Intro to Infinite Depth-and-Width Limits: Log-Gaussian Behaviour of Deep Neural Networks

See `https://youtu.be/93X0L1U5C0E` for a video tutorial of these notes!

Notation	Description	Notation	Description
$n_{in} \in \mathbb{N}$	Input Dimension	$n_{out} \in \mathbb{N}$	Output Dimension
$n \in \mathbb{N}$	Hidden layer width	$d \in \mathbb{N}$	Number of hidden layers (Depth)
$\varphi(\cdot)$	ReLU function $\varphi(x) = \max(x, 0)$		
$x \in \mathbb{R}^{n_{in}}$	Input	$W^0 \in \mathbb{R}^{n_{in} \times n}$	Weight matrix for layer 0
$z^\ell \in \mathbb{R}^n$	Hidden Layer Neurons (pre-activation)	$W^\ell \in \mathbb{R}^{n \times n}$	Weight Matrix Hidden Layer
$z^{out} \in \mathbb{R}^{n_{out}}$	Network Output	$W^{out} \in \mathbb{R}^{n \times n_{out}}$	Weight matrix for output layer
All weights initialized to iid $\mathcal{N}(0,1)$ random variables.			

1 Definitions and Main Results

Definition 1. A (fully connected) deep neural network is defined by the update rules

$$\text{First Layer:} \quad z^0 = \sqrt{\frac{1}{n_{in}}} W^0 x \tag{1}$$

$$d \text{ Hidden Layers:} \quad z^\ell = \sqrt{\frac{2}{n}} W^\ell \varphi(z^{\ell-1}) \tag{2}$$

$$\text{Last Layer:} \quad z^{out} = \sqrt{\frac{1}{n}} W^{out} z^d \tag{3}$$

Consider the limit where both the network depth $d \to \infty$ and hidden layer width $n \to \infty$ in such away that the ratio d/n converges to a non-zero constant. In this limit the output of the network is approximately **log-Gaussian scalar** times an independent **Gaussian vector**,

$$z^{out} \approx \exp\left(\frac{1}{2}G\right) \vec{W}$$

where $G \in \mathbb{R}$ is a approximately Gaussian and $\vec{W} \in \mathbb{R}^{n_{out}}$ is a Gaussian vector with iid Gaussian $\mathcal{N}(0,1)$ entries. The following theorem make this precise.

Theorem 2. *Suppose all weights are initialized to iid $\mathcal{N}(0,1)$ random variables. For **any** depth and width, the output of the network on initialization is equal in distribution to*

$$z^{out} \overset{d}{=} \frac{\|x\|}{\sqrt{n_{in}}} \exp\left(\frac{1}{2}G\right) \vec{W}, \tag{4}$$

where $\vec{W} \in \mathbb{R}^{n_{out}}$ is a Gaussian random vector with iid $\mathcal{N}(0,1)$ entries, and $G = G(n,d)$ is an independent random variable that depends only on the hidden layers of the network.

*In the **infinite depth-and-width limit** the random variable G has the following behavior in terms of a parameter β*

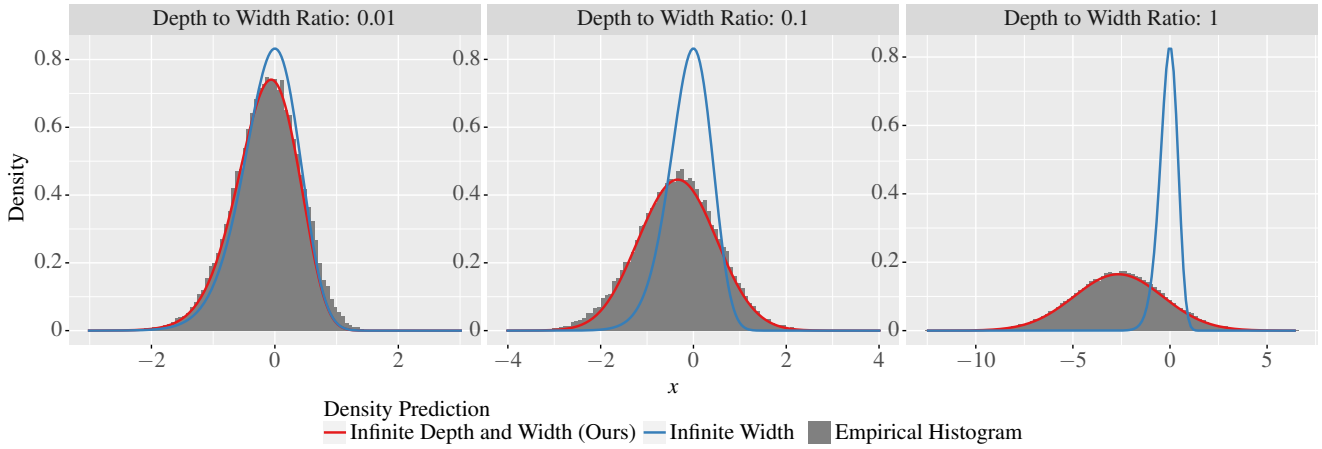
$$\beta := \frac{2}{n} + 5\frac{d}{n}$$
$$\mathbf{E}[G] = -\frac{1}{2}\beta + O\left(\frac{d}{n^2}\right)$$
$$\mathbf{Var}[G] = \beta + O\left(\frac{d}{n^2}\right)$$

*and G **converges to a Gaussian** random variable with this mean and variance.*

1.1 Comparison to the Infinite Width limit

In fixed-depth $d = cons't$ and infinite-width $n \to \infty$ limit the random variable $G \to 0$. In this infinite-width-only-limit, z^{out} is purely Gaussian Other nice properties also emerge in this limit, for instance individual neurons become asymptotically independent. These simplifications allows much more detailed information about the network and training behavior than is currently known about the infinite depth-and-width limit, for example the NTK learning regime.

However, since real networks have finite depth and finite width, the infinite width limit can be a lot less accurate than the infinite depth-and-width limit! Below is the result of a Monte Carlo simulation of 2^{15} samples comparing the theory to finite networks. Here the network widths are $n = 100$, $n_{in} = n_{out} = 10$ and at three different values of depth, $d = 1$, $d = 10$ and $d = 100$. The **probability density** of the random variable $\ln\left(\|z^{out}\|^2\right)$ is shown. The prediction of Theorem 2 is the red curve; the infinite width prediction (where $G = 0$) is the blue curve. The depth-to-width prediction is much more accurate than the infinite width prediction! For deep networks this is a big difference!



1.2 Extensions

1.2.1 Unequal layer sizes and arbitrary weights

Suppose that instead of all the layer widths being equal to n, the ℓ-th layer is width n_ℓ. A similar argument applies when the layer sizes are different to show that the effective parameter is now

$$\beta = \frac{2}{n_0} + 5\sum_{\ell=1}^{d}\frac{1}{n_\ell}.$$

The result also holds when the weights W are **any** (reasonable) distribution, not just Gaussian; although in this case the proof is much harder. This universality result is proven in [HN19b].

1.2.2 Skip connections (ResNets)

If the layer weights have skip connections in the hidden layers so that for some coefficients $\alpha, \lambda \in \mathbb{R}^+$ with $\alpha^2+\lambda^2 = 1$ so that the update rule

$$z^\ell = \alpha z^{\ell-1} + \lambda\sqrt{\frac{2}{n}} W^\ell \varphi(z^{\ell-1})$$

then a similar log-Gaussian result holds. The β parameter becomes:

$$\beta = \frac{2}{n} + \left(5\lambda^4 + 4\alpha^2\lambda^2\right)\frac{d}{n}$$

There are a few other changes in the result that happen due to the skip connections; see [LNR21] for details!.

1.2.3 Gradients

For ReLU networks, one can show that if the input is $\|x\| = 1$, then the input-output matrix has the same distribution as the original function:

$$\frac{\partial}{\partial x_i} z^{out} \stackrel{d}{=} z^{out}$$

so this log-normal result behavior holds for derivatives of the network as well. The proof can be found in [LNR21]. This connection allows one to prove similar results about quantities related to the gradients of the network in the infinite depth-and-width limit, for example corrections to the NTK in [HN19a].

2 Informal Proof Ideas

The key element is the following property of Gaussian random matrices. If W has iid $\mathcal{N}(0,1)$ entries, then for any vector x, we have

$$Wx \stackrel{d}{=} \|x\|\, g$$

where g is a vector whose entries are iid $\mathcal{N}(0,1)$ random variables. Applying this to the first and last layer of the network, we find

$$z^0 \stackrel{d}{=} \frac{\|x\|}{\sqrt{n_{in}}} g^0, \;\; z^{out} \stackrel{d}{=} \frac{\left\|z^d\right\|}{\sqrt{n}} g^{out}$$

where $g^0 \in \mathbb{R}^n$ and $g^{out} \in \mathbb{R}^{n_{out}}$ are iid collection of vectors whose entries iid $\mathcal{N}(0,1)$ random variables. In light of this, if we define

$$G := \ln\left(\frac{\left\|z^d\right\|^2/n}{\|x\|^2/n_{in}}\right),$$

then we see that G only depends on the hidden layers of the network. (G has the distribution of $\ln\left(\left\|z^d\right\|^2/n\right)$ when $z^0 = g^0$). With this definition, (4) holds!

From the construction of G, the essence of the infinite depth-and-width limit is to understand the distribution of $\left\|z^d\right\|$. To understand this, we look at the ratios $\left\|z^{\ell+1}\right\| / \left\|z^\ell\right\|$ layer by layer. By using the homogeneity property of ReLU $\varphi(|c|\, x) = |c|\, \varphi(x)$, we can divide $z^{\ell+1}$ from (1) by $\left\|z^\ell\right\|$ to obtain an expression depending only on the **unit vector** $\hat{z}^\ell = z^\ell / \left\|z^\ell\right\|$:

$$\frac{z^{\ell+1}}{\|z^\ell\|} = \sqrt{\frac{2}{n}} W^{\ell+1}\varphi(\hat{z}^\ell) \stackrel{d}{=} \sqrt{\frac{2}{n}}\left\|\varphi(\hat{z}^\ell)\right\| g^{\ell+1}.$$

In addition, this expression also shows that each hidden layer is a scalar multiple of the Gaussian vector g^ℓ, namely: $\hat{z}^\ell = \hat{g}^\ell$. Hence:

$$\frac{z^{\ell+1}}{\|z^\ell\|} \stackrel{d}{=} \sqrt{\frac{2}{n}}\left\|\varphi(\hat{g}^\ell)\right\| g^{\ell+1} \implies \frac{\left\|z^{\ell+1}\right\|}{\|z^\ell\|} \stackrel{d}{=} \sqrt{\frac{2}{n}}\left\|\varphi(\hat{g}^\ell)\right\| \left\|g^{\ell+1}\right\|$$

We can finally write $\left\|z^d\right\|$ as a telescoping product as:

$$\left\|z^d\right\| = \|z^0\| \prod_{\ell=0}^{d-1} \frac{\left\|z^{\ell+1}\right\|}{\|z^\ell\|} \stackrel{d}{=} \frac{\|x\|}{\sqrt{n_{in}}}\|g^0\| \prod_{\ell=0}^{d-1}\sqrt{\frac{2}{n}}\left\|\varphi(\hat{g}^\ell)\right\| \left\|g^{\ell+1}\right\|. \tag{5}$$

We finally use the homogeneity of $\varphi(\cdot)$ again to combine $\left\|\varphi(\hat{g}^\ell)\right\| \left\|g^\ell\right\| = \left\|\varphi\left(g^\ell\right)\right\|$ to get:

$$\left\|z^d\right\| \stackrel{d}{=} \frac{\|x\|}{\sqrt{n_{in}}} \left(\prod_{\ell=0}^{d-1} \sqrt{\frac{2}{n}} \left\|\varphi(g^\ell)\right\|\right) \left\|g^d\right\|.$$

This shows that $\left\|z^d\right\|$ is equal in distribution to a product of independent random variables! Equivalently, $G = \ln\left(\frac{\left\|z^d\right\|^2/n}{\|x\|^2/n_{in}}\right)$ is a **sum** of independent random variables, and a **central limit theorem** will apply as $d \to \infty$ to get the result. More specifically,

$$\begin{aligned} G &= \ln\left(\frac{\left\|z^d\right\|^2/n}{\|x\|^2/n_{in}}\right) \\ &= \sum_{\ell=0}^{d-1} \ln\left(\frac{2}{n}\left\|\varphi\left(g^\ell\right)\right\|^2\right) + \ln\left(\frac{1}{n}\left\|g^d\right\|^2\right) \end{aligned}$$

Each term is approximately[1] Gaussian in the limit $n \to \infty$,

$$\begin{aligned} \frac{1}{n}\|g\|^2 &\approx \mathcal{N}\left(1, \frac{2}{n}\right) \\ \frac{2}{n}\|\varphi(g)\|^2 &\approx \mathcal{N}\left(1, \frac{5}{n}\right) \end{aligned}$$

Since the variance is small, these are highly concentrated around the mean 1, by the taking the log, and using the Taylor series expansion of $\ln(1+x) \approx x - \frac{1}{2}x^2 + O(x^3)$ to get the approximation $\ln\left(\mathcal{N}(1, \frac{\sigma^2}{n})\right) \approx \mathcal{N}(-\frac{1}{2}\frac{\sigma^2}{n}, \frac{\sigma^2}{n})$ we then get

$$\begin{aligned} \ln\left(\frac{1}{n}\|g\|^2\right) &\approx \mathcal{N}\left(-\frac{1}{2}\left(\frac{2}{n}\right), \frac{2}{n}\right) \\ \ln\left(\frac{2}{n}\|\varphi(g)\|^2\right) &\approx \mathcal{N}\left(-\frac{1}{2}\left(\frac{5}{n}\right), \frac{5}{n}\right) \end{aligned}$$

Finally then,

$$\begin{aligned} G &= \ln\left(\frac{1}{n}\left\|g^d\right\|^2\right) + \sum_{\ell=0}^{d-1} \ln\left(\frac{2}{n}\left\|\varphi(g^\ell)\right\|^2\right) \\ &\approx \mathcal{N}\left(-\frac{1}{2}\left(\frac{2}{n}\right), \frac{2}{n}\right) + \sum_{\ell=0}^{d-1} \mathcal{N}\left(-\frac{1}{2}\left(\frac{5}{n}\right), \frac{5}{n}\right) \\ &\approx \mathcal{N}\left(-\frac{1}{2}\beta, \beta\right) \text{ where } \beta = \frac{2}{n} + 5\frac{d}{n} \end{aligned}$$

3 Detailed proof

3.1 Behavior of $\|g\|^2$ and $\|\varphi(g)\|^2$

Remark. $\|g\|^2 \stackrel{d}{=} \chi^2_n$ is a Chi-squred distribution with n degrees of freedom. Similarly, since $\varphi(g)_i^2 = g_i^2 \mathbb{1}\{g_i > 0\}$ and since $\mathbb{1}\{g_i > 0\}$ is a Bernoulli random variable which is independent of g_i^2, we see that $\|\varphi(g)\|^2 \stackrel{d}{=} \chi^2_{Bin(n,\frac{1}{2})}$ is a Chi-squared distribution with a number of degrees of freedom which is an independent Binomial$(n, \frac{1}{2})$ random variable. These observations can be used to prove the results in this section, but the proofs below are self contained and just use the fact that $\|g\|^2$ and $\|\varphi(g)\|^2$ are sums of n simple independent random variables.

[1] For our result, we actually only need the mean and variance of each term and that each term is concentrated around its mean, but the Gaussian approximation is an intuitive way to understand what's going on here. These approximations hold because $\|g\|^2$ and $\|\varphi(g)\|^2$ can be written as sums over the n independent components of the vectors. The next section has detailed proofs of what we precisely need.

Proposition 3. *Let $g \in \mathbb{R}^n$ be a Gaussian vector with iid $\mathcal{N}(0,1)$ entries. $\frac{1}{n}\|g\|^2 \approx \mathcal{N}(1, \frac{2}{n})$ in the sense that the following properties hold as $n \to \infty$* [2]

i) $\mathbf{E}\left[\frac{1}{n}\|g\|^2\right] = 1$

ii) $\mathbf{Var}\left[\frac{1}{n}\|g\|^2\right] = \frac{2}{n}$

iii) $\sqrt{n}\left(\frac{1}{n}\|g\|^2 - 1\right) \Rightarrow \mathcal{N}(0,2)$ *as* $n \to \infty$.

iv) $\mathbf{E}\left[\left(\frac{1}{n}\|g\|^2 - 1\right)^3\right] = O\left(\frac{1}{n^2}\right)$ *and for* $p \geq 2$ $\mathbf{E}\left[\left(\frac{1}{n}\|g\|^2 - 1\right)^{2p}\right] = O\left(\frac{1}{n^p}\right)$

Proof. All these facts are consequences of the fact that $\frac{1}{n}\|g\|^2 = \frac{1}{n}\sum_{i=1}^n g_i^2$ is an **average** of n independent random variables. i) and ii) then follow since $\mathbf{E}\left[g_i^2\right] = 1$ and $\mathbf{E}\left[g_i^4\right] = 3$. iii) can then be deduced from the central limit theorem. iv) This is a standard moment estimate which is sometimes proven in the moment method proof of the CLT. Notice that $\frac{1}{n}\|g\|^2 - 1 = \frac{1}{n}\sum_{i=1}^n \left(g_i^2 - 1\right)$ is an average of independent mean zero quantities that have finite moments of all order. Hence, $\mathbf{E}\left(\frac{1}{n}\|g\|^2 - 1\right)^3 = \frac{1}{n^2}\mathbf{E}\left[\left(g_i^2 - 1\right)^3\right]$ since all the cross terms in the expansion vanish leaving only the diagonal terms. Similar cancellations occur in the higher moments to give the bound for the $2p$-th moment.

□

Proposition 4. *Let $g \in \mathbb{R}^n$ be a Gaussian vector with iid $\mathcal{N}(0,1)$ entries. The approximations $\frac{2}{n}\|\varphi(g)\|^2 \approx \mathcal{N}(1, \frac{5}{n})$ holds in the following precise senses*

i) $\mathbf{E}\left[\frac{2}{n}\|\varphi(g)\|^2\right] = 1$

ii) $\mathbf{Var}\left[\frac{2}{n}\|\varphi(g)\|^2\right] = \frac{5}{n}$

iii) $\sqrt{n}\left(\frac{2}{n}\|\varphi(g)\|^2 - 1\right) \Rightarrow \mathcal{N}(0,5)$ *as* $n \to \infty$.

iv) $\mathbf{E}\left[\left(\frac{1}{n}\|\varphi(g)\|^2 - 1\right)^3\right] = O\left(\frac{1}{n^2}\right)$ *and for* $p \geq 2$ $\mathbf{E}\left[\left(\frac{1}{n}\|\varphi(g)\|^2 - 1\right)^{2p}\right] = O\left(\frac{1}{n^p}\right)$

Proof. As in the previous result, this is a consequence of the fact that we average iid random variables, $\frac{2}{n}\|\varphi(g)\|^2 = \frac{1}{n}\sum_{i=1}^n g_i^2 2\mathbf{1}\{g_i > 0\}$. Note additionally that g_i^2 is independent of $\mathbf{1}\{g_i > 0\}$ since g_i is symmetrically distributed which allows us to compute directly. i) and ii) follow since $\mathbf{E}\left[g_i^2 2\mathbf{1}\{g_i > 0\}\right] = 1 \cdot 2 \cdot \frac{1}{2}$ and $\mathbf{E}\left[\left(g_i^2\right)^2 \left(2\mathbf{1}\{g_i > 0\}\right)^2\right] = 3 \cdot 4 \cdot \frac{1}{2} = 6$.iii) follows by the application of the central limit theorem. iv) Follows again by the same argument as the previous lemma. □

3.2 Behavior of $\ln(\|g\|^2)$ and $\ln\left(\|\varphi(g)\|^2\right)$

Proposition 5. $\ln\left(\frac{1}{n}\|g\|^2\right)$ *obeys i)* $\mathbf{E}\left[\ln\left(\frac{1}{n}\|g\|^2\right)\right] = -\frac{1}{2}\left(\frac{2}{n}\right) + O\left(\frac{1}{n^2}\right)$, *ii)* $\mathbf{Var}\left[\ln\left(\frac{1}{n}\|g\|^2\right)\right] = \frac{2}{n} + O\left(\frac{1}{n^2}\right)$, *iii)* $\mathbf{E}\left[\left(\ln\left(\frac{1}{n}\|g\|^2\right) - \mathbf{E}\left[\ln\left(\frac{1}{n}\|g\|^2\right)\right]\right)^4\right] = O\left(\frac{1}{n^2}\right)$

Similarly $\ln\left(\frac{1}{n}\|\varphi(g)\|^2\right)$ *obeys i)* $\mathbf{E}\left[\ln\left(\frac{1}{n}\|\varphi(g)\|^2\right)\right] = -\frac{1}{2}\left(\frac{5}{n}\right) + O\left(\frac{1}{n^2}\right)$, *ii)* $\mathbf{Var}\left[\ln\left(\frac{1}{n}\|\varphi(g)\|^2\right)\right] = \frac{5}{n} + O\left(\frac{1}{n^2}\right)$, *iii)* $\mathbf{E}\left[\left(\ln\left(\frac{1}{n}\|\varphi(g)\|^2\right) - \mathbf{E}\left[\ln\left(\frac{1}{n}\|\varphi(g)\|^2\right)\right]\right)^4\right] = O\left(\frac{1}{n^2}\right)$

Proof. This follows by doing the Taylor series $\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + O(x^5)$ and plugging in $x = \frac{1}{n}\|g\|^2 - 1$ to obtain an approximation for $\ln\left(\frac{1}{n}\|g\|^2\right)$. Specifically, by Chebyshev inequaltiy we know that $\left|\frac{1}{n}\|g\|^2 - 1\right| = O\left(n^{-\frac{1}{2}+\epsilon}\right)$ with probability at least $1 - n^{\epsilon}$. On this event, the Taylor series expansion yields an error of no more that $O\left(n^{-\frac{5}{2}+5\epsilon}\right)$. The result follows by taking $\mathbf{E}$ and $\mathbf{Var}$ and using the bounds from the previous lemma to control all the higher order terms that appear. □

[2] Note that $\|g\|^2$ is a χ_n^2 distribution, so these are properties which one can look up. However, we give an elementary proof here.

3.3 Lindeberg CLT

By the work in the proof sketch section, all we have to do is apply a central limit theorem to the sum

$$G = \ln\left(\frac{\left\|g^d\right\|^2}{n}\right) + \sum_{\ell=0}^{d-1} \ln\left(\frac{\left\|\varphi\left(g^\ell\right)\right\|^2}{n}\right)$$

Since all the variables are independent, the terms of the sum form a triangular array of independent random variables. The mean and variance are as claimed by the previous mean and variance calculations. The Lindeberg CLT will then tell us that G converges to a Gaussian as claimed. The Lyapanov condition for the Lindeberg CLT for the 4th moment is verified by our previous inequalities.

References

[HN19a] Boris Hanin and Mihai Nica, *Finite depth and width corrections to the neural tangent kernel*, Int. Conf. Learning Representations (ICLR), 2019.

[HN19b] ———, *Products of many large random matrices and gradients in deep neural networks*, Communications in Mathematical Physics (2019), 1–36.

[LNR21] Mufan Li, Mihai Nica, and Daniel Roy, *The future is log-gaussian: Resnets and their infinite-depth-and-width limit at initialization*, 2021.