Precise characterization of the prior predictive distribution of deep ReLU networks

Lorenzo Noci [*1] **Gregor Bachmann** [*1] **Kevin Roth** [*1] **Sebastian Nowozin** [2] **Thomas Hofmann** [1]

Abstract

Recent works on Bayesian neural networks (BNNs) have highlighted the need to better understand the implications of using Gaussian priors in combination with the compositional structure of the network architecture. Similar in spirit to the kind of analysis that has been developed to devise better initialization schemes for neural networks (cf. He- or Xavier initialization), we derive a precise characterization of the prior predictive distribution of finite-width ReLU networks with Gaussian weights. Our analysis, based on the Meijer-G function, allows us to quantify the influence of architectural choices such as the width or depth of the network on the resulting shape of the prior predictive distribution. We also formally connect our results to previous work in the infinite width setting, demonstrating that the moments of the distribution converge to those of a normal log-normal mixture in the infinite depth limit. Finally, our results provide valuable guidance on prior design: for instance, controlling the predictive variance with depth- and width-informed priors on the weights of the network.

1. Introduction

It is well known that standard neural networks initialized with Gaussian weights tend to Gaussian processes (Rasmussen, 2003) in the infinite width limit (Neal, 1996; Lee et al., 2017; de G. Matthews et al., 2018), coined *neural network Gaussian process* (NNGP) in the literature. Although the NNGP has been derived for a number of architectures, such as convolutional (Novak et al., 2019; Garriga-Alonso et al., 2019), recurrent (Yang, 2019) and attention mechanisms (Hron et al., 2020), little is known about the finite width case.

One of the main motivations for our work is to better understand the implications of using Gaussian priors in combination with the compositional structure of the network architecture. As argued by (Wilson and Izmailov, 2020; Wilson, 2020), the prior over parameters does not carry a meaningful interpretation; the prior that ultimately matters is the *prior predictive distribution* that is induced when a prior over parameters is combined with a neural architecture (Wilson and Izmailov, 2020; Wilson, 2020).

Studying the properties of this prior predictive distribution is not an easy task, the main reason being the *compositional* structure of a neural network, which ultimately boils down to products of random matrices with a given (non-linear) activation function. The main tools to study such products are the Mellin transform and the Meijer-G function (Meijer, 1936; Springer and Thompson, 1970; Mathai, 1993; Stojanac et al., 2017), both of which will be leveraged in this work to gain theoretical insights into the inner workings of BNN priors.

Contributions Our results provide an important step towards understanding the interplay between architectural choices and the distributional properties of the prior predictive distribution, in particular:

- We characterize the prior predictive density of finite-width ReLU networks of any depth through the framework of Meijer-G functions (Section 3).
- We draw analytic insights about the shape of the distribution by studying its moments and the resulting heavy-tailedness. We disentangle the roles of width and depth, demonstrating how deeper networks become more heavy-tailed, while wider networks induce more Gaussian-like distributions. We extend this observation to the infinite-width setting, observing how different limiting behaviours emerge as we take the infinite-depth limit (Section 4).
- Finally, we introduce generalized He priors, where a desired variance can be directly specified in the function space. This allows the practitioner to make an

[*]Equal contribution [1]Dept of Computer Science, ETH Zürich [2]Microsoft Research, Cambridge, UK. Correspondence to: Lorenzo Noci <lorenzo.noci@inf.ethz.ch>, Gregor Bachmann <gregor.bachmann@inf.ethz.ch>, Kevin Roth <kevin.roth@inf.ethz.ch>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

interpretable choice for the variance, instead of implicitly tuning it through the specification of each layer variance (Section 5).

2. Background

2.1. Fully Connected Neural Network

Given an input $\boldsymbol{x} \in \mathbb{R}^d$, we define a $\boldsymbol{\theta}$-parameterized L layer fully-connected neural network $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ through the recursive equations:

- $f_k^{(l)}(\boldsymbol{x}) = \sum_{j=1}^{d_{l-1}} W_{jk}^{(l)} g_j^{(l-1)}(\boldsymbol{x})$
- $\boldsymbol{g}^{(l)}(\boldsymbol{x}) = \sigma(\boldsymbol{f}^{(l)}(\boldsymbol{x}))$

where $W_{ij}^{(l)} \sim \mathcal{N}(0, \sigma_l^2)$ are normal initialized weights and $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linearity, applied component-wise. We refer to $\boldsymbol{f}^{(l)}(\boldsymbol{x}) \in \mathbb{R}^{d_l}$ as the *pre-activations* and to $\boldsymbol{g}^{(l)}(\boldsymbol{x}) \in \mathbb{R}^{d_l}$ as the *post-activations*. When it is clear from the context, to enhance readability, we will occasionally abuse notation denoting $\boldsymbol{f}^{(l)} := \boldsymbol{f}^{(l)}(\boldsymbol{x})$ and $\boldsymbol{g}^{(l)} := \boldsymbol{g}^{(l)}(\boldsymbol{x})$.

2.2. Meijer-G function

The Meijer-G function is the central tool to our analysis of the predictive prior distribution in the finite-width regime. The Meijer-G function is a ubiquitous tool, appearing in a variety of scientific fields ranging from mathematical physics (Pishkoo and Darus, 2013) to symbolic integration software (Adamchik and Marichev, 1990) to electrical engineering (Ansari et al., 2011). Despite its high popularity in many technical fields, there have only been a handful of works in ML leveraging this elegant and convenient theoretical framework (Alaa and van der Schaar, 2019; Crabbe et al., 2020). In the following, we will introduce the Meijer-G function along with the relevant mathematical tools to develop our theory.

For $s \in \mathbb{C}$ with $\Re(s) > 0$, denote by $\Gamma(s)$ the Gamma function defined as

$$\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx \tag{1}$$

Now fix $l \in \mathbb{N}$ and consider $\boldsymbol{b} \in \mathbb{R}^l$. The Meijer-G function (Meijer, 1936) $G_{0,l}^{l,0}(z|\boldsymbol{b})$ is defined as

$$G_{0,l}^{l,0}(z|\boldsymbol{b}) = \frac{1}{2\pi i} \int_C z^{-s} \prod_{j=1}^{l} \Gamma(b_j + s) ds, \tag{2}$$

where the integration path $\mathcal{L}$ defines a suitable complex curve. For more detail on $\mathcal{L}$ and the more general Meijer-G function, we refer to the Appendix A. Small values for $l \in \mathbb{N}$ correspond to familiar functions such as the exponential $G_{0,1}^{1,0}(z|0) = e^{-z}$ and the modified Bessel function of second kind $G_{0,2}^{2,0}\left(\frac{z^2}{4}\middle|[\frac{\nu}{2}, -\frac{\nu}{2}]\right) = 2K_\nu(z)$.

The defining property of Meijer-G functions is their closure under integration, i.e. the convolution of two Meijer-G functions is again a Meijer-G function. Combined with the fact that most elementary functions can be written as a Meijer-G function, this property becomes extremely powerful to express complicated integrals neatly. Our proofs leverage this result extensively by expressing the integrands encountered in the prior predictive distribution as Meijer-G functions.

3. Predictive Priors for Neural Networks

In this section we detail our theoretical results on the predictive prior distribution implied by a fully-connected neural network with Gaussian weights and ReLU non-linearity.

The proof technique used is very similar to its linear counterpart (see Appendix C), the main difference stemming from the need to decompose the distribution over active and inactive ReLU cells. As a consequence, the resulting density is a superposition of different Meijer-G functions, each associated with a different active (linear) subnetwork. This is presented in the following:

Theorem 3.1. *Suppose $l \geq 2$, and the input has dimension d_0. Define the multi-index set $\mathcal{R} = [d_1] \times \cdots \times [d_{l-1}]$ and introduce the vector $\boldsymbol{u^r} \in \mathbb{R}^{l-1}$ through its components $\boldsymbol{u}_i^{\boldsymbol{r}} = \frac{1}{2}(d_i - r_i - d_l)$.*

$$p(\boldsymbol{f}_{ReLU}^{(l)}) = \sum_{\boldsymbol{r} \in \mathcal{R}} q_{\boldsymbol{r}} G_{0,l}^{l,0}\left(\frac{||\boldsymbol{f}_{ReLU}^{(l)}||^2}{2^l \sigma^2}\middle| \boldsymbol{u^r}\right) + q_0 \delta(\boldsymbol{f}_{ReLU}^{(l)}) \tag{3}$$

where $\sigma^2 = \prod_{i=1}^{l} \sigma_i^2$ and $q_{\boldsymbol{r}}, q_0 \in \mathbb{R}$ are weights.

We refer to Appendix D.2 for the detailed proof and the closed-form expression for $q_{\boldsymbol{r}}$, which governs how much weight is assigned to the corresponding Meijer-G function.

4. Analytic Insights into the Prior Predictive

While Meijer-G functions are not particularly suited for direct practical applications, here we highlight how one can use their mathematical machinery to derive interesting insights, relying on numerous mathematical results provided in the literature (Gradshteyn and Ryzik, 2013; Brychkov, 2008; Andrews, 2011). As a consequence, one can easily move from the rather abstract but mathematically convenient world of Meijer-G functions to very concrete results.

We demonstrate this by recovering and extending the NNGP results provided in Lee et al. (2018); Matthews et al. (2018). In particular, our analysis allows for simultaneous width and depth limits, showing how different limiting distributions emerge as a consequence of the employed growth model for L. Finally, we characterize the heavy-tailedness of the prior-predictive for any width, providing further evidence that deeper models induce distributions with heavier tails, as observed in Vladimirova et al. (2019).

We hope that this work paves the way for more theoretical progress in understanding priors, leveraging this novel connection through the rich literature on Meijer-G functions.

4.1. Infinite Width: Recovering and Extending the NNGP for a Single Datapoint

We will give an alternative proof for the Gaussian behaviour emerging as the width of the network tends to infinity, recovering the results of Lee et al. (2018); Matthews et al. (2018) in the restricted setting of having just one fixed input $\boldsymbol{x}$. Using the same technique, we extend these findings to infinite-depth. For ease of exposition, we focus on the equal width case, i.e. $d_1 = \dots = d_{l-1} = m$ with one output $d_L = 1$. To have well-defined limits, one has to resort to the so-called NTK parametrization (Jacot et al., 2018), which is achieved by setting the variances as $\sigma_1^2 = 1$ and $\sigma_i^2 = \frac{2}{m}$ for $i = 2, \dots, l$. We summarize the result in the following:

Theorem 4.1. *Consider the distribution of the output $p\left(f_{ReLU}^{(L)}\right) = p^m\left(f_{ReLU}^{(L)}\right)$, as defined in Thm. D.2. Let the depth grow as $L = c + \gamma m^\beta$ where $\beta, \gamma \geq 0$ and $c \in \mathbb{N}$ fixed. Then it holds that*

$$\mathbb{E}\left[\left(f^{(L)}\right)^{2k}\right] \xrightarrow{m \to \infty} \begin{cases} (2k-1)!! & \text{if } \beta < 1 \\ e^{\frac{5}{2}\gamma k(k-1)}(2k-1)!! & \text{if } \beta = 1 \\ \infty & \text{if } \beta > 1 \end{cases} \tag{4}$$

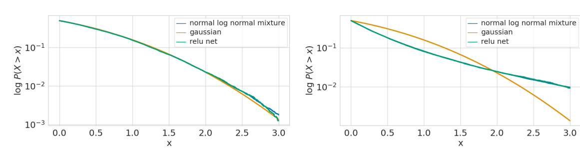
where $(2k-1)!! = (2k-1)\dots 3 \cdot 1$ denotes the double factorial (by symmetry, odd moments are zero). Moreover, for $\beta < 1$ it holds that

$$p^m(f^{(L)}) \xrightarrow{d} \mathcal{N}(0,1) \quad \text{for } m \to \infty \tag{5}$$

The proof uses the fact that we can easily calculate all the moments of $f_{\text{ReLU}}^{(L)}$ at finite width using well-known identities involving the Meijer-G function. For the limit, care has to be taken as the moments involve the moments of the Binomial distribution $\text{Bin}(m, \frac{1}{2})$ for which recursive but no analytic formulas are available. By carefully studying the coefficients of the this polynomial, we obtain the result.

This result is in stark contrast to Lee et al. (2018); Matthews et al. (2018) which cannot deal with a simultaneous depth limit due to the inductive nature of their proof, but can only describe the special case of fixed depth ($\gamma = 1, \beta = 0$). Notice that for depth growing asymptotically at a slower rate than width ($L = \gamma m^\beta$ for $\beta < 1$), we obtain the convergence in distribution since the Gaussian distribution is identified by its moments (See Theorem 30.2 in Billingsley (1986)). Denote $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{LN}(-\frac{5}{4}\gamma, \frac{5}{4}\gamma)$ for $X \perp Y$ and the resulting normal log-normal mixture by $Z = XY$. For $\beta = 1$, we can also prove convergence of the moments and identify them as arising from the normal log-normal mixture Z (Yang, 2008). Unfortunately, this does not suffice to conclude convergence in distribution as there are no known results on the identifiability of Z. Empirical evidence however, presented in Figure 3, suggests that the normal log-normal mixture captures the distribution to an excellent degree of fidelity.

The decoupling of the moments into two separate factors gives insights into the role of width and depth on the shape of the distribution. The emergence of the log-normal factor in this infinite-depth limit highlights how deeper networks encourage more heavy-tailed distribution, if not countered by a sufficient amount of width. This becomes even more drastic once depth outgrows width ($\beta > 1$), leading to divergence of all even moments. In the following section we will show that also in the finite width regime, heavy-tailedness becomes more pronounced as we increase the depth.



(a) $\gamma = 0.01$ (b) $\gamma = 0.5$

Figure 1: **Convergence to normal log-normal mixture:** We display the CDFs of a ReLU network, a Gaussian and the normal log-normal mixture. We fix $m = 100$, while L ranges from 1 to 50 using the parameter γ. The neural network CDF is constructed by drawing 10^4 samples from the prior. Note that as L increases, the CDF departs from the Gaussian but still follows the normal log-normal mixture.

4.2. Heavy-tailedness Increases with Depth

From the moments analysis, it is simple to recover a known fact about the prior distribution of neural networks, namely that deeper layers are increasingly heavy-tailed. To see this, we can derive the kurtosis, a standard measure of tailedness of the distribution (Westfall, 2014), defined as:

$$\kappa := \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^4\right]}{\mathbb{V}[X]^2}, \tag{6}$$

where X is a univariate random variable. We can calculate the kurtosis of a ReLU network analytically at any width, relying on closed-form results for the lower order moments of the Binomial distribution. We outline the exact calculation in the Appendix D and state the resulting expression:

$$\kappa_{\text{ReLU}}(m, L) = 3\left(\frac{m+5}{m}\right)^{L-1} \tag{7}$$

Note how the κ_{ReLU} increases with L and decreases with m, highlighting once again the opposite roles those two parameters take regarding the shape of the prior. As expected, for fixed depth L, the kurtosis converges to 3, which is the kurtosis of a standard Gaussian variable $\mathcal{N}(0, 1)$. For the simultaneous limit $L = \gamma m$, we find a value of $3e^{5\gamma}$, which exactly matches with the expression derived in Thm. 4.1.

5. Prior Design

In Section 4.2, we have already outlined how the choice of architecture influences the heavy-tailedness of the distribution. If a Gaussian-like output is desired, the architecture should be designed in such a way that the width m significantly exceeds the depth L (small γ), while heavy-tailed priors can be achieved by considering regimes where L exceeds m (big γ). Another consequence of our analysis is that we can now directly work with the variance in the function space, instead of implicitly tuning it by changing the variances at each layer. In this way, the variance over the weights has a clear interpretation in terms of predictive variance, making it easier for the deep learning practitioner to take an informed decision when designing a prior for BNNs. The derivation is reminiscent of the analysis for initialization schemes for neural networks, such as "He-initialization" (He et al., 2015) in the Gaussian case and "Xavier initialization" (Glorot and Bengio, 2010) for uniform weight initializations. As a consequence, we coin the resulting prior *Generalized He-prior*.

Predictive variance: We consider the parametrization $\sigma_1^2 = 1$ and $\sigma_i^2 = \frac{2t_i^2}{m}$, which, we show in Appendix D, leads to the following predictive variance:

$$\mathbb{V}[Z] = \sigma_1^2 \prod_{i=2}^{l} t_i^2 \tag{8}$$

Suppose we want a desired output variance σ^2. This can be achieved as follows: let a_i, with $i = 2, \dots, l$ be $l-1$ coefficients such that $\sum_{i=1}^{l-1} a_i = l - 1$. Then choose $t_i^2 = \left(\sigma^2\right)^{\frac{a_i}{l-1}}$, implying that the layer variances σ_i^2 are given as

$$\sigma_i^2 = \frac{2}{m}\left(\sigma^2\right)^{\frac{a_i}{l-1}}, \tag{9}$$

He-priors are a particular case when $\sigma^2 = 1$ and $a_i = 1, i = 1, \dots, l-1$, while a standard Gaussian $N(0, 1)$ prior results in $\sigma^2 = \left(\frac{m}{2}\right)^{l-1}$, $\sigma_1^2 = 1$, and $a_i = 1$, $i = 1, \dots, l-1$. Note how "far" He-priors can be from standard Gaussian for relatively deep and wide neural networks. While rather innocent-looking, using a standard Gaussian prior can lead to very high variance.

Combining with the previous insights, the practitioner can now choose a desired output variance along with a level of heavy-tailedness by specifying the architecture.

6. Discussion

We have shed light on the shape of the prior predictive distribution arising from imposing a Gaussian distribution on the weights. Leveraging the machinery of Meijer-G functions, we characterized the density in the finite width regime and derived analytic insights into its properties such as moments and heavy-tailedness. An extension to the stochastic process setting in the spirit of Lee et al. (2017) as well as to the convolutional architectures (Novak et al., 2019) could bring theory even closer to practice and we expect similar results to also hold in those cases. This is however beyond the scope of our work and we leave it as future work.

Our technique enabled us to extend the NNGP framework to infinite depth, discovering how in the more general case, the resulting distribution shares the same moments as a normal log-normal distribution. This allowed us to disentangle the roles of width and depth, where the former induces a Gaussian-like distribution while the latter encourages heavier tails. Empirically, we found that the normal log-normal mixture provides an excellent fit to the true distribution even in the non-asymptotic setting. This surprising observation begs further theoretical and empirical investigations. In particular, discovering a suitable stochastic process incorporating the normal log-normal mixture, could lend further insights into the inner workings of neural networks. Moreover, the role of heavy-tailedness regarding generalization is very intriguing, potentially being an important reason underlying the gap between infinite-width networks and their finite counterparts. This is also in-line with recent empirical works on priors, suggesting that heavy-tailed distributions can increase the performance (Fortuin et al., 2021).

Using these insights, we described how one can choose a fixed prior variance directly in the output space along with a desired level of heavy-tailedness resulting from the choice of the architecture. We leave it as future work to also consider higher moments to give a more nuanced control over the resulting prior in the function space.

Finally, we hope that the introduction of the Meijer-G function sparks more theoretical research on BNN priors and their implied inductive biases in function space.

References

Adamchik, V. S. and Marichev, O. I. (1990). The algorithm for calculating integrals of hypergeometric type functions and its realization in reduce system. In *Proceedings of the International Symposium on Symbolic and Algebraic Computation*, ISSAC '90, page 212–224, New York, NY, USA. Association for Computing Machinery.

Alaa, A. M. and van der Schaar, M. (2019). Demystifying black-box models with symbolic metamodels. In *NeurIPS*, pages 11301–11311.

Andrews, L. C. (2011). *Field guide to special functions for engineers / Larry C. Andrews.* Field guide series ; 18. SPIE, Bellingham, Wash.

Ansari, I. S., Al-Ahmadi, S., Yilmaz, F., Alouini, M.-S., and Yanikomeroglu, H. (2011). A new formula for the ber of binary modulations with dual-branch selection over generalized-k composite fading channels. *IEEE Transactions on Communications*, 59(10):2654–2658.

Beals, R. and Szmigielski, J. (2013). Meijer g-functions: a gentle introduction. *Notices of the American Mathematical Society*, 60:866–873.

Benyi, A. (2005). A recursive formula for moments of a binomial distribution. *The college mathematics journal*.

Billingsley, P. (1986). *Probability and Measure*. John Wiley and Sons, second edition.

Boros, G. and Moll, V. (2004). *Irresistible Integrals: Symbolics, Analysis and Experiments in the Evaluation of Integrals*. Cambridge University Press.

Brychkov, Y. (2008). Handbook of special functions: Derivatives, integrals, series and other formulas. *Chapman and Hall/CRC*.

Crabbe, J., Zhang, Y., Zame, W., and van der Schaar, M. (2020). Learning outside the black-box: The pursuit of interpretable models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17838–17849. Curran Associates, Inc.

de G. Matthews, A. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *ICLR*.

Fortuin, V., Garriga-Alonso, A., Wenzel, F., Rätsch, G., Turner, R., van der Wilk, M., and Aitchison, L. (2021). Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. (2019). Deep convolutional networks as shallow gaussian processes. In *ICLR*.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Gradshteyn, I. and Ryzik, I. (2013). Table of integrals, series, and products. *Elsevier/Academic Press*.

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

Hron, J., Bahri, Y., Sohl-Dickstein, J., and Novak, R. (2020). Infinite attention: Nngp and ntk for deep attention networks. In *International Conference on Machine Learning*, pages 4376–4386. PMLR.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *32rd Conference on Neural Information Processing Systems (NeurIPS)*.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2017). Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2018). Deep neural networks as gaussian processes. In *ICLR*.

Mathai, A. M. (1993). *A handbook of generalized special functions for statistical and physical sciences*. Oxford University Press, USA.

Mathai, A. M. and Saxena, R. K. (2006). *Generalized hypergeometric functions with applications in statistics and physical sciences*, volume 348. Springer.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*.

Meijer, C. (1936). Uber whittakersche bezw. besselsche funktionen und deren produkte. *Nieuw Archief voor Wiskunde*, 18(2):10–29.

Neal, R. M. (1995). *Bayesian learning for neural networks*. PhD thesis, University of Toronto.

Neal, R. M. (1996). Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer.

Novak, R., Xiao, L., Bahri, Y., Lee, J., Yang, G., Hron, J., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. (2019). Bayesian deep convolutional networks with many channels are gaussian processes. In *ICLR*.

Novak, R., Xiao, L., Lee, J., Bahri, Y., Yang, G., Hron, J., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. (2018). Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*.

Pishkoo, A. and Darus, M. (2013). Some applications of meijer g-functions as solutions of differential equations in physical models. *Journal of Mathematical Physics, Analysis, Geometry*.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer.

Springer, M. D. and Thompson, W. E. (1970). The distribution of products of beta, gamma and gaussian random variables. *SIAM Journal on Applied Mathematics*, 18(4):721–737.

Stojanac, Z., Suess, D., and Kliesch, M. (2017). On products of gaussian random variables. *arXiv preprint arXiv:1711.10516*.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. (2019). Understanding priors in bayesian neural networks at the unit level. In *International Conference on Machine Learning*, pages 6458–6467. PMLR.

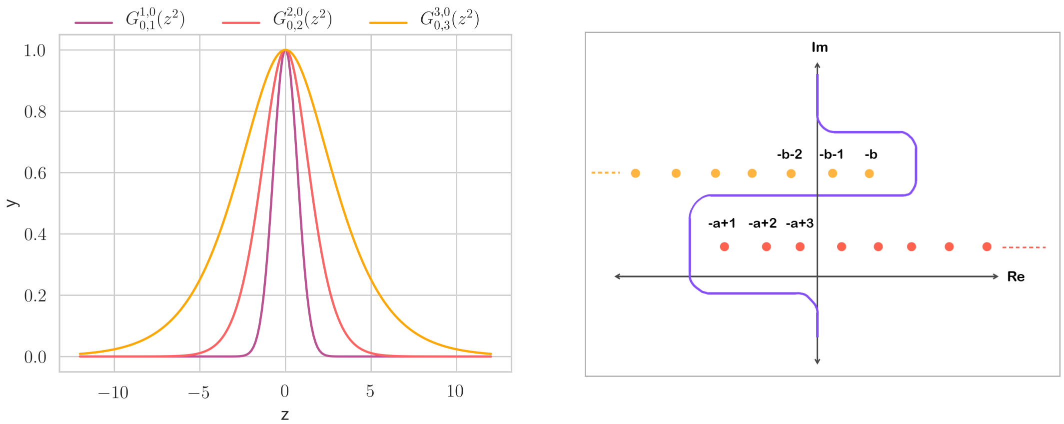
Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. rip. *The American Statistician*, 68(3):191–195.

Wilson, A. G. (2020). The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*.

Wilson, A. G. and Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*.

Yang, G. (2019). Wide feedforward or recurrent neural networks of any architecture are gaussian processes.

Yang, M. (2008). Normal log-normal mixture, leptokurtosis and skewness. *Applied Economics Letters*, 15(9):737–742.



(a) Plots of Meijer-G functions

(b) Visualization of integration path

Figure 2: (a) Plot of the Meijer-G functions $G^{l,0}_{0,l}(\cdot|\boldsymbol{b})$ for $l = 1, 2, 3$ and $\boldsymbol{b} = 0, (0, 5), (0, 5, 5)$. (b) An example path $\mathcal{L}$ in the complex plane. Notice how the orange singularities are always to the left of $\mathcal{L}$ and the red ones always to the right.

A. More Background and Additional Empirical Results

A.1. General Meijer-G Function

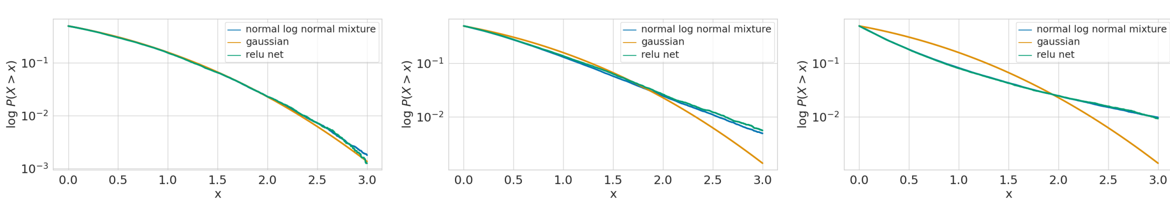
Here we introduce the general Meijer-G function for completeness. Fix $m, n, p, q \in \mathbb{N}$ such that $0 \leq m \leq q$ and $0 \leq n \leq p$ and consider $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{b} \in \mathbb{R}^q$ such that $a_i - b_j \notin \mathbb{Z}_{>0}$ $\forall i = 1, \dots, p$ and $j = 1, \dots, q$. The Meijer-G function (Meijer, 1936; Mathai, 1993; Mathai and Saxena, 2006), is defined as:

$$G^{m,n}_{p,q}\left(z \middle| \begin{smallmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{smallmatrix}\right) = \frac{1}{2\pi i} \int_{\mathcal{L}} \chi(s) z^{-s} ds, \tag{10}$$

where:

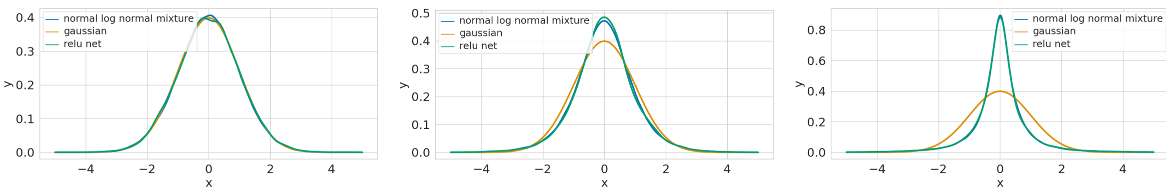
$$\chi(s) = \frac{\prod_{j=1}^{m} \Gamma(b_j + s) \prod_{j=1}^{n} \Gamma(1 - a_j - s)}{\prod_{j=m+1}^{q} \Gamma(1 - b_j - s) \prod_{j=n+1}^{p} \Gamma(a_j + s)}. \tag{11}$$

and the integration path $\mathcal{L}$ defines a suitable complex curve, described in the following. Recall that the function $\Gamma(s)$ has poles at $0, -1, -2, \dots$ all the way to $-\infty$. Hence, $\Gamma(1 - a_j - s)$ has poles $-a_j + 1, -a_j + 2, \dots,$ all the way to ∞, and $\Gamma(b_j + s)$ has poles $-b_j, -b_j - 1, \dots,$ all the way to $-\infty$. The path $\mathcal{L}$ is defined such that the poles of $\Gamma(b_j + s)$ are to the left of $\mathcal{L}$ and the ones of $\Gamma(1 - a_j - s)$ are to the right of it. The condition $a_i - b_j \notin \mathbb{Z}_{>0}$ makes sure we can find such a separation as this implies that the poles do not overlap. We illustrate an example of such a path in Figure 2b. In red we display the poles of $\Gamma(1 - a_j + s)$ and in orange the poles of $\Gamma(b_j + s)$. For interested readers we refer to Beals and Szmigielski (2013) for a more extensive introduction. and the integration path $\mathcal{L}$ defines a suitable complex curve,



(a) $\gamma = 0.01$ (b) $\gamma = 0.1$ (c) $\gamma = 0.5$

Figure 3: **Convergence to normal log-normal mixture:** We display the CDFs of a ReLU network, standard Gaussian CDF and the normal log-normal mixture. The width is fixed to 100 while depth is increased from 1 to 50 using the parameter γ. The neural network CDF is constructed empirically by drawing 10^4 samples from the prior. Note that as depth increases, the CDF departs from the Gaussian but still follows the normal log-normal mixture.



(a) $\gamma = 0.01$ (b) $\gamma = 0.1$ (c) $\gamma = 0.5$

Figure 4: **Convergence to normal log-normal mixture:** We show the density plots for the same setting as in Figure 3. Note that as depth increases, the density departs from the Gaussian but still follows the normal log-normal mixture.

described in the following. Recall that the function $\Gamma(s)$ has poles at $-1, -2, \dots$ growing to $-\infty$. Hence, $\Gamma(1 - a_j + s)$ has poles $a_j - 1, a_j - 2, \dots$, diverging to $-\infty$, and $\Gamma(b_j - s)$ has poles $b_j, b_j + 1, \dots$, diverging to ∞. The path $\mathcal{L}$ is defined such that the poles of $\Gamma(b_j - s)$ lie to the right of $\mathcal{L}$ and the ones of $\Gamma(1 - a_j + s)$ to the left of it. The condition $a_i - b_j \notin \mathbb{Z}_{>0}$ makes sure we can find such a separation as this implies that the poles do not overlap. We illustrate an example of such a path in Figure 2b. In orange we display the poles of $\Gamma(1 - a_j + s)$ diverging to $-\infty$ and in red the poles of $\Gamma(b_j - s)$. For interested readers we refer to Beals and Szmigielski (2013) for a more extensive introduction.

A.2. Prior Predictive Distribution and Infinite Width

A precise characterization of the prior predictive distribution of a neural network has been established in the so-called infinite-width setting (Neal, 1995; Lee et al., 2018; Matthews et al., 2018). By considering variances that scale inversely with the width, i.e. $\sigma_l^2 = \frac{1}{d_l}$ and fixed depth $L \in \mathbb{N}$, it can be shown that the implied prior predictive distribution converges in law to a Gaussian process:

$$f_{\boldsymbol{\theta}} \xrightarrow{d} \mathcal{GP}(0, \Sigma^{(L)}) \tag{12}$$

where $\Sigma^{(L)} : \mathbb{R}^d \times \mathbb{R}^d, \quad \boldsymbol{x}, \boldsymbol{x}' \mapsto \Sigma^{(L)}(\boldsymbol{x}, \boldsymbol{x}')$ is the NNGP kernel (Lee et al., 2017), available in closed-form through the recursion

$$\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}' \quad , \quad \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}^{(l)})}\left[\sigma(z_1)\sigma(z_2)\right], \tag{13}$$

for $l = 1, \dots, L-1$ and where $\tilde{\boldsymbol{\Sigma}}^{(l)} = \begin{pmatrix} \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) & \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \\ \Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}) & \Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}') \end{pmatrix} \in \mathbb{R}^{2 \times 2}$. The proof relies on the multivariate central limit theorem in conjunction with an inductive technique, letting hidden layer widths go to infinity in sequence. Due to the Gaussian nature of the limit, exact Bayesian inference becomes tractable and techniques from the Gaussian process literature can be readily applied (Rasmussen, 2003). Although theoretically very appealing, from a practical point of view, infinite-width networks are not as relevant due to their inferior performance on a variety of computer vision tasks (Novak et al., 2018). Gaining a better understanding of this performance gap is hence of utmost importance.

A.3. More Empirical Results

Here we present more experiments regarding the normal log-normal mixture and the prior predictive distribution arising in the finite-width regime. In Figure 3 we display the CDFs for more values of γ. Again we observe an excellent match across the different regimes, while the Gaussian approximation starts perform badly as soon as the depth is increased too much. To further assess the distributional match, we repeat the same experiment for the PDFs and observe the same the behaviour in Figure 4.

B. Properties of Meijer-G function

Here we describe the properties of Meijer-G functions which we will use extensively in the following.

The first result concerns the Mellin transform of the Meijer-G function, which will be the key to solve the integrals that we will face later.

Proposition B.1 (Mellin transform of the Meijer G function)**.**

$$\int_0^\infty x^{s-1} G^{m,n}_{p,q}(wx) = w^{-s} \frac{\prod_{j=1}^m \Gamma(b_j + s) \prod_{j=1}^n \Gamma(1 - a_j - s)}{\prod_{j=m+1}^q \Gamma(1 - b_j - s) \prod_{j=n+1}^p \Gamma(a_j + s)} \tag{14}$$

Proof. See Chapter 3.2 of (Mathai and Saxena, 2006). □

To establish the base case $m = 1$, we need the following results.

Proposition B.2. *The following identities hold:*

- $\exp(z) = G^{1,0}_{0,1}\left(-z \middle| 0\right)$
- *multiplication by power property:* $z^d G^{m,n}_{p,q}\left(z\middle|\begin{smallmatrix} a_1 & \dots & a_p \\ b_1 & \dots & b_q\end{smallmatrix}\right) = G^{m,n}_{p,q}\left(z\middle|\begin{smallmatrix} a_1+d & \dots & a_p+d \\ b_1+d & \dots & b_q+d\end{smallmatrix}\right)$

Proof. See Chapter 2.6 of (Mathai and Saxena, 2006) for the first identity. The last property follows directly from the definition of Meijer-G function. □

To perform the inductive step, we will encounter the following integral, that can be expressed in terms of the Meijer-G function.

Proposition B.3.

$$\int_1^\infty x^{-\rho}(x-1)^{\sigma-1} G^{m,n}_{p,q}\left(\alpha x\middle|\begin{smallmatrix} a_1 & \dots & a_p \\ b_1 & \dots & b_q\end{smallmatrix}\right) dx = \Gamma(\sigma) G^{m+1,n}_{p+1,q+1}\left(\alpha\middle|\begin{smallmatrix} & a_1 & \dots & a_p & \rho \\ \rho-\sigma & b_1 & \dots & b_q & \end{smallmatrix}\right) \tag{15}$$

C. Proof for Linear Networks and Derivation of their Moments

Here we collect all the results regarding linear networks, establishing the relevant technical Lemmas to derive the density and calculate the moments of the resulting distribution.

Lemma C.1. *The units of any layer are uncorrelated, i.e.*

$$Cov\left(f_k^{(l)}, f_{k'}^{(l)}\right) = 0 \tag{16}$$

for all layers l, and for all k, $k' \in [d_l]$

Proof.

$$Cov\left(f_k^{(l)}, f_{k'}^{(l)}\right) = Cov\left(\sum_{j=1}^{d_{l-1}} f_j^{(l-1)} W_{jk}^{(l)}, \sum_{j'=1}^{d_{l-1}} f_{j'}^{(l-1)} W_{j'k'}^{(l)}\right) \tag{17}$$

$$= \mathbb{E}\left[\sum_{j=1}^{d_{l-1}} f_j^{(l-1)} W_{jk}^{(l)} \sum_{j'=1}^{d_{l-1}} f_{j'}^{(l-1)} W_{j'k'}^{(l)}\right] \tag{18}$$

$$= \sum_{j=1}^{d_{l-1}} \sum_{j'=1}^{d_{l-1}} \mathbb{E}\left[f_j^{(l-1)} W_{jk}^{(l)} f_{j'}^{(l-1)} W_{j'k'}^{(l)}\right] \tag{19}$$

$$= \sum_{j=1}^{d_{l-1}} \sum_{j'=1}^{d_{l-1}} \mathbb{E}\left[f_j^{(l-1)} f_{j'}^{(l-1)}\right] \mathbb{E}\left[W_{jk}^{(l)}\right] \mathbb{E}\left[W_{j'k'}^{(l)}\right] \tag{20}$$

$$= 0 \tag{21}$$

□

However, they are **not** independent, but only conditionally independent given the previous later's units. As a remark, note that as $d_1 \to \infty$, the units $f_k^{(2)}$ approach a Gaussian distribution, for which uncorrelation implies independence.

C.1. Main technical Lemma

Here we prove the main technical Lemma that allows us to perform the inductive step.

Lemma C.2. *Let $\boldsymbol{f}^l$ and $\boldsymbol{f}^{l-1}$ be a d_l-dimensional and a d_{l-1}-dimensional vectors, respectively. Let $\sigma_w^2 > 0$, $\tilde{\sigma}^2 > 0$ be two variance parameters, and $b_1, \dots b_{l-1} \in \mathbb{R}$. Then the following integral:*

$$I := \int_{\mathbb{R}^{d_{l-1}}} \frac{1}{(||\boldsymbol{f}^{(l-1)}||^2)^{\frac{d_l}{2}}} e^{-\frac{||\boldsymbol{f}^{(l)}||^2}{2\sigma_w^2 ||\boldsymbol{f}^{(l-1)}||^2}} G^{l-1,0}_{0,l-1}\left(\frac{||\boldsymbol{f}^{(l-1)}||^2}{2^{l-1}\tilde{\sigma}^2}\middle| b_1, \dots, b_{l-1}\right) d\boldsymbol{f}^{l-1}, \tag{22}$$

has solution:

$$I = C G^{l,0}_{0,l}\left(\frac{||\boldsymbol{f}^{(l)}||^2}{2^l \sigma^2}\middle| 0, \frac{1}{2}(d_{l-1} - d_l) + b_1, \dots, \frac{1}{2}(d_{l-1} - d_l) + b_{l-1}\right), \tag{23}$$

where $\sigma^2 := \sigma_w^2 \tilde{\sigma}^2$, and $C := \frac{1}{2}\tilde{C} 2^{\frac{1}{2}(d_{l-1}-d_l)(l-1)} \tilde{\sigma}^{(d_{l-1}-d_l)}$, where $\tilde{C}$ is a constant that depends only on d_{l-1}.

Proof. The proof is based in two steps: in the first steps, we will write the integral in hyper-spherical coordinates. In the second step, we will apply a useful substitution and the properties of the Meijer-G function to solve the integral.

1. Hyper-spherical coordinates Apply the following substitution:

$$f_1^{(l-1)} = r\cos(\gamma_1) \tag{24}$$

$$f_2^{(l-1)} = r\sin(\gamma_1)\cos(\gamma_2) \tag{25}$$

$$\dots \tag{26}$$

$$f_{d_{l-1}}^{(l-1)} = r\sin(\gamma_1)\cdots\sin(\gamma_{d_{l-1}-1}), \tag{27}$$

where $r \in \mathbb{R}_{\geq 0}$ is the radius and $\gamma_1, \dots \gamma_{d_{l-1}-2} \in [0, \pi]$ and $\gamma_{d_{l-1}-1} \in [0, 2\pi]$. The Jacobian is:

$$J_n = \begin{pmatrix} \cos(\gamma_1) & -r\sin(\gamma_1) & 0 & 0 & \dots & 0 \\ \sin(\gamma_1)\cos(\gamma_2) & r\cos(\gamma_1)\cos(\gamma_2) & -r\sin(\gamma_1)\sin(\gamma_2) & 0 & \dots & 0 \\ \dots & & & & \dots & \dots \\ r\sin(\gamma_1)\cdots\sin(\gamma_{d_{l-1}-1}) & & & & & r\sin(\gamma_1)\cdots\cos(\gamma_{d_{l-1}-1}) \end{pmatrix} \tag{28}$$

, where it can be shown that its determinant is:

$$|J_n| = r^{d_{l-1}-1}\sin^{d_{l-1}-2}(\gamma_1)\sin^{d_{l-1}-3}(\gamma_2)\cdots\sin(\gamma_{d_{l-1}-2}). \tag{29}$$

Therefore:

$$\prod_{i=1}^{d_{l-1}} df_i^{(l-1)} = r^{d_{l-1}-1}\sin^{d_{l-1}-2}(\gamma_1)\sin^{d_{l-1}-3}(\gamma_2)\cdots\sin(\gamma_{d_{l-1}-2}) dr d\gamma_1 \cdots d\gamma_{d_{l-1}-1}. \tag{30}$$

By noting that the integral we are trying to solve depends only on $||\boldsymbol{f}^{(l-1)}||^2 = r^2$, we have that the density is, up to a normalization constant independent of $\boldsymbol{f}^{(l)}$:

$$I = \tilde{C}\int r^{d_{l-1}-d_l-1} \mathrm{e}^{-\frac{||\boldsymbol{f}^{(l)}||^2}{2\sigma_w^2 r^2}} G^{l-1,0}_{0,l-1}\left(\frac{r^2}{2^{l-1}\tilde{\sigma}^2}\middle| b_1, \dots, b_{l-1}\right) dr \tag{31}$$

$$= \tilde{C}\int r^{d_{l-1}-d_l-1} G^{1,0}_{0,1}\left(\frac{||\boldsymbol{f}^{(l)}||^2}{2\sigma_w^2 r^2}\middle| 0\right) G^{l-1,0}_{0,l-1}\left(\frac{r^2}{2^{l-1}\tilde{\sigma}^2}\middle| b_1, \dots, b_{l-1}\right) dr, \tag{32}$$

where we call $\tilde{C}$ the angular constant due to the integration of the angle-related terms (that do not depend on r, but only on d_{l-1}). We compute the angular constant in Lemma C.6. In the last step we have applied the identity between the exponential function and the Meijer-G function as in Proposition B.2.

2: Substitution and Meijer-G properties Defining $d = \frac{1}{2}(d_{l-1} - d_l)$, and applying the substitution $x = \frac{r^2}{2^{l-1}\tilde{\sigma}^2}$:

$$I = \tilde{C} \int (2^{\frac{l-1}{2}} \tilde{\sigma} x^{\frac{1}{2}})^{2d-1} G^{1,0}_{0,1} \left(\frac{||\boldsymbol{f}^{(l)}||^2}{2^l \sigma_w^2 \tilde{\sigma}^2 x} \middle| 0 \right) G^{l-1,0}_{0,l-1} \left(x \middle| b_1, \ldots, b_{l-1} \right) 2^{\frac{l-3}{2}} \tilde{\sigma} x^{-\frac{1}{2}} dx \tag{33}$$

$$= \frac{1}{2} \tilde{C} 2^{d(l-1)} \tilde{\sigma}^{2d} \int x^{d-1} G^{1,0}_{0,1} \left(\frac{||\boldsymbol{f}^{(l)}||^2}{2^l \sigma_w^2 \tilde{\sigma}^2 x} \middle| 0 \right) G^{l-1,0}_{0,l-1} \left(x \middle| b_1, \ldots, b_{l-1} \right) dx \tag{34}$$

Defining $\sigma^2 = \sigma_w^2 \tilde{\sigma}^2$, $a^2 := \frac{||\boldsymbol{f}^{(l)}||^2}{2^l \sigma^2}$ and $C := \frac{1}{2} \tilde{C} 2^{d(l-1)} \tilde{\sigma}^{2d}$, and expanding the $G^{1,0}_{0,1}$ term according to the definition, we get:

$$I = C \int x^{d-1} \frac{1}{2\pi i} \int \Gamma(s) \left(\frac{a^2}{x} \right)^{-s} G^{l-1,0}_{0,l-1} \left(x \middle| b_1, \ldots, b_{l-1} \right) ds dx \tag{35}$$

$$= C \frac{1}{2\pi i} \int \Gamma(s) a^{-2s} \int x^{s+d-1} G^{l-1,0}_{0,l-1} \left(x \middle| b_1, \ldots, b_{l-1} \right) dx ds \tag{36}$$

where we can change the order of integration due to the fact that the integrand is positive in the integration region (Tonelli's theorem). Now by using Proposition B.1, the inner integral has the following solution:

$$\int x^{s+d-1} G^{l-1,0}_{0,l-1} \left(x \middle| b_1, \ldots b_{l-1} \right) dx = \prod_{i=1}^{l-1} \Gamma(d + b_i + s) \tag{37}$$

$$= \prod_{i=1}^{l-1} \Gamma \left(\frac{1}{2}(d_{l-1} - d_l) + b_i + s \right) \tag{38}$$

Therefore we can conclude that:

$$I = C \frac{1}{2\pi i} \int \Gamma(s) \prod_{i=1}^{l-1} \Gamma(\frac{1}{2}(d_{l-1} - d_l) + b_i + s) a^{-2s} ds \tag{39}$$

$$= C G^{l,0}_{0,l} \left(a^2 \middle| 0, \frac{1}{2}(d_{l-1} - d_l) + b_1, \ldots, \frac{1}{2}(d_{l-1} - d_l) + b_{l-1} \right) \tag{40}$$

$$= C G^{l,0}_{0,l} \left(\frac{||\boldsymbol{f}^{(l)}||^2}{2^l \sigma^2} \middle| 0, \frac{1}{2}(d_{l-1} - d_l) + b_1, \ldots, \frac{1}{2}(d_{l-1} - d_l) + b_{l-1} \right), \tag{41}$$

where we have simply applied the definition of the Meijer-G function.

□

C.2. Probability Density Function for linear networks

We state and proof the result on the probability density function for a linear network in the following.

Theorem C.3. *Suppose $l \geq 1$, and the input has dimension d_0. Then, the joint marginal density of the random vector $\boldsymbol{f}^{(l)}$ (i.e. the density of the l-th layer pre-activations) is proportional to:*

$$p(\boldsymbol{f}^{(l)}) \propto G^{l,0}_{0,l} \left(\frac{||\boldsymbol{f}^{(l)}||^2}{2^l \sigma^2} \middle| 0, \frac{1}{2}(d_1 - d_l), \ldots, \frac{1}{2}(d_{l-1} - d_l) \right), \tag{42}$$

where $\sigma^2 = \prod_{i=1}^{l} \sigma_i^2$

Proof. We proof by induction. For the base case, consider $l = 1$. We have shown that

$$\boldsymbol{f}^{(1)} \sim \mathcal{N}(0, \sigma_1^2 I). \tag{43}$$

Therefore we can re-write its density as:

$$p(f^{(1)}, \dots f^{(d)}) = \frac{1}{(2\pi\sigma_1^2)^{\frac{d_1}{2}}} \exp\left(-\frac{||\boldsymbol{f}^{(1)}||^2}{2\sigma_1^2}\right) \tag{44}$$

$$= \frac{1}{(2\pi\sigma_1^2)^{\frac{d_1}{2}}} G^{1,0}_{0,1}\left(\frac{||\boldsymbol{f}^{(1)}||^2}{2\sigma_1^2}\middle| 0\right), \tag{45}$$

where we have used the identity between the exponential function and the Meijer-G function (Proposition B.2).

Now let $\tilde{\sigma}^2 = \prod_{i=1}^{l-1} \sigma_i^2$. Assume that $p(f_1^{(l)}, \dots, f_{d_{l-1}}^{(l-1)}) \propto G^{l-1,0}_{0,l-1}\left(\frac{||\boldsymbol{f}^{(l-1)}||^2}{2^{l-1}\tilde{\sigma}^2}\middle| 0, \frac{1}{2}(d_1 - d_{l-1}), \dots, \frac{1}{2}(d_{l-2} - d_{l-1})\right)$. Now we can use the fact that the the units of the l-th layer are conditionally independent given the previous' layer units. Furthermore the conditional distribution is Gaussian due to the fact that the weights are i.i.d Gaussian. Therefore we can write:

$$p(f_1^{(l)}, \dots f_{d_l}^{(l)}) = \int_{\mathbb{R}^{d_{l-1}}} p(f_1^{(l)}, \dots f_{d_l}^{(l)} | f_1^{(l-1)}, \dots f_{d_1}^{(l-1)}) p(f_1^{(l-1)}, \dots f_{d_1}^{(l-1)}) d\boldsymbol{f}^{(l-1)} \tag{46}$$

$$\propto \int_{\mathbb{R}^{d_{l-1}}} \frac{1}{(2\pi||\boldsymbol{f}^{(l-1)}||^2)^{\frac{d_l}{2}}} \mathrm{e}^{-\frac{||\boldsymbol{f}^{(l)}||^2}{2\sigma_l^2||\boldsymbol{f}^{(l-1)}||^2}} \tag{47}$$

$$G^{l-1,0}_{0,l-1}\left(\frac{||\boldsymbol{f}^{(l-1)}||^2}{2^{l-1}\tilde{\sigma}^2}\middle| 0, \frac{1}{2}(d_1 - d_{l-1}), \dots, \frac{1}{2}(d_{l-2} - d_{l-1})\right) d\boldsymbol{f}^{(l-1)}. \tag{48}$$

In the first step we have marginalized out the units of the $l-1$ layer, and applied the product rule of probabilities. In the second step we have applied the induction hypothesis.

The integral is in the form of Lemma C.2. For the coefficients of the Meijer-G function $b_2 = \frac{1}{2}(d_1 - d_{l-1})$, ..., $b_{l-1} = \frac{1}{2}(d_{l-2} - d_{l-1})$, note that:

$$\frac{1}{2}(d_i - d_{l-1}) + \frac{1}{2}(d_{l-1} - d_l) = \frac{1}{2}(d_i - d_l) \tag{49}$$

holds for all $i \in [d_{l-2}]$ and clearly $b_1 + \frac{1}{2}(d_{l-1} - d_l) = \frac{1}{2}(d_{l-1} - d_l)$ as $b_1 = 0$ in our case. Therefore by Lemma C.2 we can conclude that:

$$p(f_1^{(l)}, \dots f_{d_l}^{(l)}) \propto G^{l,0}_{0,l}\left(\frac{||\boldsymbol{f}^{(l)}||^2}{2^l\sigma^2}\middle| 0, \frac{1}{2}(d_1 - d_l) \dots, \frac{1}{2}(d_{l-1} - d_l)\right). \tag{50}$$

□

C.3. CDF of prior predictive

We also derive the CDF of the linear network in the following theorem and proceed to prove it.

Theorem C.4 (CDF of prior predictive)**.** *Let f^l be the output of a of a linear network of l layers. We assume the final layer is one dimensional. Then the the cdf is $F_l(t) := 1 - P(f^l > t)$, $t > 0$. We have that*

$$P(f^l > t) = \frac{t}{2C} G^{l+1,0}_{1,l+1}\left(\omega t^2 \middle| \begin{matrix} \frac{1}{2} \\ -\frac{1}{2}\ 0\ b_1\ \dots\ b_{l-1} \end{matrix}\right), \tag{51}$$

where $b_i = \frac{1}{2}(d_i - 1)$, $i \in [l-1]$, C is the normalization constant and $\omega = \frac{1}{2^l\sigma^2}$.

Proof. Let $X = f^l$.

$$P(X > t) = \frac{1}{C}\int_t^\infty G^{l,0}_{0,l}\left(\omega x^2 \middle| 0, b_1, \dots, b_{l-1}\right) dx \tag{52}$$

$$= \frac{t}{2C}\int_1^\infty y^{-\frac{1}{2}} G^{l,0}_{0,l}\left(\omega t^2 y \middle| 0, b_1, \dots, b_{l-1}\right) dx \tag{53}$$

$$= \frac{t}{2C} G^{l+1,0}_{1,l+1}\left(\omega t^2 \middle| \begin{smallmatrix} \frac{1}{2} \\ -\frac{1}{2}\ 0\ b_1\ \dots\ b_{l-1}\end{smallmatrix}\right), \tag{54}$$

where in the first step we have used the result of C.3, in the second step we have applied the substitution $y = \frac{x^2}{t^2}$, and in the last step we have used Equation 15 with $\rho = \frac{1}{2}$, $\sigma = 1$ and $\alpha = \omega t^2$ □

C.4. Resulting Moments for Linear Networks

Define $\omega = \frac{1}{2^l \sigma^2}$. Denote by $\tilde{p}$ the unnormalized measure and define the random variable

$$Z = ||\boldsymbol{f}^{(l)}||_2^2$$

We are interested in the k-th moment of Z. Using spherical coordinates and the properties of the Meijer-G function in a similar way as the proofs above, we get:

$$\begin{aligned}
\mathbb{E}\left[Z^k\right] &= \frac{1}{C}\int_{\mathbb{R}^m} ||\boldsymbol{z}||_2^{2k} G^{l,0}_{0,l}| \left(\omega ||\boldsymbol{z}||_2^2 \middle| 0, \frac{1}{2}(d_1 - d_l), \dots, \frac{1}{2}(d_{l-1} - d_l)\right) d\boldsymbol{z} \\
&= \frac{\tilde{C}_l}{C}\int_0^\infty r^{2k+d_l-1} G^{l,0}_{0,l}\left(\omega r^2 \middle| 0, \frac{1}{2}(d_1 - d_l), \dots, \frac{1}{2}(d_{l-1} - d_l)\right) dr \\
&= \frac{\tilde{C}_l}{2C}\int_0^\infty x^{k+\frac{d_l}{2}-1} G^{l,0}_{0,l}\left(\omega x \middle| 0, \frac{1}{2}(d_1 - d_l), \dots, \frac{1}{2}(d_{l-1} - d_l)\right) dx \\
&= \frac{\tilde{C}_l}{2C}\omega^{-k-\frac{d_l}{2}} \prod_{i=1}^l \Gamma\left(\frac{d_i}{2} + k\right) \\
&= \omega^{-k} \frac{\prod_{i=1}^l \Gamma\left(\frac{d_i}{2} + k\right)}{\prod_{i=1}^l \Gamma\left(\frac{d_i}{2}\right)} \\
&= \left(2^l \sigma^2\right)^k \prod_{i=1}^l \frac{\Gamma\left(\frac{d_i}{2} + k\right)}{\Gamma\left(\frac{d_i}{2}\right)}
\end{aligned}.$$

Note that it can be equivalently written as:

$$\mathbb{E}\left[Z^k\right] = (2^l\sigma^2)^{k-1} \prod_{i=1}^l \left(\frac{\Gamma\left(\frac{d_i}{2} + k - 1\right)}{\Gamma\left(\frac{d_i}{2}\right)}\right) (2^l \sigma^2) \prod_{i=1}^l \left(\frac{d_i}{2} + k - 1\right) \tag{55}$$

$$= \mathbb{E}\left[Z^{k-1}\right] \sigma^2 \prod_{i=1}^l (d_i + 2(k-1)) \tag{56}$$

so the kurtosis is:

$$\kappa = \frac{\prod_{i=1}^l (d_i + 2(2-1))}{\prod_{i=1}^l d_i} = \prod_{i=1}^l \frac{d_i + 2}{d_i} \tag{57}$$

If $d_1 = \dots = d_{l-1} = m$, and $d_l = 1$. For instance the variance ($k = 1$) is[1]:

$$(2^l\sigma^2)\left(\frac{\Gamma(\frac{m}{2}+1)}{\Gamma(\frac{m}{2})}\right)^{l-1} \frac{\Gamma(1+\frac{1}{2})}{\Gamma(\frac{1}{2})} = (2^l\sigma^2)\frac{m^{l-1}}{2^{l-1}}\frac{1}{2} = \sigma^2 m^{l-1}. \tag{58}$$

[1]By symmetry, all the odd moments are zero

C.5. Proof of Lemma C.5

We also present the infinite-width and infinite-depth result for the linear case. Due to the linear nature, the proof simplifies significantly compared to the ReLU case.

Lemma C.5. *Consider the distribution of the output $p(f^{(L)}) = p^m(f^{(L)})$, as defined in Thm. C.3. Denote $X \sim \mathcal{N}(0, 1)$, $Y \sim \mathcal{LN}(-\gamma, 2\gamma)$ and the normal log-normal mixture $Z = XY$.*

For fixed depth $L \in \mathbb{N}$, under NTK parametrization , it holds that

$$p^m(f^{(L)}) \xrightarrow{d} X \quad \text{for } m \to \infty$$

In contrast, for growing depth $L = \gamma m$, we have the following convergence of the moments

$$\mathbb{E}\left[\left(f^{(L)}\right)^{2k}\right] \xrightarrow{m\to\infty} \mathbb{E}[Z^{2k}] = e^{\gamma k(k-1)}(2k-1)!!$$

where $(2k-1)!! = (2k-1)\dots 3 \cdot 1$ denotes the double factorial.

Proof. Recall that the moments of $||\boldsymbol{f}^{(l)}||_2$ are given by

$$\mathbb{E}\left[||\boldsymbol{f}^{(l)}||_2^{2k}\right] = \left(2^l\sigma^2\right)^k \prod_{i=1}^{l} \frac{\Gamma\left(\frac{d_i}{2}+k\right)}{\Gamma\left(\frac{d_i}{2}\right)}$$

where $\sigma^2 = \prod_{i=1}^{l} \sigma_i^2$. Assuming $d_1 = \dots d_{l-1} = m$ and $d_l = 1$ and the NTK parametrization $\sigma_1^2 = 1$ and $\sigma_2^2 = \dots = \sigma_l^2 = \frac{1}{m}$ simplifies this to

$$\begin{aligned}
\mathbb{E}\left[\left(f^{(l)}\right)^{2k}\right] &= \left(\frac{2^l}{m^{l-1}}\right)^k \frac{\Gamma(\frac{m}{2}+k)^{l-1}}{\Gamma(\frac{m}{2})^{l-1}} \frac{\Gamma(\frac{1}{2}+k)}{\Gamma(\frac{1}{2})} \\
&= \left(\frac{2^l}{m^{l-1}}\right)^k \frac{\Gamma(\frac{m}{2}+k)^{l-1}}{\Gamma(\frac{m}{2})^{l-1}} 2^{-k}(2k-1)!! \\
&= \left(\frac{2^k}{m^k} \frac{\Gamma(\frac{m}{2}+k)}{\Gamma(\frac{m}{2})}\right)^{l-1} (2k-1)!! \\
&= \left(\frac{2^k}{m^k} \left(\frac{m}{2}+k-1\right) \dots \left(\frac{m}{2}+1\right) \frac{m}{2}\right)^{l-1} (2k-1)!!
\end{aligned}$$

Define the k-th order polynomial $p(m) = \left(\frac{m}{2}+k-1\right)\dots\left(\frac{m}{2}+1\right)\frac{m}{2}$. Denote its coefficients by α_i for $i = 1, \dots, k$. We know that $\alpha_k = \frac{1}{2^k}$ and from Lemma D.4 that

$$\alpha_{k-1} = \frac{1}{2^{k-1}} \sum_{i=1}^{k} (k-i) = \frac{1}{2^{k-1}} \left(k^2 - \frac{k(k+1)}{2}\right) = \frac{k^2-k}{2^k}$$

Assuming constant depth, performing the division by m^k thus leads to

$$\begin{aligned}
\left(2^k\left(\alpha_k + \alpha_{k-1}\frac{1}{m} + \dots + \mathcal{O}\left(\frac{1}{m^2}\right)\right)\right)^{l-1} (2k-1)!! &= \left(1 + \frac{k(k-1)}{m} + \mathcal{O}\left(\frac{1}{m^2}\right)\right)^{l-1} (2k-1)!! \\
&= \left(1 + \frac{(l-2)\left((k-1)k\right)}{m} + \mathcal{O}\left(\frac{1}{m^2}\right)\right) (2k-1)!!
\end{aligned}$$

Now we can easily see that

$$\mathbb{E}\left[\left(f^{(l)}\right)^{2k}\right] \xrightarrow{m\to\infty} (2k-1)!!$$

Recall that for $X \sim \mathcal{N}(0,1)$ we have the same moments $\forall k \in \mathbb{N}$: $\mathbb{E}\left[X^{2k}\right] = (2k-1)!!$, whereas the odd moments vanish for both distribution due to symmetry. The convergence of the moments, due to Billingsley (1986) and the identifiability of the Gaussian distribution implies convergence in distribution.

On the other hand, if we assume that depth grows proportional to width, i.e. $l-1 = \gamma m$ for $\gamma > 0$, we arrive at a different limit given by

$$\left(1 + \frac{k(k-1)}{m} + \mathcal{O}\left(\frac{1}{m^2}\right)\right)^{\gamma m} (2k-1)!! \xrightarrow{m \to \infty} e^{\gamma k(k-1)}(2k-1)!!$$

Consider the random variable $Z = XY$ where $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{LN}(s, t^2)$ are two independent variables. For $k \in \mathbb{N}$, we can compute the moments as

$$\mathbb{E}\left[Z^k\right] = \mathbb{E}\left[X^k Y^k\right] = \mathbb{E}\left[X^k\right]\mathbb{E}\left[Y^k\right] = \begin{cases} 0 & \text{k odd} \\ (2k-1)!! e^{ks + \frac{k^2 t^2}{2}} & \text{k even} \end{cases}$$

Choosing $s = -\gamma$ and $t^2 = 2\gamma$ hence recovers the moments exactly. □

C.6. Normalization Constant and Angular Constant

We complete the picture by calculating the normalization constant of the resulting distribution.

Lemma C.6 (normalization constant)**.** *Under the conditions of Theorem C.3, the normalization constant C for the density of the l-th layer can be computed as:*

$$C = \frac{1}{2}\tilde{C}_l \left(\frac{1}{2^l \sigma^2}\right)^{-\frac{d_l}{2}} \prod_{i=1}^{l} \Gamma\left(\frac{d_i}{2}\right), \tag{59}$$

or, expanding $\tilde{C}_l$ according to Lemma C.7:

$$\frac{\pi^{\frac{d_l}{2}}}{\Gamma\left(\frac{d_l}{2}\right)} \left(\frac{1}{2^l \sigma^2}\right)^{-\frac{d_l}{2}} \prod_{i=1}^{l} \Gamma\left(\frac{d_i}{2}\right) \tag{60}$$

proof of lemma C.6. The normalization constant has the following form:

$$\begin{aligned}
C = \int_{\mathbb{R}^{d_l}} \tilde{p}(\boldsymbol{f}^{(l)}) d\boldsymbol{f}^{(l)} &= \int_{\mathbb{R}^{d_l}} G^{l,0}_{0,l}\left(\frac{||\boldsymbol{f}^{(l)}||^2}{2^l\sigma^2}\middle| 0, \frac{1}{2}(d_1 - d_l)\dots, \frac{1}{2}(d_{l-1} - d_l)\right) d\boldsymbol{f}^{(l)} && (61)\\
&= \tilde{C}_l \int_0^\infty r^{d_l - 1} G^{l,0}_{0,l}\left(\frac{r^2}{2^l\sigma^2}\middle| 0, \frac{1}{2}(d_1 - d_l)\dots, \frac{1}{2}(d_{l-1} - d_l)\right) dr && (62)\\
&= \frac{1}{2}\tilde{C}_l \int_0^\infty x^{\frac{d_l}{2}-1} G^{l,0}_{0,l}\left(\frac{x}{2^l\sigma^2}\middle| 0, \frac{1}{2}(d_1 - d_l)\dots, \frac{1}{2}(d_{l-1} - d_l)\right) dx && (63)\\
&= \frac{1}{2}\tilde{C}_l \left(\frac{1}{2^l\sigma^2}\right)^{-\frac{d_l}{2}} \prod_{i=1}^{l} \Gamma\left(\frac{d_i}{2}\right) && (64)
\end{aligned}$$

where we used spherical coordinates and the substitution $x = r^2$. We denote the angular constant by $\tilde{C}_l$, and according to Lemma C.7 has solution: $\tilde{C}_l = \frac{2\pi^{\frac{d_l}{2}}}{\Gamma(\frac{d_l}{2})}$.

□

C.7. Proof of Lemma C.7

Lemma C.7. *The angular constant:*

$$\tilde{C}_l = \int_0^{2\pi} d\gamma_{d_l-1} \int_0^{\pi} \sin^{d_l-2}(\gamma_1) d\gamma_1 \int_0^{\pi} \sin^{d_l-3}(\gamma_2) d\gamma_2 \cdots \int_0^{\pi} \sin(\gamma_{d_l-2}) d\gamma_{d_l-2} \tag{65}$$

has solution:

$$\tilde{C}_l = \frac{2\pi^{\frac{d_l}{2}}}{\Gamma\left(\frac{d_l}{2}\right)} \tag{66}$$

Proof. The angular constant $\tilde{C}_l$ can be calculated as follows (for $d_l \geq 2$):

$$\begin{aligned}
\tilde{C}_l &= \int_0^{2\pi} d\gamma_{d_l-1} \int_0^{\pi} \sin^{d_l-2}(\gamma_1) d\gamma_1 \int_0^{\pi} \sin^{d_l-3}(\gamma_2) d\gamma_2 \cdots \int_0^{\pi} \sin(\gamma_{d_l-2}) d\gamma_{d_l-2} \\
&= 2\pi \prod_{k=1}^{d_l-2} \int_0^{\pi} \sin^{d_l-k-1}(\gamma_k) d\gamma_k \\
&= 2\pi \prod_{k=1}^{d_l-2} \frac{\Gamma(d_l-k-1)}{2^{d_l-k-1}\Gamma\left(\frac{d_l-k-1}{2}\right)\Gamma\left(\frac{d_l-k-1}{2}+1\right)}(2\pi) \\
&= \frac{(2\pi)^{d_l-1}}{2^{\frac{1}{2}(d_l-2)(d_l-1)}} \prod_{k=1}^{d_l-2} \frac{\Gamma(d_l-k-1)}{\Gamma\left(\frac{d_l-k-1}{2}\right)\Gamma\left(\frac{d_l-k-1}{2}+1\right)}
\end{aligned}$$

where we have used Lemma C.8 to compute the integrals. If $d_l = 1$, then there is no need to write the integral in spherical coordinates and we can simply set $\tilde{C} = 1$.

Now we can apply the Legendre duplication formula:

$$\Gamma(z)\Gamma(z+\frac{1}{2}) = 2^{1-2z}\sqrt{\pi}\Gamma(2z) \tag{67}$$

to the numerator term $2z = d_l - k - 1$ and get:

$$\tilde{C}_l = \frac{(2\pi)^{d_l-1}}{\pi^{\frac{d_l-2}{2}} 2^{\frac{1}{2}(d_l-2)(d_l-1)}} \prod_{k=1}^{d_l-2} \frac{\cancel{\Gamma\left(\frac{d_l-k-1}{2}\right)}\Gamma(\frac{d_l-k-1}{2}+\frac{1}{2})2^{d_l-k-2}}{\cancel{\Gamma\left(\frac{d_l-k-1}{2}\right)}\Gamma\left(\frac{d_l-k-1}{2}+1\right)}. \tag{68}$$

Note that that the product:

$$\prod_{k=1}^{d_l-2} \frac{\Gamma(\frac{d_l-k-1}{2}+\frac{1}{2})}{\Gamma\left(\frac{d_l-k-1}{2}+1\right)} = \frac{\Gamma(\frac{d_l-1}{2})}{\Gamma(\frac{d_l}{2})} \cdot \frac{\Gamma(\frac{d_l-2}{2})}{\Gamma(\frac{d_l-1}{2})} \cdots \frac{1}{\Gamma(\frac{1}{2})} = \frac{1}{\Gamma(\frac{d_l}{2})} \tag{69}$$

Finally, we have the product

$$\prod_{k=1}^{d_l-2} 2^{d_l-k-2} = 2^{d_l(d_l-2)-\frac{(d_l-2)(d_l-1)}{2}-2(d_l-2)} \tag{70}$$

From which we can conclude, after some elementary algebraic manipulations:

$$\tilde{C}_l = \frac{2\pi^{\frac{d_l}{2}}}{\Gamma\left(\frac{d_l}{2}\right)} \tag{71}$$

□

Finally we prove the technical Lemma that we used in the previous proof.

Lemma C.8.

$$\int_0^\pi \sin^k(x)dx = \frac{\Gamma(k)}{2^k\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{k}{2}+1\right)}(2\pi) \tag{72}$$

Proof. By integrating by parts, and using some algebraic manipulation, it is easy to see that:

$$\int \sin^k(x)dx = -\frac{1}{k}\sin^{k-1}x\cos x + \frac{k-1}{k}\int \sin^{k-2}x dx. \tag{73}$$

Evaluating the integral between 0 and π, we get:

$$\int_0^\pi \sin^k(x)dx = \frac{k-1}{k}\int_0^\pi \sin^{k-2}x dx. \tag{74}$$

By unrolling the recursion:

$$\int_0^\pi \sin^k(x)dx = \frac{(k-1)(k-3)\cdots}{k(k-2)\cdots}\begin{cases}\int_0^\pi dx = \pi & \text{if } k \text{ is even}\\ \int_0^\pi \sin(x)dx = 2 & \text{if } k \text{ is odd}\end{cases}. \tag{75}$$

The following expression includes both the even and the odd case:

$$\int_0^\pi \sin^k(x)dx = \frac{\Gamma(k)}{2^k\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{k}{2}+1\right)}(2\pi) \tag{76}$$

In fact, if k is even, then:

$$\begin{aligned}\frac{\Gamma(k)}{2^k\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{k}{2}+1\right)}(2\pi) &= \frac{(k-1)(k-2)\cdots}{2^k\frac{k}{2}(\frac{k}{2}-1)^2(\frac{k}{2}-2)^2\cdots}(2\pi)\\ &= \frac{(k-1)(k-3)\cdots}{k(k-2)\cdots}\pi\end{aligned}$$

If k is odd, we use the identity:

$$\Gamma\left(\frac{k}{2}\right) = \frac{(k-2)!!\sqrt{\pi}}{2^{\frac{k-1}{2}}}, \tag{77}$$

where $k!!$ is the double factorial. Following a very similar procedure, we get the desired result. □

Important Remark: In the ReLU case, we will see that that the integral is from 0 to $\frac{\pi}{2}$. In that case, we get:

$$\int_0^{\frac{\pi}{2}} \sin^k(x)dx = \frac{\Gamma(k)}{2^k\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{k}{2}+1\right)}\pi \tag{78}$$

Therefore the angular constant is:

$$\tilde{C}_l = \int_0^{\frac{\pi}{2}} d\gamma_{d_l-1} \int_0^{\frac{\pi}{2}} \sin^{d_l-2}(\gamma_1)d\gamma_1 \int_0^{\frac{\pi}{2}} \sin^{d_l-3}(\gamma_2)d\gamma_2 \cdots \int_0^{\frac{\pi}{2}} \sin(\gamma_{d_l-2})d\gamma_{d_l-2} \tag{79}$$

$$= \frac{\pi}{2}\prod_{k=1}^{d_l-2}\int_0^{\frac{\pi}{2}} \sin^{d_l-k-1}(\gamma_k)d\gamma_k \tag{80}$$

$$= \frac{\pi^{d_l-1}}{2}\prod_{k=1}^{d_l-2}\frac{1}{2^{d_l-k-1}}\frac{\Gamma(d_l-k-1)}{\Gamma\left(\frac{d_l-k-1}{2}\right)\Gamma\left(\frac{d_l-k-1}{2}+1\right)} \tag{81}$$

$$= \frac{\pi^{d_l-1}}{2\pi^{\frac{d_l-2}{2}}}\prod_{k=1}^{d_l-2}\frac{\Gamma(\frac{d_l-k-1}{2}+\frac{1}{2})2^{-1}}{\Gamma\left(\frac{d_l-k-1}{2}+1\right)} \tag{82}$$

$$= \frac{\pi^{d_l-1}}{2\pi^{\frac{d_l-2}{2}}}\frac{2^{-(d_l-2)}}{\Gamma(\frac{d_l}{2})} \tag{83}$$

$$= \frac{\pi^{\frac{d_l}{2}}}{2^{d_l-1}\Gamma(\frac{d_l}{2})} \tag{84}$$

C.8. Kurtosis of Linear Networks

Using the closed form expressions from Section C.4, we can describe the kurtosis of the output $f^{(l)}$ as

$$\kappa_{\text{lin}} := \frac{3\sigma_w^{2l}(m+2)^{l-1}}{\sigma_w^{2l}m^{l-1}} = 3\left(\frac{m+2}{m}\right)^{l-1}. \tag{85}$$

In particular, the distribution is always more heavy-tailed than a Gaussian, for which $\kappa = 3$ (i.e. the distribution is leptokurtic). The second obvious conclusion is that depth increases the heavy-tailedness exponentially, which is in-line with the theoretical results of (Vladimirova et al., 2019). On the contrary, the width has the effect of ”normalizing” the distribution, in particular in the limit of large width we have that:

$$\lim_{m\to\infty}\kappa = 3, \tag{86}$$

which is the kurtosis of the Gaussian distribution, as anticipated from Lemma C.5.

D. Proofs for ReLU networks and Derivation of their Moments

Now, we extend these results to ReLU networks. We need the following additional notation: we call $\delta(x-x_0)$ the Dirac delta function centered at x_0, and $\mathbb{1}_A := \begin{cases}1 & x\in A\\ 0 & \text{else}\end{cases}$ the indicator function. Also, we indicate with $\mathcal{S}$ the set of indices $1,\ldots,d$ that index d random variables, and with Ω its power set, i.e., the set of all possible subsets of $\mathcal{S}$. Note that $|\Omega| = 2^d$. We will use the following lemma, which explains what happens to a joint density when the marginals are transformed by the ReLU function.

D.1. Effect of ReLU activation function on the joint density

Lemma D.1. *Let $p(f_1,\ldots f_d)$ be the joint density of the d random variables $f_1,\ldots f_d$.[a] Assume $p(f_1,\ldots f_d)$ is symmetric around zero. When we apply the transformation $g_i = ReLU(f_i)$ to each $i\in[d]$, then the joint density of the transformed variables has the following form:*

$$p_{ReLU}(g_1,\ldots g_d) = \sum_{A\in\Omega}\frac{1}{2^{|A|}}\prod_{i\in A}\delta(g_i)p(\boldsymbol{g}_{\mathcal{S}\setminus A})\prod_{j\in\mathcal{S}\setminus A}\mathbb{1}_{g_j>0}, \tag{87}$$

where $p(\boldsymbol{g}_{\mathcal{S}\setminus A})$ is the marginal density of the random variables whose indexes are in $\mathcal{S}\setminus A$.

[a] We use the same notation for the random variable and the corresponding dummy variable in the density function.

Proof. Again we use conditional independence: the activations $g_1, \dots, g_d$ are independent given the pre-activations $f_1, \dots, f_d$. So we can write:

$$p_{\text{ReLU}}(g_1, \dots, g_d) = \int_{\mathbb{R}^d} p(g_1, \dots, g_d | f_1, \dots, f_d) p(f_1, \dots, f_d) d\boldsymbol{f} \tag{88}$$

$$= \int_{\mathbb{R}^d} \prod_{i=1}^{d} p(g_i | f_i) p(f_1, \dots, f_d) d\boldsymbol{f} \tag{89}$$

Now, $p(g_i|f_i) = \begin{cases} \delta(g_i) & f_i < 0 \\ \delta(g_i - f_i) & f_i \geq 0 \end{cases} = \delta(g_i)\mathbb{1}_{f_i<0} + \delta(g_i - f_i)\mathbb{1}_{f_i \geq 0}$. So we can write:

$$p_{\text{ReLU}}(g_1, \dots, g_d) = \int_{\mathbb{R}^d} \prod_{i=1}^{d} \left(\delta(g_i)\mathbb{1}_{f_i<0} + \delta(g_i - f_i)\mathbb{1}_{f_i \geq 0}\right) p(f_1, \dots, f_d) d\boldsymbol{f} \tag{90}$$

$$= \int_{\mathbb{R}^d} \prod_{i=1}^{d} \left(\delta(g_i)\mathbb{1}_{f_i<0} + \delta(g_i - f_i)\mathbb{1}_{f_i \geq 0}\right) p(f_1, \dots, f_d) d\boldsymbol{f} \tag{91}$$

$$= \int_{\mathbb{R}^d} \sum_{A \in \Omega} \left(\prod_{i \in A} \delta(g_i)\mathbb{1}_{f_i<0} \prod_{j \in \mathcal{S} \setminus A} \delta(g_j - f_j)\mathbb{1}_{f_j \geq 0} \right) p(f_1, \dots, f_d) d\boldsymbol{f} \tag{92}$$

$$= \sum_{A \in \Omega} \int_{\mathbb{R}^{|A|}_{<0}} \int_{\mathbb{R}^{d-|A|}} \prod_{i \in A} \delta(g_i) \prod_{j \in \mathcal{S} \setminus A} \delta(g_j - f_j)\mathbb{1}_{f_j \geq 0} p(f_1, \dots, f_d) d\boldsymbol{f} \tag{93}$$

$$= \sum_{A \in \Omega} \prod_{i \in A} \delta(g_i) \int_{\mathbb{R}^{|A|}_{<0}} \int_{\mathbb{R}^{d-|A|}} \prod_{j \in \mathcal{S} \setminus A} \delta(g_j - f_j)\mathbb{1}_{f_j \geq 0} p(f_1, \dots, f_d) d\boldsymbol{f} \tag{94}$$

$$= \sum_{A \in \Omega} \prod_{i \in A} \delta(g_i) \int_{\mathbb{R}^{|A|}_{<0}} p(\boldsymbol{f}_A, \boldsymbol{g}_{\mathcal{S} \setminus A}) d\boldsymbol{f}_A \prod_{j \in \mathcal{S} \setminus A} \mathbb{1}_{g_j > 0} \tag{95}$$

$$= \sum_{A \in \Omega} \frac{1}{2^{|A|}} \prod_{i \in A} \delta(g_i) p(\boldsymbol{g}_{\mathcal{S} \setminus A}) \prod_{j \in \mathcal{S} \setminus A} \mathbb{1}_{g_j > 0}, \tag{96}$$

where in the second to last step we have used the well known property of the Delta function $\int_{-\infty}^{\infty} f(x)\delta(x - x_0)dx = f(x_0)$ and in the last step we used the fact that the density p is symmetric around 0. □

D.2. proof of Theorem D.2

Theorem D.2. *Suppose $l \geq 2$, and the input has dimension d_0. Define the multi-index set $\mathcal{R} = [d_1] \times \cdots \times [d_{l-1}]$*[a] *and introduce the vector $\boldsymbol{u}^{\boldsymbol{r}} \in \mathbb{R}^{l-1}$ through its components $\boldsymbol{u}^{\boldsymbol{r}}_i = \frac{1}{2}(d_i - r_i - d_l)$.*

$$p(\boldsymbol{f}^{(l)}_{ReLU}) = \sum_{\boldsymbol{r} \in \mathcal{R}} q_{\boldsymbol{r}} G^{l,0}_{0,l} \left(\frac{||\boldsymbol{f}^{(l)}_{ReLU}||^2}{2^l \sigma^2} \middle| 0, \boldsymbol{u}^{\boldsymbol{r}} \right) + q_0 \delta(\boldsymbol{f}^{(l)}_{ReLU}), \tag{97}$$

where $\sigma^2 = \prod_{i=1}^{l} \sigma_i^2$ and the individual weights are given by

$$q_{\boldsymbol{r}} = \pi^{-\frac{d_l}{2}} 2^{-\frac{l}{2} d_l} (\sigma^2)^{-\frac{d_l}{2}} \prod_{i=1}^{l-1} \binom{d_i}{r_i} \frac{1}{2^{d_i} \Gamma\left(\frac{d_i - r_i}{2}\right)}, \tag{98}$$

and

$$q_0 = 1 - \prod_{i=1}^{l-1} \frac{2^{d_i} - 1}{2^{d_i}} \tag{99}$$

[a] Here $[d_i] := \{0, \dots, d_i - 1\}$

Proof. Before starting the proof, note that there is a special case the has to be handled separately: the case in which *all units* are inactive, i.e., the ReLU activation sets to zero all the pre-activation in a layer. This will be handled at end of the proof. First, let's assume that there is at least one active unit per layer.

The proof is again by induction. The base case ($l = 2$) is stated in Lemma D.3. For the general case, we use again an identical approach as in Theorem C.3. We expand the coefficients and write $c^l_{r_l} := \binom{d_l}{r_l}\frac{1}{2^{d_l}\Gamma\left(\frac{d_l-r_l}{2}\right)}$, where $l > 0$ is the layer index. Induction step: assume that the pre-nonlinearities have the following form:

$$p(f_1^{(l-1)}, \dots, f_{d_{l-1}}^{(l-1)}) = \sum_{r_1=0}^{d_1-1} c^1_{r_1} \cdots \sum_{r_{l-2}=0}^{d_{l-2}-1} c^{l-2}_{r_{l-2}} \pi^{-\frac{d_{l-1}}{2}} 2^{-\frac{(l-1)d_{l-1}}{2}} (\tilde{\sigma}^2)^{-\frac{d_{l-1}}{2}} \tag{100}$$

$$G^{l-1,0}_{0,l-1}\left(\frac{||\boldsymbol{f}^{(l-1)}||^2}{2^{l-1}\tilde{\sigma}^2}\middle| 0, \frac{1}{2}\left(d_1 - r_1 - d_{l-1}\right), \dots, \frac{1}{2}\left(d_{l-2} - r_{l-2} - d_{l-1}\right)\right), \tag{101}$$

where $\tilde{\sigma}^2 = \prod_{i=1}^{l-1}\sigma_i^2$. We know from Lemma D.1, that the activations $g_1^{(l-1)}, \dots, g_{d_{l-1}}^{(l-1)}$ have the following density:

$$p_{\text{ReLU}}(g_1^{(l-1)}, \dots, g_{d_{l-1}}^{(l-1)}) = \sum_{A\in\Omega} \frac{1}{2^{|A|}} \prod_{i\in A} \delta(g_i^{(l-1)}) p(\boldsymbol{g}_{\mathcal{S}\setminus A}^{(l-1)}) \prod_{j\in\mathcal{S}\setminus A} \mathbb{1}_{g_j>0} \tag{102}$$

$$= \sum_{A\in\Omega} \frac{1}{2^{|A|}} \sum_{r_1=0}^{d_1-1} c^1_{r_1} \cdots \sum_{r_{l-2}=0}^{d_{l-2}-1} c^{l-2}_{r_{l-2}} \pi^{-\frac{d_{l-1}}{2}} 2^{-\frac{(l-1)d_{l-1}}{2}} (\tilde{\sigma}^2)^{-\frac{d_{l-1}}{2}} \tag{103}$$

$$G^{l-1,0}_{0,l-1}\left(\frac{||\boldsymbol{g}_{\mathcal{S}\setminus A}^{(l-1)}||^2}{2^{l-1}\tilde{\sigma}^2}\middle| 0, \frac{1}{2}\left(d_1 - r_1 - d_A\right), \dots, \frac{1}{2}\left(d_{l-2} - r_{l-2} - d_A\right)\right) \tag{104}$$

$$\prod_{i\in A} \delta(g_i^{(l-1)}) \prod_{j\in\mathcal{S}\setminus A} \mathbb{1}_{g_j>0}, \tag{105}$$

where $d_A := |\mathcal{S} \setminus A| = d_{l-1} - |A|$. Also, here we abuse the notation and consider that $\mathcal{S}$ is not in the power set, i.e., $\mathcal{S} \notin \Omega$. This is to be consistent with the fact that we are handling the case in which at least one unit is active after the ReLU activation is applied. Now following a similar procedure as in Lemma D.3, we have

$$p(f_1^{(l)}, \dots f_{d_l}^{(l)}) = \int_{\mathbb{R}^{d_{l-1}}_{\geq 0}} p(f_1^{(l)}, \dots f_{d_l}^{(l)} | g_1^{(l-1)}, \dots g_{d_1}^{(l-1)}) p_{\text{ReLU}}(g_1^{(l-1)}, \dots g_{d_1}^{(l-1)}) d\boldsymbol{g}^{(l-1)}, \tag{106}$$

which is equal to:

$$\int_{\mathbb{R}_{\geq 0}^{d_{l-1}}} \frac{1}{(2\pi\sigma_l^2 ||\boldsymbol{g}^{(l-1)}||^2)^{\frac{d_l}{2}}} \mathrm{e}^{-\frac{||\boldsymbol{f}^{(l)}||^2}{2\sigma_l^2 ||\boldsymbol{g}^{(l-1)}||^2}} \sum_{A\in\Omega} \frac{1}{2^{|A|}} \sum_{r_1=0}^{d_1-1} c_{r_1}^1 \cdots \sum_{r_{l-2}=0}^{d_{l-2}-1} c_{r_{l-2}}^{l-2} \tag{107}$$

$$\pi^{-\frac{d_{l-1}-|A|}{2}} 2^{-\frac{(l-1)(d_{l-1}-|A|)}{2}} (\tilde{\sigma}^2)^{-\frac{d_{l-1}-|A|}{2}} \tag{108}$$

$$G_{0,l-1}^{l-1,0}\left(\frac{||\boldsymbol{g}_{\mathcal{S}\setminus A}^{(l-1)}||^2}{2^{l-1}\tilde{\sigma}^2}\middle| 0, \frac{1}{2}\left(d_1 - r_1 - d_A\right), \ldots, \frac{1}{2}\left(d_{l-2} - r_{l-2} - d_A\right)\right) \tag{109}$$

$$\prod_{i\in A} \delta(g_i^{(l-1)}) \prod_{j\in\mathcal{S}\setminus A} \mathbb{1}_{g_j>0} d\boldsymbol{g}^{(l-1)} \tag{110}$$

$$= \sum_{A\in\Omega} \frac{\pi^{-\frac{d_{l-1}-|A|}{2}} 2^{-\frac{(l-1)(d_{l-1}-|A|)}{2}} (\tilde{\sigma}^2)^{-\frac{d_{l-1}-|A|}{2}}}{2^{|A|}(2\pi\sigma_l^2)^{\frac{d_l}{2}}} \sum_{r_1=0}^{d_1-1} c_{r_1}^1 \cdots \sum_{r_{l-2}=0}^{d_{l-2}-1} c_{r_{l-2}}^{l-2} \tag{111}$$

$$\int_{\mathbb{R}_{>0}^{d_{l-1}}} \frac{1}{(||\boldsymbol{g}_{\mathcal{S}\setminus A}^{(l-1)}||^2)^{\frac{d_l}{2}}} \mathrm{e}^{-\frac{||\boldsymbol{f}^{(l)}||^2}{2\sigma_l^2 ||\boldsymbol{g}_{\mathcal{S}\setminus A}^{(l-1)}||^2}} \tag{112}$$

$$G_{0,l-1}^{l-1,0}\left(\frac{||\boldsymbol{g}_{\mathcal{S}\setminus A}^{(l-1)}||^2}{2^{l-1}\tilde{\sigma}^2}\middle| 0, \frac{1}{2}\left(d_1 - r_1 - d_A\right), \ldots, \frac{1}{2}\left(d_{l-2} - r_{l-2} - d_A\right)\right) d\boldsymbol{g}_{\mathcal{S}\setminus A}^{(l-1)} \tag{113}$$

For each set A we have a $(d_{l-1} - |A|)$- dimensional integral that can be solved using once again Lemma C.2 [2]. Note that for the new Meijer-G coefficients of Lemma C.2:

$$\frac{1}{2}\left(d_i - r_i - d_{l-1} + |A|\right) + \frac{1}{2}\left(d_{l-1} - |A| - d_l\right) = \frac{1}{2}\left(d_i - r_i - d_l\right) \tag{114}$$

holds for all $i \in [d_{l-2}]$. Therefore the solution of each integral is equal to

$$\frac{1}{2}\tilde{C}_A 2^{\frac{1}{2}(d_{l-1}-|A|-d_l)(l-1)} \tilde{\sigma}^{(d_{l-1}-|A|-d_l)} \tag{115}$$

$$G_{0,l}^{l,0}\left(\frac{||\boldsymbol{f}^{(l)}||^2}{2^l\tilde{\sigma}^2}\middle| 0, \frac{1}{2}\left(d_1 - r_1 - d_l\right), \ldots, \frac{1}{2}\left(d_{l-1} - r_{l-1} - d_l\right)\right). \tag{116}$$

The new coefficient for every set A is:

$$c_A = \frac{\pi^{-\frac{d_{l-1}-|A|}{2}} 2^{-\frac{(l-1)(d_{l-1}-|A|)}{2}} (\tilde{\sigma}^2)^{-\frac{d_{l-1}-|A|}{2}}}{2^{|A|}(2\pi\sigma_l^2)^{\frac{d_l}{2}}} \frac{1}{2}\tilde{C}_A 2^{\frac{1}{2}(d_{l-1}-|A|-d_l)(l-1)} \tilde{\sigma}^{(d_{l-1}-|A|-d_l)} \tag{117}$$

$$= \frac{\cancel{\pi^{-\frac{d_{l-1}-|A|}{2}}} 2^{-\frac{(l-1)(d_{l-1}-|A|)}{2}} (\tilde{\sigma}^2)^{-\frac{d_{l-1}-|A|}{2}}}{\cancel{2^{|A|}}(2\pi\sigma_l^2)^{\frac{d_l}{2}}} \frac{1}{\cancel{2}} \frac{\cancel{\pi^{\frac{d_{l-1}-|A|}{2}}}}{2^{d_{l-1}-\cancel{|A|}-\cancel{1}}\Gamma\left(\frac{d_{l-1}-|A|}{2}\right)} \tag{118}$$

$$\tilde{C}_A 2^{\frac{1}{2}(\cancel{d_{l-1}-|A|}-d_l)(l-1)} \tilde{\sigma}^{(\cancel{d_{l-1}-|A|}-d_l)} \tag{119}$$

$$= \frac{\pi^{-\frac{d_l}{2}} 2^{-\frac{l d_l}{2}} (\sigma^2)^{-\frac{d_l}{2}}}{2^{d_{l-1}}\Gamma\left(\frac{d_{l-1}-|A|}{2}\right)} \tag{120}$$

Therefore, because the dependence on A is only through its cardinality $r_{l-1}^{l-1} := |A|$, we define:

$$c_{r_{l-1}}^{l-1} := \binom{d_{l-1}}{r_{l-1}} \frac{1}{2^{d_{l-1}}\Gamma\left(\frac{d_{l-1}-r_{l-1}}{2}\right)} \tag{121}$$

[2] see proof of Lemma D.3 for a small but important detail of this integral

So the solution is:

$$p(f_1^{(l)},\ldots,f_{d_l}^{(l)}) = \sum_{r_1=0}^{d_1-1} c_{r_1}^1 \cdots \sum_{r_{l-1}=0}^{d_{l-1}-1} c_{r_{l-1}}^{l-1} \pi^{-\frac{d_l}{2}} 2^{-\frac{l d_l}{2}} (\sigma^2)^{-\frac{d_l}{2}} \tag{122}$$

$$G_{0,l}^{l,0}\left(\frac{||\boldsymbol{f}^{(l)}||^2}{2^l\sigma^l}\middle| 0, \frac{1}{2}\left(d_1 - r_1 - d_l\right), \ldots, \frac{1}{2}\left(d_{l-1} - r_{l-1} - d_l\right)\right). \tag{123}$$

Special case: all units are inactive If at least in one layer it happens that all post-activations are zero, then the distribution of the network is a point mass at 0. Let's call this event E, and its probability q_0. The probability of its complement $\bar{E}$ is the probability that for all the intermediate layers, at least one unit is active. These are $l-1$ independent events, the probability of each being $\frac{2^{d_i}-1}{2^{d_i}}$ (one unit is active in $2^{d_i}-1$ cases out of the all possible combinations of units). Therefore we can conclude that:

$$q_0 = 1 - \prod_{i=1}^{l-1} \frac{2^{d_i - 1}}{2^{d_i}} \tag{124}$$

□

D.3. Base case for ReLU nets

Lemma D.3 (second layer pre-activations density). *Let $\sigma = \sigma_1\sigma_2$. Conditioned on the event that at least one unit of the first layer is active ($g_j^{(1)} \neq 0$ for at least one $j \in [d_1]$), the density of the second layer's pre-activations $f_1^{(2)},\ldots f_{d_2}^{(2)}$ is the following linear combination of Meijer-G functions:*

$$p(f_1^{(2)},\ldots f_{d_2}^{(2)}) = \sum_{r=0}^{d_1-1} c_r 2^{-d_2} (\sigma^2)^{-\frac{d_2}{2}} \pi^{\frac{-d_2}{2}} G_{0,2}^{2,0}\left(\frac{||\boldsymbol{f}^{(2)}||^2}{4\sigma^2}\middle| 0, \frac{1}{2}((d_1 - r) - d_2)\right), \tag{125}$$

where $c_r := \binom{d_1}{r}\frac{1}{2^{d_1}\Gamma\left(\frac{d_1-r}{2}\right)}$.

Proof.

$$p(f_1^{(2)},\ldots f_{d_2}^{(2)}) = \int_{\mathbb{R}_{\geq 0}^{d_1}} p(f_1^{(2)},\ldots f_{d_2}^{(2)}|g_1^{(1)},\ldots g_{d_1}^{(1)}) p_{\text{ReLU}}\left(g_1^{(1)},\ldots g_{d_1}^{(1)}\right) d\boldsymbol{g}^{(1)} \tag{126}$$

$$= \int_{\mathbb{R}_{\geq 0}^{d_1}} \prod_{k=1}^{d_2} p(f_k^{(2)}|\boldsymbol{g}^{(1)}) \prod_{k'=1}^{d_1} p_{ReLU}\left(g_{k'}^{(1)}\right) d\boldsymbol{g}^{(1)} \tag{127}$$

$$= \int_{\mathbb{R}_{\geq 0}^{d_1}} \frac{1}{(2\pi\sigma_2^2||\boldsymbol{g}^{(1)}||^2)^{\frac{d_2}{2}}} \exp\left(-\frac{||\boldsymbol{f}^{(2)}||^2}{2\sigma_2^2||\boldsymbol{g}^{(1)}||^2}\right). \tag{128}$$

$$\sum_{A\in\Omega} \frac{1}{2^{|A|}} \prod_{i\in A} \delta(g_i^{(1)}) \frac{1}{(2\pi\sigma_1^2)^{\frac{d_1-|A|}{2}}} \exp\left(-\frac{\sum_{i\in\mathcal{S}\setminus A}(g_i^{(1)})^2}{2\sigma_1^2}\right) \prod_{i\in\mathcal{S}\setminus A} \mathbb{1}_{g_i^{(1)}>0} d\boldsymbol{g}^{(1)} \tag{129}$$

$$= \sum_{A\in\Omega} \frac{1}{2^{|A|}} \frac{1}{(2\pi\sigma_1^2)^{\frac{d_1-|A|}{2}} (2\sigma_2^2)^{\frac{d_2}{2}}} \int_{\mathbb{R}_{\geq 0}^{d_1}} \frac{1}{(||\boldsymbol{g}^{(1)}||^2)^{\frac{d_2}{2}}} \exp\left(-\frac{||\boldsymbol{f}^{(2)}||^2}{2\sigma_2^2||\boldsymbol{g}^{(1)}||^2}\right) \tag{130}$$

$$\prod_{i\in A} \delta(g_i^{(1)}) \exp\left(-\frac{\sum_{i\in\mathcal{S}\setminus A}(g_i^{(1)})^2}{2\sigma_1^2}\right) \prod_{i\in\mathcal{S}\setminus A} \mathbb{1}_{g_i^{(1)}>0} d\boldsymbol{g}^{(1)}, \tag{131}$$

where we can exchange sum and integration due to non-negativeness of the integration variables (Tonelli's theorem). Also, here we abuse the notation and consider that $\mathcal{S}$ is not in the power set, i.e., $\mathcal{S} \notin \Omega$. This is to be consistent with the fact that we are conditioning on the event in which at least one unit is active after the ReLU activation is applied in the first layer.

Now we can use the property of the delta function $\int f(x)\delta(x-x_0)dx = f(x_0)$ and the property of the indicator function $\int_A f(x)\mathbb{1}_{x\in B}dx = \int_B f(x)dx$ and get:

$$\sum_{A\in\Omega}\frac{1}{2^{|A|}}\frac{1}{(2\pi\sigma_1^2)^{\frac{d_1-|A|}{2}}(2\sigma_2^2)^{\frac{d_2}{2}}} \tag{132}$$

$$\int_{\mathbb{R}^{d_1}_{>0}}\frac{1}{(||\boldsymbol{g}^{(1)}_{\mathcal{S}\setminus A}||^2)^{\frac{d_2}{2}}}\exp\left(-\frac{||\boldsymbol{f}^{(2)}||^2}{2\sigma_2^2||\boldsymbol{g}^{(1)}_{\mathcal{S}\setminus A}||^2}\right)\exp\left(-\frac{||\boldsymbol{g}^{(1)}_{\mathcal{S}\setminus A}||^2}{2\sigma_1^2}\right)d\boldsymbol{g}^{(1)}_{\mathcal{S}\setminus A}. \tag{133}$$

Note that the above integral is $(d_1-|A|)$ dimensional due to the effect of the delta. Now the integral(s) above can be solved in an equivalent manner as in the previous section using Lemma C.2[3], and they are equal to

$$\frac{1}{2}\tilde{C}_A(2\sigma_1^2)^{\frac{1}{2}(d_1-|A|-d_2)}G^{2,0}_{0,2}\left(\frac{||\boldsymbol{f}^{(2)}||^2}{4\sigma^2}\middle|0,\frac{1}{2}((d_1-|A|)-d_2)\right) \tag{134}$$

So we can conclude that:

$$p(f_1^{(2)},\dots f_{d_2}^{(2)}) = \sum_{A\in\Omega}\frac{1}{2^{|A|}}\frac{1}{2}\tilde{C}_A(2\sigma_1^2)^{\frac{1}{2}(d_1-|A|-d_2)}\frac{1}{(2\pi\sigma_1^2)^{\frac{d_1-|A|}{2}}(2\sigma_2^2)^{\frac{d_2}{2}}} \tag{135}$$

$$G^{2,0}_{0,2}\left(\frac{||\boldsymbol{f}^{(2)}||^2}{4\sigma^2}\middle|0,\frac{1}{2}((d_1-|A|)-d_2)\right) \tag{136}$$

$$= \pi^{-\frac{d_1}{2}}\sum_{r=0}^{d_1-1}\binom{d_1}{r}\frac{\tilde{C}_r}{2^{r+1}(2\sigma_1^2)^{\frac{d_2}{2}}(2\sigma_2^2)^{\frac{d_2}{2}}\pi^{\frac{-r+d_2}{2}}} \tag{137}$$

$$G^{2,0}_{0,2}\left(\frac{||\boldsymbol{f}^{(2)}||^2}{4\sigma^2}\middle|0,\frac{1}{2}((d_1-r)-d_2)\right), \tag{138}$$

where we have used the fact that the expression depends on the set $A\in\Omega$ only through $|A|$, and therefore we can use the fact that the number of subsets with r elements is given by the binomial coefficient $\binom{d_1}{r}$. Define:

$$c_r := \binom{d_1}{r}\pi^{-\frac{d_1}{2}}\frac{\tilde{C}_r}{2^{r+1}}\pi^{\frac{r}{2}} \tag{139}$$

$$= \binom{d_1}{r}\pi^{-\frac{d_1}{2}}\frac{\pi^{\frac{d_1-r}{2}}}{2^{d_1-r-1}\Gamma\left(\frac{d_1-r}{2}\right)}\frac{1}{2^{r+1}}\pi^{\frac{r}{2}} \tag{140}$$

$$= \binom{d_1}{r}\frac{1}{2^{d_1}\Gamma\left(\frac{d_1-r}{2}\right)}. \tag{141}$$

So we can conclude:

$$p(f_1^{(2)},\dots f_{d_2}^{(2)}) = \sum_{r=0}^{d_1-1}c_r 2^{-d_2}(\sigma^2)^{\frac{-d_2}{2}}\pi^{\frac{-d_2}{2}}G^{2,0}_{0,2}\left(\frac{||\boldsymbol{f}^{(2)}||^2}{4\sigma^2}\middle|0,\frac{1}{2}((d_1-r)-d_2)\right). \tag{142}$$

□

Remark Any non empty subset of $d < d_2$ units has the same distribution (with terms involving d_2 replaced by d).

[3] Note that the integral is only for the positive reals. Lemma C.2 can still be used because when switching to spherical coordinates, we are interested in the radius part, while the angular constant can still be calculated, but now we the angles are all from 0 to $\frac{\pi}{2}$

D.4. Resulting moments

Let $d_l = 1, d_1, \ldots, d_{l-1} = m$, and $b_i = \frac{1}{2}(m - r_i - 1)$, $i = 1, \ldots, l-1$.

$$\begin{aligned}
\mathbb{E}\left[Z^{2k}\right] &= \int_{\mathbb{R}} z^{2k} \sum_{r_1=0}^{m} c_{r_1}^1 \cdots \sum_{r_{l-1}=0}^{m} c_{r_{l-1}}^{l-1} \pi^{-\frac{1}{2}} 2^{\frac{l}{2}} (\sigma^2)^{-\frac{1}{2}} G_{0,l}^{l,0}\left(\frac{z^2}{2^l\sigma^2}\middle| 0, b_1, \ldots, b_{l-1}\right) dz \\
&= \sum_{r_1=0}^{m} c_{r_1}^1 \cdots \sum_{r_{l-1}=0}^{m} c_{r_{l-1}}^{l-1} \pi^{-\frac{1}{2}} 2^{\frac{l}{2}} (\sigma^2)^{-\frac{1}{2}} \int_0^\infty r^{2k} G_{0,l}^{l,0}\left(\frac{r^2}{2^l\sigma^2}\middle| 0, b_1, \ldots, b_{l-1}\right) dr \\
&= \sum_{r_1=0}^{m} c_{r_1}^1 \cdots \sum_{r_{l-1}=0}^{m} c_{r_{l-1}}^{l-1} \pi^{-\frac{1}{2}} 2^{\frac{l}{2}} (\sigma^2)^{-\frac{1}{2}} \int_0^\infty x^{k-\frac{1}{2}} G_{0,l}^{l,0}\left(\frac{x}{2^l\sigma^2}\middle| 0, b_1, \ldots, b_{l-1}\right) dx \\
&= \sum_{r_1=0}^{m} c_{r_1}^1 \cdots \sum_{r_{l-1}=0}^{m} c_{r_{l-1}}^{l-1} \pi^{-\frac{1}{2}} 2^{\frac{l}{2}} (\sigma^2)^{-\frac{1}{2}} \left(\frac{1}{2^l\sigma^2}\right)^{-k-\frac{1}{2}} \Gamma\left(k+\frac{1}{2}\right) \prod_{i=1}^{l-1} \Gamma\left(k+\frac{1}{2}+b_i\right) \\
&= \pi^{-\frac{1}{2}} 2^{\frac{l}{2}} (\sigma^2)^{-\frac{1}{2}} \left(\frac{1}{2^l\sigma^2}\right)^{-k-\frac{1}{2}} \Gamma\left(k+\frac{1}{2}\right) \sum_{r_1=0}^{m} c_{r_1}^1 \cdots \sum_{r_{l-1}=0}^{m} c_{r_{l-1}}^{l-1} \prod_{i=1}^{l-1} \Gamma\left(k+\frac{1}{2}+b_i\right) \\
&= (2k-1)!! 2^{k(l-1)} \sigma^{2k} \left(\frac{1}{2^m} \sum_{r=0}^{m} \binom{m}{r} \frac{\Gamma\left(k+\frac{m-r}{2}\right)}{\Gamma\left(\frac{m-r}{2}\right)}\right)^{l-1}
\end{aligned}$$

For instance, for the variance ($k = 1$) the sum becomes:

$$\sum_{r_1=0}^{m} c_{r_1}^1 \cdots \sum_{r_{l-1}=0}^{m} \binom{m}{r_{l-1}} \frac{1}{2^m \Gamma\left(\frac{m-r_{l-1}}{2}\right)} \prod_{i=1}^{l-1} \frac{1}{2}(m-r_i)\Gamma\left(\frac{1}{2}(m-r_i)\right) \tag{143}$$

$$= \sum_{r_1=0}^{m} \binom{m}{r_1} \frac{(m-r_1)}{2^{m+1}} \cdots \sum_{r_{l-1}=0}^{m} \binom{m}{r_{l-1}} \frac{(m-r_{l-1})}{2^{m+1}} \tag{144}$$

Now each sum can be solved independently:

$$\begin{aligned}
\sum_{r_i=0}^{m} \binom{m}{r_i} \frac{(m-r_i)}{2^{m+1}} &= \frac{1}{2^{m+1}} \left[\sum_{r_i=0}^{m} \binom{m}{r_i} m - \sum_{r_i=0}^{m} \binom{m}{r_i} r_i\right] &(145)\\
&= \frac{1}{2^{m+1}} \left[m 2^m - m 2^{m-1}\right] &(146)\\
&= \frac{m}{4} &(147)
\end{aligned}$$

Therefore the variance is:

$$\begin{aligned}
\mathbb{V}[Z] &= \pi^{-\frac{1}{2}} 2^{\frac{l}{2}} (\sigma^2)^{-\frac{1}{2}} \left(\frac{1}{2^l\sigma^2}\right)^{-1-\frac{1}{2}} \Gamma\left(1+\frac{1}{2}\right) \frac{m^{l-1}}{2^{2(l-1)}} &(148)\\
&= \frac{1}{2} 2^{\frac{l}{2}} (\sigma^2)^{-\frac{1}{2}} \left(\frac{1}{2^{-\frac{3}{2}l}(\sigma^2)^{-\frac{3}{2}}}\right) \frac{m^{l-1}}{2^{2(l-1)}} &(149)\\
&= \frac{1}{2} 2^l \sigma^2 \frac{m^{l-1}}{2^{2(l-1)}} &(150)\\
&= \frac{\sigma^2 m^{l-1}}{2^{l-1}} &(151)
\end{aligned}$$

Note how the variance of a ReLU net is significantly reduce if compared with the variance of a linear network of the same depth (compare with Eq. 58). Similarly, one can get the fourth moment:

$$\mathbb{E}[Z^4] = \frac{3(\sigma^2)^2(m+5)^{l-1} m^{l-1}}{2^{2(l-1)}} \tag{152}$$

Therefore the kurtosis is:

$$\kappa = 3\left(\frac{m+5}{m}\right)^{l-1} \tag{153}$$

Note how ReLU nets are more heavy-tailed than linear nets.

To calculate the asymptotic moments we need three technical Lemmas that express the quantities encountered in a better form. First we describe the coefficients of a factorized polynomial:

Lemma D.4. *Consider coefficients $a_1, \dots, a_m \in \mathbb{R}$. Define the polynomial*

$$p(x) = \prod_{i=1}^{m}(x + a_i) = \sum_{i=1}^{m} \alpha_i x^i$$

Then it holds that $\alpha_m = 1$ and $\alpha_{m-1} = \sum_{i=1}^{m} a_i$.

Next we use Lemma D.4 to write the ratio of Gamma functions as a polynomial:

Lemma D.5. *Fix $k \in \mathbb{N}$ and $x \in \mathbb{R}$. Then we can express the fraction of Gamma functions as follows:*

$$\frac{\Gamma(k + \frac{x}{2})}{\Gamma(\frac{x}{2})} = P_k(x) = \sum_{i=0}^{k} \alpha_i x^i$$

where P_k is a k-th order polynomial with coefficients $\alpha_k = 2^{-k}$ and $\alpha_{k-1} = \frac{k^2-k}{2^k}$.

Proof. The leading coefficient can easily be obtained from multiplying together the terms $\frac{m}{2}$. From Lemma D.4 we conclude that

$$\alpha_{k-1} = \frac{1}{2^{k-1}}\sum_{i=1}^{k}(k-i) = \frac{1}{2^{k-1}}\left(k^2 - \frac{k(k+1)}{2}\right) = \frac{k^2-k}{2^k}$$

□

Next we need to control the sums involving the factorials. Since we just expressed the ratio of Gamma functions as a polynomial, we essentially need to know how to control sums of the type

$$\frac{1}{2^m}\sum_{r=0}^{m}\binom{m}{r} r^k$$

which amounts to controlling the moments of a binomial distribution with fault probability $p = \frac{1}{2}$. We do this as follows:

Lemma D.6. *Fix $k, m \in \mathbb{N}$. Then we can express the following sum as a polynomial $\forall k \in \mathbb{N}$:*

$$\frac{1}{2^m}\sum_{r=0}^{m}\binom{m}{r} r^k = \frac{m}{2^k} Q_{k-1}(m)$$

where Q_{k-1} is a $k-1$-th order polynomial. Moreover, writing Q_l in monomial basis

$$Q_l(m) = \sum_{i=0}^{l} \alpha_i m^i$$

it holds that $\alpha_l = 1$ and $\alpha_{l-1} = \frac{l(l+1)}{2}$ $\forall l \in \mathbb{N}$.

Proof. For a proof of the recursion, we refer to Boros and Moll (2004); Benyi (2005). Moreover the polynomials satisfy the recursion

$$Q_k(m) = 2mQ_{k-1}(m) - (m-1)Q_{k-1}(m-1)$$

Denote by α^k the leading coefficient of Q_k. Using the recursion and performing a comparison of coefficients we see that

$$\alpha^k = 2\alpha^{k-1} - \alpha^{k-1} = \alpha^{k-1}$$

Using the fact that for $k = 1$

$$\frac{1}{2^m}\sum_{r=0}^{m}\binom{m}{r}r = \frac{m}{2} = \frac{m}{2^1}Q_0(m)$$

we conclude that $\alpha_0 = 1$ and thus $\alpha_k = 1 \; \forall k \in \mathbb{N}$. For the second coefficient, we will again use the recursion. Let us first again express the polynomials in monomial bases, i.e.

$$Q_k(m) = \sum_{i=0}^{k}\alpha_i^{(k)}m^i \quad , \qquad Q_{k-1}(m) = \sum_{i=0}^{k-1}\alpha_i^{(k-1)}m^i$$

Using the recursion we thus see that

$$\sum_{i=0}^{k}\alpha_i^{(k)}m^i = \sum_{i=0}^{k-1}2\alpha_i^{(k-1)}m^{i+1} - \sum_{i=0}^{k-1}\alpha_i^{(k-1)}(m-1)^{i+1}$$

We have to understand the terms involving m^{k-1}. Thus we need to expand $(m-1)^k$ which we can do with the help of Lemma D.4:

$$(m-1)^k = m^k - km^{k-1} + \dots$$

We also need to expand the next polynomial as follows:

$$(m-1)^{k-1} = m^{k-1} + \dots$$

Collecting all the coefficients, we end up with the following recursion for the second coefficient (which we denote by α_{-2}:

$$\begin{aligned}\alpha_{-2}^{(k)} &= 2\alpha_{-2}^{(k-1)} - \alpha_{-2}^{(k-1)} + k\alpha_{k-1}^{(k-1)} \\ &= \alpha_{-2}^{(k-1)} + k\end{aligned}$$

Using the fact that

$$Q_1(m) = m + 1$$

thus $\alpha_{-1}^{(1)} = 1$, we conclude that

$$\alpha_{-1}^{(k)} = 1 + \sum_{i=2}^{k} i = \frac{k(k+1)}{2}$$

□

Finally, we need a result on exponential functions and their limit definition:

Lemma D.7. *Fix $c \in \mathbb{R}$ and $\gamma \in \mathbb{R}_+$. Then we have the following limit:*

$$\lim_{m\to\infty}\left(1 + \frac{c}{m} + \mathcal{O}\left(\frac{1}{m^2}\right)\right)^{(\gamma m)} = e^{\gamma c}$$

We can now prove the convergence of the moments as follows.

Lemma D.8. *Consider a ReLU network of equal width $d_1 = \dots = d_{l-1} = m$ and its one dimensional output $f^{(l)}$. Then, under NTK parametrization $\sigma_i^2 = \sigma_w^2 = \frac{2}{m}$, $i = 2, \dots, l-1$ and $\sigma_1 = 1$, it holds that*

$$\mathbb{E}\left[\left(f^{(l)}\right)^{2k}\right] \xrightarrow{m\to\infty} (2k-1)!!$$

Proof. Recall that we arrived at

$$\mathbb{E}\left[\left(f^{(l)}\right)^{2k}\right] = (2k-1)!!2^{k(l-1)}\sigma_w^{2kl}\left(\frac{1}{2^m}\sum_{r=0}^{m}\binom{m}{r}\frac{\Gamma\left(k+\frac{m-r}{2}\right)}{\Gamma\left(\frac{m-r}{2}\right)}\right)^{l-1}$$

Using the NTK parametrization for ReLU, i.e. $\sigma_1^2 = 1$ and $\sigma_2^2 = \dots = \sigma_l^2 = \frac{2}{m}$, this amounts to

$$\mathbb{E}\left[\left(f^{(l)}\right)^{2k}\right] = (2k-1)!!\left(\frac{2^{2k}}{m^k 2^m}\sum_{r=0}^{m}\binom{m}{r}\frac{\Gamma\left(k+\frac{m-r}{2}\right)}{\Gamma\left(\frac{m-r}{2}\right)}\right)^{l-1}$$

We thus essentially need to understand the term

$$M(m) = \frac{2^{2k}}{2^m m^k}\sum_{r=0}^{m}\binom{m}{r}\frac{\Gamma\left(k+\frac{m-r}{2}\right)}{\Gamma\left(\frac{m-r}{2}\right)} = \frac{2^{2k}}{2^m m^k}\sum_{s=0}^{m}\binom{m}{s}\frac{\Gamma\left(k+\frac{s}{2}\right)}{\Gamma\left(\frac{s}{2}\right)}$$

We first use Lemma D.5 to expand the ratio $\frac{\Gamma\left(k+\frac{m-r}{2}\right)}{\Gamma\left(\frac{m-r}{2}\right)}$ as a polynomial. Denote the coefficients by β_i for $i = 1, \dots, k$ ($i \neq 0$ because the polynomial has no intercept). We then swap the two sums:

$$M(m) = \frac{2^{2k}}{2^m m^k}\sum_{s=0}^{m}\binom{m}{s}\sum_{i=1}^{k}s^i\beta_i = \frac{2^{2k}}{m^k}\sum_{i=1}^{k}\beta_i\frac{1}{2^m}\sum_{s=0}^{m}\binom{m}{s}s^i$$

Now we can apply Lemma D.6 to expand the inner sum for each i, denoting the corresponding polynomials again by Q_i:

$$M(m) = \frac{2^{2k}}{m^k}\sum_{i=1}^{k}\beta_i\frac{m}{2^i}Q_{i-1}(m)$$

Notice that $mQ_{i-1}(m)$ is a polynomial of order i. For large m, the factor $\frac{1}{m^k}$ dominates all such polynomials except for the one with $i = k$. Thus in the large-width limit it holds

$$M(m) \xrightarrow{m\to\infty} 2^{2k}\beta_k\frac{1}{2^k} = 1$$

where we used that the leading coefficient of Q_{k-1} is 1. For fixed depth $l \in \mathbb{N}$, we can pull the limit $\lim_{m\to\infty} M(m)^{l-1}$ inside and conclude that

$$\mathbb{E}\left[\left(f^{(l)}\right)^{2k}\right] \xrightarrow{m\to\infty} (2k-1)!!$$

For growing depth $l - 1 = \gamma m$, we have to be a bit more careful since we need to compute the coefficient in front of $\frac{1}{m}$, similarly as in the linear case. We now need to collect all the polynomial terms in $M(m)$ giving rise to a $\frac{1}{m}$ factor. First recall that

$$M(m) = \frac{2^{2k}}{m^k}\sum_{i=1}^{k}\beta_i\frac{m}{2^i}Q_{i-1}(m)$$

The only coefficients contributing to $\frac{1}{m}$ are the second highest coefficient of Q_{k-1} and the highest coefficient of Q_{k-2}. Using Lemma D.6 and Lemma D.5, we hence find that

$$\begin{aligned}M(m) &= 1 + 2^{2k}\left(\beta_k\frac{k(k-1)}{2}\frac{1}{2^k} + \beta_{k-1}\frac{1}{2^{k-1}}\right)\frac{1}{m} + \mathcal{O}\left(\frac{1}{m^2}\right)\\ &= 1 + 2^{2k}\left(\frac{1}{2^{2k+1}}(k-1)k + \frac{1}{2^{2k-1}}(k-1)k\right)\frac{1}{m} + \mathcal{O}\left(\frac{1}{m^2}\right)\\ &= 1 + \left(\frac{5((k-1)}{2}\right)\frac{1}{m} + \mathcal{O}\left(\frac{1}{m^2}\right)\end{aligned}$$

Applying D.7 concludes that

$$\mathbb{E}\left[\left(f^{(l)}\right)^{2k}\right] = (2k-1)!!\left(1+\left(\frac{5k(k-1)}{2}\right)\frac{1}{m}+\mathcal{O}\left(\frac{1}{m^2}\right)\right)^{\gamma m}$$
$$\xrightarrow{m\rightarrow\infty} (2k-1)!!e^{\frac{5\gamma k(k-1)}{2}}$$

Finally, taking $X\sim\mathcal{N}(0,1), Y\sim\mathcal{LN}(-\frac{5}{2}\gamma, 5\gamma)$ and defining $Z = XY$, we can easily see that

$$\mathbb{E}\left[Z^k\right] = \begin{cases} 0 & \text{k odd} \\ (2k-1)!!e^{ks+\frac{k^2t^2}{2}} & \text{k even} \end{cases}$$

Matching the two moments expression concludes the proof. □

E. Additonal results and lemmas

Here we list some of the moments arising from a Binomial distribution of the form $U\sim\text{Bin}(n,\frac{1}{2})$. We invite the reader to sanity-check our results in Lemma D.6 regarding the coefficients of Q_k.

Lemma E.1. *Consider the random variable* $U\sim\text{Bin}(m,\frac{1}{2})$. *We can calculate its first* 5 *moments as*

- $\mathbb{E}[U] = \sum_{i=0}^m \binom{m}{i} i = \frac{m}{2}Q_0(m) = \frac{m}{2}$
- $\mathbb{E}[U^2] = \sum_{i=0}^m \binom{m}{i} i^2 = \frac{m}{2^2}Q_1(m) = \frac{m}{2^2}(m+1)$
- $\mathbb{E}[U^3] = \sum_{i=0}^m \binom{m}{i} i^3 = \frac{m}{2^3}Q_2(m) = \frac{m}{2^3}\left(m^2+3m\right)$
- $\mathbb{E}[U^4] = \sum_{i=0}^m \binom{m}{i} i^4 = \frac{m}{2^4}Q_3(m) = \frac{m}{2^4}\left(m^3+6m^2+3m+4\right)$