Improving the Gaussian Approximation in Neural Networks: Para-Gaussians and Edgeworth Expansions

Mihai Nica
Department of Mathematics and Statistics
University of Guelph
`nicam@uoguelph.ca`

Janosch Ortmann
Département AOTI
Université du Québec à Montéal
`ortmann.janosch@uqam.ca`

Abstract

Gaussian approximations are often used for developing the theory of how neural networks scale as the number of neurons grows large. However, it is known that these approximations break down as depth increases due to the accumulation of approximation errors. To remedy this, we provide a new family of distributions that appear naturally in neural networks and provide more accurate approximations than the usual Gaussian approximation. We develop a method for obtaining the probability density function via Hermite polynomials and connect this to the classical Edgeworth expansion.

1 Introduction

One mathematical method to advance the theoretical understanding of neural networks is through neural scaling limits. In the same way that statistical mechanics understands physical phenomena by considering large numbers of particles, this area of research understands neural networks by considering networks with large numbers of neurons. A common theoretical tool in this area are Gaussian limit, which approximate behaviour when the number of neurons in each layer grows to infinity [MRH$^+$18, JGH18, LXS$^+$19, Yan19, GHLG23].

However, for finite networks, the Gaussian approximation is only approximate! For a fully connected network with n neurons in each layer, microscopic errors of size $1/n$ in each approximation can accumulate through the L layers of a deep neural network and yield macroscopic effects when the depth-to-width ratio L/n is large. Many recent authors have noticed this effect and explored different ways to obtain corrections when network depth is on the same order as the network width [HN20a, HN20b, Yai20, RYH22, SK22a, SK22b, JN24].

In this article, we investigate a method to understand these $1/n$-sized fluctuations away from the simple Gaussian approximation. We develop a general family of distributions, which we call para-Gaussian distributions, that capture the non-Gaussian corrections via the characteristic function, see Section 2. We also show how to use the Hermite polynomials to explicitly evaluate the density function that arises for the special case of ReLU networks with one input in Section 3 and two inputs in Section 4. The error achieved by these methods is lower order than the Gaussian approximation, and is small enough to maintain accuracy as one passes through many layers of the neural network. Thus, it is possible to use the techniques developed here to precisely analyse how non-Gaussian distributions arise in deep neural networks as depth increases.

2 Para-Gaussian Distributions

Definition 1. Let I be some index set. We say that the I-indexed random vector $\{z_\alpha\}_{\alpha \in I} \in \mathbb{R}^I$ is ***para-Gaussian*** (with associated kernels K and C and scaling parameter n) if there exist symmetric

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

positive-definite $\{K(\alpha,\beta)\}_{\alpha,\beta\in I}$ and collection $\{C(\alpha,\beta;\gamma,\delta)\}_{\alpha,\beta,\gamma,\delta\in I}$ such that as $n\to\infty$ we have the following form of the characteristic function for z for any $\vec{\lambda}\in\mathbb{R}^{|I|}$:

$$\mathbb{E}\left[\exp\sum_{\alpha\in I} z_\alpha\lambda_\alpha\right] = \exp\left(-\frac{1}{2}\vec{\lambda}^T K\vec{\lambda}\right)\left(1+\frac{1}{8n}\sum_{\alpha,\beta,\gamma,\delta\in I}\lambda_\alpha\lambda_\beta\lambda_\gamma\lambda_\delta C(\alpha,\beta;\gamma,\delta)+O\left(\frac{1}{n^2}\right)\right). \tag{1}$$

Note that when $C\equiv 0$, this is simply the characteristic function of a I-indexed Gaussian random vector. The factor C therefore creates a $1/n$-sized perturbation around the ordinary Gaussian distribution. Our main general theorem is the following result, which shows that fully connected neural networks naturally create para-Gaussian distributions on initialisation.

Theorem 2. *Let I be an index set and let $\{z_\alpha\}_{\alpha\in I}$ be any I-indexed random vector. Assume that distribution $\{z_\alpha\}_{\alpha\in I}$ has finite exponential moments of all order.*

Suppose that $z_{\cdot,1}, z_{\cdot,2},\ldots,z_{\cdot,n}$ are n independent and identically distributed (iid) copies of $\{z_\alpha\}_{\alpha\in I}$. (We think of α as indexing various inputs to the neural network, and $1\le i\le n$ as indexing neurons in a layer of the network.) Define a new collection of random variables $\{z'_{\alpha,i}\}_{\alpha\in I,i\in[n]}$ by passing through one layer of a neural net with non-linearity φ on initialization as follows:

$$z'_{\alpha,i} := \frac{1}{\sqrt{n}}\sum_{j=1}^{n} W_{ij}\varphi\left(z_{\alpha,j}\right),$$

where W_{ij} iid standard $N(0,1)$ Gaussians. Then the output $z'_{\alpha,1}, z'_{\alpha,2},\ldots,z'_{\alpha,n}$ are iid and the distribution $\{z'_\alpha\}_{\alpha\in I}$ of each element is para-Gaussian with kernels K' and C' given explicitly by:

$$K'(\alpha,\beta)=\mathbb{E}\left[\varphi(z_\alpha)\varphi(z_\beta)\right],\ \ C'(\alpha,\beta,\gamma,\delta)=\mathbb{E}\left[\varphi\left(z_\alpha\right)\varphi\left(z_\beta\right)\varphi\left(z_\gamma\right)\varphi\left(z_\delta\right)\right]-K'(\alpha,\beta)K'(\gamma,\delta).$$

Theorem 2 shows that if we have an index set I of inputs to a neural network, then the operation of moving from one layer of i.i.d. neuron values z to the next layer of neuron values z' results in a para-Gaussian distribution. By iterating this, we see that *all* the layers of the neural network will be para-Gaussian on initialization, with various kernels K, C depending on the layer depth. To understand the dependence on depth therefore, one has only to analyze the functions K and C, which will evolve through the layers of the network. This presents a way of understanding the evolution of the distribution in deep networks through the characteristic function. This is in contrast to previous work that has analyzed the non-Gaussian phenomenon through the lens of the moments of the distributions [HN20b, Yai20, RYH22]. More recent work [DHM$^+$24] uses techniques from field theory in physics to also understand non-Gaussian phenomenon more precisly as a series expansion. The main contribution of our work is to explicitly recover the probability density using the Hermite orthogonal polynomials. For multiple points, where $|I|>2$, multivariable generalizations of the Hermite polynomials are necessary which significantly complicate the calculations.

3 A Single Input: Connection to Edgeworth Expansions

In the special case that the index set $I=\{1\}$ is a single point, we are dealing with simple $\mathbb{R}$ valued random variables. The law of $z'\in\mathbb{R}$ is simply a random sum:

$$z'=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i\varphi(z_i), \tag{2}$$

where W_i are iid standard Gaussian random variables. The fact that z' converges to a Gaussian random variable as $n\to\infty$ is the classical Central Limit Theorem (CLT). The para-Gaussian correction term in this case is $C(1,1;1,1)=\mathbb{E}\left[\varphi(z)^4\right]-\mathbb{E}\left[\varphi(z)^2\right]^2$. For the special case that $\varphi(x)=\sqrt{2}(x)_+$ is the calibrated ReLU non-linearity, and the input z is a standard $N(0,1)$ Gaussian, one can calculate by elementary means that $C(1,1;1,1)=5$. In Appendix A.3, we show how to compute the density of this para-Gaussian distribution in this case, and find that the density by:

$$\rho_{\text{paraGaussian}}(x)=e^{-\frac{1}{2}x^2}\left(1+\frac{5}{8n}H_4(x)+O\left(\frac{1}{n^2}\right)\right), \tag{3}$$

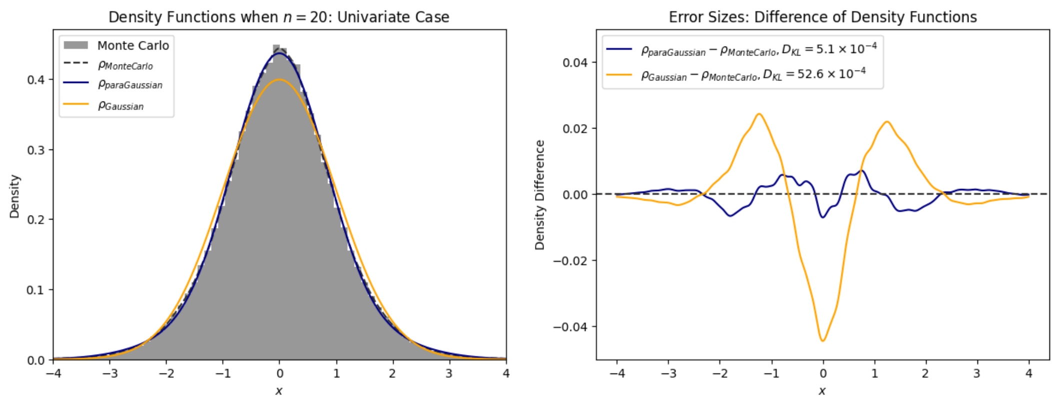

Figure 1: Monte Carlo simulations comparing the 1D para-Gaussian distribution prediction (3) in the case of a single input $|I| = 1$ when $n = 20$, compared to the pure Gaussian density $(2\pi)^{-1/2}e^{-x^2/2}$. 2^{18} Monte Carlo samples of (2) are simulated and Kernel Density Estimation is used to obtain $\rho_{MonteCarlo}$. $D_{KL} = D_{KL}(\rho_{Approx}, \rho_{MonteCarlo})$ is calculated with respect to reference probability measure $\rho_{MonteCarlo}$.

where $H_4(x) = x^4 - 6x^2 + 3$ is the 4th order Hermite polynomial.

In Figure 1 we empirically show that the $1/n$ correction term in the para-Gaussian distribution matches Monte Carlo simulations much more accurately than the pure Gaussian approximation. This is comparable to similar plots in [Yai20] which are derived using the moments of the distribution instead of through developing a formula for the density function with the Hermite polynomials.

The formula for the density for the para-Gaussian distribution of this special case matches the classical Edgeworth expansion for the sum (2) [Hal13]. Edgeworth expansions provide a series expansion in $1/n$ that better approximate random sums; the classical CLT is the 0-th order term only. The para-Gaussian distribution in this case is the distribution which includes 1st correction term of the Edgeworth expansion. By this connection, one can think of the idea of a para-Gaussian distribution from Section 2 as a generalization of the classical Edgeworth expansion to multi-variable situations.

4 Two Inputs: Multivariate Hermite Polynomials

In the case where the index set is two points, $I = \{1, 2\}$, we are looking at vectors with two components. For concreteness, we again consider the case where $\varphi(x) = \sqrt{2}(x)_+$ is the calibrated ReLU. By scaling z_1, z_2 by constant factors, one can assume without loss of generality that $K(1,1) = \mathbb{E}[z_1^2] = 1, K(2,2) = \mathbb{E}[z_2^2] = 1$ and find $\theta \in (0, \pi)$ so that $K(1,2) = \mathbb{E}[z_1 z_2] = \cos(\theta)$. In this case, we develop the following formula for the 2D probability density function for the output z'. Figure 4 shows Monte Carlo simulations of this denisty with input $\theta = 0.55$ to confirm again that the para-Gaussian distribution is more accurate than the pure Gaussian distribution. The probability density function we find is:

$$\rho_{\text{paraGaussian}}(x_1, x_2) = \rho_{\text{Gaussian},\theta'}(x_1, x_2)\left(1 + \frac{1}{n}\mathcal{H}^\theta(x_1, x_2) + O\left(\frac{1}{n^2}\right)\right) \tag{4}$$

where θ' is the angle between the outputs z_1', z_2' given in (6), and $\rho_{\text{Gaussian},\theta'}$ is the 2d Gaussian density with covariance structure $K = \begin{pmatrix} 1 & \cos(\theta') \\ \cos(\theta') & 1 \end{pmatrix}$, and where the Hermite correction $\mathcal{H}^\theta$ is:

$$\mathcal{H}^\theta(x_1, x_2) = \frac{1}{8\sin^4(\theta)} \sum_{a=1}^{4} \widehat{C}(a) H_4^{(-\cos\theta)}\left(\underbrace{\left(\sin\theta + \frac{\cos^2(\theta)}{\sin(\theta)}\right)x_1 - \frac{\cos(\theta)}{\sin(\theta)}x_2,}_{a \text{ times}} \underbrace{\frac{1}{\sin(\theta)}x_2 - \frac{\cos(\theta)}{\sin(\theta)}x_1}_{4-a \text{ times}}\right)$$

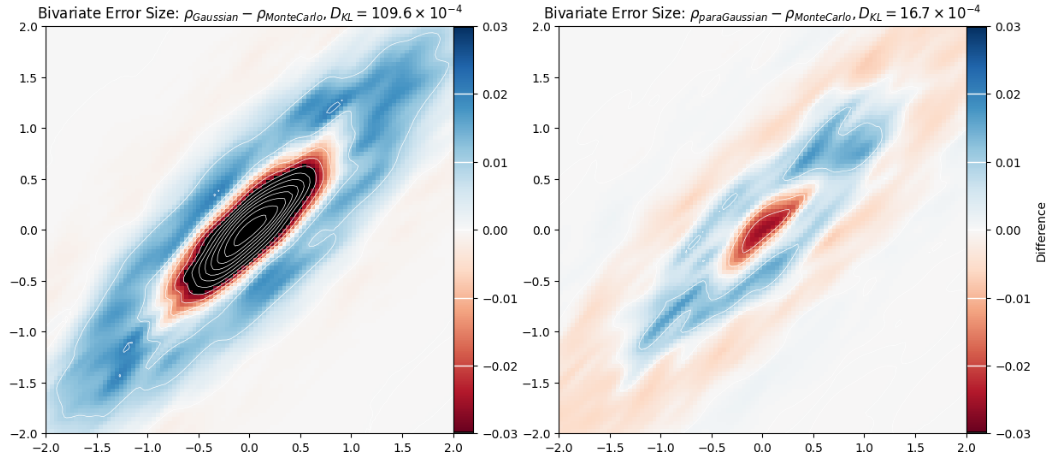

Figure 2: Comparing the error sizes of the Gaussian and para-Gaussian approximation (4) in the two variable case $I = \{1, 2\}$. For this experiment, $\theta = 0.55$, $n = 20$, and $\rho_{MonteCarlo}$ estimated from 2^{18} Monte Carlo samples by using Kernel Density Estimation. $D_{KL} = D_{KL}(\rho_{Approx}, \rho_{MonteCarlo})$ is calculated with respect to reference probability measure $\rho_{MonteCarlo}$.

where $\widehat{C}(a)$ is the sum of $C(\alpha, \beta, \gamma, \delta)$ over indices with a number of 1s and $H_4^{(B)}$ are 4^{th} order multidimensional Hermite polynomial given explicitly as follows:

$$
\begin{aligned}
&H_4^{(B)}(U,U,U,U) = U^4 - 6U^2 + 3, \quad H_4^{(B)}(V,V,V,V) = V^4 - 6V^2 + 3 \\
&H_4^{(B)}(U,U,U,V) = U^3V - 3UV - 3BU^2 + 3B, \quad H_4^{(B)}(U,V,V,V) = UV^3 - 3UV - 3BV^2 + 3B \\
&H_4^{(B)}(U,U,V,V) = U^2V^2 - V^2 - U^2 - 4BUV + 2B^2 + 1.
\end{aligned}
$$

These Hermite polynomials are a special case of the more general multi-variable Hermite polynomials [Sob63], which we define combinatorially in terms of a collection of 4 Gaussian random variables X_1, X_2, X_3, X_4 and the correlation function $\{\mu_{a,b}\}_{1 \le a,b \le 4}$, $\mu_{a,b} = \mathbb{E}[X_a X_b]$:

$$
\begin{aligned}
H_4^{(\mu_{a,b})}(x_1, x_2, x_3, x_4) = {} & x_1x_2x_3x_4 - x_1x_2\mu_{3,4} - x_1x_3\mu_{2,4} - x_1x_4\mu_{2,3} \\
& - x_2x_3\mu_{1,4} - x_2x_4\mu_{1,3} - x_2x_4\mu_{1,3} + \mu_{1,2}\mu_{3,4} + \mu_{1,3}\mu_{2,4} + \mu_{1,4}\mu_{2,3}.
\end{aligned} \tag{5}
$$

The classical 4th order Hermite polynomial $H_4(x)$ corresponds to the case where all $X_1 = X_2 = X_3 = X_4$ so that $\mu_{a,b} = 1$ always. In our use case, the random variables are repeated (the first one a times and the second one $4 - a$ times) and the correlation between them is taken to be $-\cos\theta$, which yields the formulas given in (5).

4.1 Computing K' and C'

We now compute the kernels K' and C' in the bivariate case for the ReLU when the input z is itself para-Gaussian and assume by rescaling the z_α that $K = \begin{pmatrix} 1 & \cos(\theta) \\ \cos(\theta) & 1 \end{pmatrix}$ for some $\theta \in [0, \pi]$. By iterating these calculations, one can compute in principle the para-Gaussian kernels for any layer of a deep neural network

Theorem 3. *With the setup of Theorem 2, in the case that* $I = \{1, 2\}$*, we have* $K' = \begin{pmatrix} 1 & \cos(\theta') \\ \cos(\theta') & 1 \end{pmatrix}$*, where*

$$
\cos(\theta') = \frac{\sin(\theta) + (\pi - \theta)\cos(\theta)}{\pi} \tag{6}
$$

$$
+ \frac{1}{n} \cdot \frac{1}{4\sin(\theta)} \left(C(1,1,1,1)\cos(2\theta) - 4C(1,1;1,2)\cos(\theta) + [2C(1,1;2,2) + C(1,2;1,2)] \right) \tag{7}
$$

Moreover, let r be the number of 1's in $(\alpha, \beta, \gamma, \delta)$. Then

$$C'(\alpha, \beta, \gamma, \delta) = \Gamma_r(\theta) - \cos(\theta')^{1_{\alpha\neq\beta}+1_{\gamma\neq\delta}}$$

where

$$\Gamma_r(\theta) = \begin{cases} 6 & \text{if } r \in \{0, 4\} \\ \frac{9\sin(\theta)+\sin(3\theta)+12(\pi-\theta)\cos(\theta)}{2\pi} & \text{if } r \in \{1, 3\} \\ \frac{6\sin(2\theta)+4(\pi-\theta)(\cos(2\theta)+2)}{2\pi} & \text{if } r = 2. \end{cases} \tag{8}$$

In order to prove Theorem 3, we need to compute expressions of the form $\mathbb{E}\left[f(\vec{z}')\right]$, for different functions f. The following result gives us a recipe of doing so.

Proposition 4. *Let $\vec{z}$ be a random vector in $\mathbb{R}^2$ with a para-Gaussian distribution with kernels $K = \begin{pmatrix} 1 & \cos(\theta) \\ \cos(\theta) & 1 \end{pmatrix}$ and $C \in \mathbb{R}^{4\times 4}$. For any function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$,*

$$\begin{aligned} \mathbf{E}\left[f(\vec{z})\right] &= \mathbf{E}\left[f(\vec{\zeta})\right] + o(\frac{1}{n}) \\ &+ \frac{1}{8n\sin(\theta)^4} \sum_{\alpha,\beta,\gamma,\delta} C(\alpha, \beta; \gamma, \delta) \mathbf{E}\left[H_4\left(\underbrace{R\sin(\theta-\alpha)}_{a \text{ times}}, \underbrace{R\sin(\alpha)}_{b \text{ times}}\right) f\left(R\cos(\alpha), R\cos(\theta-\alpha)\right)\right] \end{aligned} \tag{9}$$

where a and b count the number of occurrences of 1 and 2 in $(\alpha, \beta, \gamma, \delta)$ respectively, $\vec{\zeta}$ is a centred two-dimensional Gaussian with covariance matrix K and R and α are independent random variables with Raleigh (scale 1) and uniform $[0, 2\pi]$ distribution respectively.

Equation (4) follows from Proposition 4 by choosing f to be a suitable indicator function. A sketch of the proof of Proposition 4 is given in Appendix A.2. The remainder of this paper sketches the proof of Theorem 3 by applying Proposition 4 to compute K' and C'.

4.1.1 Formula for K'

In order to compute the kernel K, we first compute the diagonal (Lemma 5) and then the off-diagonal terms (Lemma 6).

Lemma 5. *We have $K'(1, 1) = K'(2, 2) = 1$*

Proof. The fact that $K'(1, 1) = K'(2, 2)$ follows from symmetry. To compute $K'(1, 1)$ we apply Proposition 4 to the function $f(z_1, z_2) = \varphi(z_1)^2$ First note that by calibration of the ReLU φ,

$$\mathbf{E}[\varphi(\zeta_1)^2] = 1.$$

For the order-$\frac{1}{n}$ corrections, we first observe that the only non-zero term in (9) is when $\alpha = \beta = \gamma = \delta = 1$ (i.e. $a = 4$) (for example by appealing to Proposition 7 in Appendix A.2 and a limiting argument). For the remaining term, (9) yields

$$\begin{aligned} \mathbf{E}\left[H_4(\zeta_1) f(\zeta_1)\right] &= \mathbf{E}\left[H_4(\zeta_1)\varphi^2(\zeta_1)\right] = 2\mathbf{E}\left[H_4(|\zeta_1|)|\zeta_1|^2 1_{\zeta_1>0}\right] \\ &= 2\mathbf{E}[1_{\zeta_1>0}]\mathbf{E}\left[H_4(|\zeta_1|)|\zeta_1|^2\right] = 0 \end{aligned}$$

where we have used the fact that $|\zeta_1|$ and $1_{\zeta_1>0}$ are independent, that x^2 lies in the linear span of $\{H_0(x), H_2(x)\}$ and that the Hermite polynomials are orthogonal with respect to the Gaussian measure. $\square$

Lemma 6. *We have $K'(1, 2) = K'(2, 1) = \cos(\theta')$ with θ' as in* (6).

Proof. For the constant term, using the fact that $\vec{\zeta} = (R\cos(\alpha), R\sin(\alpha))$ and that $\mathbf{E}R^2 = 2$,

$$\begin{aligned} \mathbf{E}[\varphi(\zeta_1)\varphi(\zeta_2)] &= \mathbf{E}_{R,\alpha}\left[\varphi(R\cos(\alpha))\varphi\left(R\cos\left(\theta-\alpha\right)\right)\right] \\ &= 2\mathbf{E}[R^2]\mathbf{E}_\alpha\left[\cos\left(\alpha\right)\cos\left(\theta-\alpha\right) 1_{\{-\pi/2+\theta<\alpha<\pi/2\}}\right] \\ &= \frac{2}{\pi}\int_{-\pi/2+\theta}^{\pi/2} \cos(\alpha)\cos(\theta-\alpha)\mathrm{d}\alpha = J_{1,1}(\theta). \end{aligned}$$

For the order-$\frac{1}{n}$ term, we need to consider three cases: $a = 0$, $a = 1$ and $a = 2$. The remaining two cases follow by symmetry. We show the case $a = 0$, the others are by a similar computation. In the following, we use the fact that $(\mathbf{E}R^2, \mathbf{E}R^4, \mathbf{E}R^6) = (2, 8, 48)$ for the Raleigh-distributed random variable R and that R and α are independent:

$$\begin{aligned}
&\mathbf{E}\left[H_4\left(R\sin(\alpha)\right)\varphi\left(R\cos(\alpha)\right)\varphi\left(R\cos(\theta-\alpha)\right)\right] \\
&= \sqrt{2}^2\mathbf{E}\left[\left(R^6\sin(\alpha)^4 - 6R^4\sin(\alpha)^2 + 3R^2\right)\cos(\alpha-\theta)\cdot\cos\left(\alpha\right)\mathbb{1}_{\{-\frac{\pi}{2}+\theta<\alpha<\frac{\pi}{2}\}}\right] \\
&= 2\mathbf{E}_\alpha\left[\left(48\sin(\alpha)^4 - 6\cdot 8\sin(\alpha)^2 + 3\cdot 2\right)\cos(\alpha-\theta)\cdot\cos\left(\alpha\right)\mathbb{1}_{\{-\frac{\pi}{2}+\theta<\alpha<\frac{\pi}{2}\}}\right] \\
&= \frac{1}{\pi}\int_{-\frac{\pi}{2}+\theta}^{\pi/2}\cos(4\alpha)\cos\left(\alpha-\theta\right)\cos\left(\alpha\right)\mathrm{d}\alpha = \frac{\sin^3(\theta)\cos(2\theta)}{\pi}.
\end{aligned}$$

Using a similar computation we obtain $-\frac{\cos(\theta)\sin^3(\theta)}{\pi}$ for $a = 1$ and $\frac{\sin^3(\theta)}{\pi}$ for $a = 2$. □

4.1.2 Formula for C'

To compute the $C'(\alpha, \beta; \gamma, \delta)$ terms, it remains to compute $\mathbf{E}\left[\varphi\left(z'_\alpha\right)\varphi\left(z'_\beta\right)\varphi\left(z'_\gamma\right)\varphi\left(z'_\delta\right)\right]$ for all $\alpha, \beta, \gamma, \delta \in \{1, 2\}$. By symmetry, this only depends on the number of indices equal to 1. Let us denote that number by r, so that there are r 1's and $4 - r$ 2's in $(\alpha, \beta, \gamma, \delta)$.

By Proposition 4 applied to the function $f(z) = \varphi(z_1)^r\varphi(z_2)^{4-r}$,

$$\mathbf{E}\left[\varphi\left(z'_\alpha\right)\varphi\left(z'_\beta\right)\varphi\left(z'_\gamma\right)\varphi\left(z'_\delta\right)\right] = \Gamma_r(\theta) + O(1/n),$$

where $\Gamma_r(\theta) = \mathbf{E}\left[\varphi\left(\zeta_1\right)^r\varphi\left(\zeta_2\right)^{4-r}\right]$ (recall that $\vec{\zeta}$ is a centred Gaussian with correlation matrix K).

When $r = 0$ or $r = 4$, we get $\Gamma_r(\theta) = \mathbf{E}[\varphi(\zeta_1)^4] = 6$. For $r \in \{1, 2, 3\}$, by writing $(\zeta_1, \zeta_2) = (W_1, \cos(\theta)W_1 + \sin(\theta)W_2)$ for a standard Gaussian (W_1, W_2) so that $(W_1, W_2) = R(\cos\alpha, \sin\alpha)$ for R, α as in Proposition 4 (and arguing similarly to the proof of Lemma 6),

$$\begin{aligned}
\Gamma_r(\theta) &= \mathbf{E}\left[\phi(R\cos\alpha)^r\phi\left(R\left(\cos(\theta)\cos(\alpha) + \sin(\theta)\sin(\alpha)\right)\right)^{4-r}\right] \\
&= \frac{16}{\pi}\int_{-\pi/2+\theta}^{\pi/2}\cos(\alpha)^r\cos\left(\theta-\alpha\right)^{4-r}\,d\alpha
\end{aligned}$$

Evaluating each this integrals for each $r \in \{1, 2, 3\}$ yields (8).

5 Conclusion

We have developed a framework using para-Gaussian distributions for approximating the behaviour of neurons in deep neural networks that takes into account small $1/n$-sized corrections beyond the Gaussian law. We have also demonstrated how one can recover the density of these distributions by use of the Hermite polynomials in simple cases. We believe that this provides a first step towards using this method to understanding what happens in networks as depth increases, by iterating the evolution $C, K \to K', C'$ over many layers. By generalising these results to larger input index sets, $|I| \geq 3$, one may also understand how the joint distribution for many points evolves.

Acknowledgements

The authors wish to extend thanks to Boris Hanin for showing us the trick with Hermite polynomials for inverting the characteristic function. This is what first set us down the path of this article!

A Proof Ideas

A.1 Sketch of Proof of Theorem 2

Sketch Proof of Theorem 2. In the following argument we omit the bounds required to deal with the CLT approximation. Let $\mathcal{G}$ denote the σ-algebra generated by the $\vec{z}$. Then, using the iterated conditional expectation,

$$\mathbf{E}\left[\exp\left(\imath\left\langle\vec{\lambda},\vec{z}'\right\rangle\right)\right]=\mathbb{E}\left[\mathbb{E}\left[\exp\left\{\imath\left\langle\vec{\lambda},\vec{z}'\right\rangle\right\}|\mathcal{G}\right]\right].$$

Now, conditional on $\mathcal{G}$, the z'_α are mean zero Gaussian with covariance

$$\Sigma_{\alpha\beta}:=\mathbb{E}\left[z'_\alpha z'_\beta|\mathcal{G}\right]=\frac{1}{n}\varphi(z_\alpha)\varphi(z_\beta).$$

Thus the conditional expectation on the right-hand side is given by

$$\mathbb{E}\left[\exp\left\{\imath\left\langle\vec{\lambda},\vec{z}'\right\rangle\right\}|\mathcal{G}\right]=\exp\left\{-\frac{1}{2}\lambda^T\Sigma\lambda\right\}=\exp\left\{-\frac{1}{2}\cdot\frac{1}{n}\sum_{(\alpha,\beta)}\left(\lambda_\alpha\lambda_\beta\right)\varphi(z_\alpha)\varphi(z_\beta)\right\},$$

where we have used the fact that $\mathbb{E}e^{i\lambda\cdot X}=\exp\left\{-\frac{1}{2}\lambda\Sigma\lambda\right\}$ for a multivariate centred Gaussian X with covariance matrix Σ. Now by the CLT applied to the z (which are sums of unconditionally i.i.d. random variables):

$$\frac{1}{n}\varphi(z_\alpha)\cdot\varphi(z_\beta)\approx\mathbb{E}\left[\varphi(z_{\alpha,1})\varphi(z_{\beta,1})\right]+\frac{1}{\sqrt{n}}G_{\alpha,\beta}=K(\alpha,\beta)+\frac{1}{\sqrt{n}}G_{\alpha,\beta}$$

where $\approx$ means equality up to lower order terms in n and the $G_{\alpha,\beta}$ are centred normals with the covariance structure $\mathbf{Cov}(G_{\alpha,\beta},G_{\gamma,\delta})=C(\alpha,\beta,\gamma,\delta)$. Therefore, up to corrections of order $\frac{1}{n}$,

$$\begin{aligned}\mathbf{E}\left[\exp\left(\imath\left\langle\vec{\lambda},\vec{z}'\right\rangle\right)\right]&\approx\mathbb{E}\left[\exp\left\{-\frac{1}{2}\cdot\sum_{(\alpha,\beta)}\left(\lambda_\alpha\lambda_\beta\right)\left(K(\alpha,\beta)+\frac{1}{\sqrt{n}}G_{\alpha,\beta}\right)\right\}\right]\\&=\exp\left\{-\frac{1}{2}\cdot\lambda^TK\lambda\right\}\mathbb{E}\left[\exp\left\{\sum_{\alpha,\beta}M_{\alpha,\beta}G_{\alpha,\beta}\right\}\right],\end{aligned}$$

where $M_{\alpha,\beta}=-\frac{1}{2\sqrt{n}}\lambda_\alpha\lambda_\beta$. Now, using the fact that $\mathbb{E}\left[\exp\left\{G\right\}\right]=\exp\left\{\frac{1}{2}\mathbf{Var}(G)\right\}$ for a centred Gaussian G,

$$\begin{aligned}\mathbf{E}\left[\exp\left(\imath\left\langle\vec{\lambda},\vec{z}'\right\rangle\right)\right]&\approx\exp\left\{-\frac{1}{2}\cdot\lambda^TK\lambda\right\}\exp\left\{\frac{1}{2}\mathbf{Var}\left[\sum_{\alpha,\beta}M_{\alpha,\beta}G_{\alpha,\beta}\right]\right\}\\&=\exp\left\{-\frac{1}{2}\cdot\lambda^TK\lambda\right\}\exp\left\{\frac{1}{2}\sum_{\alpha,\beta}\sum_{\gamma,\delta}M_{\alpha,\beta}C(\alpha,\beta,\gamma,\delta)M_{\gamma,\delta}\right\}\\&=\exp\left\{-\frac{1}{2}\cdot\lambda^TK\lambda\right\}\exp\left\{\frac{1}{2}\cdot\frac{1}{4n}\sum_{\alpha,\beta,\gamma,\delta}C\left(\alpha,\beta,\gamma,\delta\right)\lambda_\alpha\lambda_\beta\lambda_\gamma\lambda_\delta\right\}\end{aligned}$$

as claimed. ∎

A.2 Sketch proof of Proposition 4

The proof is split into two parts. The first is a formula for the expectation of $f(\vec{z})$ when f is four times differentiable:

Proposition 7. *Let I $\vec{z}$ be a random vector in $\mathbb{R}^I$ with a para-Gaussian distribution with kernels K and C. For any function $f \in C^4(\mathbb{R}^I)$,*

$$\mathbf{E}\left[f(\vec{z})\right] = \mathbf{E}\left[f(\vec{\zeta})\right] + \frac{1}{8n} \sum_{\alpha,\beta,\gamma,\delta \in I^4} C(\alpha,\beta;\gamma,\delta)\mathbf{E}\left[\left(\partial_\alpha\partial_\beta\partial_\gamma\partial_\delta f\right)(\vec{\zeta})\right] + O(\frac{1}{n^2})$$

where $\vec{\zeta}$ is an I-indexed centred Gaussian with covariance kernel K.

Proof. For any function, let $\widehat{f}$ denote the Fourier transform of f and write $d = |I|$. In particular, the characteristic function ϕ of $\vec{z}'$ can then be written $\phi = \widehat{g}_{\vec{z}}$ where $g_{\vec{z}}$ is the density of the law of $\vec{z}'$ with respect to the d-dimensional Lebesgue measure. Now Theorem 2 gives us a formula for the characteristic function ϕ of $\vec{z}'$. In order to use it to compute the expectation of $f(\vec{z}')$ we use the Parseval formula and the Taylor expansion $\exp(\frac{A}{2n}) = 1 + \frac{A}{2n} + O(\frac{1}{n^2})$ to obtain (denoting by $\approx$ equality up $O(\frac{1}{n^2})$ terms),

$$\begin{aligned}
\mathbf{E}f(\vec{z}') &= \int_{\mathbb{R}^d} \widehat{f}(\vec{\lambda})\phi(\vec{\lambda})\, d\vec{\lambda} \\
&\approx \int_{\mathbb{R}^d} \hat{f}(\vec{\lambda}) \exp\left(-\frac{(2\pi)^d}{2}\vec{\lambda}^T K\lambda\right) \mathrm{d}\vec{\lambda} \\
&\quad + \frac{1}{8n} \sum_{\alpha,\beta,\gamma,\delta \in I} C'(\alpha,\beta;\gamma,\delta) \int (2\pi)^{2d} \hat{f}(\vec{\lambda})\lambda_\alpha\lambda_\beta\lambda_\gamma\lambda_\delta \exp\left(-\frac{(2\pi)^d}{2}\lambda^T K\lambda\right) \mathrm{d}\lambda_I
\end{aligned}$$

The result now follows by recognising $\lambda_\alpha\lambda_\beta\lambda_\gamma\lambda_\delta\widehat{f}(\vec{\lambda})$ as the Fourier transform of $\partial_\alpha\partial_\beta\partial_\gamma\partial_\delta f$ and then applying the Parseval formula again. □

Unfortunately the expression in Proposition 7 is not sufficient for us because the ReLU function ϕ is not differentiable. In order to get rid of the derivatives, we use a multidimensional integration-by-parts formula for the Gaussian law (see Lemma 9 below). It is this formula that leads to the appearance of the Hermite polynomials and leads us to the following result:

Proposition 8. *For $\alpha, \beta, \gamma, \delta \in \{1, 2\}$ let a be the number of 1's in $(\alpha, \beta, \gamma, \delta)$. Then,*

$$\mathbf{E}\left[\left(\partial_\alpha\partial_\beta\partial_\gamma\partial_\delta f\right)(\vec{\zeta})\right] = \frac{1}{\sin^4(\theta)}\mathbf{E}\left[H_4\left(\underbrace{\sin\theta W_1 - \cos\theta W_2}_{a\ \textit{times}},\ \underbrace{W_2}_{4-a\ \textit{times}}\right) f\left(W_1, \cos(\theta)W_1 + \sin(\theta)W_2\right)\right],$$

where (W_1, W_2) is a two-dimensional mean zero, identity covariance matrix Gaussian.

By combining Propositions 7 and 8, we obtain Proposition 4 for four times differentiable functions. This can then be extended to all measurable functions for which the expectations in (9) are finite by a density argument.

Thus complete this section by sketching the proof of Proposition 8. A key ingredient is the following integration by parts formula.

Lemma 9. *For any $a, b \in \mathbb{R}$, any suitable function $g\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ and any polynomial $\pi \in \mathbb{R}[x_{1,}, x_2]$ we have*

$$\mathbf{E}\left[\pi\left(W_1, W_2\right)\left(a\partial_1 + b\partial_2\right) g(W_1, W_2)\right] = \mathbf{E}\left[\left(aT_1 + bT_2\right)(\pi)[W_1, W_2] \cdot g(W_1, W_2)\right],$$

and hence, choosing $\pi = 1$,

$$\mathbf{E}\left[\left(a\partial_1 + b\partial_2\right) g(W_1, W_2)\right] = \mathbf{E}\left[\left(aT_1 + bT_2\right)(1)[W_1, W_2] \cdot g(W_1, W_2)\right] = \mathbf{E}\left[H(aW_1 + bW_2)g(W_1, W_2)\right]$$

where $(T_\alpha f)(w_{1,} w_2) = w_\alpha f(w_1, w_2) - \partial_\alpha f(w_1, w_2)$.

Proof. The second equality follows from the definition of the multivariable Hermite polynomials. For the first equality we may take without loss of generality $a = 1$ and $b = 0$ (by linearity and symmetry).

Let ρ denote the density of (W_1, W_2). Applying integration by parts, the chain rule and then the identity $\partial_1 \rho(w_1, w_2) = -w_1 \rho(w_1, w_2)$, we obtain

$$\begin{aligned}
\mathbf{E}\left[\partial_1 g(W_1, W_2)\right] &= \int_{\mathbb{R}^2} \pi(w_1, w_2) \left(\partial_1 g\right)(w_1, w) \rho(w_1, w_2) \, \mathrm{d}w \\
&= -\int_{\mathbb{R}^2} \partial_1 \left(\pi \rho\right)(w_1, w_2) \, g\left(w_1, w_2\right) \, \mathrm{d}w \\
&= -\int_{\mathbb{R}^2} \left[\left(\partial_1 \pi\right) \rho + \pi \partial_1 \rho\right](w_1, w_2) \; g\left(w_1, w_2\right) \, \mathrm{d}w \\
&= \int_{\mathbb{R}^2} \left(w_1 - \partial_1 \pi(w_1, w_2)\right) g\left(w_1, w_2\right) \rho\left(w_1, w_2\right) \, \mathrm{d}w \\
&= \mathbb{E}\left[T_1(\pi)\left(W_1, W_2\right) g\left(W_1, W_2\right)\right].
\end{aligned}$$

□

The proof of Proposition 8 now follows by repeatedly applying Lemma 9 and by using the interaction between the multidimensional Hermite polynomials and the operators T.

A.3 Proof of (2)

To prove (2), we appeal to Proposition 7 with $I = \{1\}$. As mentioned in the paragraph after (2), we have $C(1,1,1,1) = 5$ in that case. An application of the one-dimensional Gaussian integration by parts formula (which Lemma 9 generalises) allows to remove the four derivatives and to obtain the fourth Hermite polynomial. The density of the law is then computed by choosing f a suitable indicator function.

References

[DHM+24] Mehmet Demirtas, James Halverson, Anindita Maiti, Matthew D Schwartz, and Keegan Stoner. Neural network field theories: non-gaussianity, actions, and locality. *Machine Learning: Science and Technology*, 5(1):015002, jan 2024.

[GHLG23] Tianxiang Gao, Xiaokai Huo, Hailiang Liu, and Hongyang Gao. Wide neural networks as gaussian processes: Lessons from deep equilibrium models. *Advances in Neural Information Processing Systems*, 36:54918–54951, 2023.

[Hal13] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

[HN20a] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[HN20b] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8580–8589, Red Hook, NY, USA, 2018. Curran Associates Inc.

[JN24] Cameron Jakub and Mihai Nica. Depth degeneracy in neural networks: Vanishing angles in fully connected relu networks on initialization. *Journal of Machine Learning Research*, 25(239):1–45, 2024.

[LXS+19] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[MRH+18] Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

[RYH22] Daniel A. Roberts, Sho Yaida, and Boris Hanin. *The Principles of Deep Learning Theory*. Cambridge University Press, 2022. `https://deeplearningtheory.com`.

[SK22a] Mariia Seleznova and Gitta Kutyniok. Analyzing finite neural networks: Can we trust neural tangent kernel theory? In *Mathematical and Scientific Machine Learning*, pages 868–895. PMLR, 2022.

[SK22b] Mariia Seleznova and Gitta Kutyniok. Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19522–19560. PMLR, 17–23 Jul 2022.

[Sob63] Milton Sobel. Multivariate Hermite polynomials, Gram-Charlier expansions and Edgeworth expansions. Technical report, University of Minnesota, 1963.

[Yai20] Sho Yaida. Non-Gaussian processes and neural networks at finite widths. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 165–192. PMLR, 20–24 Jul 2020.

[Yan19] Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *Advances in Neural Information Processing Systems*, 32, 2019.