Characteristics of Monte Carlo Dropout in Wide Neural Networks

Joachim Sicking [*1] **Maram Akila** [*1] **Tim Wirtz** [*1 2] **Sebastian Houben** [1] **Asja Fischer** [3]

Abstract

Monte Carlo (MC) dropout is one of the state-of-the-art approaches for uncertainty estimation in neural networks (NNs). It has been interpreted as approximately performing Bayesian inference. Based on previous work on the approximation of Gaussian processes by wide and deep neural networks with random weights, we study the limiting distribution of wide untrained NNs under dropout more rigorously and prove that they as well converge to Gaussian processes for fixed sets of weights and biases. We sketch an argument that this property might also hold for infinitely wide feed-forward networks that are trained with (full-batch) gradient descent. The theory is contrasted by an empirical analysis in which we find correlations and non-Gaussian behaviour for the pre-activations of finite width NNs. We therefore investigate how (strongly) correlated pre-activations can induce non-Gaussian behavior in NNs with strongly correlated weights.

1. Introduction

Despite the huge success of deep neural networks (NNs), robustly quantifying their prediction uncertainty is still an open problem. One of the most popular methods for uncertainty prediction is Monte Carlo (MC) dropout (Gal & Ghahramani, 2016). MC dropout has been proven practically successful in many applications, such as different regression tasks (Kendall & Gal, 2017), natural language processing (Press & Wolf, 2017), and object detection (Miller et al., 2018). However, providing only a rough approximation of Bayesian inference, MC dropout also carries theoretical and practical drawbacks. For example, as Gal et al. (2017) point out, the dropout rate has to be carefully tuned in order to correctly estimate the uncertainty at hand and Osband (2016) showed that it does not converge to concentrated distributions in the infinite data limit.

We therefore revisit the characteristics of the output distribution of NNs under dropout in this paper. To do so, we (i) present a rigorous theoretical analysis building on previous work for wide feed-forward NNs. This work has established that an i.i.d. prior over their weights leads to approximating Gaussian processes, i.e., drawing one set of weights yields a NN prediction that approximates the realization of one drawn function from a corresponding Gaussian process initially been shown for NNs with only one hidden layer (Neal, 1994) and was recently generalized to deep NNs (Lee et al., 2018; de G. Matthews et al., 2018). We adapt the proof strategy of the latter work to show that performing dropout during inference in wide NNs with fixed weights also leads to an approximation of Gaussian processes (Sec. 2). (ii) Since the applicability of the central limit theorem to the pre-activations (i.e. the inputs) of the neurons plays a central role in our proof, we investigate the pre-activations in randomly initialized and trained NNs under dropout and find that the former are approximately normal distributed while the latter may follow non-Gaussian distributions (Sec. 3). (iii) We then demonstrate that such non-Gaussian distributions can be induced by correlations of the pre-activations in the previous layer, by showing that these lead to distributions with exponential tails in a toy model and NNs with strongly correlated weights (Sec. 4).

2. Theoretical Analysis of Random Neural Networks

As a first step towards a characterization of the limiting process of NNs under MC dropout at inference time we rigorously study a random NN. We consider a feed-forward NN with $k+2$ layers which, for an input vector $x \in \mathbb{R}^d$, is parameterized as follows:

$$f_i^{(\nu)}(x) = \sum_{j=1}^{h_{\nu-1}(n)} \frac{w_{ij}^{(\nu)} g_j^{(\nu-1)}(x)}{\sqrt{h_{\nu-1}(n)}} + b_i^{(\nu)} , \qquad (1)$$

where $g_i^{(\nu)}(x) = z_i^{(\nu)} \phi(f_i^{(\nu)}(x))$ for $\nu = 1, \dots, k$, $g_i^{(0)}(x) = x_i$. ϕ is the activation function, z_i are the dropout Bernoulli random variables (RVs) with keep rate $q = 1 - p$,

[*]Equal contribution [1]Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Sankt Augustin, Germany [2]Fraunhofer Center for Machine Learning [3]Faculty of Mathematics, Ruhr-University Bochum, Bochum, Germany. Correspondence to: J. Sicking <joachim.sicking@iais.fraunhofer.de>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

and $h_\nu(n)$ is the neuron count in layer ν.[1]

In previous work, see de G. Matthews et al. (2018); Lee et al. (2018); Wu et al. (2019); Tsuchida et al. (2019), the authors considered NNs with independent prior distributions on the NN parameters. The resulting summands in Eq. (1) are independent RVs such that a central limit theorem can readily be applied. In contrast we examine a NN with a *fixed set* of parameters and dropout acting as independent prior distribution on the activations. As a consequence, the resulting summands in Eq. (1) are independent RVs only in the second layer ($\nu = 2$) and are dependent RVs for all later ones. This follows from an inspection of the covariance of the pre-activations yielding

$$\begin{aligned}\mathrm{Cov}\left(f_i^{(\nu)}, f_j^{(\nu)}\right) &= \frac{q}{h_{\nu-1}} \sum_{k,l=1}^{h_{\nu-1}(n)} w_{il}^{(\nu)} w_{jk}^{(\nu)} \\ &\times \Big(\left(\delta_{kl} + q\left(1-\delta_{kl}\right)\right) \mathbf{E}_Z\left\{\phi_k^{(\nu-1)}\phi_l^{(\nu-1)}\right\} \\ &- q\, \mathbf{E}_Z\left\{\phi_k^{(\nu-1)}\right\} \mathbf{E}_Z\left\{\phi_l^{(\nu-1)}\right\}\Big) \ ,\end{aligned} \tag{2}$$

where we dropped the arguments of $f_i^{(\nu)}$ and $\phi_l^{(\nu-1)}$ for simplicity of notation. For finite $h_{\nu-1}(n)$ and $i \neq j$ the expression above is in general non-zero. Thus, we cannot apply standard central limit theorems. Nonetheless, we are able to show that a random NN under MC dropout converges to a Gaussian process in the limit of infinite width. Our analysis facilitates the property of weights and biases being realizations of independent Gaussian distributed RVs in order to bound their behavior if the width of the NN approaches infinity. We summarize our findings and main result of this section in the following theorem.

Theorem 1. *Let $f^{(k)}(x) \in \mathbb{R}^L$ be the pre-activation vector of the $k+2$-th layer of a NN, as defined by Eq. (1) , where weights and biases follow a Gaussian distribution with zero mean and unit variance. For the widths of the NN layers going simultaneously to infinity, for each bounded $x \in \mathbb{R}^d$ with $||x||_2^2 \leq \alpha\, d$, and for a fixed set of weight and bias variables, the pre-activations of all layers $\nu = 2, \ldots, k+1$ converge in distribution over MC dropout to an uncorrelated Gaussian process.*

The theorem is a direct consequence of corollary 2 and is proven in Appendix A. The following observation regarding the mutual correlations between $f_i^{(\nu)}(x)$ and $f_j^{(\nu)}(x)$ is central to the proof: since weights and biases are i.i.d. Gaussian RVs, the correlations between the pre-activations are rather weak and vanish for $n \to \infty$ whenever $i \neq j$. Extending theorem 1 to trained NNs can be achieved using that for (full-batch) gradient descent and rather weak assumptions on the loss function, the by $1/\sqrt{n}$ rescaled distance (n being a control parameter of the width, see Appendix A) of weights at initialization time and at (a finite) optimization step t, $||W(0) - W(t)||_{op}/\sqrt{n}$, converges uniformly to zero for $n \to \infty$ on $t \in [0, T]$ (Du et al.; Jacot et al., 2018). This indicates that we can expect to find Gaussian processes also for infinitely wide, *trained* NNs with MC dropout (we leave a rigorous proof of this sketched argument for future work). This theoretical insight stands in contrast to the non-Gaussian behavior of trained wide NNs we obtain in the experiments outlined in the subsequent section. We therefore will focus the investigation of the implications (strongly) correlated (pre-)activations in Sec. 4.

3. Ambiguous Empirical Observations

To substantiate our discussions in Secs. 2 and 4 empirically, we study the pre-activations of a fully connected NN of the form given in Eq. 1. We train it for classification on FashionMNIST (Xiao et al., 2017) using a cross-entropy loss and hyperbolic tangent (tanh) non-linearities. We employ a *narrow* network $\mathcal{H}_{\text{narrow}}$ with $k = 9$ hidden layers of width $h_\nu(n) = h = 100$ each as well as a *wide* network $\mathcal{H}_{\text{wide}}$ with $k = 7$ and $h = 1000$. All network weights are initialized with i.i.d. random values. We train for 100 epochs using the Adam (Kingma & Ba, 2015) optimizer and an initial learning rate of 0.001 (see Appendix B for more implementation details). Bernoulli dropout with $p = 0.2$ is applied to the activations of all hidden network layers.

First, we study the pre-activation distributions of the untrained $\mathcal{H}_{\text{narrow}}$ and $\mathcal{H}_{\text{wide}}$. These distributions are generated by running 30,000 forward passes with dropout for a fixed test image. We randomly pick one neuron from the last hidden layer of each network and visualize its pre-activation distribution in Fig. 1 (top row). To ease visual comparison between different neurons, we report normalized pre-activations. The resulting distributions are clearly Gaussian, compare our discussion in Sec. 2. The weight correlations and pre-activation correlations of these untrained networks are studied in Appendix B.

Analyzing the trained networks $\mathcal{H}_{\text{narrow}}$ and $\mathcal{H}_{\text{wide}}$ yields more complex results. The pre-activation distributions for exemplarily selected hidden units from the last hidden layer of the trained $\mathcal{H}_{\text{narrow}}$ and $\mathcal{H}_{\text{wide}}$ are shown in Fig. 1 (middle and bottom row). The neurons are handpicked to illustrate the variety of distributions we encounter in this layer. We observe Gaussian as well as non-Gaussian pre-activation distributions that range over skewed Gaussians to clearly non-Gaussian, exponential ones [2]. For $\mathcal{H}_{\text{wide}}$, approximately 40% of all neurons have Gaussian distributions, 40% skewed Gaussians and 20% exponential ones. Each

[1] A more detailed overview of the structure and notation of this feed-forward NN is summarized in Appendix A.

[2] Technically also the tails of a Gaussian decay exponentially, however with $\exp(-\xi^2)$. Throughout this text we refer to exponential tails only for the cases of $\exp(-|\xi|)$ or slower decay.

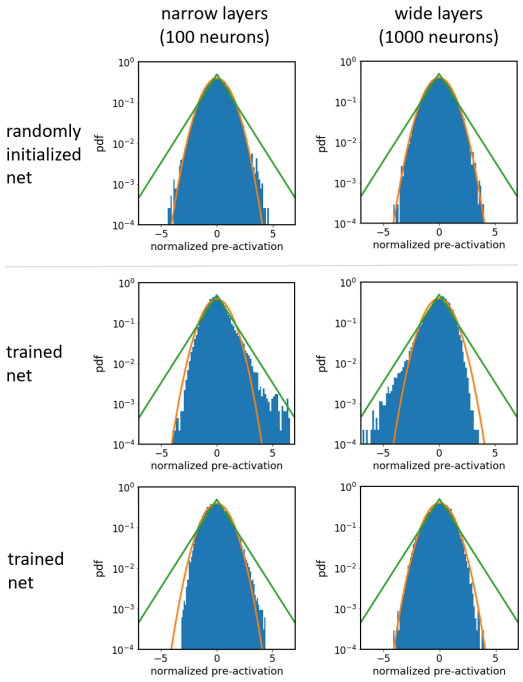

Figure 1. Normalized pre-activation distributions of selected neurons from different networks. We consider randomly initialized (top row) and trained (middle and bottom row) nets that are either narrow ($h = 100$, left column) or wide ($h = 1000$, right column). We handpicked the neurons from the trained networks (see text). Standard normal (orange) and double exponential (green) densities are given for comparison.

histogram in Fig. 1 is based on 30,000 forward passes with dropout.

Moreover, we find these observations to depend on the input image: training inputs yield less tailed distributions than test inputs that in turn lead to less pronounced tails than artificial ones that are superpositions of two training images. However, there are large differences within those input categories, e.g., between test images. Shuffling the trained weight matrices yields results that are similar to the randomly initialized case. This indicates that weight and pre-activation dependencies are important while changes of the marginal distributions due to model training are not. While these results apply for the *last* hidden layer, we observe mostly Gaussian or slightly skewed Gaussians in earlier layers. In Sec. 4, we demonstrate on a toy model that non-Gaussian properties accumulate during the propagation through the network. Further empirical observations and an analysis of the weight correlations and pre-activation correlations of the trained networks can be found in Appendix B.

These complex dependence structures are the result of an interplay of weight distributions, input-immanent structures, and network non-linearities that are orchestrated by network training. A theoretical foundation for the Gaussian results was laid out in the previous section. For the deviations we take a closer look at correlations in the next section.

4. Completing the Picture - Strongly Correlated Systems

As we have seen in Eq. (2) the (pre-)activations from any layer ν are, in general, not independent from one another. We explore possible consequences of this observation in terms of consecutive toy experiments. For this, the product of *two* RVs, $Z = XY$, is central. On an abstract level, Y represents a random activation from a previous layer and X the product $w_{ij}z_j$. For later simplicity, we model both terms as Gaussian, $X, Y \sim \mathcal{N}(\mu, \sigma)$. In the case of vanishing mean, $\mu = 0$, one obtains the well known analytic form for the PDF of Z,

$$\mathrm{PDF}_Z(\xi) = \frac{1}{\pi} K_0(|\xi|) \qquad (\sigma = 1) \ , \tag{3}$$

with the modified Bessel function K_0. Derivations and further comments are relegated to Appendix C. In contrast to the normal distributions its asymptotic,

$$\mathrm{PDF}_Z(\xi) \sim \frac{1}{\sqrt{2\pi|\xi|}} \mathrm{e}^{-|\xi|} \qquad (\sigma = 1,\ |\xi| \gg 1) \ , \tag{4}$$

reveals exponentially decaying tails.

For any given neuron we encounter sums over h terms, where h denotes the layer width, instead of single products. If these summands $X_i Y_i$ were independent as argued for in the first section, one would recover a normal distributed outcome for the neuron by the central limit theorem. But if they are correlated, the result may differ drastically. The outcome strongly depends on the respective covariance matrices of X_i, Y_i, for details see Appendix C. As a rule of thumb, larger correlations favor non-Gaussian results. This can be illustrated in terms of a toy extension choosing

$$Z = \sum_{i=1}^{h} X_i Y_i \,, \quad X_i = c\, x_0 + (1-c)\, x_i \tag{5}$$

and similarly for Y_i with Gaussian RVs x_γ, y_γ for $\gamma = 0, 1, \dots h$. The factor c controls a global correlation among the entries. For this case we find the decomposition

$$\begin{aligned} \sum_{i=1}^{h} X_i Y_i =& (1-c)^2 \sum_{i=1}^{h} x_i\, y_i + c^2 h x_0\, y_0 \\ &+ c\,(1-c) \left(x_0 \sum_{i=1}^{h} y_i + y_0 \sum_{i=1}^{h} x_i \right) \ . \end{aligned} \tag{6}$$

Heuristically speaking, the first term is of Gaussian nature while the second (and later ones) contribute exponential tails,

see the limits for $c = 0$ or 1, respectively. An important observation is that both terms are on similar footing with respect to h, i.e., also in the large h limit the Gaussian term will not suppress the first one. This finding is due to the choice of a global correlation present in all h terms of x_i and y_i.

Assuming that the input Y has exponential tails, we can investigate how those propagate for the $Z = XY$ model (given $\mu_x = 0$). For this, we take

$$\mathrm{PDF}_Y(\xi) \propto \frac{1}{|\xi|^{(n-1)/n}} \exp\left(-\alpha|\xi|^{2/n}\right) \qquad n \in \mathbb{N} \tag{7}$$

as an analytically tractable approximation capturing the tail behavior. For $n = 1$, i.e., Gaussian-like decay, we obtain a Bessel result for Z comparable to Eq. (3), in which the tail is modeled by $n = 2$, see Eq. (4). More generally, for small values of n we obtain the explicit asymptotics for Z as:

$$\mathrm{PDF}_Z(\xi) \sim \frac{1}{|\xi|^{n/(n+1)}} \exp\left(-\kappa^{1/(n+1)}|\xi|^{2/(n+1)}\right) \tag{8}$$

with $|\xi| \gg 1$ and some positive constant $\kappa > 0$ depending on α^n/σ_X^2. As the decay slows down with each iteration, this effect might contribute to an "accumulation" of tails.

Inspired by the presented heuristics, we revisit the randomly initialized NN introduced in Sec. 3. However, we initialize the weight matrices in a *correlated* fashion similar to the toy model in Eq. (5), see Appendix C. The resulting distributions for the pre-activations given random input are shown in Fig. 2. With $c = 0.1$ the correlation is deliberately small and the dropout rate $p = 0.2$ set as before, but we investigate deeper networks of $L = 50$ layers. The left panel shows that for $h = 1000$ only the initial layer follows a Gaussian behavior and all later ones exhibit exponential tails of increasing strength. A closer look reveals that the decay, especially for later layers, is slightly weaker than exponential, which might be caused by the effects presented in Eq. (8). On the right hand side we compare the pre-activations of the 13$^{\text{th}}$ layer for different network widths, from shallow ($h = 100$) to wide ($h = 2000$). Except for small fluctuations of the tails, which besides statistics are caused by choosing different (fixed) random inputs for each network, the result is stable with h, suggesting that also the infinite width limit will not lead to a Gaussian outcome.

Figs. 3 and 4 in Appendix B show the empirical correlations of the trained network from Sec. 3. While their structure is more involved, we find that the correlation of the pre-activations increases for deeper layers. More importantly, both correlations do not decrease to zero with increased layer width, which satisfies a central assumption of the toy model here. It might therefore be less surprising that the trained network can exhibit exponentially tailed pre-activations, cf. Fig. 1.

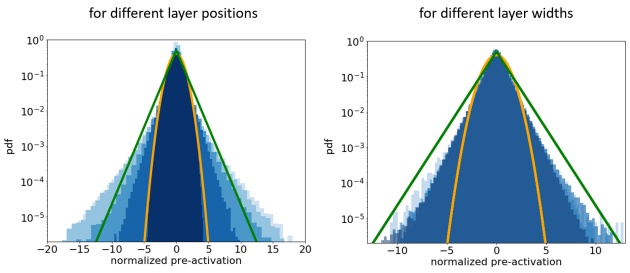

Figure 2. Normalized pre-activation distributions from networks with correlated random weights, details in text. On the l.h.s. we color-code hidden layer position, first (dark blue) to last (light blue). The r.h.s. shows different network widths (narrow (light blue) to wide (dark blue)) for layer 13. Gaussian (orange) and exponential tailed distribution (green) are shown for comparison.

5. Discussion

We rigorously proved that random neural networks with fixed parameters under dropout converge to Gaussian processes in the limit of infinitely wide layers. A sketched argumentation fuels hopes that a similar behavior can be shown for weakly correlated, infinitely wide networks that are trained with (full-batch) gradient descent.

Empirically studying wide (trained) networks with a rich dependence structure reveals a more complex picture: the coexistence of Gaussian and non-Gaussian, exponential, latent distributions. We shed light on this observation using a simple toy model and a network with correlated random initialization. These two systems indicate the existence of (at least) two regimes: one of *weakly* correlated pre-activations with Gaussian distributions and another one of *strongly* correlated pre-activations with exponentially tailed limiting distribution functions. Future work will investigate these two regimes in more detail to understand how to deliberately manipulate the properties of these limiting distributions under MC dropout. Searching for the borders of these regimes and further classes of limiting distributions is another research path.

Our empirical observation of Gaussianity that decreases from training data over test data to out-of-distribution data, gives rise to both theoretical and applied future work: firstly, to better understand the generalization properties of the *learned* Gaussian or non-Gaussian behavior and, secondly, to employ this knowledge to construct more robust uncertainty quantifications.

Acknowledgement — The research of the authors M. Akila, S. Houben and J. Sicking was funded by the German Federal Ministry for Economic Affairs and Energy within the project "KI Absicherung – Safe AI for Automated Driving". Said authors would like to thank the consortium for

the successful cooperation. The work (T. Wirtz) was developed by the Fraunhofer Center for Machine Learning within the Fraunhofer Cluster for Cognitive Internet Technologies. This work of A. Fischer was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2092 CASA – 390781972.

References

Billingsley, P. *Probability and Measure - 3rd ed.* Wiley-Interscience, 1995.

Dashti, M. and Stuart, A. M. The Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification*. Springer International Publishing, 2017.

de G. Matthews, A. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 1675–1685.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Gal, Y., Hron, J., and Kendall, A. Concrete dropout. In *Advances in Neural Information Processing Systems*, pp. 3581–3590, 2017.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pp. 8571–8580, 2018.

Kendall, A. and Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, pp. 5574–5584, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015.

Lee, J., Sohl-Dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

Miller, D., Nicholson, L., Dayoub, F., and Sünderhauf, N. Dropout sampling for robust object detection in open-set conditions. In *International Conference on Robotics and Automation*, pp. 1–7. IEEE, 2018.

Neal, R. M. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, Dept.of Computer Science, 1994.

Osband, I. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *Neural Information Processing Systems Workshop on Bayesian Deep Learning*, volume 192, 2016.

Paszke, A., Gross, S., and Massa, F. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8026–8037. 2019.

Press, O. and Wolf, L. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL, Volume 2: Short Papers*, pp. 157–163. Association for Computational Linguistics, 2017.

Tsuchida, R., Roosta, F., and Gallagher, M. Richer priors for infinitely wide multi-layer perceptrons. *arXiv preprint arXiv:1911.12927*, 2019.

Wu, A., Nowozin, S., Meeds, E., Turner, R. E., Hernández-Lobato, J. M., and Gaunt, A. L. Deterministic variational inference for robust Bayesian neural networks. In *7th International Conference on Learning Representations, ICLR*, 2019.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

A. Central Limit Theorem for Random Neural Networks

The network architectures considered here, are fully connected and therefore defined by the width of the hidden layers, which we will denote by $h_0, \ldots, h_{k+1}$. The depth and the size of the input and the output layer of the NN, $h_0 = d$ and $h_{k+1} = L$, respectively, are kept fix during the analysis. More formally, we define a set of *width functions* $h_\nu : \mathbb{N} \to \mathbb{N},\ n \mapsto h_\nu(n)$, for $\nu = 1, \ldots, k$ in which n is a control parameter meaningless for actual applications and $h_\nu(n)$ specifies the number of hidden neurons at layer ν for a given input n. We consider the case where $n \to \infty$ implies $h_\nu(n) \to \infty$ and $h_\nu(n)/n < \infty$ for all n. Without loss of generality we assume the same non-linear activation function in each layer, denoted by $\phi : \mathbb{R} \to \mathbb{R}$, ϕ to be (piecewise) Lipschitz continuous, i.e. $|\phi(u) - \phi(v)| \leq K|u - v|$ for a fixed $K \in \mathbb{R}_+$ and all $u, v \in \mathbb{R}$, and that the weights $w_{ij}^{(\nu)}$ and biases $b_{ij}^{(\nu)}$, are drawn from a standard normal distribution $\mathcal{N}(0, 1)$. Let $x \in \mathbb{R}^d$ be a bounded input vector to the NN, that we parametrize as follows

$$f_i^{(\nu)}(x) = \frac{1}{\sqrt{h_{\nu-1}(n)}} \sum_{j=1}^{h_{\nu-1}(n)} w_{ij}^{(\nu)} g_j^{(\nu-1)}(x) + b_i^{(\nu)} \,, \tag{9}$$

$$g_i^{(\nu)}(x) = \begin{cases} z_i^{(\nu)} \phi\left(f_i^{(\nu)}(x)\right) & ,\nu = 1, \ldots, k-1 \\ x_i & ,\nu = 0 \end{cases} , \tag{10}$$

where $i = 1, \ldots, h_\nu(n)$, $\nu = 0, \ldots, k$, and $z_i^{(\nu)} \in \{0, 1\}$ are independent binary random variables with $\mathrm{P}(z_j^{(\nu)} = 1) = q$ for all i, j, ν. For analytic tractability, we do not consider dropout in the input layer because its dimension is kept fixed for $n \to \infty$.

The proof that random neural networks with MC dropout converge to a Gaussian Processes relies on the same mathematical model as proposed in (de G. Matthews et al., 2018), which we summarize for completeness. In order to study the limiting distribution of the pre-activations, in layer ν, we set its width to infinity, $h_\nu(n) = \infty$ and embed it into $\mathbb{R}^\infty$. Convergence in distribution in this abstract space can be studied only with respect to a certain topology. Following (Billingsley, 1995) this topology is generated by a metric ρ:

$$\rho(u, v) = \sum_{i=1}^{\infty} \frac{\min\{1, |v_i - u_i|\}}{2^i}$$

According to (Billingsley, 1995) this metric metricise the product topology of the product of countable many copies of $\mathbb{R}$ with the usual Euclidean topology (Dashti & Stuart, 2017). Moreover, to prove weak convergence in such an abstract space, it is sufficient to prove weak convergence for each finite dimensional marginal. To study the convergence behavior of the finite dimensional marginals simultaneously (de G. Matthews et al., 2018) suggested to use the Cramér-Wold device.

Theorem 2. *For random vectors $X_n = (X_{n1}, \ldots, X_{nM})$ and $Y = (Y_1, \ldots, Y_M)$, a necessary and sufficient condition for a X_n converging in distribution to Y ($X_n \Rightarrow Y$) is that*

$$\sum_u t_u X_{nu} \Rightarrow \sum_u t_u Y_u \,, \quad \forall\, t \in \mathbb{R}^M \,. \tag{11}$$

That is, if every linear combination of the coordinates of X_n converges in distribution to the correspondent linear combination of coordinates of Y.

This theorem reduces the problem of weak convergence of the distribution of the finite marginals, to the convergence of the distribution of a one-dimensional random variable. Based on the Cramér-Wold theorem, we will show that

$$\psi_\nu(t) = \frac{1}{s_n} \sum_{u=0}^{M} t_u \left(f_u^{(\nu)}(x) - \mathbf{E}_Z\left\{f_u^{(\nu)}(x)\right\}\right) \tag{12}$$

$$= \frac{1}{s_n} \sum_{j=1}^{h_{\nu-1}(n)} \frac{1}{\sqrt{h_{\nu-1}(n)}} \sum_{u=0}^{M} t_u w_{uj}^{(\nu)} \times \left(z_j^{(\nu)} g_j^{(\nu-1)}(x) - q\mathbf{E}_Z\left\{g_j^{(\nu-1)}(x)\right\}\right) \tag{13}$$

$$\equiv \frac{1}{s_n} \sum_{j=1}^{h_{\nu-1}(n)} \gamma_{j,n}^{(\nu)} \,, \tag{14}$$

where $\mathbf{E}_Z\{\psi_\nu(t)\} = 0$ and

$$s_n^{(\nu)2} = \sum_{j=1}^{h_{\nu-1}(n)} \sigma_{jn}^{(\nu)2} \tag{15}$$

$$= \sum_{j=1}^{h_{\nu-1}(n)} \sum_{u=0,v=0}^{M} \frac{t_u t_v w_{uj}^{(\nu)} w_{vj}^{(\nu)}}{h_{\nu-1}(n)} \times \left(q\mathbf{E}_Z\left\{\left(g_j^{(\nu-1)}(x)\right)^2\right\} - q^2\mathbf{E}_Z^2\left\{g_j^{(\nu-1)}(x)\right\}\right) . \tag{16}$$

In the expression above we set $\mathbf{E}\{{\gamma_{j,n}^{(\nu)}}^2\} \equiv \sigma_{jn}^{(\nu)2}$. As defined above, $\psi_\nu(t)$ is a random variable with mean zero and unit variance. As mentioned earlier, see Eq. (2), the pre-activations have a non-trivial correlation structure. Thus the activations are dependent random variables. Since the sums in the pre-activations are over dependent random variables, standard central limit theorems can not be applied. However, due to the *self-averaging* property of the weights of the

random neural network, we can study the behavior of the variables when the width goes to infinity. We summarize our observations in the following lemma.

Lemma 1. *Let the pre-activations $\{f_u^{(\nu-1)}(x)\}_{u\in\mathbb{N}}$ converge, for $n\to\infty$ in distribution, to independently distributed random variables, let $f_i^{(\nu)}(x)$, $g_i^{(\nu)}(x)$, $s_n^{(\nu)2}$ and $\gamma_{j,n}^{(\nu)}$ be as defined in Eq. (9), (10), (15) and (14), respectively. Then it follows that $\psi^{(\nu)}(t)$ converges in distribution to the distribution of $\psi^{(\nu)}(t)$ with $\gamma_{j,n}^{(\nu)}$ replaced by a random variable drawn from the* marginalized distribution *of $\gamma_{j,n}^{(\nu)}$.*

Proof. According to the discussion above we prove that $\psi^{(\nu)}(t)$, as, defined in Eq. (13), converges to a Gaussian distributed random variable with mean zero and variance one. We first show that $\psi^{(\nu)}(t)$ converges in distribution to the distribution of $\psi^{(\nu)}(t)$ with $\gamma_{j,n}^{(\nu)}$ replaced by a random variable drawn from the *marginalized distribution* of $\gamma_{j,n}^{(\nu)}$. To prove this, we have to show that

$$\left|\mathbf{E}_Z\left\{\prod_{j=1}^{h_{\nu-1}(n)} e^{\mathrm{i}\lambda\gamma_{j,n}^{(\nu)}/s_n}\right\} - \prod_{j=1}^{h_{\nu-1}(n)}\mathbf{E}_Z\left\{e^{\mathrm{i}\lambda\gamma_{j,n}^{(\nu)}/s_n}\right\}\right| \tag{17}$$

converges to zero for $n\to\infty$. To simplify the expression above, we combine the following estimate (Billingsley, 1995) of the exponential function

$$|(1+\mathrm{i}\lambda x - \frac{\lambda^2}{2}x^2) - e^{\mathrm{i}\lambda x}| \leq \min\{|x\lambda|^2, |x\lambda|^3\} \tag{18}$$

with the following observation

$$\begin{aligned}&\left|\mathbf{E}_Z\left\{\prod_{j=1}^{h_{\nu-1}(n)} e^{\mathrm{i}\lambda\gamma_{j,n}^{(\nu)}/s_n}\right\} - c\right| \leq \\ &\mathbf{E}_Z\left\{\left|\left(1+\frac{\mathrm{i}\lambda}{s_n}\gamma_{i,n}^{(\nu)} - \frac{\lambda^2}{s_n^2 2}{\gamma_{i,n}^{(\nu)}}^2\right) - e^{\mathrm{i}\lambda\gamma_{i,n}^{(\nu)}/s_n}\right|\right\} + \\ &\left|\mathbf{E}_Z\left\{\left(1+\frac{\mathrm{i}\lambda}{s_n}\gamma_{i,n}^{(\nu)} - \frac{\lambda^2}{s_n^2 2}{\gamma_{i,n}^{(\nu)}}^2\right)\prod_{j\neq i}^{h_{\nu-1}(n)} e^{\mathrm{i}\lambda\gamma_{j,n}^{(\nu)}/s_n}\right\} - c\right|\end{aligned} \tag{19}$$

where $c\in\mathbb{C}$. It holds for all $i\leq h_{\nu-1}(n)$ and is a simple consequence of the fact that the characteristic function has absolute value one. Applying Eq. (19) and (18) to both terms in expression (17) multiple times, we end up with the following upper bound on Eq. (17),

$$\begin{aligned}&\left|\mathbf{E}_Z\left\{\prod_{j=1}^{h_{\nu-1}(n)} e^{\mathrm{i}\lambda\gamma_{j,n}^{(\nu)}/s_n}\right\} - \prod_{j=1}^{h_{\nu-1}(n)}\mathbf{E}_Z\left\{e^{\mathrm{i}\lambda\gamma_{j,n}^{(\nu)}/s_n}\right\}\right| \\ &\leq \sum_{j=1}^{h_{\nu-1}(n)} 2\mathbf{E}_Z\left\{\min\left\{\left|\frac{\gamma_{j,n}^{(\nu)}\lambda}{s_n}\right|^2, \left|\frac{\gamma_{j,n}^{(\nu)}\lambda}{s_n}\right|^3\right\}\right\} \qquad (20) \\ &+ \left|\mathbf{E}_Z\left\{\prod_{j=1}^{h_{\nu-1}(n)}\left(1+\frac{\mathrm{i}\lambda\gamma_{j,n}^{(\nu)}}{s_n} - \frac{\lambda^2}{s_n^2 2}{\gamma_{j,n}^{(\nu)}}^2\right)\right\}\right. \\ &\left. - \prod_{j=1}^{h_{\nu-1}(n)}\mathbf{E}_Z\left\{\left(1+\frac{\mathrm{i}\lambda\gamma_{j,n}^{(\nu)}}{s_n} - \frac{\lambda^2}{2s_n^2}{\gamma_{j,n}^{(\nu)}}^2\right)\right\}\right| . \qquad (21)\end{aligned}$$

The sum (20) in the expression above, appears also in the proof of the Lindeberg central limit theorem. Following (Billingsley, 1995) let $\epsilon>0$ be positive, then due to min-function in the expectation, the sum (20) is at most

$$\begin{aligned}&\frac{|\lambda|^2}{s_n^2}\sum_{j=1}^{h_{\nu-1}(n)}\int_{|\gamma_{j,n}^{(\nu)}|>\epsilon}\mathrm{dP}_j(\gamma_{j,n}^{(\nu)})\,\gamma_{j,n}^{(\nu)2} \\ &+\frac{|\lambda|^3}{s_n^3}\sum_{j=1}^{h_{\nu-1}(n)}\int_{|\gamma_{j,n}^{(\nu)}|<\epsilon}\mathrm{dP}_j(\gamma_{j,n}^{(\nu)})\left|\gamma_{j,n}^{(\nu)}\right|^3 \qquad (22) \\ &\leq\epsilon\frac{|\lambda|^3}{s_n} + \frac{|\lambda|^2}{s_n^2}\sum_{j=1}^{h_{\nu-1}(n)}\int_{|\gamma_{j,n}^{(\nu)}|>\epsilon}\mathrm{dP}_j(\gamma_{j,n}^{(\nu)})\,\gamma_{j,n}^{(\nu)2}, \qquad (23)\end{aligned}$$

where $\mathrm{dP}_j(\gamma_{j,n}^{(\nu)})$ is the marginal probability measure of $\gamma_{j,n}^{(\nu)}$. Setting $\epsilon\to\epsilon s_n$ the second term in Eq. (23), is exactly the Lindeberg condition. In Lindeberg-CLT a necessary condition for the in distribution convergence to a Gaussian distributed random variable is that the Lindeberg-Condition vanishes for $n\to\infty$. A necessary condition for the Lindeberg-condition to be satisfied is that the Lyapunov-condition (26) holds (Billingsley, 1995). The latter is proven to be satisfied in lemma 7. Thus, since second term in expression (23) converges to zero for $n\to\infty$ and that ϵ is arbitrary, the sum (20) converges to zero.

To conclude the proof, it remains to show that expression (21) converges to zero for $n\to\infty$. First of all, we observe that $\mathbf{E}_Z\{1+\mathrm{i}\lambda\gamma_{j,n}^{(\nu)}/s_n - \lambda^2{\gamma_{j,n}^{(\nu)}}^2/2s_n^2\} = 1-\lambda^2\sigma_{jn}^{(\nu)2}/2s_n$. Expanding the products in Eq. (21), it turns out that we have to show that the absolute values of

$$\sum_{i\in D}\mathbf{E}_Z\left\{\prod_{j=1}^{l_1}\gamma_{i_j,n}^{(\nu)}\prod_{k=l_1+1}^{l_1+l_2}\gamma_{i_k,n}^{(\nu)2}\right\}, \tag{24}$$

where $D=\{i=(i_1,\ldots,i_{l_1+l_2})|1\leq i_1<\cdots<i_{l_1+l_2}\leq h_{\nu-1}\}$, converges to zero for $n\to\infty$, for all l_1,l_2 such that

$l_1 > 1$ and $l_1 + l_2 \leq h_{\nu-1}$ as well as that the absolute value of

$$\sum_{i \in D'} \left(\mathbf{E}_Z \left\{ \prod_{j=1}^{l} \gamma_{i_j,n}^{(\nu)2} \right\} - \prod_{j=1}^{l} \sigma_{jn}^{(\nu)2} \right) , \qquad (25)$$

where $D' = \{i = (i_1, \ldots, i_l) | 1 \leq i_1 < \cdots < i_l \leq h_{\nu-1}\}$, converges to zero for $n \to \infty$, for all l such that $l \leq h_{\nu-1}$. Both claims follow similar argumentation. Due to the fact that the weights are Gaussian distributed random variables with zero mean and unit variance, it immediately follows that the mean with respect to the weights of layer ν is zero. Thus, we can apply Chebyshev's inequality $\mathbf{P}_W(|f(W)| \leq \sqrt{\text{Var}(f(W))/\delta}) \geq 1-\delta$ to obtain a high probability bound on both expressions. In both cases the variance becomes a double sum over either D or D' which is at most of order $\mathcal{O}(h_{\nu-1}^{2l_1+2l_2})$, the summands contributed a factor of order $\mathcal{O}(h_{\nu-1}^{-l_1-2l_2})$ times the following average over the weights of layer ν

$$\mathbf{E}_W \left\{ \prod_{o=1}^{l_1} w_{u_o i_o}^{(\nu)} w_{u'_o j_o}^{(\nu)} \prod_{b=l_1+1}^{l_1+l_2} w_{u_b i_b}^{(\nu)} w_{u'_b i_b}^{(\nu)} w_{u''_o j_b}^{(\nu)} w_{u'''_o j_b}^{(\nu)} \right\} ,$$

where $1 \leq i_1 < \cdots < i_{l_1+l_2} \leq h_{\nu-1}$ and $1 \leq j_1 < \cdots < j_{l_1+l_2} \leq h_{\nu-1}$. Due to the Gaussian nature of the weights, the leading order contributions of the expression above are due to the second are of the form $\prod_{o=1}^{l_1} \delta_{i_o j_o}$. The Kronecker-Deltas lead to a reduction of the order of the sum, namely the reduce contributions of the double sum by a factors of $\mathcal{O}(h_{\nu-1}^{l_1})$. Thus, there exist a constant C such that the variance of expression (24) can be bounded, with at least probability $1 - \delta$, by

$$\begin{aligned} &\frac{C}{\sqrt{\delta}} \left(\frac{||t||_2}{s_n} \right)^{2(l_1+2l_2)} \\ &\times \max_{i,k \in D} \left[\mathbf{E}_Z \left\{ \prod_{j=1}^{l_1} \left(g_{i_j}^{(\nu-1)} - \mathbf{E}_Z \left\{ g_{i_j}^{(\nu-1)} \right\} \right) \right. \right. \\ &\left. \left. \times \prod_{b=l_1+1}^{l_1+l_2} \left(g_{i_b}^{(\nu-1)} - \mathbf{E}_Z \left\{ g_{i_b}^{(\nu-1)} \right\} \right)^2 \right\} \right. \\ &\left. \times (i_j, i_b) \text{ replaced by } (k_i, k_b) \right] + \mathcal{O}(h_\nu^{-1}) . \end{aligned}$$

However, due to the assumptions, s_n converges to a non-zero limit for $n \to \infty$ and the expectations asymptotically decompose into a product over the is and js which are zero. Thus the expression above and the expression (24) converge to zero for all $\delta \in (0, 1]$. Similar, we find that expression (25) can be bounded, with at least probability $1 - \delta$, by

$$\begin{aligned} &\frac{C}{\sqrt{\delta}} \left(\frac{||t||_2}{s_n} \right)^{4l} \\ &\times \max_{i,j \in D'} \left[\left(\mathbf{E}_Z \left\{ \prod_{b=1}^{l} \left(g_{i_b}^{(\nu-1)} - \mathbf{E}_Z \left\{ g_{i_b}^{(\nu-1)} \right\} \right)^2 \right\} \right. \right. \\ &\left. \left. - \prod_{b=1}^{l} \mathbf{E}_Z \left\{ \left(g_{i_b}^{(\nu-1)} - \mathbf{E}_Z \left\{ g_{i_b}^{(\nu-1)} \right\} \right)^2 \right\} \right) \right. \\ &\left. \times i_b \text{ replaced by } j_b \right] + \mathcal{O}(h_\nu^{-1}) . \end{aligned}$$

As above, by the assumptions, s_n converges to a non-zero limit for $n \to \infty$ and the expectations asymptotically decompose into a product expectations over the is and js. However, the latter is exactly what is subtracted and thus the expression above and the expression in Eq.(25) go to zero for all $\delta \in (0, 1]$. Thus in summary, we finally showed that expression (17) converges to zero for $n \to \infty$ with probability of at least one over the weights. □

To finally conclude that of $\gamma_{j,n}^{(\nu)}$ in each layer $\nu = 2, \ldots, k$ become independent random variables, we have to show that this is the case for $\nu = 2$.

Corollary 1. *Let $f_i^{(2)}(x)$, $g_i^{(2)}(x)$, $s_n^{(2)2}$ and $\gamma_{j,n}^{(2)}$ be as defined in Eq. (9), (10), (15) and (14), respectively. Then it follows that $\gamma_{j,n}^{(\nu)}$ are independent random variables.*

Proof. This follows from the definition of the pre-activations and the fact that the preceding layer is deterministic and does not have dropout. □

Since $\psi^{(\nu)}(t)$ and $\psi^{(\nu)}(t)$ with $\gamma_{j,n}^{(\nu)}$ replaced by a random variable drawn from the *marginalized distribution* of $\gamma_{j,n}^{(\nu)}$, have the same limiting distribution, we can study the limiting distribution of the form by studying the limiting distribution of the latter, which is a simple consequence of Lyapunov's central limit theorem.

Lemma 2. *Let the pre-activations $\{f_u^{(\nu-1)}(x)\}_{u \in \mathbb{N}}$ converge, for $n \to \infty$ in distribution, to independently distributed random variables, let $f_i^{(\nu)}(x)$, $g_i^{(\nu)}(x)$, $s_n^{(\nu)2}$ and $\gamma_{j,n}^{(\nu)}$ be as defined in Eq. (9), (10), (15) and (14), respectively. Then $\psi^{(\nu)}(t)$ converges, for $\nu = 2, \ldots, k$, in distribution to a Gaussian distributed random variable with zero mean and unit variance*

Proof. Due to lemma 1 it is enough to consider $\psi^{(\nu)}(t)$ with $\gamma_{j,n}^{(\nu)}$ replaced by a random variable drawn from the *marginalized distribution* of $\gamma_{j,n}^{(\nu)}$. The random variables drawn from the marginalized distribution are independent by definition. From lemma 7 it follows that they satisfy

the Lyapunov-Condition (26), so we can apply Lyapuno's central limit theorem 3 to proof that $\psi^{(\nu)}(t)$ converges in distribution to a Gaussian distributed random variable. Since it has by construction zero mean and unit variance, the claim follows immediately. □

To finally prove that the pre-activation in layer ν have Gaussian distributed is a corollary of the results obtained so far.

Corollary 2. *Under the assumptions of lemma 2, the pre-activations of layer ν, $\{f_u^{(\nu)}(x)\}_{u\in\mathbb{N}}$, converge in distribution, for $n \to \infty$, to independent Gaussian random variables.*

Proof. From Lemma 2, Lemma 1 and the Cramér-Wold theorem, it follows that every finite marginal of $f^{(\nu)}(x) \in \mathbb{R}^\infty$ converges in distribution to a Gaussian distributed random variable. From which follows, that $\{f_u^{(\nu)}(x)\}_{u\in\mathbb{N}}$ converges in distribution to a Gaussian process. Due to lemma 3 for all finite index-sets $M \subset \mathbb{N}$ the covariance the pre-activations $f_v^{(\nu)}(x)$ and $f_u^{(\nu)}(x)$ for $u \neq v$ and $u, v \in M$ converges to zero almost surely. Because the pre-activations converges to a uncorrelated Gaussian process, it follows, that $\{f_u^{(\nu)}(x)\}_{u\in\mathbb{N}}$ converge to independent random variables. □

A.1. Limiting Covariance Structure of Layer ν

In this section we prove that the covariance structure of a finite sub set of the pre-activations in each layer becomes diagonal, in the limit of infinite width, i.e. for $n \to \infty$.

Lemma 3. *Let the pre-activations $\{f_u^{(\nu-1)}(x)\}_{u\in\mathbb{N}}$ converge, for $n \to \infty$ in distribution to independently distributed random variables, $f_i^{(\nu)}(x)$ and $g_i^{(\nu)}(x)$ be as defined in Eq. (9), (10) . Then, for $\nu = 2, \ldots, k$ and all finite index-sets $M \subset \mathbb{N}$,*

$$Cov\left(f_u^{(\nu)}, f_v^{(\nu)}\right) = \boldsymbol{E}_Z\left\{f_u^{(\nu)} f_v^{(\nu)}\right\} - \boldsymbol{E}_Z\left\{f_u^{(\nu)}\right\}\boldsymbol{E}_Z\left\{f_v^{(\nu)}\right\}$$

converges (almost surely) to zero for $n \to \infty$, whenever $u \neq v$, where $u, v \in M$.

Proof. To show that the covariance vanishes for $u \neq v$, we use high probability bounds on the weights. Since the expectation with respect to the weights of layer ν vanishes, we can bound the expression above using Chebyshev's inequality $\mathbf{P}_W(|f(W)| \leq \sqrt{\mathrm{Var}_W(f(W))/\delta}) \geq 1 - \delta$. The variance is our case is simply the second moment $\mathbf{E}_W\left\{f^2(W)\right\}$. With probability of at least $1 - \delta$, we find that

$$\mathrm{Cov}\left(f_v^{(\nu)}, f_u^{(\nu)}\right) \leq \frac{1}{h_{\nu-1}^2\sqrt{\delta}} \sum_{l\neq k=1}^{h_{\nu-1}} \Big(\mathbf{E}_Z\left\{g_k^{(\nu-1)} g_l^{(\nu-1)}\right\} - \mathbf{E}_Z\left\{g_k^{(\nu-1)}\right\}\mathbf{E}_Z\left\{g_l^{(\nu-1)}\right\}\Big)^2 + \mathcal{O}(h_{\nu-1}^{-1}) .$$

For $\nu = 2$, the leading term in the expression above is zero, because by definition, see Eq. (10), $\mathbf{E}\{g_k^{(\nu-1)} g_l^{(\nu-1)}\} = \mathbf{E}\{g_k^{(\nu-1)}\}\mathbf{E}\{g_l^{(\nu-1)}\}$. For $\nu > 2$ the sum is of order $\mathcal{O}(h_{\nu-1}^2)$ such that the leading contribution with respect to $h_{\nu-1}$ of expression on the right hand side, is at most of order $\mathcal{O}(1)$. However, since the pre-activations of the proceeding layer converge to independent random variables as $n \to \infty$, i.e. $\mathbf{E}\{g_k^{(\nu-1)} g_l^{(\nu-1)}\} \to \mathbf{E}\{g_k^{(\nu-1)}\}\mathbf{E}\{g_l^{(\nu-1)}\}$ for $l \neq k$, the expression above converges to zero for all $\delta \in (0, 1]$.

For $\nu = 2, \ldots, k$ we rescale δ by a factor of $2/(|M|(|M| - 1))$ and apply the union bound over $u \neq v$ where $u, v \in M$ to find that $\mathrm{Cov}\left(f_v^{(\nu)}, f_u^{(\nu)}\right)$ converges to zero for all $u \neq v$, where $u, v \in M$, with at least probability $1 - \delta$. Since $\delta \in (0, 1]$ is arbitrary the claim follows. □

A.2. Lyapunov Condition for Random Neural Networks

We prove that a neural network, given by Eq. (9) and (10) satisfies the Lyapunov condition,

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{r_n} \mathbf{E}_Z\left\{|X_{nk} - \mu_{nk}|^{2+\delta}\right\} = 0 , \qquad (26)$$

where $\mu_{nk} = \mathbf{E}\{X_{nk}\}$ and $s_n^2 = \sum_j \mathbf{E}\left\{(X_{nj} - \mu_{nj})^2)\right\}$. The Lyapunov condition is a necessary condition of Lyapunovs central limit theorem (Billingsley, 1995), which we state for completeness.

Theorem 3. *Suppose that for each n the sequence $X_{n1}, \ldots, X_{nr_n}$ is independent and satisfies $\mu_{nk} = \boldsymbol{E}_Z\{X_{nk}\}$, $\sigma_{nk}^2 = \boldsymbol{E}_Z\left\{(X_{nk} - \mu_{nk})^2\right\}$, $s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2$ and let $S_n = X_{n1} + \cdots + X_{nr_n}$. If the Lyapunov-condition (26) holds for some positive δ, then $(S_n - \sum_{k=1}^{r_n} \mu_{nk})/s_n \Rightarrow N$.*

The main argument of the proof relies on the fact, that the activations of the neural network in the numerator and the denominator of condition (26), can be bounded independent from the width of the neural network.

Due to the assumptions on the activation functions of being Lipschitz continuous with Lipschitz constant L, we find that

$$|\phi(u)| \leq |\phi(v)| + K|u - v| , \qquad (27)$$
$$|\phi(u)| \geq |\phi(v)| - K|u - v| \qquad (28)$$

holds for all $u, v \in \mathbb{R}$. These bounds will allow us to bound the activation functions. First, we need a simple lemma bounding the activation functions of a neural network with weights drawn form a normal distribution. To formulate the next lemma we make the following assumption. According to the idea of dropout, we set a part of the activation in each

layer to zero. We define the set $U^{(\nu)}$ to be the index set of activations in layer ν not set to zero. The collection of index sets $U^{(\nu)}$, for $\nu = 1, \ldots, k$, is called an *activation pattern*.

Lemma 4. *Let $\alpha > 0$, $x \in \mathbb{R}^N$ be fixed such that $||x||_2^2 \leq N\alpha$ and the bias b as well as the weights w_i of a network pre-activation be drawn from $\mathcal{N}(0,1)$. Then for each finite integer k, independent of N and all $U \subseteq U \in \mathcal{P}(\{1, ..., N\})$ there exists a constant C independent of N such that*

$$\boldsymbol{E}_W \left\{ \left| b + \frac{1}{\sqrt{N}} \sum_{j \in U} w_j x_j \right|^k \right\} \leq C \, . \tag{29}$$

Proof. Let $k = 2m$, $x_\perp = (x_i)_{i \in U}$ and $w_\perp = (w_i)_{i \in U}$ such that $\sum_{j \in U} w_j x_j = \langle w_\perp, x_\perp \rangle$ and $||x_\perp||_2^2 \leq ||x||_2^2 \leq \alpha N$. Without loss of generality we assume that $1 \in U$. We rotate $w_\perp$ such that $x_\perp$ is given by a vector with $||x_\perp||_2$ in the first element and zero elsewhere. This simplifies the expression (29) of the lemma to

$$\begin{aligned} &\sum_{j=1}^{m} \binom{2m}{2j} \left(\frac{||x_\perp||_2}{\sqrt{N}} \right)^{2(m-j)} c_{2(m-j)} c_{2j} \\ \leq &\sum_{j=0}^{m} \binom{2m}{2j} \alpha^{(m-j)} c_{2(m-j)} c_{2j} \leq C \, , \end{aligned}$$

where c_n is the nth moment of a Gaussian random variable with zero mean and unit variance. The last inequality in the estimate above is due to the fact that the sum as well as the moments and coefficients are independent of N, proving the claim for even k. Let X be a random variable with $\mathbf{E}\left\{|X|^{2k}\right\} < \infty$, then it follows that

$$\begin{aligned} \mathbf{E}\left\{|X|^{2m+1}\right\} &= \sqrt{\mathbf{E}\left\{|X|^{4m+2}\right\} - \mathrm{Var}\left(|X|^{2m+1}\right)} \\ &\leq \sqrt{\mathbf{E}\left\{|X|^{4m+2}\right\}} \, . \end{aligned} \tag{30}$$

Since all finite moments of a Gaussian random variable exist, the claim follows for odd k from Eq. (30). □

We now apply lemma 4 recursively to the neural network to make the following observation. Since it would only slightly change the absolute values of the constants bounding the expressions in the following and not change their dependence on the layer width, we will keep all activations in the networks, i.e. set $U^{(\nu)} = [h_\nu(n)]$, without explicitly mentioning it.

Lemma 5. *Let $\alpha > 0$, $x \in \mathbb{R}^N$ be fixed such that $||x||_2^2 \leq N\alpha$, $g_i^{(\nu)}(x)$ be as defined in Eq. (10) and the bias $b_i^{(\nu)}$ as well as the weights $w_{ij}^{(\nu)}$, of a network pre-activation, be drawn from $\mathcal{N}(0,1)$. Then there exists a constant $C < \infty$ such that*

$$\boldsymbol{E}_W \left\{ \left| g_i^{(\nu)}(x) \right|^k \right\} \leq C \quad \forall \nu = 1, \ldots, L \, ,$$

for each finite, positive integer k, where the expectation is over the weights and the bias.

Proof. We prove the statement by induction over ν. For $\nu = 1$, this follows immediately from lemma 4 and Eq. (27) with $v = 0$. Thus, the induction hypothesis is that this is true for ν. For $\nu + 1$ we find

$$\begin{aligned} \mathbf{E}_W \left\{ \left| g_i^{(\nu+1)}(x) \right|^k \right\} &\leq \mathbf{E}_W \left\{ \left(K \left| f_i(x)^{(\nu+1)} \right| + |\phi(0)| \right)^k \right\} \\ &= \sum_{j=0}^{k} \binom{k}{j} K^j \, |\phi(0)|^{k-j} \, \mathbf{E}_W \left\{ \left| f_i(x)^{(\nu+1)} \right|^j \right\} \, . \end{aligned}$$

As $f_i^{(\nu+1)}(x)$ is an affine-linear combination of $g_j^{(\nu)}$ which is bounded according to the induction hypothesis, all absolute moments of $f_i^{(\nu+1)}(x)$ are bounded. This concludes the induction step. □

We facilitate this bound on the absolute moments of the individual moments, to find a high probability bound on polynomials of the absolute value of the activation functions.

Lemma 6. *Let $p(x)$ be a polynomial of degree k in x. With probability of at least $1 - \delta$ it follows, under the assumption of Lemma 5, that there exists a constant C, independent of $h_\nu(n)$ such that*

$$\left| p\left(\left| g_i^{(\nu)}(x) \right| \right) - \boldsymbol{E}_W \left\{ p\left(\left| g_i^{(\nu)}(x) \right| \right) \right\} \right| \leq \sqrt{\frac{h_\nu(n)}{\delta}} C \, ,$$

where the expectations are over the weights and bias, holds for all $i = 1, \ldots, h_\nu(n)$ and any finite integer k.

Proof. We have that $\mathrm{Var}(X) \leq \sqrt{\mathbf{E}X^2}$ for any random variable X with finite second moment. From this inequality, an expansion of p in $\mathbf{E}|g_i^{(\nu)}(x)|$ and lemma 5 we find that there exists a constant C' such that

$$\mathrm{Var}_W \left(p\left(\left| g_i^{(\nu)}(x) \right| \right) \right) \leq C' \, ,$$

where the variance is with respect to the weights. This estimate in combination with Chebyshev's inequality, yields that with probability of at least $1 - \delta / h_\nu(n)$

$$\left| p\left(\left| g_i^{(\nu)}(x) \right| \right) - \mathbf{E}_W \left\{ p\left(\left| g_i^{(\nu)}(x) \right| \right) \right\} \right| \leq \sqrt{\frac{h_\nu(n)}{\delta}} C \, .$$

Applying the union bound over i to the expression above, we find that it holds for all $i = 1, \ldots, h_\nu(n)$ with probability of at least $1 - \delta$. □

The following generalization of the lemmas above, will simplify the proof of the final lemma in this section.

Corollary 3. *Let $g_{i,l}^{(\nu)}(x)$ be the activation function with the activation pattern $U_l^{(\nu')}$ for $\nu' = 1, \ldots, \nu-1$ and $l \in \mathcal{L}$, where $\mathcal{L}$ is a finite index set. Then lemma 5 and lemma 6 hold also for polynomials of finite degree in $\{|g_{i,l}^{(\nu)}(x) \mid l \in \mathcal{L}\}$.*

Proof. The claim follows immediately form the fact that lemma 4 holds for all activation patterns and the application of it in the proofs of lemma 5 and lemma 6. □

With lemma 6 and corollary 3 at hand, we can easily prove that the condition of the Lyapunov CLT is satisfied. In contrast to the lemmas above, the expectation in the following lemma will be over the dropout random variables.

Lemma 7. *Let $\alpha > 0$, $x \in \mathbb{R}^d$ be fixed such that $||x||_2^2 \leq d\alpha$, $\gamma_{j,n}^{(\nu)}(t)$ and $s_n^{(\nu)}(x)$ be as defined in Eq. (14) and (15) respectively. We assume that $s_n^{(\nu)}$ does not converge to zero. Then we have that*

$$\lim_{n\to\infty} \left(\frac{1}{s_n^{(\nu)}(x)}\right)^4 \sum_{j=1}^{h_{\nu-1}} \boldsymbol{E}_Z \left\{ \left| \gamma_{j,n}^{(\nu)}(t) \right|^4 \right\} = 0 \,,$$

almost surely over the weights.

Proof. We consider the following sequences

$$a_n = \sqrt{\frac{\delta}{h_{\nu-1}(n)}} \left(s_n^{(\nu)}\right)^2$$

and

$$b_n = \frac{\delta}{h_{\nu-1}(n)} \sum_{j=1}^{h_{\nu-1}} \mathbf{E}_Z \left\{ \left| \gamma_{j,n}^{(\nu)}(t) \right|^4 \right\} \,.$$

To see that a_n converges we make the following estimate

$$a_n \leq C \frac{1}{h_{\nu-1}(n)} \sum_{j=1}^{h_{\nu-1}} \sum_{u,v\in M} w_{uj}^{(\nu)} w_{vj}^{(\nu)} t_u t_v \,,$$

where we made use of (16), lemma 6 and corollary 3 as follows. The expectation value in Eq. (16) is over the space of dropout configurations and can be interpreted as average over activation patterns. Since the corollary 3 holds for all activation patterns, we arrive at the expression above. The remainder converges to $C||t||_2^2$ for $n \to \infty$ with probability 1, for all values of $\delta \in [0, 1]$.

The convergence of b_n to zero is due to the following observations. By definition we have that $b_n \geq 0$. Moreover, the summands of the sum over j are polynomials of degree 4. These polynomials consists of terms of the form

$$(-1)^{4-a} \left| \mathbf{E}_Z \left\{ g_j^{(\nu-1)}(x) \right\} \right|^{4-a} \mathbf{E}_Z \left\{ \left| g_j^{(\nu-1)}(x) \right|^a \right\} \,,$$

where $a = 0, \ldots, 4$. The first two factors in the expression above can be upper bounded by $\mathbf{E}\{|g_j^{(\nu-1)}(x)|^{4-a}\}$. Due to corollary 3 and lemma 6 there exists a constant C, such that with probability of at least $1-\delta$ the resulting polynomial can be bounded by $\sqrt{h_{\nu-1}(n)/\delta}\, C$ for all $j = 1, \ldots, h_{\nu-1}(n)$. This leads us to the upper bound

$$b_n \leq \sqrt{\frac{\delta}{h_{\nu-1}(n)}} \frac{C}{h_{\nu-1}^2(n)} \times \sum_{j=1}^{h_\nu} \sum_{u,v,l,o\in M} w_{uj}^{(\nu)} w_{vj}^{(\nu)} w_{lj}^{(\nu)} w_{oj}^{(\nu)} t_u t_v t_l t_o \,.$$

An inspection of an expectation of the expression above, with respect to the weights, in combination with the Markov inequality shows that it converges to zero for $n \to \infty$ with probability 1 for all values of $\delta \in [0, 1]$ and $t \in \mathbb{R}^{|M|}$.

By assumption of the lemma, $s_n(\nu)$ does not converge to zero, neither does a_n. Thus since a_n and b_n converge, we can find, applying the union bound, that the limit of b_n/a_n^2 as quotient of their limits and thus find that it converges to zero for $n \to \infty$. However, b_n/a_n^2 is nothing else then the sequence of the claim. Since we showed that it converges with probability of at least $1 - \delta$ over the weights, for all $\delta \in [0, 1]$, we proved the claim. □

B. Empirical Observations

Further implementation details All biases of the networks are set to zero. Training is done on shuffled mini-batches of size 100 using the standard train-test data split without any further preprocessing. All experiments were run in pytorch (Paszke et al., 2019) using a `Intel(R) Xeon(R) Gold 6126` CPU and a `NVidia GeForce GTX 1080ti` GPU.

Correlations in random networks We analyze the weight correlations and pre-activation correlations of the untrained $\mathcal{H}_{\text{narrow}}$ and $\mathcal{H}_{\text{wide}}$. The (row-wise) Pearson correlations of the weight matrices (Fig. 3, top row) are centered around zero with estimation errors that are determined by the width of the respective network. The column-wise weight correlations look the same. To calculate pre-activation correlations (see Eq. 9), we run 10,000 forward passes with dropout for a fixed test image. Fig. 4 (top row) shows that these correlations are largely similar to the weight correlations—as can be expected.

Correlations in trained networks For $\mathcal{H}_{\text{narrow}}$, we find similar weight correlations for all hidden layers with correlations coefficients that range from -1 to 1 (Fig. 3, bottom left). While the distributions for $\mathcal{H}_{\text{wide}}$ are even more homogeneous across layers, they span only from approximately -0.5 to 0.5 (Fig. 3, bottom right). More importantly, these distributions resemble the central part of the $\mathcal{H}_{\text{narrow}}$ distributions and do by no means collapse to zero, i.e., although network width varies by one order of magnitude, the global dependence structures are similar in both networks. Next, we study the dependencies between pre-activations that are (mainly) induced by the weight dependencies. Fig. 4 shows roughly the same dependence pattern as Fig. 3, however, for $\mathcal{H}_{\text{wide}}$ the pre-activation correlation distribution gets broader with layer depth (Fig. 4, bottom right). Intuitively, we can make sense of this observation: the network iteratively withdraws input information to finally only keep what is useful for classification, and less information distributed over a fixed number of neurons means stronger correlations. Moreover, not only pre-activation correlations increase with layer depth but also their variances and thus covariances.

Further observations for trained networks We find that sigmoid activations suppress non-normal distributions, which is in contrast to tanh, ReLU, and linear activation functions. Further experiments with a custom non-linearity that is a proxy to sigmoid suggest that the combination of being constrained and mapping onto only one half-axis might be a critical condition for a Gaussian inducing non-linearity.

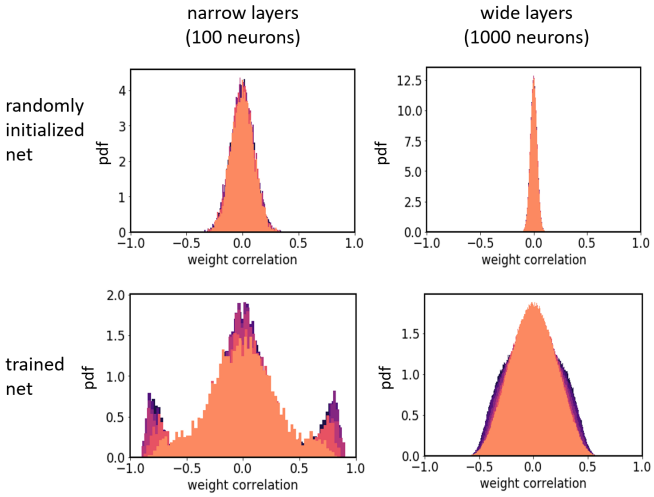

Figure 3. Weight correlations of randomly initialized (top row) and trained (bottom row) networks that are either narrow ($h = 100$, left column) or wide ($h = 1000$, right column). The different hidden layers are color-coded: from first (melon) to last (purple) hidden layer. The Pearson correlation coefficient is used.

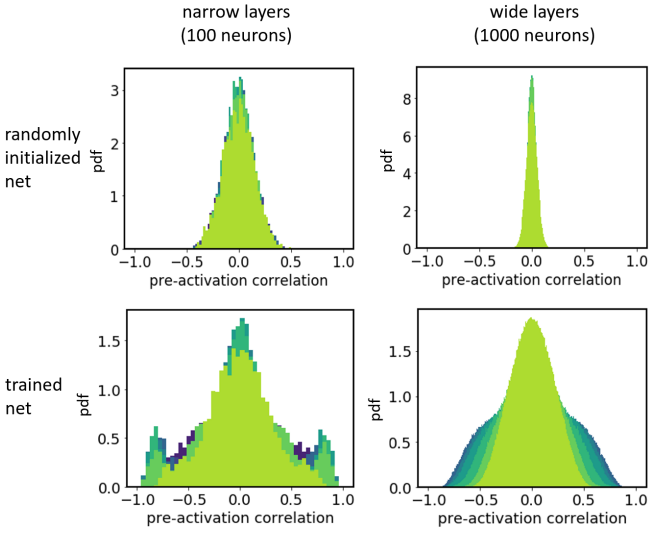

Figure 4. Pre-activation correlations of randomly initialized (top row) and trained (bottom row) networks that are either narrow ($h = 100$, left column) or wide ($h = 1000$, right column). The different hidden layers are color-coded: from first (green) to last (dark blue) hidden layer. The Pearson correlation coefficient is used.

C. On Correlated Systems

Here, we provide additional information accompanying the calculations in Sec. 4. For instance, the PDF in Eq. (3) can be calculated using a filter integral:

$$\begin{aligned}\mathrm{PDF}_{\tilde{f}}(\xi) &\propto \int \mathrm{d}x\mathrm{d}y\, \delta(\xi - xy)\mathrm{e}^{-x^2/2-y^2/2} && (31)\\ &\propto \int \mathrm{d}x\mathrm{d}y\mathrm{d}k\, \mathrm{e}^{-\mathrm{i}k(\xi-xy)}\mathrm{e}^{-x^2/2-y^2/2} && (32)\\ &= \int \mathrm{d}k\, \mathrm{e}^{-\mathrm{i}k\xi} \int \mathrm{d}x\mathrm{d}y\, \exp\left(-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{A}\,\boldsymbol{b}\right) && (33)\end{aligned}$$

with the shorthand notations

$$\boldsymbol{b} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{and} \quad \boldsymbol{A} = \begin{pmatrix} 1 & -\mathrm{i}k \\ -\mathrm{i}k & 1 \end{pmatrix} . \qquad (34)$$

Evaluating the inner integrals thus leads to

$$\mathrm{PDF}_{\tilde{f}}(\xi) \propto \int_{-\infty}^{+\infty} \mathrm{d}k\, \frac{\mathrm{e}^{-\mathrm{i}k\xi}}{\sqrt{1+k^2}} = 2K_0(|\xi|) . \qquad (35)$$

The differing prefactor in Eq. (3) follows from the normalization condition for a PDF. For convenience, we omitted any prefactors here.

A direct extension of this calculation to non-zero mean is challenging. Instead we provide a heuristic explanation where we include this aspect via

$$X = \mu_X + \sigma_X\epsilon_X \quad \text{with } \epsilon_X \sim \mathcal{N}(0,1) \quad (Y \text{ analogous}) , \qquad (36)$$

leading to

$$XY = \underbrace{\mu_X\mu_Y}_{\text{const.}} + \underbrace{\mu_X\sigma_Y\,\epsilon_Y + \mu_Y\sigma_X\,\epsilon_X}_{\text{“Gaussian”}} + \underbrace{\sigma_X\sigma_Y\,\epsilon_X\epsilon_Y}_{\text{“tail”}} . \qquad (37)$$

Neglecting the first term as constant, the two middle summands follow a Gaussian behavior while the last one contains a product giving rise to exponential tails. While this heuristic breakdown ignores the correlations of the twice occurring ϵ's it qualitatively captures the behavior of XY. Indeed, we find a stronger emphasis towards exponential tails for $\sigma > \mu$ and for Gaussian behavior the other way around. For illustration, Fig. 5 shows the empirical distribution of Z for $\mu_X = \mu_Y = 0$ (left) and $\mu_X = \mu_Y = 10$ (right). As a guide we added a numerically obtained PDF as well as an approximation following the logic of our explanation for Eq. (37). In this second case we used the addition of two independent random variables $\tilde{Z} = \tilde{X} + \tilde{Y}$, where $\tilde{X} \sim N(\mu_X\mu_Y, \sqrt{(\sigma_X\mu_Y)^2 + (\sigma_Y\mu_X)^2})$ is a Gaussian which models the first three terms in Eq. (37). For $\tilde{Y}$ we use an exponential distribution,

$$\mathrm{PDF}_{\tilde{Y}}(\xi) = \frac{1}{2\sigma_X\sigma_Y} \exp\left(-\frac{|\xi|}{\sigma_X\sigma_Y}\right) , \qquad (38)$$

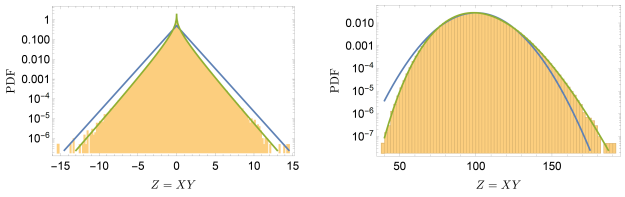

Figure 5. Logarithmic visualization of the PDF for the product of two Gaussian random variables X, Y. Shown in blue and green are an approximation to the PDF (see text) and an exact numerical result respectively. Left side shows $\mu_X = \mu_Y = 0$ and right $\mu_X = \mu_Y = 10$, in both panels $\sigma_X = \sigma_Y = 1$.

designed to capture the tail properties of the Bessel function. Here, we neglected the additional $|\xi|^{-1/2}$ dependence of the asymptotic, Eq. (8), to keep the resulting integrals of a Gaussian type. This allows us to give an explicit expression for $\tilde{Z}$,

$$\mathrm{PDF}_{\tilde{Z}}(\xi) = \frac{\mathrm{e}^{\frac{\sigma_1^2}{2\sigma_2^2}}}{4\sigma_2}\left(\mathrm{e}^{\frac{\mu-\xi}{\sigma_2}}\,\mathrm{Erfc}\left(\frac{\sigma_1^2 + (\mu-\xi)\sigma_2}{\sqrt{2}\sigma_1\sigma_2}\right) + \mathrm{e}^{\frac{\xi-\mu}{\sigma_2}}\,\mathrm{Erfc}\left(\frac{\sigma_1^2 + (\xi-\mu)\sigma_2}{\sqrt{2}\sigma_1\sigma_2}\right)\right) . \qquad (39)$$

Therein we used the short hand notations $\mu = \mu_x\mu_Y$, $\sigma_1 = \sqrt{(\sigma_X\mu_Y)^2 + (\sigma_Y\mu_X)^2)}$ and $\sigma_2 = \sigma_X\sigma_Y$. Also,

$$\mathrm{Erfc}(\zeta) = 1 - \mathrm{Erf}(\zeta) = 1 - \frac{2}{\sqrt{\pi}}\int_0^{\zeta} \mathrm{d}t\, \mathrm{e}^{-t^2} \qquad (40)$$

denotes the complementary error function. As can be seen in Fig. 5 this approximation roughly captures the exponential tails. Furthermore, we find that for $\mu > \sigma$ the tail behavior is suppressed and the distribution becomes asymmetric, a property we similarly observe in the real data, see Fig. 1.

The extension of the $Z = XY$ toy model to a sum of random variables $Z = \sum_{i=1}^{h} X_iY_i$ can be treated, at least theoretically, in the same way as the original problem:

$$\begin{aligned}\mathrm{PDF}_Z(\xi) &\propto \int \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}\, \delta(\xi - \boldsymbol{x}^T\boldsymbol{y})\mathrm{e}^{-\boldsymbol{x}^T\boldsymbol{\Sigma}_X^{-1}\boldsymbol{x}/2 - \boldsymbol{y}^T\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{y}/2}\\ &= \int \mathrm{d}k\, \mathrm{e}^{-\mathrm{i}k\xi} \int \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}\, \exp\left(-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{b}\right) , \end{aligned} \qquad (41)$$

where in this case $\boldsymbol{b} \in \mathbb{R}^{2h}$ is instead a stacked vector of both $\boldsymbol{x}$ and $\boldsymbol{y}$, compare Eq. (34). Furthermore,

$$A = \begin{pmatrix} \Sigma_X^{-1} & -\mathrm{i}k\mathbb{1} \\ -\mathrm{i}k\mathbb{1} & \Sigma_Y^{-1} \end{pmatrix} . \qquad (42)$$

Evaluating the inner integrals gives rise to

$$\mathrm{PDF}_Z(\xi) \propto \int_{-\infty}^{+\infty} \mathrm{d}k \, \frac{\mathrm{e}^{-\mathrm{i}k\xi}}{\sqrt{\det\left(\Sigma_{\tilde{z}}^{-1}\Sigma_{\tilde{g}}^{-1} + k^2 \mathbb{1}\right)}} \ , \quad (43)$$

where we used the identity

$$\det\begin{pmatrix} E & B \\ C & D \end{pmatrix} = \det\left(ED - BC\right) \quad (44)$$

valid for arbitrary square matrices B, C, D, E as long as $CD - DC = 0$ holds. Assuming that the matrix $\Sigma_{\tilde{z}}^{-1}\Sigma_{\tilde{g}}^{-1}$ is diagonalizable with (positive) eigenvalues σ_i^2 leads to

$$\mathrm{PDF}_Z(\xi) \propto \int_{-\infty}^{+\infty} \mathrm{d}k \, \frac{\mathrm{e}^{-\mathrm{i}k\xi}}{\sqrt{\prod_{i=1}^{h}\left(\sigma_i^2 + k^2\right)}} \ . \quad (45)$$

Depending on the eigenvalue spectrum the resulting function can exhibit quite different tail behaviors. To briefly illustrate this, let us make two remarks: If we assume doubly degenerate eigenvalues we can "ignore" the square root and Eq. (45) instead has poles of *first* order at the positions $k = \pm\mathrm{i}\sigma_i$. Based on the residue theorem we then find

$$\mathrm{PDF}_Z(\xi) \propto \sum_{i=1}^{h} \frac{\mathrm{e}^{-\sigma_i|\xi|}}{2\sigma_i \prod_{j\neq i}^{h}(\sigma_j^2 + \sigma_i^2)} \, . \quad (46)$$

In a loose sense this might be seen as a discrete variant of a Laplace transform and therefore has a similar variety of outcomes. Those could be largely restricted if one assumes the spectrum of the σ_i to be bounded. In the body of the paper we omitted this discussion and instead chose an easier example closer to the intended application.

The correlation expressed in Eq. (5) can be seen as stemming from a multivariate Gaussian with zero mean and covariance matrix

$$\boldsymbol{\Sigma}_{ij} = \begin{cases} 1 & i = j \\ c & i \neq j \end{cases} \ . \quad (47)$$

While this matrix has a very simple structure with only two distinct eigenvalues of $1 + (h-1)c$ and $1 - c$, the latter is $h - 1$ fold degenerate. Therefore, our considerations from Eq. (46) do not directly apply. Instead, we use this model to build the weight matrices for the random network shown in Fig. 2. To this end, we draw for each $\boldsymbol{W}_\nu$ two sets of such multivariate Gaussian distributed vectors $\{\boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{R}^h\}$. These vectors form two matrices $\boldsymbol{A} = (a_1, \ldots, a_h) \in \mathbb{R}^{h\times h}$ ($\boldsymbol{B}$ analogous) with correlated rows such that each $\boldsymbol{W}$ is given by $\boldsymbol{W}_\nu = (\boldsymbol{A}_\nu + \boldsymbol{B}_\nu^T)/(2\sqrt{qh})$, where $q = 1 - p$ denotes the keep rate. This way we ensure independent correlations both among rows and columns of the matrices.