N}(0,\sigma^2 I_d)$, we have $\mathbb{P}(\phi_0(X)=0) = \mathbb{P}(\max(X,0)=0) = \prod_{i=1}^n \mathbb{P}(X_i \le 0) = \frac{1}{2^d}$ due to the symmetry of the i-th component X_i with respect to the origin, independent of the choice of $\sigma>0$. The output of the k-th layer is actually not Gaussian, nevertheless exploiting its symmetry properties and piecewise linearity of the ReLU activations, the probability that the output $x^{(k)}$ will be zero can be computed with a similar calculation as follows and this probability is independent of the choice of σ.

Lemma 17. *Under Gaussian initialization **(A1)**–**(A2)** with ReLU activation, i.e. when $a=0$, for any $\sigma>0$ given, $\mathbb{P}(x^{(k)}=0) = 1-(1-\frac{1}{2^d})^k$.*

Proof. Consider the first layer

$$x^{(1)} = [x_1^{(1)}, x_2^{(1)}, \dots, x_d^{(1)}]^T,$$

where $x_i^{(1)} = \phi(\sum_{j=1}^d W_{ij}^{(1)} x_j^{(0)})$. According to the assumption, $W_{ij}^{(1)}$ is normally distributed with a zero mean. Then, $\sum_{j=1}^d W_{ij}^{(1)} x_j^{(0)}$ is also normally distributed with zero mean $\mathbb{P}(\sum_{j=1}^d W_{ij}^{(1)} x_j^{(0)} \ge 0) = \frac{1}{2}$. Therefore, $\mathbb{P}(x_i^{(1)} \neq 0) = 1-\frac{1}{2} = \frac{1}{2}$ and $\mathbb{P}(x^{(1)} \neq 0) = \frac{1}{2^d}$. If consider the k-th layer, we can get similarly

$$\mathbb{P}(x^{(k)} \neq 0 | x^{(k-1)} \neq 0) = 1 - \frac{1}{2^d}.$$

Since

$$\begin{aligned}\mathbb{P}(x^{(k)} \neq 0) =& \mathbb{P}(x^{(k)} \neq 0 | x^{(k-1)} \neq 0) \\ & \mathbb{P}(x^{(k-1)} \neq 0 | x^{(k-2)} \neq 0) \dots \mathbb{P}(x_i^{(1)} \neq 0) \\ =& (1-\frac{1}{2^d})^k,\end{aligned}$$

we can obtain the result

$$\mathbb{P}(x^{(k)}=0) = 1-\mathbb{P}(x^{(k)} \neq 0) = 1-(1-\frac{1}{2^d})^k.$$

□

E Proof of Theorem 5

Proof. By the same argument given in the proof of Theorem 1, for any $k \ge 0$, if $x^{(k)} \neq 0$, we have

$$\frac{\|x^{(k+1)}\|}{\|x^{(k)}\|} \sim \|\phi_0(z)\|, \tag{E.1}$$

where $z \sim \mathcal{N}(0, I_d)$ is a d-dimensional standard normal random vector, and in particular $\frac{\|x^{(k+1)}\|}{\|x^{(k)}\|}$ is independent from the choice of $x^{(k)}$ and the past history $x^{(j)}$ for $j<k$. Let A_k be the event that $x^{(k)} \neq 0$. We note that

$$A_k = \cap_{j=0}^k A_j, \tag{E.2}$$

that is $x^{(k)} \neq 0$ if and only if $x^{(j)} \neq 0$ for $j \le k$. This fact follows simply from the piecewise linear structure of the ReLU activation function. Conditioning on the event A_k, we can write

$$\frac{1}{k}\left(\log \frac{\|x^{(k)}\|}{\|x^{(0)}\|} | A_k\right) = \frac{1}{k}\sum_{j=0}^{k-1}\left(\frac{1}{2}\log \frac{\|x^{(j+1)}\|^2}{\|x^{(j)}\|^2} | A_j\right), \tag{E.3}$$

where[2] the logarithm is well-defined as the ratio $\frac{\|x^{(j+1)}\|}{\|x^{(j)}\|} > 0$ conditional on A_j. Due to (E.1) and (E.2), the right-hand side of (E.3) can be viewed as an average of i.i.d. random variables with mean

$$\begin{aligned}m_1 =& \frac{1}{2}\mathbb{E}\log\left(\frac{\|x^{(1)}\|^2}{\|x^{(0)}\|^2} | A_0\right) \\ =& \frac{1}{2}\mathbb{E}\log\left(\|\phi_0(\sigma z)\|^2 \,|\, z \notin \mathbb{R}_-^d\right),\end{aligned}$$

where $\mathbb{R}_-^d = \{x \in \mathbb{R}^d | x_i \le 0 \text{ for } i=1,2,\dots,d\}$ denotes the (closed) non-positive orthant of vectors and variance

$$\begin{aligned}m_2 =& \mathrm{var}\left(\frac{1}{2}\log\left(\|\phi_0(\sigma z)\|^2 \,|\, z \notin \mathbb{R}_-^d\right)\right) \\ =& \frac{1}{4}\mathrm{var}\left(\log\left(\|\phi_0(z)\|^2 \,|\, z \notin \mathbb{R}_-^d\right)\right).\end{aligned} \tag{E.4}$$

In the rest of the proof, we compute m_1 and m_2 explicitly showing them that they are finite; then by the central limit theorem and the law of large numbers, the theorem will hold with

$$\mu_0(\sigma) = m_1 \quad \text{and} \quad s_0^2 = m_2. \tag{E.5}$$

[2] Here, the equality is to be understood in the sense of distributions, i.e. the left-hand side and the right-hand side have the same distribution.

We note that

$$m_1 = \log(\sigma) + \mathbb{E}\left(\log\|\phi_0(z)\| \,\middle|\, z \notin \mathbb{R}^d_-\right) \quad \text{(E.6)}$$

$$= \log(\sigma) + \frac{1}{2}\mathbb{E}\left(\log\|\phi_0(z)\|^2 \,\middle|\, z \notin \mathbb{R}^d_-\right) \quad \text{(E.7)}$$

By (B.3) and following the same proof technique in Theorem 1, we can show that given that $z \notin \mathbb{R}^d_-$,

$$\left(\|\phi_0(z)\|^2 \,\middle|\, z \notin \mathbb{R}^d_-\right) \sim Y_n \quad \text{(E.8)}$$

with probability

$$\pi_d(n) = \frac{p_d(n)}{\sum_{n=1}^d p_d(n)} = \binom{d}{n}\frac{1}{2^d - 1},$$

for $n \geq 1$ where Y_n is a chi-square distribution with n degrees of freedom where $p_d(n)$ is given by (B.5). We have also

$$m_1 = \log(\sigma) + \mathbb{E}\log\left(\|\phi_0(z)\|^2 \,\middle|\, z \notin \mathbb{R}^d_-\right)$$
$$= \log(\sigma) + \sum_{n=1}^d \pi_d(n)\left[\mathbb{E}\log[Y_n]\right].$$

Using the mixture representation (E.8) and according to Lemma 23, we have

$$\operatorname{var}(\log\|\phi_0(z)\|^2 \,|\, z \notin \mathbb{R}^d_-)$$
$$= \sum_{n=0}^d \pi_d(n)\operatorname{var}(\log(Y_n)) + \sum_{n=0}^d \pi_d(n)(\mathbb{E}\log(Y_n))^2$$
$$- \left(\sum_{n=0}^d \pi_d(n)\mathbb{E}\log(Y_n)\right)^2.$$

Logarithmic moments of chi-square distributions are explicitly available as

$$\mathbb{E}\log[Y_n] = \log(2) + \Psi\left(\frac{n}{2}\right),$$

and

$$\operatorname{var}(\log(Y_n)) = \psi_1(n/2),$$

where $\psi_1(z)$ is the tri-gamma function (see (Cohen and Newman, 1984, Lemma 2.3)). Therefore, from (E.7), we obtain

$$m_1 = \log(\sigma) + \frac{1}{2}\sum_{n=1}^d \pi_d(n)\left[\log(2) + \Psi\left(\frac{n}{2}\right)\right].$$

Then, from (E.4) we get,

$$m_2 = \frac{1}{4}\operatorname{var}(\log\|\phi_0(z)\|^2 \,|\, z \notin \mathbb{R}^d_-)$$
$$= \frac{1}{4}\left(\sum_{n=1}^d \pi_d(n)\psi_1(n/2)\right)$$
$$+ \frac{1}{4}\left(\sum_{n=1}^d \pi_d(n)\left[\log(2) + \Psi\left(\frac{n}{2}\right)\right]^2\right)$$
$$- \frac{1}{4}\left(\sum_{n=1}^d \pi_d(n)\left[\log(2) + \Psi\left(\frac{n}{2}\right)\right]\right)^2.$$

We conclude from (E.5). □

F Proof of Theorem 7

Proof. The approach is similar to the proof of Theorem 1. We first note that for $x^{(k)} \neq 0$,

$$\frac{\|x^{(k+1)}\|}{\|x^{(k)}\|} = \frac{\|\phi_a(W^{(k+1)}x^{(k)})\|}{\|x^{(k)}\|}$$
$$= \left\|\phi_a\left(W^{(k+1)}\frac{x^{(k)}}{\|x^{(k)}\|}\right)\right\|,$$

which we used the equality $\phi_a(cy) = c\phi_a(y)$ for any given $c > 0$ and arbitrary vector y (with the choice of $y = W^{(k)}x^{(k)}$ and $c = 1/\|x^{(k)}\|$). By a similar reasoning to (B.1)–(B.2), we obtain

$$\mathbb{E}\left[\left(\frac{\|x^{(k)}\|}{\|x^{(0)}\|}\right)^s\right] = (\sigma^s\mathbb{E}\|\phi_a(z)\|^s)^k, \quad \text{(F.1)}$$

where z is a d-dimensional random vector with standard normal distribution $\mathcal{N}(0, I_d)$. In the rest of the proof, we compute the term $\mathbb{E}\|\phi_a(z)\|^s$ explicitly and establish that it is equal to $I_a(s,d)$ where $I_a(s,d)$ is given by (4.2). Consider

$$\mathbb{E}\|\phi_a(z)\|^s = \mathbb{E}\left[\phi_a^2(z_1) + \phi_a^2(z_2) + \cdots + \phi_a^2(z_d)\right]^{s/2},$$

where $z = (z_1, z_2, \ldots, z_d)$ is a d-dimensional standard normal random vector. We first note that the function $\phi_a(x) : \mathbb{R}^d \to \mathbb{R}$ has a piecewise linear structure on $\mathbb{R}^d$ depending on the sign of the components x_i of a vector x. In particular, we observe that by the definition of the ϕ_a function,

$$\|\phi_a(z)\|^2 = \phi_a^2(z_1) + \phi_a^2(z_2) + \cdots + \phi_a^2(z_d)$$
$$= \sum_{i:z_i>0}(z_i)^2 + a^2\sum_{i:z_i<0}(z_i)^2, \quad \text{(F.2)}$$

which depends on the orthant that the vector z resides in $\mathbb{R}^d$. In particular, there are 2^d (open) orthants in dimension d, where each orthant is defined by a system of inequalities:

$$\varepsilon_1 x_1 > 0, \quad \varepsilon_2 x_2 > 0, \quad \varepsilon_3 x_3 > 0, \quad \ldots \varepsilon_n x_n > 0,$$

where each ε_i is 1 or -1. Therefore, we can identify each orthant from an element of the set $\{+,-\}^d$. For example, the non-negative (open) orthant corresponds to $\{+,+,\ldots,+\}$ whereas the non-positive (open) orthant corresponds to $\{-,-,\ldots,-\}$. On every quadrant that corresponds to n plus signs and $d-n$ minus signs (with arbitrary order of the signs), the distribution of (F.2) is the same as the distribution of

$$X_n := Y_n + Z_n, \quad \text{(F.3)}$$

where

$$Y_n = \chi^2(n) \quad \text{and} \quad Z_n = a^2\chi^2(d-n).$$

$\chi^2(v)$ denotes a chi-squared distribution with v degrees of freedom. In this representation, Y_n and Z_n are independent as they are related to i.i.d. entries of the z vector. If we choose a random quadrant; with probability

$$\mathbb{P}(B_d = n) = p_d(n) = \binom{d}{n}\frac{1}{2^d} \tag{F.4}$$

we will be in such a quadrant where B_d is a Binomial random variable defined in (C.5). Therefore, we can write

$$\mathbb{E}\left\|\phi_a(z)\right\|^s = \sum_{n=0}^{d} p_d(n)\mathbb{E}(X_n^{s/2}). \tag{F.5}$$

In the special case $s = 2$, we have

$$\begin{aligned}
\mathbb{E}\left\|\phi_a(z)\right\|^2 &= \sum_{n=0}^{d} p_d(n)\mathbb{E}(X_n) \\
&= \sum_{n=0}^{d} p_d(n)(\mathbb{E}(Y_n) + \mathbb{E}(Z_n)) \\
&= \sum_{n=0}^{d} p_d(n)(n + a^2(d-n)) \\
&= (1+a^2)\frac{d}{2} \\
&= I_a(2,d),
\end{aligned}$$

where we used $\mathbb{E}(B_d) = \sum_{n=0}^{d} p_d(n)n = \frac{d}{2}$ and (F.1) implies directly that (4.1) holds for the $s = 2$ case. Next, we consider the case $s < 2$ where we compute $\mathbb{E}\left\|\phi_a(z)\right\|^s$ through moment generating function techniques. We will show that it is equal to $I_a(s,d)$ defined by (4.2).

Let $M_X(t) = \mathbb{E}(e^{tX})$ denote the moment generating function (MGF) of a random variable X. If we consider arbitrary moments $\alpha > 0$ (where α is not necessarily a positive integer) of a non-negative random variable X; we have

$$\mathbb{E}\left[X^\alpha\right] = D^\alpha M_X(0), \tag{F.6}$$

where D^α denotes the fractional derivative of order α in the Riemann-Louiville sense (Cressie and Borkent, 1986).[3] It is well-known that

$$M_{Y_n}(t) = \frac{1}{(1-2t)^{n/2}}, \quad M_{Z_n}(t) = \frac{1}{(1-2a^2t)^{(d-n)/2}},$$

(see e.g. (Bulmer, 1979)). By independence of Y_n and Z_n, we have also

$$\begin{aligned}
M_{X_n}(t) =& M_{Y_n}(t)M_{Z_n}(t) \\
=& \frac{1}{(1-2t)^{n/2}} \cdot \frac{1}{(1-2a^2t)^{(d-n)/2}}.
\end{aligned} \tag{F.7}$$

Hence,

$$\begin{aligned}
\frac{d}{dt}M_{X_n}(t) =& \frac{n}{(1-2t)^{\frac{n}{2}+1}} \cdot \frac{1}{(1-2a^2t)^{\frac{d-n}{2}}} \\
&+ \frac{1}{(1-2t)^{\frac{n}{2}}} \cdot \frac{a^2(d-n)}{(1-2a^2t)^{\frac{d-n}{2}+1}},
\end{aligned} \tag{F.8}$$

and by (Cressie and Borkent, 1986, eqn. (7)), for $\alpha \in (0,1)$, we have also

$$\begin{aligned}
D^\alpha M_{X_n}(0) &= \frac{1}{\Gamma(1-\alpha)}\int_{-\infty}^{0}(-z)^{-\alpha}\frac{dM_{X_n}(z)}{dz}dz \\
&= \frac{1}{\Gamma(1-\alpha)}\int_{0}^{\infty}(z)^{-\alpha}\frac{dM_{X_n}(-z)}{dz}dz.
\end{aligned} \tag{F.9}$$

Evaluating this integral requires computing integrals of the form

$$J_{m,\ell}(\alpha) = \int_0^\infty z^{-\alpha}\frac{1}{(1+2z)^{m/2}} \cdot \frac{1}{(1+2a^2z)^{\frac{\ell}{2}}}dz\,,$$

for integer values of m and ℓ satisfying $m + \ell = d + 2$. If we substitute $u = 1 - \frac{1}{2z+1}$, then $dz = \frac{1}{2(1-u)^2}du$ which leads to

$$\begin{aligned}
&J_{m,l}(\alpha) = \\
&\frac{1}{2^{-\alpha+1}}\int_0^1 u^{-\alpha}(1-u)^{\frac{m+\ell}{2}+\alpha-2}\left(1-(1-a^2)u\right)^{-\frac{\ell}{2}}du.
\end{aligned} \tag{F.10}$$

Using the binomial series

$$(1+x)^{-n} = \sum_{k=0}^{\infty}(-1)^k\binom{n+k-1}{k}x^k \tag{F.11}$$

for $|x| < 1$, we obtain

$$\begin{aligned}
J_{m,l}(\alpha) =& \frac{1}{2^{-\alpha+1}}\sum_{k=0}^{1}\binom{\frac{\ell}{2}+k-1}{k}(1-a^2)^k \\
&\int_0^1 u^{-\alpha+k}(1-u)^{\frac{m+\ell}{2}+\alpha-2}du \\
=& \frac{1}{2^{-\alpha+1}}\sum_{k=0}^{\infty}\binom{\frac{\ell}{2}+k-1}{k}(1-a^2)^k \\
&\mathrm{B}(k+1-\alpha, \frac{m+\ell}{2}+\alpha-1),
\end{aligned} \tag{F.12}$$

[3] In the special case when α is a positive integer, the fractional derivative reduces to the ordinary derivative and we obtain $\mathbb{E}\left[X^\alpha\right] = D^\alpha M_x(0) = \frac{d^\alpha M_X(t)}{dt^\alpha}|_{t=0}$.

where

$$\mathrm{B}(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}\,dt$$

is the Beta function. From (F.6), (F.8) and (F.9); we have

$$\begin{aligned}&\mathbb{E}(X_n^\alpha)\\ =&D^\alpha M_{X_n}(0)\\ =&\frac{1}{\Gamma(1-\alpha)}(nJ_{n+2,d-n}(\alpha)+a^2(d-n)J_{n,d-n+2}(\alpha)).\end{aligned}\tag{F.13}$$

From (F.5), choosing $\alpha=s/2$ for $s\in(0,2)$, we conclude that

$$\begin{aligned}&\mathbb{E}\left\|\phi_a(z)\right\|^s=\sum_{n=1}^d p_d(n)\mathbb{E}(X_n^{s/2})\\ =&\frac{1}{\Gamma(1-s/2)}\sum_{n=1}^d p_d(n)\\ &\left(nJ_{n+2,d-n}(\frac{s}{2})+a^2(d-n)J_{n,d-n+2}(\frac{s}{2})\right)\\ =&\frac{1}{2^{-s/2}}\frac{1}{\Gamma(1-s/2)}\sum_{n=1}^d p_d(n)\sum_{k=0}^\infty w_{k,n}\mathrm{B}(k+1-\frac{s}{2},\frac{d}{2}+\frac{s}{2})\\ =&I_a(d,s),\end{aligned}\tag{F.14}$$

where $I_a(s,d)$ is as in (4.2) and

$$\begin{aligned}w_{k,n}=&\frac{1}{2}(1-a^2)^k\\ &\left[\binom{\frac{d-n}{2}+k-1}{k}n+a^2(d-n)\binom{\frac{d-n}{2}+k}{k}\right].\end{aligned}\tag{F.15}$$

This proves (4.1). The proofs of remaining parts of the theorem follow with a similar reasoning to the proof of Theorem 1 and are omitted.

□

G Proof of Corollary 9

Proof. First, we consider the case of fixed a and s where we vary d. Note that, by definition $\bar\sigma_a(s,d)=\frac{1}{\sqrt[s]{I_a(s,d)}}$ where

$$I_a(s,d)=\mathbb{E}\left[\phi_a^2(z_1)+\phi_a^2(z_2)+\cdots+\phi_a^2(z_d)\right]^{s/2},$$

and z_i are i.i.d. standard normal random variables. Clearly,

$$\begin{aligned}I_a(s,d+1)=&\mathbb{E}\left[\phi_a^2(z_1)+\phi_a^2(z_2)+\cdots+\phi_a^2(z_{d+1})\right]^{s/2}\\ >&I_a(s,d).\end{aligned}$$

where the strict inequality stems from the fact that $\phi_a^2(z_{d+1})>0$ for $z_{d+1}>0$. Since $\bar\sigma_a(s,d)=\frac{1}{\sqrt[s]{I_a(s,d)}}$, we can conclude that $\bar\sigma_a(s,d+1)<\bar\sigma_a(s,d)$.

Secondly, we consider the case of fixed s and d and vary a. According to (F.2), for every $a\in[0,1]$ we have

$$\begin{aligned}I_a(d,s)=&\mathbb{E}\left[\sum_{i:z_i\geq 0}(z_i)^2+a^2\sum_{i:z_i<0}(z_i)^2\right]^{s/2}\\ \leq&\mathbb{E}\left[\sum_{i:z_i\geq 0}(z_i)^2+\sum_{i:z_i<0}(z_i)^2\right]^{s/2}\\ =&I_1(d,s)\end{aligned}$$

Differentiating the left hand side with respect to a, for $a>0$ we obtain

$$\frac{d}{da}I_a(d,s)\tag{G.1}$$

$$=\mathbb{E}\left[\frac{d}{da}\left(\sum_{i:z_i\geq 0}(z_i)^2+a^2\sum_{i:z_i<0}(z_i)^2\right)\right]^{s/2}\tag{G.2}$$

$$=\mathbb{E}\left[2a\sum_{i:z_i<0}(z_i)^2\right]^{s/2}>0,\tag{G.3}$$

where the interchangeability of the differentiation and expectation in (G.2) follows from the fact that both $I_a(d,s)$ and the expectation in (G.3) are finite. This proves that $I_a(d,s)$ is monotonically strictly increasing in a. Since $\bar\sigma_a(s,d)=\frac{1}{\sqrt[s]{I_a(s,d)}}$, this implies that $\bar\sigma_a(s,d)$ is (monotonically) strictly decreasing in a.

Finally, we consider fixed a and d and consider the monotonicity of $\bar\sigma_a(s,d)$ with respect to s for $s>0$. By the definition of $\bar\sigma_a(s,d)$, $\sigma=\bar\sigma_a(s,d)$ solves the implicit equation

$$F(s,\sigma)=\sigma^s I_a(s,d)=1\tag{G.4}$$

where $\bar\sigma_a(s,d)>0$ for $s>0$. Differentiating both sides with respect to s, by the chain rule,

$$\frac{dF}{d\sigma}(s,\bar\sigma_a(s,d))\frac{d\bar\sigma_a(s,d)}{ds}+\frac{dF}{ds}(s,\bar\sigma_a(s,d))=0\,,$$

where the derivatives exist as the function F is continuously differentiable in s and σ. This is equivalent to

$$\frac{s}{\bar\sigma_a(s,d)}\frac{d\bar\sigma_a(s,d)}{ds}+\frac{dF}{ds}(s,\bar\sigma_a(s,d))=0.\tag{G.5}$$

Note that we have also

$$F(s,\bar\sigma_a(s,d))=1.\tag{G.6}$$

For $\sigma=\bar\sigma_a(s,d))$, consider the function

$$\kappa(\tilde s):=F(\tilde s,\sigma)=\sigma^{\tilde s}I_a(\tilde s,d)=\mathbb{E}\|\phi_a(We_1)\|^{\tilde s}.$$

Clearly, $\kappa(\tilde s)$ is continuously differentiable with respect to $\tilde s$. It is also known that $\kappa(\tilde s)$ is a log-convex function

of $\tilde{s}$ for $\tilde{s} > 0$ (see e.g. (Buraczewski et al., 2014)), a fact which follows from the non-negativity of the second derivative of $\log \kappa(\tilde{s})$. Therefore, $\kappa(\tilde{s})$ is convex in $\tilde{s}$. If we consider the tangent line to the function $\kappa(\tilde{s})$ at $\tilde{s} = 0$ and $\tilde{s} = s$, by convexity of the function κ, we have

$$\kappa(\tilde{s}) \geq \kappa(s) + \kappa'(s)(\tilde{s} - s), \tag{G.7}$$
$$\kappa(\tilde{s}) \geq \kappa(0) + \kappa'(0)\tilde{s}, \tag{G.8}$$

for any $\tilde{s} \geq 0$ where $\kappa'(\tilde{s}) := \frac{d\kappa}{d\tilde{s}}(\tilde{s})$. Noticing that $\kappa(0) = \kappa(s) = 1$ and plugging in $\tilde{s} = 0$ in (G.7) and plugging in $\tilde{s} = s$ in (G.8), we obtain

$$1 \geq 1 - s\kappa'(s), \tag{G.9}$$
$$1 \geq 1 + \kappa'(0)s. \tag{G.10}$$

Since $s > 0$, we conclude that we have necessarily $\kappa'(0) \leq 0$ and $\kappa'(s) \geq 0$. Assume $\kappa'(s) = 0$. Then (G.7) would imply $\kappa(\tilde{s}) \geq \kappa(s) = 1$ for $\tilde{s} \geq 0$ and we would obtain $\kappa'(0) = 0$ and $\kappa(\tilde{s}) = 1$ for $\tilde{s} \in [0, s]$ which would be a contradiction. Therefore, we have necessarily

$$\kappa'(s) = \frac{dF}{ds}(s, \bar{\sigma}_a(s, d)) > 0.$$

Then, this implies that

$$\frac{d\bar{\sigma}_a(s,d)}{ds} = -\left(\frac{\bar{\sigma}_a(s,d)}{s}\right)\frac{dF}{ds}(s, \bar{\sigma}_a(s,d)) < 0\,, \tag{G.11}$$

for $s > 0$ and therefore $\bar{\sigma}_a(s, d)$ is a monotonically (strictly) decreasing function of s. □

H Proof of Corollary 10

Proof. For a linear activation function, we have $a = 1$. In this case, $w_{k,n} = d/2$ for $k = 0$ and $w_{k,n} = 0$ for $k > 0$. Then, it follows that

$$I_1(s,d) = 2^{s/2}\sum_{n=1}^{d} p_d(n)\frac{\Gamma(\frac{d}{2}+\frac{s}{2})}{\Gamma(\frac{d}{2})} = 2^{s/2}\frac{\Gamma(\frac{d}{2}+\frac{s}{2})}{\Gamma(\frac{d}{2})},$$

where we used $\mathrm{B}(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ and the fact that $\Gamma(\frac{d}{2}+1) = \frac{d}{2}\Gamma(\frac{d}{2})$. This yields $\bar{\sigma}_1(s,d) = \frac{1}{\sqrt{2}}\left(\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+\frac{s}{2})}\right)^{1/s}$. In the special case with $s = 2$, using the identity $\Gamma(\frac{d}{2}+1) = \frac{d}{2}\Gamma(\frac{d}{2})$ again, we obtain $\bar{\sigma}_1(2,d) = \frac{1}{\sqrt{d}}$ which recovers the results of LeCun et al. (1998b) for linear activations and is the basis for Lecun initialization.

The rest of the proof follows a similar approach to the proof of Corollary 4. From (C.2) and (C.1), we obtain

$$I_1(s,d) = 2^{s/2}(\frac{d}{2})^{s/2}\left(1 + \frac{\frac{s}{2}(\frac{s}{2}-1)}{d} + \mathcal{O}(\frac{1}{d^2})\right).$$

This implies that

$$\begin{aligned}\bar{\sigma}_1(s,d) &= \frac{1}{\sqrt[s]{I_1(s,d)}} \\ &= \frac{1}{\sqrt{d}}\left[1 + \frac{\frac{s}{2}(\frac{s}{2}-1)}{d} + \mathcal{O}(\frac{1}{d^2})\right]^{-1/s} \\ &= \frac{1}{\sqrt{d}} - \frac{(\frac{s}{2}-1)}{2d\sqrt{d}} + \mathcal{O}(\frac{1}{d\sqrt{d}}),\end{aligned}$$

where we used $(1+x)^s = 1 + sx + \mathcal{O}(x^2)$. Taking square of both sides, we obtain

$$\bar{\sigma}_1^2(s,d) = \frac{1}{\sqrt[s/2]{I_1(s,d)}} = \frac{1}{d} + \frac{2-s}{2d^2} + \mathcal{O}(\frac{1}{d^2\sqrt{d}}).$$

Next, we approximate $\bar{\sigma}_a^2(s,d)$ for $a > 0$ small. Following the notation in the proof of Theorem 7, from (F.13) we have,

$$\mathbb{E}(X_n^\alpha) = \frac{1}{\Gamma(1-\alpha)}(nJ_{n+2,d-n}(\alpha) + a^2(d-n)J_{n,d-n+2}(\alpha)). \tag{H.1}$$

For $m + \ell = d + 2$, from (F.10), we have

$$\begin{aligned}&J_{m,\ell}(\alpha) \\ =&\frac{1}{2^{-\alpha+1}}\int_0^1 u^{-\alpha}(1-u)^{m/2+\alpha-2}\big(1+\frac{a^2u}{1-u}\big)^{-\ell/2}du \\ =&\frac{1}{2^{-\alpha+1}}\int_0^1 u^{-\alpha}(1-u)^{m/2+\alpha-2} \\ &\big(1 - \frac{\ell}{2}\frac{a^2u}{1-u} + \frac{\frac{\ell}{2}(\frac{\ell}{2}+1)}{2}\frac{a^4u^2}{(1-u)^2} + \mathcal{O}(a^6)\big)du \\ =&J_{m,\ell}|_{a=0} - \frac{a^2}{2^{-\alpha+1}}\frac{\ell}{2}B(2-\alpha, \frac{m}{2}+\alpha-2) \\ &+ \frac{\ell(\ell+2)}{2^{-\alpha+4}}a^4B(3-\alpha, \frac{m}{2}+\alpha-3) + \mathcal{O}(a^6),\end{aligned} \tag{H.2}$$

where we used the Binomial formula and (F.11). Plugging $a = 0$ in (H.2),

$$\begin{aligned}J_{m,\ell}|_{a=0} =&\frac{1}{2^{-\alpha+1}}\int_0^1 u^{-\alpha}(1-u)^{m/2+\alpha-2}du \\ =&\frac{1}{2^{-\alpha+1}}B(1-\alpha, m/2+\alpha-1).\end{aligned}$$

Therefore, from (H.1),

$$\begin{aligned}
&\mathbb{E}(X_n^\alpha)\\
=&\mathbb{E}(X_n^\alpha)|_{a=0} + \frac{1}{\Gamma(1-\alpha)}\mathbb{E}\big[\\
&-\frac{a^2(d-n)n}{2^{-\alpha+2}}B(2-\alpha,\frac{n}{2}+\alpha-1)\\
&+\frac{n(d-n)(d-n+2)a^4}{2^{-\alpha+4}}B(3-\alpha,\frac{n}{2}+\alpha-2)\\
&+\frac{a^2(d-n)}{2^{-\alpha+1}}B(1-\alpha,\frac{n}{2}+\alpha-1)\\
&-\frac{a^4(d-n)(d-n+2)}{2^{-\alpha+2}}B(2-\alpha,\frac{n}{2}+\alpha-2)+\mathcal{O}(a^6)\big]\\
=&\mathbb{E}(X_n^\alpha)|_{a=0} + \mathbb{E}\left[\frac{\alpha a^2(d-n)}{2^{-\alpha+1}\Gamma(1-\alpha)}B(1-\alpha,\frac{n}{2}+\alpha-1)\right]\\
&-\mathbb{E}\left[\frac{(d-n)(d-n+2)a^4\alpha}{2^{-\alpha+3}\Gamma(1-\alpha)}B(2-\alpha,\frac{n}{2}+\alpha-2)\right]\\
&+\mathcal{O}(a^6),
\end{aligned} \tag{H.3}$$

where we used the identities $\mathrm{B}(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y)$ and $\Gamma(x+1)=x\Gamma(x)$ for $x,y>0$. We denote

$$\begin{aligned}
T_1(n,\alpha) &:= \frac{\alpha a^2(d-n)}{2^{-\alpha+1}\Gamma(1-\alpha)}B(1-\alpha,\frac{n}{2}+\alpha-1),\\
T_2(n,\alpha) &:= \frac{(d-n)(d-n+2)a^4\alpha}{2^{-\alpha+3}\Gamma(1-\alpha)}B(2-\alpha,\frac{n}{2}+\alpha-2).
\end{aligned}$$

We notice from (F.14) that

$$I_a(s,d) = \mathbb{E}\left((X_{B_d})^{s/2}\right) = \mathbb{E}\left((X_{B_d})^{s/2}\right), \tag{H.4}$$

where B_d follows a binomial distribution with $\mathbb{P}(B_d=n)=p_d(n)$. From (H.3), it follows that

$$\begin{aligned}
I_a(s,d) =& I_0(s,d) + \mathbb{E}[T_1(B_d,s/2)] - \mathbb{E}[T_2(B_d,s/2)]\\
&+\mathcal{O}(a^6 d^{s/2})
\end{aligned}$$

Recall that from (C.7) we have

$$Z_d = \frac{B_d-\mathbb{E}(B_d)}{\sqrt{\mathrm{var}B_d}} = \frac{B_d-\frac{d}{2}}{\sqrt{d}/2} \longrightarrow \mathcal{N}(0,I).$$

Similar to (C.8), we consider

$$H(\frac{s}{2}) := (1+\frac{1}{\sqrt{d}}Z_d)^{s/2}S(\frac{d}{4}+\frac{\sqrt{d}}{4}Z_d, s/2),$$

which admits the expansion

$$\begin{aligned}
\mathbb{E}[H(\frac{s}{2})] =& \mathcal{O}(e^{-d/2}) + 1 + \binom{s/2}{2}\frac{4}{d} + \binom{s/2}{2}\frac{1}{d}\\
&+\binom{s/2}{2}\frac{4}{d^2}(\frac{s^2}{8}-\frac{3s}{4}+1)+o(\frac{1}{d^2}).
\end{aligned}$$

If we let $\alpha=\frac{s}{2}$, we also have

$$\begin{aligned}
T_1(B_d,\frac{s}{2}) &= \frac{a^2 s(\frac{d}{2}-\frac{\sqrt{d}}{2}Z_d)}{2^{-\frac{s}{2}+2}}\frac{\Gamma(\frac{d}{4}+\frac{\sqrt{d}}{4}Z_d+\frac{s}{2}-1)}{\Gamma(\frac{d}{4}+\frac{\sqrt{d}}{4}Z_d)}\\
&= \frac{a^2 s(\frac{d}{2}-\frac{\sqrt{d}}{2}Z_d)}{2^{-\frac{s}{2}+2}}\left(\frac{d}{4}\right)^{\frac{s}{2}-1}H(\frac{s}{2}-1)\\
&= \left(\frac{d}{2}\right)^{\frac{s}{2}}\frac{a^2 s(1-\frac{1}{\sqrt{d}}Z_d)}{2}H(\frac{s}{2}-1).
\end{aligned}$$

According to (C.9), we have

$$\begin{aligned}
\mathbb{E}[T_1(B_d,\frac{s}{2})] = &\left(\frac{d}{2}\right)^{\frac{s}{2}}\frac{a^2 s}{2}\\
&\left[1+(\frac{5}{8}s-3)(s-2)\frac{1}{d}+o(\frac{1}{d})\right].
\end{aligned}$$

Similarly, we can write

$$\begin{aligned}
T_2(B_d,\frac{s}{2}) = &\left(\frac{d}{2}\right)^{\frac{s}{2}}\frac{a^4(2-s)s}{2}H(\frac{s}{2}-2)\\
&\left(\frac{1}{4}-\frac{1}{2\sqrt{d}}Z_d-\frac{1}{d\sqrt{d}}Z_d+\frac{1}{d}+\frac{1}{4d^2}Z_d^2\right),
\end{aligned}$$

and we have

$$\begin{aligned}
\mathbb{E}[T_2(B_d,\frac{s}{2})] = &\left(\frac{d}{2}\right)^{\frac{s}{2}}\frac{a^4(2-s)s}{2}\\
&\left[\frac{1}{4}+(\frac{5}{32}s^2-\frac{33}{16}s+7)\frac{1}{d}+o(\frac{1}{d})\right].
\end{aligned}$$

Therefore, we can calculate

$$\begin{aligned}
&I_a(s,d)\\
=&I_0(s,d)+\mathbb{E}[T_1(B_d,\frac{s}{2})]-\mathbb{E}[T_2(B_d,\frac{s}{2})]+O(a^6 d^{s/2})\\
=&\left(\frac{d}{2}\right)^{\frac{s}{2}}K\Big[1+\frac{1}{K}(s-2)\Big[\frac{5s}{8}+\frac{a^2 s}{2}(\frac{5}{8}s-3)\\
&+\frac{a^4 s}{2}(\frac{5}{32}s^2-\frac{33}{16}s+7)\Big]\frac{1}{d}+O(a^6)+o(\frac{1}{d})\Big],
\end{aligned}$$

where K is defined as

$$K := 1+\frac{a^2 s}{2}+\frac{a^4 s(s-2)}{8}.$$

Then we obtain

$$\begin{aligned}
&\bar{\sigma}_a^2(s,d) = I_a^{-2/s}(s,d)\\
=&\frac{2}{d}K^{-2/s}\Big[1+\frac{1}{K}(s-2)\Big[\frac{5s}{8}+\frac{a^2 s}{2}(\frac{5}{8}s-3)\\
&+\frac{a^4 s}{2}(\frac{5}{32}s^2-\frac{33}{16}s+7)\Big]\frac{1}{d}+\mathcal{O}(a^6)+o(\frac{1}{d})\Big]^{-2/s}\\
=&\frac{2}{d}K^{-2/s}\Big[1+\frac{1}{K}(2-s)\Big[\frac{5}{4}+a^2(\frac{5}{8}s-3)\\
&+a^4(\frac{5}{32}s^2-\frac{33}{16}s+7)\Big]\frac{1}{d}+\mathcal{O}(a^6)+o(\frac{1}{d})\Big]
\end{aligned}$$

If a is small, we can write

$$\begin{aligned}&\bar{\sigma}_a^2(s,d)\\ =&\frac{2}{d}\left(1+\frac{a^2 s}{2}+\mathcal{O}(a^4)\right)^{-2/s}\\ &\left[1+\frac{2}{2+a^2 s}(2-s)(\frac{5}{4}+a^2(\frac{5}{8}s-3))\frac{1}{d}+\frac{\mathcal{O}(a^4)}{d}\right]\\ =&\frac{2}{1+a^2}\frac{1}{d}+(2-s)\frac{(5s-24)a^2+10}{2(s+2)a^2+4}\frac{1}{d^2}\\ &+\mathcal{O}(\frac{a^4}{d})+o(\frac{1}{d^2})\end{aligned}$$

which is equivalent to the claimed result for $\bar{\sigma}_a^2(s,d)$. This completes the proof. □

I Proof of Theorem 11

Proof. The proof follows by a similar reasoning to the proof of Theorem 5. The same proof technique applies where we can show that the theorem holds with constants

$$\begin{aligned}\mu_a(\sigma) =&\frac{1}{2}\mathbb{E}\log\|\phi_a(\sigma z)\|^2\\ =&\log(\sigma)+\frac{1}{2}\mathbb{E}\left(\log\|\phi_a(z)\|^2\right),\end{aligned}\tag{I.1}$$

and

$$s_a^2=\frac{1}{4}\mathrm{var}\left(\log\left(\|\phi_a(z)\|^2\right)\right).\tag{I.2}$$

We also recall from (F.2)–(F.4) that

$$\|\phi_a(z)\|^2\sim X_n\quad\text{with probability}\quad p_d(n),\tag{I.3}$$

for $n\geq 1$ where X_n is defined by (F.3) and $p_d(n)$ is defined by (F.4). In the rest of the proof we compute $\mathbb{E}(\log(X_n))$ and $\mathrm{var}(\log(X_n))$ for every $n\geq 1$ and then use the identities (I.1), (I.2) and (I.3) to obtain an explicit formula for $\mu_a(\sigma)$ and s_a^2.

Note that X_n is non-negative, and we have

$$\mathbb{E}(\log(X_n))=\frac{d}{d\alpha}\mathbb{E}(X_n^\alpha)|_{a=0},$$

and

$$\begin{aligned}\mathrm{var}(\log(X_n)) =&\frac{d^2}{d\alpha^2}\left(\log\mathbb{E}(X_n^\alpha)\right)|_{a=0}\\ =&\frac{d}{d\alpha}\left(\frac{\frac{d}{d\alpha}\mathbb{E}(X_n^\alpha)}{\mathbb{E}(X_n^\alpha)}\right)|_{a=0},\end{aligned}$$

provided that the expectations are finite (see e.g. (Cohen and Newman, 1984)). For computing these expectations, we calculate

$$\begin{aligned}\frac{d}{d\alpha}\mathbb{E}(X_n^\alpha) =&\frac{d}{d\alpha}\Bigg(\frac{1}{\Gamma(1-\alpha)}(nJ_{n+2,d-n}(\alpha)\\ &+a^2(d-n)J_{n,d-n+2}(\alpha))\Bigg).\end{aligned}\tag{I.4}$$

By the product rule for derivatives, for an integer $m>0$,

$$\begin{aligned}&\frac{d}{d\alpha}J_{m,d+2-m}(\alpha)=\log(2)J_{m,d+2-m}(\alpha)\\ &+\frac{1}{2^{-\alpha+1}}\sum_{k=0}^{\infty}\binom{\frac{d-m}{2}+k}{k}(1-a^2)^k\frac{d}{d\alpha}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha)\end{aligned}\tag{I.5}$$

We also have

$$\begin{aligned}\frac{d}{d\alpha}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha) =&\frac{d}{d\alpha}\frac{\Gamma(k+1-\alpha)\Gamma(\frac{d}{2}+\alpha)}{\Gamma(\frac{d}{2}+k+1)}\\ =&b_k\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha),\end{aligned}$$

where

$$b_{k,\alpha}=\psi_0(\frac{d}{2}+\alpha)-\psi_0(k+1-\alpha),$$

and we used the fact $\mathrm{B}(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y)$ for real scalars $x,y>0$. Inserting this formula into (I.5),

$$\begin{aligned}&\frac{d}{d\alpha}J_{m,d+2-m}(\alpha)\\ =&\log(2)J_{m,\ell}(\alpha)+\frac{1}{2^{-\alpha+1}}\sum_{k=0}^{\infty}\binom{\frac{d-m}{2}+k}{k}(1-a^2)^k\\ &b_{k,\alpha}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha).\end{aligned}$$

From (I.4), we also get

$$\begin{aligned}&\frac{d}{d\alpha}\mathbb{E}(X_n^\alpha)\\ =&\frac{1}{\Gamma(1-\alpha)}\frac{d}{d\alpha}(nJ_{n+2,d-n}(\alpha)+a^2(d-n)J_{n,d-n+2}(\alpha))\\ &+\frac{\Gamma'(1-\alpha)}{\Gamma^2(1-\alpha)}(nJ_{n+2,d-n}(\alpha)+a^2(d-n)J_{n,d-n+2}(\alpha))\\ =&\frac{1}{\Gamma(1-\alpha)}\frac{d}{d\alpha}(nJ_{n+2,d-n}(\alpha)+a^2(d-n)J_{n,d-n+2}(\alpha))\\ &+\frac{\psi_0(1-\alpha)}{\Gamma(1-\alpha)}(nJ_{n+2,d-n}(\alpha)+a^2(d-n)J_{n,d-n+2}(\alpha))\\ =&\frac{1}{\Gamma(1-\alpha)}\frac{1}{2^{-\alpha}}\sum_{k=0}^{\infty}w_{k,n}b_{k,\alpha}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha)\\ &+[\log(2)+\psi_0(1-\alpha)]\,\mathbb{E}(X_n^\alpha),\end{aligned}\tag{I.6}$$

where we used (F.12), (F.13) and $w_{k,n}$ is defined by (F.15). Therefore,

$$\begin{aligned}\frac{\frac{d}{d\alpha}\mathbb{E}(X_n^\alpha)}{\mathbb{E}(X_n^\alpha)} =&\frac{\sum_{k=0}^{\infty}w_{k,n}b_{k,\alpha}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha)}{\sum_{k=0}^{\infty}w_{k,n}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha)}\\ &+\log(2)+\psi_0(1-\alpha),\end{aligned}$$

where we used (F.13) again. Differentiating with respect to α, we find

$$\begin{aligned}&\frac{d}{d\alpha}\left(\frac{\frac{d}{d\alpha}\mathbb{E}(X_n^\alpha)}{\mathbb{E}(X_n^\alpha)}\right)\\ =&\frac{\sum_{k=0}^\infty w_{k,n}(b_{k,\alpha}^2+\frac{d}{d\alpha}b_{k,\alpha})\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha)}{\sum_{k=0}^\infty w_{k,n}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha)}\\ &-\left(\frac{\sum_{k=0}^\infty w_{k,n}b_{k,\alpha}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha)}{\sum_{k=0}^\infty w_{k,n}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha)}\right)^2\\ &-\psi_1(1-\alpha),\end{aligned} \tag{I.7}$$

where

$$\frac{d}{d\alpha}b_{k,\alpha}=\psi_1(\frac{d}{2}+\alpha)+\psi_1(k+1-\alpha).$$

Evaluating the expression (I.6) at $\alpha=0$, we find

$$\begin{aligned}m_n:=&\mathbb{E}(\log(X_n))=\frac{d}{d\alpha}\mathbb{E}(X_n^\alpha)|_{\alpha=0}\\ =&\sum_{k=0}^\infty w_{k,n}b_{k,0}\mathrm{B}(k+1,\frac{d}{2})+[\log(2)-\gamma]\\ =&\sum_{k=0}^\infty w_{k,n}\left(\psi_0(\frac{d}{2})-\psi_0(k+1)\right)\mathrm{B}(k+1,\frac{d}{2})\\ &+[\log(2)-\gamma]\end{aligned} \tag{I.8}$$

In the last two steps, we used the fact that $\psi_0(1)=\gamma$ where γ is the Euler–Mascheroni constant. Similarly,

$$\begin{aligned}v_n:=&\mathrm{var}(\log(X_n))\\ =&\frac{d^2}{d\alpha^2}\mathbb{E}(X_n^\alpha)|_{\alpha=0}\\ =&\frac{\sum_{k=0}^\infty w_{k,n}(b_{k,0}^2+\frac{d}{d\alpha}b_{k,0})\mathrm{B}(k+1,\frac{d}{2})}{\sum_{k=0}^\infty w_{k,n}\mathrm{B}(k+1,\frac{d}{2})}\\ &-\left(\frac{\sum_{k=0}^\infty w_{k,n}b_{k,0}\mathrm{B}(k+1,\frac{d}{2})}{\sum_{k=0}^\infty w_{k,n}\mathrm{B}(k+1,\frac{d}{2})}\right)^2-\psi_1(1).\end{aligned} \tag{I.9}$$

On the other hand, by (F.13) and (F.12), we have

$$\mathbb{E}(X_n^\alpha)=\frac{1}{\Gamma(1-\alpha)}\frac{1}{2^{-\alpha}}\sum_{k=0}^\infty w_{k,n}\mathrm{B}(k+1-\alpha,\frac{d}{2}+\alpha). \tag{I.10}$$

We note that $X_n^\alpha\le S_n:=1+X_n$ for $\alpha\in[0,1]$ where $\mathbb{E}(S_n)<\infty$. Therefore, by the dominated convergence theorem we have

$$\lim_{\alpha\to0}\mathbb{E}(X_n^\alpha)=\mathbb{E}(X_n^0)=1.$$

Taking limits in (I.10) as $\alpha\to0$,

$$1=\lim_{\alpha\to0}\mathbb{E}(X_n^\alpha)=\frac{1}{\Gamma(1)}\sum_{k=0}^\infty w_{k,n}\mathrm{B}(k+1,\frac{d}{2}).$$

Since $\Gamma(1)=1$, this is equivalent to

$$\sum_{k=0}^\infty w_{k,n}\mathrm{B}(k+1,\frac{d}{2})=1\quad\text{for every}\quad n\ge1.$$

Plugging this identity into (I.9),

$$\begin{aligned}v_n=&\sum_{k=0}^\infty w_{k,n}(b_{k,0}^2+\frac{d}{d\alpha}b_{k,0})\mathrm{B}(k+1,\frac{d}{2})\\ &-\left(\sum_{k=0}^\infty w_{k,n}b_{k,0}\mathrm{B}(k+1,\frac{d}{2})\right)^2-\psi_1(1)\\ =&\psi_1(\frac{d}{2})+\sum_{k=0}^\infty\left[\psi_1(k+1)-\psi_1(1)\right]w_{k,n}\mathrm{B}(k+1,\frac{d}{2})\\ &+\sum_{k=0}^\infty\left[\psi_0(\frac{d}{2})-\psi_0(k+1)\right]^2w_{k,n}\mathrm{B}(k+1,\frac{d}{2})\\ &-\left[\sum_{k=0}^\infty\left(\psi_0(\frac{d}{2})-\psi_0(k+1)\right)w_{k,n}\mathrm{B}(k+1,\frac{d}{2})\right]^2.\end{aligned}$$

We conclude that

$$\mu_a(\sigma)=\log(\sigma)+\frac{1}{2}\mathbb{E}\log X_n=\log(\sigma)+\frac{1}{2}\sum_{n=0}^d p_d(n)m_n \tag{I.11}$$

and

$$\begin{aligned}s_a^2=&\frac{1}{4}\left[\sum_{n=0}^d p_d(n)v_n+\sum_{n=0}^d p_d(n)(m_n)^2\right.\\ &\left.-\left(\sum_{n=0}^d p_d(n)m_n\right)^2\right]\end{aligned} \tag{I.12}$$

where $p_d(n)$ is defined by (B.5). This completes the proof. □

Remark 18. ***(First-order stochastic dominance property compared to Kaiming's method)*** *Figure 4 illustrates Theorem 11, showing the pdf and cdf of $R_{k,a}$ for linear activations ($a=1$) and Leaky ReLU activations with $a=0.01$ after $k=100$ layers with two choices of σ according to Kaiming initialization and our initialization technique which preserves approximately the fractional moment of order $s=1$. We observe that the distribution of $R_{k,a}$ is similar to a Gaussian distribution, and with our initialization, the network output $R_{k,a}$ possesses a first-order stochastic dominance property in the sense of Hadar and Russell (1969) (Remark 6). This dominance property will hold for large enough k, as our initialization can choose a larger σ and hence results in a larger mean value $\mu_a(\sigma)$ in the setting of Theorem 7 and as the results also admit non-asymptotic versions according to Remark 12.*

J Extensions of results to dropout

In this section, we consider extensions of our results reported in the main text to dropout which is a mechanism where some neurons are removed randomly to prevent overfitting (see Remark 13 in the main text for more details).

J.1 ReLU activation with dropout

Theorem 19. ***(Explicit characterization of the critical variance*** $\sigma_0^2(s,d)$ ***with dropout)*** *Consider a fully connected network with an input* $x^{(0)} \in \mathbb{R}^d$ *and Gaussian initialization satisfying* ***(A1)-(A2)*** *with ReLu activation function* $\phi_0(x) = \max(x,0)$ *with dropout where the probability to keep the neurons is given by* $q \in (0,1]$*. Let* $s > 0$ *be a given real scalar. The s-th moment of the output of the k-th layer is given by*

$$\mathbb{E}\left[\|x^{(k)}\|^s\right] = \|x^{(0)}\|^s(\sigma^s I_{0,q}(s,d))^k,$$
$$I_{0,q}(s,d) = \frac{1}{q^s}2^{s/2}\sum_{n=0}^{d} q_d(n)\frac{\Gamma(n/2+s/2)}{\Gamma(n/2)}, \tag{J.1}$$

where

$$q_d(n) = \binom{d}{n}(\frac{q}{2})^n(1-\frac{q}{2})^{d-n}, \tag{J.2}$$

and Γ *is the Gamma function. Then, it follows that we have three possible cases:*

(i) If $\sigma = \bar{\sigma}_{0,q}(s,d)$ *where* $\bar{\sigma}_{0,q}(s,d) := \frac{1}{\sqrt[s]{I_{0,q}(s,d)}}$*, then the network preserves the s-th moment of the layer outputs, i.e. for every* $k \geq 1$*,* $\mathbb{E}\left[\|x^{(k)}\|^s\right] = \|x^{(0)}\|^s$*, whereas for any* $p > s$*,* $\mathbb{E}\|x^{(k)}\|^p \to \infty$ *exponentially fast in k.*

(ii) If $\sigma < \bar{\sigma}_{0,q}(s,d)$*, then* $\mathbb{E}\left[\|x^{(k)}\|^s\right] \to 0$ *exponentially fast in k.*

(iii) If $\sigma > \bar{\sigma}_{0,q}(s,d)$*, then* $\mathbb{E}\left[\|x^{(k)}\|^s\right] \to \infty$ *exponentially fast in k.*

Proof. The proof follows from minor adaptations to the proof of Theorem 1. In the proof of Theorem 1, it suffices to replace $p_d(n)$ with

$$q_d(n) := \binom{d}{n}(\frac{q}{2})^n(1-\frac{q}{2})^{d-n}$$

and X_n with X_n/q^2 as the effect of dropout is to scale the network output and change the mixing probabilities of the chi-square distributions arising in the proof of Theorem 1. This yields

$$\mathbb{E}\left[\|x^{(k)}\|^s\right] = \|x^{(0)}\|^s(\sigma^s I_{0,q}(s,d))^k,$$
$$I_{0,q}(s,d) = \frac{1}{q^s}2^{s/2}\sum_{n=0}^{d} q_d(n)\frac{\Gamma(n/2+s/2)}{\Gamma(n/2)}. \tag{J.3}$$

The proofs of remaining parts follow from a reasoning similar to the proof of Theorem 1 and are omitted. □

Corollary 20. ***(Critical variance*** $\bar{\sigma}_{0,q}(d,s)$ ***when*** d ***is large with dropout)*** *For fixed width* d *and* $s \in (0,2]$*, we have*

$$\bar{\sigma}^2_{0,q}(s,d) = \frac{2q}{d} + \frac{(2-s)(6-q)}{2d^2} + o(\frac{1}{d^2}),$$

Therefore, it follows from Theorem 1 that if $\sigma^2 = \frac{2q}{d} + \frac{(2-s)(6-q)}{2d^2}$*, then the network will preserve the moment of order* $s + o(\frac{1}{d})$ *of the network output if dropout is used.*

Proof. The proof follows from minor modifications to the proof of Corollary 4. Following the proof technique of Corollary 4, we can write

$$\frac{(\frac{d}{2})^{s/2}}{I_{0,q}(s,d)} = \frac{q^s}{\mathbb{E}(F_d(B_{d,q}))}, \tag{J.4}$$

where $B_{d,q}$ is a Binomial random variable, i.e.

$$\mathbb{P}(B_{d,q} = n) = q_d(n) = \binom{d}{n}(\frac{q}{2})^n(1-\frac{q}{2})^{d-n} \tag{J.5}$$

for $n = 0, 1, \ldots, d$, where F_d is defined by (C.6). By the normal approximation to binomial distribution, we have

$$Z_{d,q} := \frac{B_d - \mathbb{E}(B_d)}{\sqrt{\mathrm{var}B_d}} = \frac{B_d - \frac{dq}{2}}{\frac{\sqrt{d}}{2}\sqrt{2q-q^2}} \longrightarrow \mathcal{N}(0,I) \tag{J.6}$$

which is similar to (C.7). Then, we follow similar computations to the proof of Corollary 4:

$$\begin{aligned}
&\mathbb{E}(F_d(B_{d,q}))\\
=&\mathbb{E}\left(F_d(\frac{dq}{2} + \frac{\sqrt{d}}{2}\sqrt{2q-q^2}Z_d)\right)\\
=&2^{s/2}\mathbb{E}\left[\frac{(\frac{dq}{2} + \frac{\sqrt{d}}{2}\sqrt{2q-q^2}Z_d)^{s/2}}{d^{s/2}}\right.\\
&\left.S(\frac{dq}{2} + \frac{\sqrt{d}}{2}\sqrt{2q-q^2}Z_d, s/2)\right]\\
=&q^{s/2}\mathbb{E}\left[(1 + \frac{1}{\sqrt{d}}\frac{\sqrt{2-q}}{\sqrt{q}}Z_d)^{s/2}\right.\\
&\left.S(\frac{dq}{2} + \frac{\sqrt{d}}{2}\sqrt{2q-q^2}Z_d, s/2)\right].
\end{aligned}$$

Using the Binomial expansion,

$$(1+x)^{s/2} = \sum_{k=0}^{\infty} \binom{s/2}{k} x^k \quad \text{for} \quad |x| < 1.$$

Therefore, for $Z_{d,q} < \sqrt{d}\left(\frac{\sqrt{q}}{\sqrt{2-q}}\right)$, we can write

$$\begin{aligned}
&(1+\frac{1}{\sqrt{d}}\frac{\sqrt{2-q}}{\sqrt{q}}Z_{d,q})^{s/2} S(\frac{dq}{2}+\frac{\sqrt{d}}{2}\sqrt{2q-q^2}Z_{d,q}, s/2) \\
=&\left[\sum_{k=0}^{\infty}\binom{s/2}{k}\frac{\sqrt{2-q}^k}{(\sqrt{dq})^k}Z_{d,q}^k\right]\left[\sum_{m=0}^{M} A_m(s/2) \right. \\
&\left.\left(\frac{2}{\frac{dq}{2}+\frac{\sqrt{d}}{2}\sqrt{2q-q^2}Z_{d,q}}\right)^m\right] \\
=&\left[\sum_{k=0}^{\infty}\binom{s/2}{k}\frac{\sqrt{2-q}^k}{(\sqrt{dq})^k}Z_{d,q}^k\right]\left[\sum_{m=0}^{M} A_m(s/2)\frac{2^m}{(dq)^m} \right. \\
&\left.\left(\frac{2}{1+\frac{1}{\sqrt{d}}\frac{\sqrt{2-q}}{\sqrt{q}}Z_{d,q}}\right)^m\right] \\
=&\left[\sum_{k=0}^{\infty}\binom{s/2}{k}\frac{\sqrt{2-q}^k}{(\sqrt{dq})^k}Z_{d,q}^k\right]\left[\sum_{m=0}^{M} A_m(s/2)\frac{4^m}{d^m q^m} \right. \\
&\left.\left(\sum_{\ell=0}^{\infty}\frac{1}{\sqrt{d}^\ell}\frac{\sqrt{2-q}^\ell}{\sqrt{q}^\ell}Z_{d,q}^\ell\right)^m\right] \\
=&\left(1+\binom{s/2}{1}\frac{\sqrt{2-q}}{\sqrt{dq}}Z_{d,q}+\binom{s/2}{2}\frac{2-q}{dq}Z_{d,q}^2+\dots\right) \\
&\cdot\left(1+\binom{s/2}{2}\frac{4}{dq}\left(\sum_{\ell=0}^{\infty}\frac{1}{\sqrt{d}^\ell}\frac{\sqrt{2-q}^\ell}{\sqrt{q}^\ell}Z_{d,q}^\ell\right)+\dots\right) \\
=&1+\binom{s/2}{1}\frac{1}{\sqrt{d}}Z_{d,q}+\binom{s/2}{2}\frac{6-q}{dq}Z_d^2+\dots,
\end{aligned}$$

where we used the identity $A_1(s/2) = \binom{s/2}{2} = \frac{\frac{s}{2}(\frac{s}{2}-1)}{2}$. Since $\mathbb{P}(Z_{d,q} \geq \sqrt{d}) = \mathcal{O}(e^{-d/2})$ and the function S is non-negative and bounded by 1 according to (C.3), we have

$$\begin{aligned}
&\mathbb{E}\left[(1+\frac{1}{\sqrt{d}}\frac{\sqrt{2-q}}{\sqrt{q}}Z_{d,q})^{\frac{s}{2}} S(\frac{dq}{2}+\frac{\sqrt{d}}{2}\sqrt{2q-q^2}Z_{d,q}, \frac{s}{2})\right] \\
=&\mathbb{E}\left[1+\binom{s/2}{1}\frac{1}{\sqrt{d}}Z_d+\binom{s/2}{2}\frac{(6-q)}{dq}Z_d^2+\dots\right] \\
&+\mathcal{O}(e^{-\frac{d}{2}(2q-q^2)}) \\
=&1+\binom{s/2}{2}\frac{6-q}{dq}+o(\frac{1}{d}),
\end{aligned}$$

where we used the fact that $\mathbb{E}(Z_{d,q}^k) \to \mathbb{E}(Z^k)$ as $d \to \infty$ for any fixed k implied by (C.7) where Z is a standard-normal variable in $\mathbb{R}$ with the property that $\mathbb{E}(Z) = 0$ and $\mathbb{E}(Z^2) = 1$. Then, it follows from (J.4) that

$$\begin{aligned}
\frac{(\frac{d}{2})^{s/2}}{I_{0,q}(d,s)} =& q^{s/2}\left[1-\binom{s/2}{2}\frac{6-q}{dq}+o(\frac{1}{d})\right] \\
=& q^{s/2}\left[1-\frac{(6-q)s(s-2)}{8dq}+o(\frac{1}{d})\right],
\end{aligned} \tag{J.7}$$

which implies

$$\begin{aligned}
\sigma_{0,q}^2(d,s) =& (\frac{1}{I_{0,q}(d,s)})^{2/s} \\
=& \frac{2q}{d}\left[1-\frac{(6-q)s(s-2)}{8dq}+o(\frac{1}{d})\right]^{2/s} \\
=& \frac{2q}{d}\left[1-\frac{(6-q)(s-2)}{4dq}+o(\frac{1}{d})\right] \\
=& \frac{2q}{d}-\frac{(6-q)(s-2)}{2d}+o(1/d^2).
\end{aligned} \tag{J.8}$$

This completes the proof. □

J.2 Parametric ReLU activation with dropout

Theorem 21. ***(Explicit characterization of the critical variance $\sigma_{a,q}^2(s,d)$ with dropout)*** *Consider a fully connected network with an input $x^{(0)} \in \mathbb{R}^d$ and Gaussian initialization satisfying **(A1)**–**(A2)** with activation function $\phi_a(x)$ for any choice of $a \in (0,1]$ fixed and with dropout where the probability to keep a neuron is $q \in (0,1]$. Then, for any $s \in (0,2]$, the output of the k-th layer satisfies*

$$\mathbb{E}\left[\|x^{(k)}\|^s\right] = \|x^{(0)}\|^s (\sigma^s I_{a,q}(s,d))^k \tag{J.9}$$

with $I_{a,q}(s,d)$ defined as:

$$\begin{aligned}
I_{a,q}(s,d) =& \frac{1}{q^s} 2^{s/2} \frac{1}{\Gamma(1-s/2)} \sum_{n=0}^{d}\sum_{m=0}^{d-n} q_d(n,m) \\
&\sum_{k=0}^{\infty} w_{k,n,m} \mathrm{B}(k+1-\frac{s}{2}, \frac{n+m}{2}+\frac{s}{2})
\end{aligned}$$

for $s \in (0,2)$, especially if $s=2$

$$I_{a,q}(s,d) = \frac{1}{q^2}(1+a^2)\frac{d}{2}$$

where $q_d(n,m)$ is defined by (J.11), $\mathrm{B}(\cdot,\cdot)$ is the Beta function and

$$w_{k,n,m} = \frac{1}{2}(1-a^2)^k\left[\binom{\frac{m}{2}+k-1}{k}n + a^2 m\binom{\frac{m}{2}+k}{k}\right]. \tag{J.10}$$

Let $\bar{\sigma}_{a,q}(s,d) = \frac{1}{\sqrt[s]{I_{a,q}(s,d)}}$. We have three possible cases:

(*i*) *If* $\sigma = \bar{\sigma}_{a,q}(s,d)$ *where then the network preserves the s-th moment of the layer outputs, i.e. for every* $k \geq 1$, $\mathbb{E}\left[\|x^{(k)}\|^s\right] = \|x^{(0)}\|^s$, *whereas for any* $p > s$, $\mathbb{E}\|x^{(k)}\|^p \to \infty$ *exponentially fast in* k.

(*ii*) *If* $\sigma < \bar{\sigma}_{a,q}(s,d)$, *then* $\mathbb{E}\left[\|x^{(k)}\|^s\right] \to 0$ *exponentially fast in* k.

(*iii*) *If* $\sigma > \bar{\sigma}_{a,q}(s,d)$, *then* $\mathbb{E}\left[\|x^{(k)}\|^s\right] \to \infty$ *exponentially fast in* k.

Proof. The proof follows from minor changes to the proof of Theorem 7. In the abscence of dropout (i.e. when $q = 1$), the quantity defined in the proof of Theorem 7, $I_a(d,s)$ has the distribution

$$X_n := \chi_1^2(n) + a^2\chi_2^2(d-n),$$

with probability $p_d(n)$ where $\chi_1^2(n)$ and $\chi_2^2(d-n)$ are independent chi-square distributions with degrees of freedom n and $d-n$ respectively. When there is dropout, the distribution of X_n and the corresponding binomial probabilities will be subject to change because now there is the possibility of zero output from some neurons due to dropout and scaling the neuron outputs. The corresponding probabilities will come from the trinomial distribution instead. More specifically, it suffices to replace X_n with

$$X_{n,m} = \frac{1}{q^2}\left(\chi_1^2(n) + a^2\chi_2^2(m)\right)$$

with probabilities from the trinomial distribution

$$q_d(n,m) = \frac{d!}{n!m!(d-n-m)!}(\frac{q}{2})^{n+m}(1-q)^{d-n-m}. \tag{J.11}$$

Moreover, we can compute $\mathbb{E}X_{n,m}^{s/2}$ by simply replacing d with $n+m$ in the formula for $\mathbb{E}((X_n)^{s/2})$ we obtained in (F.13). After following similar steps to the proof of Theorem 7, we obtain the desired result. □

Corollary 22. ***(Critical variance*** $\bar{\sigma}_{1,q}(d,s)$ ***when*** d ***is large with dropout)*** *For fixed width* d *and* $s \in (0,2]$, *we have*

$$\bar{\sigma}_{1,q}^2(s,d) = \frac{q}{d} + \frac{(3-q)(2-s)}{4d^2} + o(\frac{1}{d^2})$$

with $\bar{\sigma}_{1,q}^2(2,d) = \frac{q}{d}$. *Therefore, it follows from Theorem 7 that if* $\sigma^2 = \frac{1}{d} + \frac{(3-q)(2-s)}{4d^2}$, *then the network with linear activation will preserve the moment of order* $s + o(\frac{1}{d})$ *of the network output.*

Proof. In the linear activation function case, we have $a = 1$, then we obtain $w_{k,n,m} = \frac{m+n}{2}$ for $k = 0$, $w_{k,n,m} = 0$ for $k > 0$. Then it follows that

$$\begin{aligned}
&I_{1,q}(s,d)\\
=&\frac{1}{q^s}2^{s/2}\frac{1}{\Gamma(1-\frac{s}{2})}\sum_{n=0}^{d}\sum_{m=0}^{d-n} q_d(n,m)\frac{m+n}{2}\\
&B(1-\frac{s}{2}, \frac{m+n}{2}+\frac{s}{2})\\
=&\frac{1}{q^s}2^{s/2}\sum_{n=0}^{d}\sum_{m=0}^{d-n} q_d(n,m)\frac{\Gamma(\frac{m+n}{2}+\frac{s}{2})}{\Gamma(\frac{m+n}{2})}.
\end{aligned}$$

where we use the identity $\Gamma(\frac{m+n}{2}+1) = \frac{m+n}{2}\Gamma(\frac{m+n}{2})$ and the fact that $\mathrm{B}(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ for $x, y > 0$. Then denote $t = m+n$, we get

$$I_{1,q}(s,d) = \frac{1}{q^s}2^{s/2}\sum_{t=0}^{d} h_d(t)\frac{\Gamma(\frac{t}{2}+\frac{s}{2})}{\Gamma(\frac{t}{2})}, \tag{J.12}$$

where

$$h_d(t) := \binom{d}{t}q^t(1-q)^{d-t}.$$

We see that in the special case $q = 1/2$, this formula reduces to the analysis provided in Corollary 4. The proof will be similar where we will follow a similar approach to the proof of Corollary 4. Similar to the proof technique of Corollary 4, We write

$$I_{1,q}(s,d) = \frac{1}{q^s}2^{s/2}\sum_{t=1}^{d} h_d(t)(\frac{t}{2})^{s/2}S(t/2, s/2).$$

Note that

$$\frac{(\frac{d}{2})^{s/2}}{I_{1,q}(s,d)} = \frac{q^s}{\mathbb{E}(F_d(H_d))}, \tag{J.13}$$

where H_d is a Binomial random variable, i.e.

$$\mathbb{P}(H_d = n) = \binom{d}{t}q^t(1-q)^{d-t} \quad \text{for} \quad t = 0, 1, \ldots, d,$$

where F_d is defined by (C.6). By the normal approximation of the binomial distribution, we also have

$$\xi_{d,q} := \frac{H_d - \mathbb{E}(H_d)}{\sqrt{\mathrm{var}H_d}} = \frac{H_d - dq}{\sqrt{dq(1-q)}} \longrightarrow \mathcal{N}(0,1) \tag{J.14}$$

in distribution. We also have

$$\begin{aligned}
&\mathbb{E}(F_d(H_d))\\
=&\mathbb{E}\left(F_d(dq+\sqrt{dq(1-q)}\xi_{d,q})\right)\\
=&2^{s/2}\mathbb{E}\Bigg[\frac{(dq+\sqrt{dq(1-q)}\xi_{d,q})^{s/2}}{d^{s/2}}\\
&S(\frac{1}{2}(dq+\sqrt{dq(1-q)}\xi_{d,q}),s/2)\Bigg]\\
=&(2q)^{s/2}\mathbb{E}\Bigg[(1+\sqrt{\frac{1-q}{dq}}\xi_{d,q})^{s/2}\\
&S(\frac{1}{2}(dq+\sqrt{dq(1-q)}\xi_{d,q}),s/2)\Bigg].
\end{aligned}$$

Recall the Binomial expansion formula,

$$(1+x)^{s/2}=\sum_{k=0}^{\infty}\binom{s/2}{k}x^k \quad \text{for} \quad |x|<1.$$

For $\xi_{d,q}<\sqrt{\frac{dq}{1-q}}$, we can write

$$\begin{aligned}
&\left[(1+\sqrt{\frac{1-q}{dq}}\xi_{d,q})^{s/2}S(\frac{1}{2}(dq+\sqrt{dq(1-q)}\xi_{d,q}),s/2)\right]\\
=&\left[\sum_{k=0}^{\infty}\binom{s/2}{k}(\frac{1-q}{dq})^{k/2}\xi_{d,q}^k\right]\left[\sum_{m=0}^{M}\right.\\
&\left.A_m(s/2)\left(\frac{2}{dq+\sqrt{dq(1-q)}\xi_{d,q}}\right)^m\right]\\
=&\left[\sum_{k=0}^{\infty}\binom{s/2}{k}(\frac{1-q}{dq})^{k/2}\xi_{d,q}^k\right]\left[\sum_{m=0}^{M}A_m(s/2)(\frac{2}{dq})^m\right.\\
&\left.\left(\frac{1}{1+\sqrt{\frac{1-q}{dq}}\xi_{d,q}}\right)^m\right]\\
=&\left[\sum_{k=0}^{\infty}\binom{s/2}{k}(\frac{1-q}{dq})^{k/2}\xi_{d,q}^k\right]\left[\sum_{m=0}^{M}A_m(s/2)(\frac{2}{dq})^m\right.\\
&\left.\left(\sum_{\ell=0}^{\infty}(-1)^l(\sqrt{\frac{1-q}{dq}})^\ell\xi_{d,q}^\ell\right)^m\right]\\
=&1+\binom{s/2}{2}\frac{2}{dq}+\left[\binom{s/2}{1}\sqrt{\frac{1-q}{qd}}-\binom{s/2}{2}\frac{2}{dq}\sqrt{\frac{1-q}{dq}}\right.\\
&\left.+\binom{s/2}{1}\binom{s/2}{2}\frac{2}{dq}\sqrt{\frac{1-q}{dq}}\right]\xi_{d,q}\\
&+\binom{s/2}{2}\left[\frac{2(1-q)}{d^2q^2}+\frac{1-q}{dq}+\binom{s/2}{2}\frac{2(1-q)}{d^2q^2}\right]\xi_{d,q}^2\\
&-\binom{s/2}{1}\binom{s/2}{2}\frac{2(1-q)}{d^2q^2}\xi_{d,q}^2+\cdots.
\end{aligned}$$

where we used the identity $A_1(s/2)=\frac{\frac{s}{2}(\frac{s}{2}-1)}{2}$. Since $\mathbb{P}(\xi_{d,q}\geq\sqrt{\frac{dq}{1-q}})=\mathcal{O}(e^{-dq/2(1-q)})$ and the function S is non-negative and bounded by 1 according to (C.3), we have

$$\begin{aligned}
&\mathbb{E}\left[(1+\sqrt{\frac{1-q}{dq}}\xi_{d,q})^{s/2}S(\frac{1}{2}(dq+\sqrt{dq(1-q)}\xi_{d,q}),s/2)\right]\\
=&\mathcal{O}(e^{-dq/2(1-q)})+1+\binom{s/2}{2}\frac{2}{dq}\\
&+\binom{s/2}{2}\left[\frac{2}{dq}\frac{1-q}{dq}+\frac{1-q}{dq}+\binom{s/2}{2}\frac{2}{dq}\frac{1-q}{dq}\right.\\
&\left.-\binom{s/2}{1}\frac{2}{dq}\frac{1-q}{dq}\right]+\cdots.\\
=&1+\binom{s/2}{2}\frac{3-q}{dq}+o(\frac{1}{d}),
\end{aligned}$$

where we used the fact that $\mathbb{E}(\xi_{d,q}^k)\to\mathbb{E}(Z^k)$ as $d\to\infty$ for any fixed k implied by (C.7) where Z is a standard-normal variable in $\mathbb{R}$ which satisfies $\mathbb{E}(Z)=0$ and $\mathbb{E}(Z^2)=1$. Then, it follows from (J.13) that

$$\begin{aligned}
\frac{(\frac{d}{2})^{s/2}}{I_{1,q}(s,d)} &= q^s\frac{1}{\mathbb{E}(F_d(B_d))}\\
&= (\frac{q}{2})^{s/2}(1-\binom{s/2}{2}\frac{3-q}{dq}+o(\frac{1}{d}))\\
&= (\frac{q}{2})^{s/2}\left(1-\frac{(3-q)s(s-2)}{8dq}+o(\frac{1}{d})\right),
\end{aligned}$$

which implies

$$\begin{aligned}
\bar{\sigma}_{1,q}^2(s,d) &= \frac{1}{\sqrt[s/2]{I_{1,q}(s,d)}}\\
&= \frac{q}{d}\left(1+\frac{(3-q)(2-s)}{4dq}+o(\frac{1}{d})\right),
\end{aligned}$$

which completes the proof for the case $s\in(0,2]$. For $s=2$, (J.12) simplifies to

$$I_{1,q}(2,d)=\frac{1}{q^2}2\sum_{t=0}^{d}h_d(t)\frac{t}{2}=\frac{d}{q} \tag{J.15}$$

where we used $\Gamma(\frac{t}{2}+1)=\frac{t}{2}\Gamma(\frac{t}{2})$ and the fact that $\mathbb{E}(H_d)=qd$. This leads to $\bar{\sigma}_{1,q}(2,s)=q/d$ as desired. □

K Proof of Theorem 14

Proof. We first consider the ReLU case where $a=0$. In this case, the fact that $x^{(k)}$ goes to zero a.s. follows from a relatively simple argument. After a simple computation (see Lemma 17), we find that $\mathbb{P}(x^{(k)}=0)=1-(1-\frac{1}{2^d})^k$ regardless of the choice of $\sigma>0$. This implies that for all $\varepsilon>0$,

$$\sum_{k\geq1}\mathbb{P}(|x^{(k)}|>\varepsilon)\leq\sum_{k\geq1}(1-\frac{1}{2^d})^k<\infty.$$

Therefore, $x^{(k)} \to 0$ almost surely. We next consider the $a \in (0,1]$ case and build on the theory of iterated random Lipschitz maps. Recall that the layer outputs obey the stochastic recursion

$$x^{(k+1)} = F^{(k+1)}(x^{(k)}) = M_{W^{(k+1)},a}(x^{(k)}) \qquad \text{(K.1)}$$

where $M_{W,a}(x) := \phi_a(Wx)$. We also note that for a non-negative random variable

$$\mathbb{E}\log(M) = \frac{d}{ds}\mathbb{E}(M^s)|_{s=0}\,,$$

when the expectations are finite. Therefore, choosing $M = \|\phi_a(We_1)\|$,

$$\mu_a(\sigma) = \frac{d}{ds}\mathbb{E}\|\phi_a(We_1)\|^s|_{s=0}\,. \qquad \text{(K.2)}$$

Then, similar to the proof of Corollary 9, we consider $\kappa(s) = \sigma^s I_a(s,d) = \mathbb{E}\|\phi_a(We_1)\|^s$ where $I_a(s,d)$ is defined by (4.2). The function $\kappa(s)$ is convex and continuously differentiable (see the proof of Corollary 9). Notice that $\mu_a(\sigma) = \kappa'(0)$ and $\kappa(0) = 1$. If $\mu_a(\sigma) < 0$, then $\kappa(s) < 1$ for $s > 0$ small enough. Since $\kappa(s) = \mathbb{E}\|\phi_a(We_1)\|^s$ goes to infinity as s goes to infinity, we conclude that there exists $s_* > 0$ such that $\kappa(s_*) = 1$. From the definition of the κ function, this is equivalent to saying $\sigma = \frac{1}{\sqrt[s_*]{I_a(s_*,d)}} = \bar{\sigma}_a(s_*,d)$ for some $s_* > 0$. Correspondingly, if $\sigma = \frac{1}{\sqrt[s_*]{I_a(s_*,d)}} = \bar{\sigma}_a(s_*,d)$ for some $s_* > 0$, then $\kappa(s_*) = 1$ and since $\kappa(0) = 1$, by convexity of κ we find that $\kappa(s) < 1$ for $s \in (0, s_*)$ which implies $\mu_a(\sigma) = \kappa'(0) < 0$. For $s \in (0, s_*)$, by Corollary 9 we have $\bar{\sigma}_a(s_*,d) < \bar{\sigma}_a(s,d)$. If apply part (ii) of Theorem 7 with the fact that $\sigma = \bar{\sigma}_a(s_*,d) < \bar{\sigma}_a(s,d)$, then we obtain $\mathbb{E}(\|x^{(k)}\|^s) \to 0$, i.e. $x^{(k)}$ converges to zero in the space L_s.

We next prove that $x^{(k)}$ has a subsequence that converges to zero a.s. when $\mu_a(\sigma) < 0$. From Theorem 11, we see that for any constant $C > 0$, we have

$$\begin{aligned}
&\lim_{k\to\infty} \mathbb{P}(\|x^{(k)}\| > C) \\
&= \lim_{k\to\infty} \mathbb{P}(\log\|x^{(k)}\| > \log(C)) \\
&= \lim_{k\to\infty} \mathbb{P}\left(\frac{\log\|x^{(k)}\| - \mu_a(\sigma)k}{\sqrt{k}} > \frac{\log(C)}{\sqrt{k}} - \mu_a(\sigma)\sqrt{k}\right) \\
&= \begin{cases} 0 & \text{if} \quad \mu_a(\sigma) < 0, \\ 1 & \text{if} \quad \mu_a(\sigma) > 0. \end{cases}
\end{aligned} \qquad \text{(K.3)}$$

We have two cases, depending on the sign of $\mu_a(\sigma)$.

(i) $(\mu_a(\sigma) < 0)$: In this case, for $C = 1/2$, based on (K.3), we can choose n_1 large enough so that

$$\mathbb{P}(\|x^{(n_1)}\| > \frac{1}{2}) \le \frac{1}{2}.$$

Continuing by a recursive fashion choose n_k large enough such that

$$\mathbb{P}(\|x^{(n_k)}\| > \frac{1}{2^k}) \le \frac{1}{2^k}$$

with $n_1 < n_2 < \cdots < n_k$. Then the event $A_k = \{\|x^{(n_k)}\| > \frac{1}{2^k}\}$ is such that $\sum_k \mathbb{P}(A_k) < \infty$. By the Borel-Cantelli lemma, we find that

$$\mathbb{P}\left(\limsup_{k\to\infty}\{\|x^{(n_k)}\| > \frac{1}{2^k}\}\right) = 0.$$

This proves that for any $\varepsilon > 0$ given $\mathbb{P}(\|x^{(n_k)}\| \ge \varepsilon \text{ infinity often}) = 0$ which is equivalent to saying $\mathbb{P}(\|x^{(n_k)}\| < \varepsilon) = 1$ or yet equivalently $x^{(n_k)} \to 0$ almost surely.

In the special case when $s_* > 1$, we can stronger results. In particular, we can consider

$$\begin{aligned}
\sum_{j=0}^{\infty} \mathbb{E}\|x^{(j+1)} - x^{(j)}\| &\le \sum_{j=0}^{\infty} \left(\mathbb{E}\|x^{(j+1)}\| + \mathbb{E}\|x^{(j)}\|\right) \\
&\le 2\sum_{j=0}^{\infty} \mathbb{E}\|x^{(j)}\| < \infty
\end{aligned} \qquad \text{(K.4)}$$

where we applied part (ii) of Theorem 7 with the fact that $\sigma = \bar{\sigma}_a(s_*,d) < \bar{\sigma}_a(1,d)$. Then, by (Steinsaltz, 1999, Lemma 1), $x^{(k)}$ converges almost surely to a limit. Since the subsequence $x^{(n_k)}$ converges to zero, we obtain that $x^{(k)}$ converges a.s. to zero.

(ii) $(\mu_a(\sigma) > 0)$: The proof follows from a similar approach to part (i).When $\mu_a(\sigma) > 0$, we see from Theorem 7 that all the moments $\mathbb{E}(\|x^{(k)}\|^\alpha)$ diverges for any $\alpha > 0$ (because if it were not, then $\sigma = \sigma_a(s,d)$ for some $s > 0$ which would imply $\mu_a(\sigma) < 0$ by the discussion above). Furthermore, based on (K.3), $x^{(k)}$ diverges to infinity in probability and we can choose a subsequence $\bar{n}_k$ such that

$$\mathbb{P}(\|x^{(\bar{n}_k)}\| > 2^k) \ge 1 - \frac{1}{2^k}$$

with $\bar{n}_1 < \bar{n}_2 < \cdots < \bar{n}_k$. Then the event $\bar{A}_k = \{\|x^{(n_k)}\| < 2^k\}$ is such that $\sum_k \mathbb{P}(\bar{A}_k) < \infty$. By the Borel-Cantelli lemma, we find that

$$\mathbb{P}\left(\limsup_{k\to\infty}\{\|x^{(n_k)}\| < 2^k\}\right) = 0.$$

This proves that for any $\varepsilon > 0$ given $\mathbb{P}(\|x^{(n_k)}\| \le \varepsilon \text{ infinity often}) = 0$ which is equivalent to saying $\mathbb{P}(\|x^{(n_k)}\| > \varepsilon) = 1$ or yet equivalently $x^{(n_k)} \to \infty$ almost surely.

□

L Proof of Theorem 15

Proof. Due to the addition of Gaussian noise to post-activations, we have the recursion over the layers

$$\tilde{x}^{(k+1)} = M_{W^{(k+1)},\xi^{(k+1)}}(\tilde{x}^{(k)}) := W^{(k+1)}\tilde{x}^{(k)} + \xi^{(k+1)} \tag{L.1}$$

where $\tilde{x}^{(k)}$ denotes the input to the $(k+1)$-st layer and $\xi^{(k)}$ is a random vector with components $\xi_i^{(k)}$ that are i.i.d. mean zero random variables. The map $M_{W^{(k)},\xi^{(k)}}$ is a random Lipschitz (linear) map whose convergence behavior has been studied in the literature. If the following conditions hold

$$\mathbb{E}\left[\max\left(0, \log(\|W^{(k+1)}\|)\right)\right] < \infty, \tag{L.2}$$

$$\mathbb{E}\left[\max\left(0, \log(\|\xi^{(k+1)}\|)\right)\right] < \infty, \tag{L.3}$$

$$c_1 = \inf_k \frac{1}{k}\mathbb{E}\log\|W^{(k)}W^{(k-1)}\dots W^{(1)}\| < \infty, \tag{L.4}$$

then it is known that $\tilde{x}^{(k)}$ admits an almost sure limit $\tilde{x}^{(\infty)}$ in which case the limit is given by the formula

$$x^{(\infty)} = \sum_{j=1}^{\infty}\left(\prod_{i=1}^{j-1} W^{(i)}\right)\xi^{(j)}, \tag{L.5}$$

(see e.g. (Diaconis and Freedman, 1999, Thm. 2.1)). We check the conditions in (L.3) and (L.4). The second condition in (L.3) is satisfied by the assumption on the noise $\xi^{(k)}$, and the first condition in (L.3) is satisfied as

$$\begin{aligned}&\mathbb{E}\left[\max\left(0, \log(\|W^{(k+1)}\|)\right)\right]\\ \le&\mathbb{E}\left[\max\left(0, \frac{1}{2}\log\left(\sum_{i,j=1}^{d}(W_{ij}^{(k+1)})^2\right)\right)\right] < \infty,\end{aligned} \tag{L.6}$$

where we used the fact that $\sum_{i,j=1}^{d}(W_{ij}^{(k+1)})^2$ is a chi-square distribution with d^2 degrees of freedom. Finally, the condition (L.4) is equivalent to

$$c_1 = \inf_k \frac{1}{k}\mathbb{E}\log\frac{\|x^{(k)}\|}{\|x^{(0)}\|}, \tag{L.7}$$

where $x^{(k)}$ are the iterations without noise, i.e. $x^{(k)}$ satisfies $x^{(k+1)} = W^{(k+1)}x^{(k)}$ starting from $x^{(0)}$. It follows from the analysis of Theorem 11 that we have also

$$c_1 = \mu_1. \tag{L.8}$$

Due to the choice of $\sigma = \bar{\sigma}_1(s,d)$, by Theorem 14, we have also $\mu_1 < 0$. We conclude from (L.8) that $c_1 < 0$ and (L.4) is also satisfied. Hence, having checked that assumptions (L.3)–(L.4) hold, we conclude that the limit $\tilde{x}^{(\infty)}$ exists, it is non-zero and is given by the series sum (L.5). With the addition of i.i.d. noise to activations, moments cannot grow slower; i.e. it is not hard to show that

$$\mathbb{E}(\|\tilde{x}^{(k)}\|^p) \ge \mathbb{E}(\|x^{(k)}\|^p)$$

with the same initialization i.e. $x^{(0)} = \tilde{x}^{(0)}$. By Theorem 7, we also know that $\mathbb{E}(\|x^{(k)}\|^p) \to \infty$ for any $p > s$ as $k \to \infty$. Therefore we conclude that

$$\mathbb{E}(\|\tilde{x}^{(k)}\|^p) \to \infty, \quad \text{for any } p > s,$$

as $k \to \infty$. Then, we have necessarily $\mathbb{E}(\|\tilde{x}^{(\infty)}\|^p) = \infty$ because otherwise $\tilde{x}^{(k)}$ would converge to $\tilde{x}^{(\infty)}$ in L_p which would be a contradiction as $\mathbb{E}(\|\tilde{x}^{(k)}\|^p) \to \infty$. This proves that the limit $\tilde{x}^{(\infty)}$ is heavy tailed in the sense that its moments of order p are infinite for any $p > s$. In particular for $s < 2$, this implies that the variance of the limit $\tilde{x}^{(\infty)}$ is infinite. This completes the proof. □

M A Supporting Lemma

Lemma 23. *Let M_i be the random variables and p_i be the constant weights. Let M be the mixture distribution $M := \sum_i p_i M_i$. We have*

$$\mathrm{var}(M) = \sum_i p_i \mathrm{var}(M_i) + \sum_i p_i(\mathbb{E}[M_i])^2 - \left(\sum_i p_i \mathbb{E}[M_i]\right)^2$$

Proof. Let $\mu^{(r)}$ denote the r-th (raw) moment of M, and $\mu_i^{(r)}$ the r-th moment of M_i. Then we obtain

$$\mu^{(r)} = \sum_i p_i \mathbb{E}[M_i^r] = \sum_i p_i \mu_i^{(r)}.$$

The variance of M can be written as

$$\mathrm{var}(M) = \mu^{(2)} - \left(\mu^{(1)}\right)^2 = \sum_i p_i \mu_i^{(2)} - \left(\sum_i p_i \mu_i^{(1)}\right)^2.$$

Since $\mu_i^{(2)} = \mathrm{var}(M_i) + (\mu_i^{(1)})^2$, we have

$$\begin{aligned}\mathrm{var}(M) &= \sum_i p_i(\mathrm{var}(M_i) + (\mu_i^{(1)})^2) - \left(\sum_i p_i \mu_i^{(1)}\right)^2\\ &= \sum_i p_i \mathrm{var}(M_i) + \sum_i p_i(\mathbb{E}[M_i])^2\\ &\quad - \left(\sum_i p_i \mathbb{E}[M_i]\right)^2.\end{aligned}$$

□

N Extensions of results to Convolutional networks

For a convolutional layer, we can write the process as

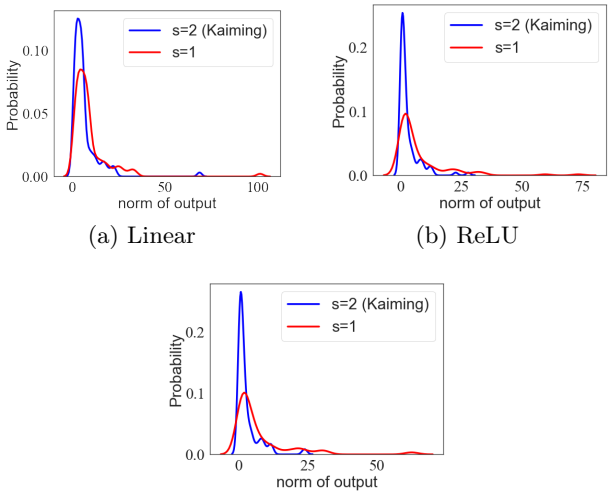
$$x^{(k+1)} = \phi_a(W^k x^k + b^k),$$

where x^k is a $m_k^2 c_k \times 1$ vector which represents co-located $m_k \times m_k$ pixels in c_k input channels, where m_k here is the spatial filter size of the layer k. If we introduce the quantities $d_k = m_k^2 c_k$, and n_k as the number of filters in layer k then W^k is a $n_k \times d_k$ matrix and each row of W^k represents the weights of one filter. Moreover, we also have $c_{k+1} = n_k$ by the definition. Therefore, we can use our method to initialize the convolutional neural networks where we take $d_k = m_k^2 c_k$.

O Further Numerical Experiments and Illustrations

O.1 Numerical Illustrations

In this section, we present additional figures and numerical experiments that were not part of the main text due to space considerations.



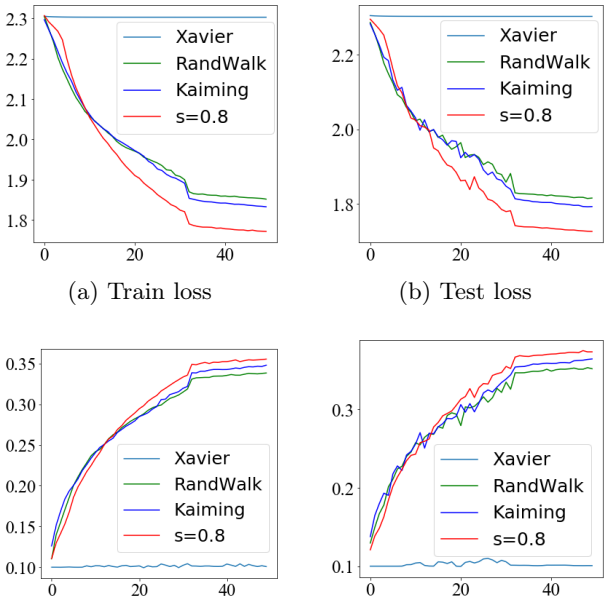
(a) Linear (b) ReLU

(c) Leaky-ReLU

Figure 8: Distribution of norm of the output $\|x^{(k)}\|$ through $k = 100$ layers. 8a: Probability density of $\|x^{(k)}\|$ for linear activation, where we set $\sigma^2 = \frac{1}{d} + \frac{1}{2d^2} \approx \bar{\sigma}_a^2(s,d)$ with $a = 0$ and $s = 1$ with our initialization. 8b: Probability density of $\|x^{(k)}\|$ for ReLU activation, where we set $\sigma^2 = \frac{2}{d} + \frac{5}{2d^2} \approx \bar{\sigma}_a^2(s,d)$ with $a = 0$ and $s = 1$ with our initialization. 8c: Probability density of $\|x^{(k)}\|$, where we set $\sigma^2 \approx \bar{\sigma}_a^2(s,d)$ with $a = 0.01$ and $s = 1$ with our initialization. Kaiming initialization corresponds to $\sigma^2 = \bar{\sigma}_a^2(s,d)$ for $s = 2$.

Distribution of the network output. The distribution of the natural logarithm of the norm of the output $R_{k,0}$ is plotted in Figure 2 in the main text. Figure 8 illustrates the distribution the norm of the k-th layer output for linear, ReLU and Leaky ReLU activations which supplements Figures 2 and 4. The distribution is obtained from the samples by standard kernel density estimation methods provided in the Python package `seaborn`.[1] We observe that our initialization leads to heavier tails compared to Kaiming initialization, where the frequency of small outputs is less frequent in our method compared to Kaiming initialization.

O.2 Numerical Experiments



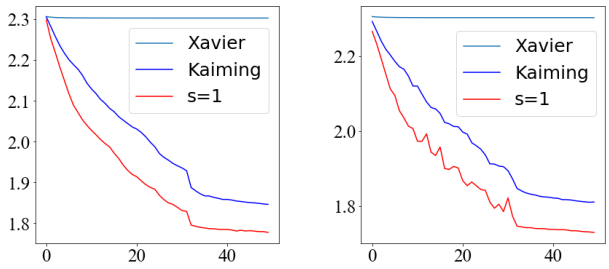
(a) Train loss (b) Test loss

(c) Train accuracy (d) Test accuracy

	train loss		test loss	
	mean	std	mean	std
Xavier	2.3026	1.1405	2.3026	1.1985
Randwalk	1.8519	0.1231	1.8158	0.1367
Kaiming	1.833	0.098	1.793	0.109
s=0.8	**1.772**	0.1294	**1.7264**	0.1469
	train acc		test acc	
	mean(%)	std	mean(%)	std
Xavier	10.07	0.0006	10.07	0.0014
Randwalk	33.84	0.0346	35.21	0.0386
Kaiming	34.8	0.0304	36.46	0.0329
s=0.8	**35.55**	0.0431	**37.37**	0.05

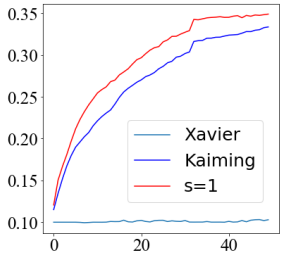
Figure 9: Fully connected network with ReLU activation on CIFAR-10. The results are the *average* over 10 samples. The x-axis is epoch number.

[1] This package is publicly available at https://seaborn.pydata.org/.

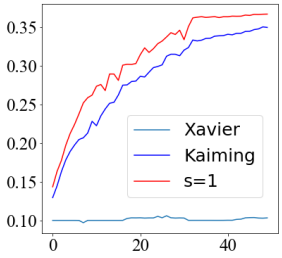


(a) Train loss

(b) Test loss



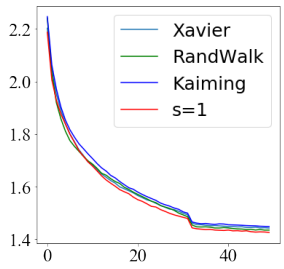
(c) Train accuracy



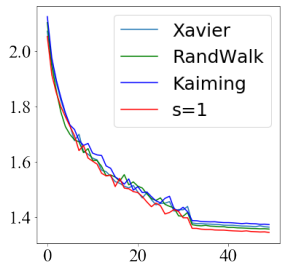
(d) Test accuracy

	train loss		test loss	
	mean	std	mean	std
Xavier	2.3016	0.0001	2.3026	0.0001
Kaiming	1.8459	0.1266	1.8108	0.1383
s=1	**1.7771**	0.1668	**1.7298**	0.1796
	train acc		test acc	
	mean(%)	std	mean(%)	std
Xavier	10.29	0.0147	10.32	0.0163
Kaiming	33.36	0.0376	34.93	0.0433
s=1	**34.9**	0.0512	**36.65**	0.0568

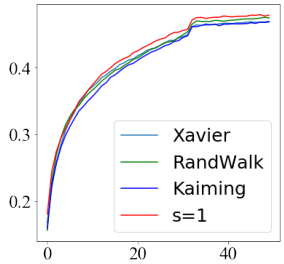
Figure 10: Fully connected network with Leaky ReLU on CIFAR-10. The plots are averages over 10 runs, where mean and standard deviation (std) are also reported. The x-axis is the epoch number.



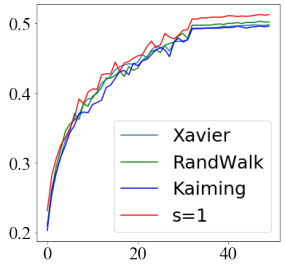
(a) Train loss



(b) Test loss



(c) Train accuracy



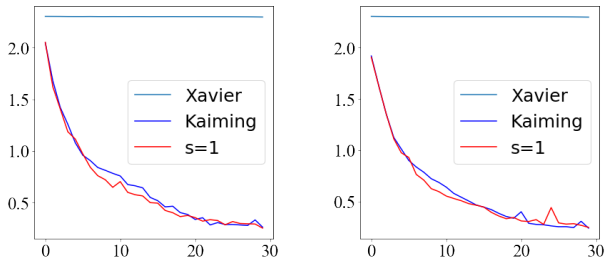
(d) Test accuracy

	train loss		test loss	
	mean	std	mean	std
Xavier	1.442	0.0169	1.3636	0.0103
Randwalk	1.4344	0.0128	1.3565	0.0104
Kaiming	1.4479	0.0124	1.3727	0.0088
s=1	**1.4255**	0.0236	**1.3445**	0.0239
	train acc		test acc	
	mean(%)	std	mean(%)	std
Xavier	46.92	0.0059	49.79	0.0043
Randwalk	47.43	0.0033	50.18	0.0031
Kaiming	46.82	0.0053	49.61	0.0066
s=1	**47.87**	0.0096	**51.22**	0.0114

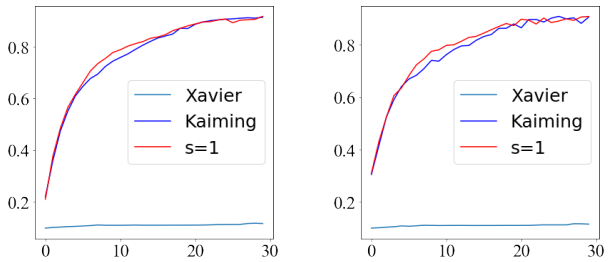
Figure 11: Linear fully connected network on CIFAR-10. The results are the *average* over 10 samples. The x-axis is epoch number.

CIFAR-10 dataset.[2] CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. For ReLU activations, we compare our initializaion method with Kaiming and Xavier method as well as with the random walk initialization. However for the Leaky ReLU activation, we compare our new method with Kaiming and Xavier method only as the parameters of random walk initialization are not available for Leaky ReLU initialization. For understanding the effect of initialization on training, we report the first 50 epochs in the training process where we train our networks with stochastic gradient descent (SGD) using a constant stepsize. We tuned the SGD stepsize and used the same stepsize for each method.

Figure 9 shows the results of a fully connected network with ReLU activation. For our method in the ReLU case, we set $\sigma^2 = \frac{2}{d} + \frac{3}{d^2}$, which preserves the moment $s \approx 0.8$ according to Corollary 4. Figure 11 displays the results of network with linear activation on CIFAR-10 with a similar setup where we set $\sigma^2 = \frac{1}{d} + \frac{1}{2d^2}$ which corresponds to the choice of $s \approx 1$. Similarly, Figure 12 reports the corresponding results for Leaky ReLU. In all cases (linear, ReLU and Leaky ReLU activations), we use two convolutional layers, 20 fully-connected layers and $d = 64$ for all hidden layers in this network. We consider four criteria for comparison: train loss, test loss, train accuracy, and test accuracy. We observe that our method performs no worse than other methods (Xavier initialization, Kaiming initialization, and Random walk initialization) and in many cases leads



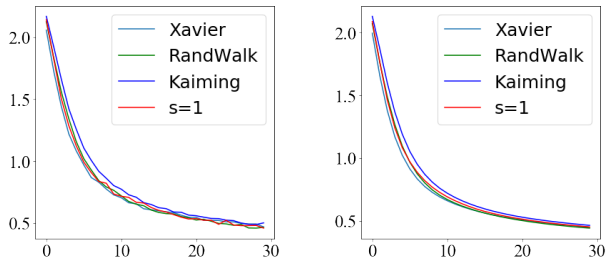
(a) Train loss (b) Test loss



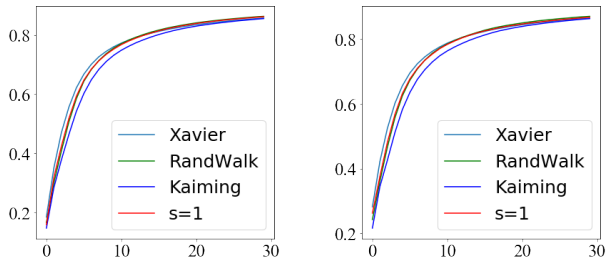
(c) Train accuracy (d) Test accuracy

	train loss		test loss	
	mean	std	mean	std
Xavier	2.2975	0.0157	2.2966	0.0196
Kaiming	0.2613	0.1027	0.2394	0.1096
s=1	**0.2518**	0.1459	**0.2464**	0.1496
	train acc		test acc	
	mean(%)	std	mean(%)	std
Xavier	11.63	0.0502	11.57	0.0452
Kaiming	91.4	0.0446	90.55	0.0479
s=1	**91.78**	0.0575	**90.82**	0.0578

Figure 12: Fully connected network with Leaky ReLU on MNIST. The results are the *average* over 20 samples. The x-axis is epoch number.



(a) Train loss (b) Test loss



(c) Train accuracy (d) Test accuracy

	train loss		test loss	
	mean	std	mean	std
Xavier	0.4686	0.0567	0.4485	0.0311
Randwalk	0.4652	0.042	**0.4414**	0.0317
Kaiming	0.5011	0.0519	0.4634	0.0399
s=1	**0.4586**	0.0371	0.451	0.0348
	train acc		test acc (%)	
	mean(%)	std	mean(%)	std
Xavier	**85.71**	0.0084	86.52	0.0084
Randwalk	86.29	0.0096	**87.22**	0.0104
Kaiming	85.53	0.0121	86.33	0.12
s=1	86.16	0.008	86.74	0.0089

Figure 13: Linear fully connected network on MNIST. The results are the *average* loss over 20 runs. The x-axis is epoch number.

to an improvement.

MNIST dataset.[3] MNIST database is a database of handwritten digits with a training set of 60,000 examples, and a test set of 10,000 examples. The setup is similar to our experiments for CIFAR-10. In our results, we consider 20 runs. Figure 5 (reported in the main text), Figure 12 and Figure 13 show the performance of our method of the fully connected network with ReLU, Leaky ReLU, and linear activations respectively in the first 30 epochs. For the ReLU and Leaky ReLU case, we use 20 layers with $d = 64$. For the linear case, we use 30 layers with $d = 64$. Similar to the CIFAR-10 experiments, we set $s \approx 0.8$ for ReLU, $s = 1$ for Leaky ReLU and linear case.

Both MNIST and CIFAR-10 experiments are implemented by the Python package `torch`.[4] Our experiments are trained on Nvidia GTX 1080Ti GPU. Each experiment of MNIST takes around 3-4 hours, and each experiment of CIFAR-10 takes around 6-7 hours.

[2]This dataset can be downloaded from https://www.cs.toronto.edu/~kriz/cifar.html.

[3]This dataset can be downloaded from http://yann.lecun.com/exdb/mnist/.

[4]This package is publicly available at https://pytorch.org/.