Deep neural networks with dependent weights: Gaussian Process mixture limit, heavy tails, sparsity and compressibility

Hoil Lee
KAIST
Daejeon, South Korea
hoil.lee@kaist.ac.kr

Fadhel Ayed
Huawei Technologies
Paris, France
fadhel.ayed@gmail.com

Paul Jung
Fordham University
New York City, USA
pjung3@fordham.edu

Juho Lee
KAIST
Daejeon, South Korea
juholee@kaist.ac.kr

Hongseok Yang
KAIST
Daejeon, South Korea
hongseok.yang@kaist.ac.kr

François Caron
University of Oxford
Oxford, UK
caron@stats.ox.ac.uk

Abstract

This work studies the infinite-width limit of deep feedforward neural networks whose weights are dependent, and modelled via a mixture of Gaussian distributions. Under this model, we show that each layer of the infinite-width neural network can be characterised by two simple quantities: a non-negative scalar parameter and a Lévy measure on the positive reals. If the scalar parameters are strictly positive and the Lévy measures are trivial at all hidden layers, then one recovers the classical Gaussian process (GP) limit, obtained with iid Gaussian weights. More interestingly, if the Lévy measure of at least one layer is non-trivial, we obtain a mixture of Gaussian processes (MoGP) in the large-width limit. The behaviour of the neural network in this regime is very different from the GP regime. One obtains correlated outputs, with non-Gaussian distributions, possibly with heavy tails. We illustrate some of the benefits of the MoGP regime over the GP regime in terms of representation learning and compressibility on simulated, MNIST and Fashion MNIST datasets.

1 Introduction

Two decades after the seminal work of Radford Neal [1996], the last few years have seen a renewed and growing interest in the analysis of (deep) neural networks, with random weights, in the infinite-width limit. When the weights are independent and identically distributed (iid), and suitably scaled Gaussian random variables, the random function corresponding to this random neural network converges to a Gaussian process [Neal, 1996, Lee et al., 2018, Matthews et al., 2018, Yang, 2019, Bracale et al., 2021]. The connection to Gaussian processes has deepened our understand-

Presented at the workshop on Heavy Tails in ML at NeurIPS 2023. This is an extended abstract of Lee et al. [2022].

ing of large-width neural networks, and motivated both the use of Bayesian or kernel regression inference methods [Lee et al., 2018] as well as the development of kernel methods for training via gradient descent in the infinite-width limit [Jacot et al., 2018].

While insightful, the Gaussian process connection also highlights some of the limitations of large-width neural networks with iid Gaussian weights. As already noted by Neal [1995], *"with Gaussian priors the contributions of individual hidden units are all negligible, and consequently, these units do not represent 'hidden features' that capture important aspects of the data."* Moreover, the different dimensions of the output of the neural network become independent Gaussian processes in the infinite-width limit, which is generally undesirable. Finally, from a Bayesian perspective, the Gaussian independence assumption on weights is often seen as unrealistic: estimated weights of deep neural networks generally exhibit dependencies and heavy tails [Martin and Mahoney, 2019, Wenzel et al., 2020, Fortuin et al., 2021], and thus a family of prior distributions which allow for heavy tails is desirable. To alleviate some of these limitations, iid non-Gaussian random weights have been considered, either assuming stable [Neal, 1996, Der and Lee, 2006, Favaro et al., 2020], or more generally light-tailed/heavy-tailed distributions [Jung et al., 2023]. However, due to the same iid assumption, some of the above limitations pertain, such as independence of the dimensions of the output.

We consider a more structured distribution on the weights of the neural network. We assume that weights emanating from a given node are dependent, where the dependency is captured via a scale mixture of Gaussians. More precisely, for a weight $W_{jk}^{(l+1)}$ between node $j = 1, \dots, p_l$ at hidden layer l and node $k = 1, \dots, p_{l+1}$ at hidden layer $l+1$, we assume that

$$W_{jk}^{(l+1)} = \sqrt{\lambda_{p_l,j}^{(l)}} V_{jk}^{(l+1)} \tag{1}$$

where $\lambda_{p_l,j}^{(l)}$, for $j = 1, \dots, p_l$, are nonnegative iid random variance parameters, one for each node $j = 1, \dots, p_l$ at layer l, and $V_{jk}^{(l+1)}$ are iid centred Gaussian random variables with variance $\sigma_v^2 > 0$. The per-node variance term $\lambda_{p_l,j}^{(l)}$ induces some dependency over the weights $W_{j1}^{(l+1)}, \dots, W_{jp_{l+1}}^{(l+1)}$ connected to node j. This assumption has been considered by a number of authors for training (finite) neural networks either (i) as a prior for Bayesian learning and pruning of neural networks, or (ii) as an implicit prior where a regularised empirical risk minimiser with group-sparse penalty is interpreted as a maximum a posteriori estimator, or (iii) as a random weight initialisation scheme for stochastic gradient descent.

2 Main contributions

The objective of this work is to analyse the infinite-width properties of feedforward neural networks with dependent weights of the form in Equation (1). Our work shows that the choice of the distribution of the per-node variance is crucial and can lead to fundamentally different infinite-width limits. Our main assumption is that, at each hidden layer l,

$$\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \xrightarrow{\mathtt{d}} \Lambda^{(l)} \quad \text{as the width } p_l \to \infty, \tag{2}$$

where $\xrightarrow{\mathtt{d}}$ refers to convergence in distribution and $\Lambda^{(l)}$ is some nonnegative random variable, which may be constant. This assumption is natural as it implies that the activations and outputs of the neural network are almost surely finite in the infinite-width limit. Note that $\sum_{j=1}^{p_l} \mathrm{Var}\left(W_{jk}^{(l+1)} \middle| (\lambda_{p_l,j}^{(l)})_{j \geq 1} \right) = \sigma_v^2 \sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)}$. Hence, the assumption in Equation (2) is simi-

lar to the commonly made assumption, in the iid case, that the sum of the variances of the incoming weights to a node converges to a constant in the infinite-width limit [Glorot and Bengio, 2010, He et al., 2015]. The iid Gaussian case indeed arises as a special case by setting $\lambda_{p_l,j}^{(l)} = \frac{c}{p_l}$ for all $j = 1, \ldots, p_l$ for some $c > 0$. Note that $\Lambda^{(l)} = c$ is deterministic in this particular case.

The nonnegative random variable $\Lambda^{(l)}$ is necessarily infinitely divisible, and parameterised by

(i) a location parameter $a^{(l)} \geq 0$ and
(ii) a Lévy measure $\rho^{(l)}$ on $(0, \infty)$.

We prove that, if $a^{(l)} > 0$ and the Lévy measures are trivially zero (that is $\int_0^\infty \rho^{(l)}(dx) = 0$) at all hidden layers l, then the limit is a Gaussian process (GP), as in the iid Gaussian case. As a consequence, all weights are uniformly small, with $\max_{j=1,\ldots,p_l} |W_{jk}^{(l+1)}| \to 0$ in probability. We show that this GP limit arises with a few models proposed in the literature, such as the group lasso [Scardapane et al., 2017, Wang et al., 2017] and inverse gamma [Ober and Aitchison, 2021] priors. These neural network models therefore are asymptotically equivalent to a model with iid Gaussian weights in the infinite-width limit.

More interestingly, if at least one of the Lévy measures is non-trivial, we obtain a very different behaviour, and the limit is now a *mixture of Gaussian processes* (MoGP), with a given random covariance kernel. In other words, any m-dimensional output at layer l has the limiting form

$$\mathbf{E}\left[\bigotimes_{k=1,\ldots,m} \mathcal{N}(0, \Sigma^{(l)})\right]$$

where the expectation is taken over the randomness of a random covariance matrix $\Sigma^{(l)}$ which is determined by a random kernel. Here, the size of $\Sigma^{(l)}$ is $n \times n$, where n is the number of inputs. Under the MoGP regime, we show that the following results hold in the infinite-width limit, none of which hold for the iid Gaussian case.

- $\max_{j=1,\ldots,p_l} |W_{jk}^{(l+1)}|$ converges in probability to a random variable which is not degenerately 0. That is, some weights remain non-negligible asymptotically. It is natural to interpret this as being connected to nodes representing important hidden features.
- The different dimensions of the output are dependent.
- The outputs are non-Gaussian, and may exhibit heavy tails depending on the behaviour of the Lévy measures at infinity.
- Pruning the network according to the variance parameter $\lambda_{p_l,j}^{(l)}$ at some level $\epsilon > 0$ sufficiently small, provides a finite, non-empty neural network with positive probability.[1] The resulting error associated to the pruned network can be related to the location parameter and the behaviour of the Lévy measure near 0.
- If in addition, the location parameters satisfy $a^{(l)} = 0$ for all l, then the network is compressible: when pruning the network by removing a fixed proportion $(1 - \kappa) \in (0, 1)$ of nodes at each layer according to the variance parameter $\lambda_{p_l,j}^{(l)}$, the difference between the outputs of the pruned and unpruned networks converges to 0 in probability in the infinite-width limit.
- The random kernels which determine the $\Sigma^{(l)}$ come from a Markov sequence, in l. Moreover, the distribution of the lth kernel can be recursively defined.

[1] Note that there is always some small probability of pruning everything and leaving an empty network.

Model	Limit process	Depend. outputs	Distribution of $Z_k^{(2)}(\mathbf{x}, p_1)$	Tail of $Z_k^{(2)}(\mathbf{x}, p_1)$	Number of active nodes	$\max \lvert W_{jk}^{(2)} \rvert \xrightarrow{\text{pr}} 0$	Tail of of $W_{jk}^{(2)}$	$(W_{(j)k}^{(2)})^2$ decrease in	Compressible
iid	GP	No	Gaussian	Expon.	∞	Yes	Expon.	–	No
(a)	GP	No	Gaussian	Expon.	∞	Yes	Expon.	–	No
(b)	MoGP	Yes	Compound Poisson	Expon.	$\text{Poisson}(2)$	No	Expon.	–	Yes
(c)	MoGP	Yes	Normal-gamma	Expon.	∞	No	Expon.	$O(e^{-cj})$	Yes
(d)	MoGP	Yes	Cauchy	Power-law	∞	No	Power-law	$O(j^{-2})$	Yes

Table 1: Summary of the properties of the neural network models for four different distributions on the per-node variances.

3 Some illustrative examples

We now briefly present some illustrative examples in the case of a simple feedforward neural network with one hidden layer, d_{in}-dimensional input $\mathbf{x} = (x_1, \ldots, x_{d_{\text{in}}})^T$, 2-dimensional output $(Z_1^{(2)}(\mathbf{x}; \mathbf{p}), Z_2^{(2)}(\mathbf{x}; \mathbf{p}))^T$, no bias, $\sigma_v = 1$ and rectified linear unit (ReLU) activation function. For $k = 1, 2$, the output is such that

$$Z_k^{(2)}(\mathbf{x}; p_1) = \sum_{j=1}^{p_1} \sqrt{\lambda_{p_1,j}^{(1)}} V_{jk}^{(2)} \max\left(0, \frac{1}{\sqrt{d_{\text{in}}}} \sum_{i=1}^{d_{\text{in}}} V_{ij}^{(1)} x_i\right).$$

More general deep neural networks and other examples are considered in the full paper [Lee et al., 2022]. As mentioned above, it is well known (see for instance [Lee et al., 2018]) that, if $\lambda_{p_1,j} = \frac{2}{p_1}$ (iid Gaussian weights, or He initialisation [He et al., 2015]), the outputs are asymptotically independent Gaussian processes with, for $k = 1, 2$,

$$\begin{pmatrix} Z_k^{(2)}(\mathbf{x}; p_1) \\ Z_k^{(2)}(\mathbf{x}'; p_1) \end{pmatrix} \xrightarrow{\text{d}} \mathcal{N}\left(0, \begin{pmatrix} \mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}) & \mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}') \\ \mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}') & \mathcal{K}^{(2)}(\mathbf{x}', \mathbf{x}') \end{pmatrix}\right) \text{ as } p_1 \to \infty \tag{3}$$

where the (deterministic) covariance kernel $\mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}')$ is defined by

$$\mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}') = \frac{\|\mathbf{x}\|\|\mathbf{x}'\|}{d_{\text{in}}} \times \frac{1}{\pi}\left(\sqrt{1 - \rho_{\mathbf{x},\mathbf{x}'}^2} + \left(\frac{\pi}{2} + \arcsin \rho_{\mathbf{x},\mathbf{x}'}\right)\rho_{\mathbf{x},\mathbf{x}'}\right), \tag{4}$$

with correlation $\rho_{\mathbf{x},\mathbf{x}'} = \frac{\mathbf{x}^T\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}$.

Consider now the following models for $p_1 \geq 2$:

(a) $\lambda_{p_1,j}^{(1)} \sim \text{IG}\left(2, \frac{2}{p_1}\right)$ (b) $\lambda_{p_1,j}^{(1)} \sim \text{Bernoulli}\left(\frac{2}{p_1}\right)$

(c) $\lambda_{p_1,j}^{(1)} \sim \text{Beta}\left(\frac{1}{p_1}, \frac{1}{2}\right)$ (d) $\lambda_{p_1,j}^{(1)} = \pi^2 \frac{U_j^2}{p_1^2}$ where $U_j \sim \text{Cauchy}_+(0, 1)$

where $\text{IG}(\beta_1, \beta_2)$ denotes the inverse gamma distribution with shape $\beta_1 > 0$ and scale $\beta_2 > 0$, and $\text{Cauchy}_+(0, 1)$ denotes the half-Cauchy distribution with pdf

$$f(u) = \frac{2}{\pi(1 + u^2)} \times \mathbf{1}_{\{u>0\}}. \tag{5}$$

For all the above models (a-d), we have $\lambda_{p_1,j}^{(1)} \to 0$ in probability as $p_1 \to \infty$. For (a-c), $\mathbf{E}[\sum_j \lambda_{p_1,j}^{(1)}] \to 2$ as $p_1 \to \infty$ (the expectation is infinite for example (d)), as in the iid Gaussian case. However, the infinite-width limits are all very different.

Under the inverse gamma model (a), the infinite-width limit is the same as the iid Gaussian case. Under models (b-d), the infinite-width limit is a mixture of Gaussian processes, i.e. a Gaussian process with a random covariance kernel. These models illustrate some of the benefits of the MoGP regime.

The outputs are now dependent in the infinite-width limit. The models (b-d) are compressible in the sense that the difference between the output of the pruned network and the output of the unpruned network vanishes in the infinite-width limit. This is not the case for the iid Gaussian model, nor for model (a). The weights as well as the outputs can have an exponential tail (b-c) or a power-law tail (d). The properties of the different models are summarised in Table 1.

3.1 Examples from related works

More general deep neural networks and other examples are considered in the full paper [Lee et al., 2022]. Let us briefly point out some examples of distributions for the random variance $\lambda_{p_l,j}^{(l)}$ that have already appeared in the literature. Examples include the Bernoulli [Jantre et al., 2021], the horseshoe [Louizos et al., 2017, Ghosh et al., 2018, 2019, Popkes et al., 2019], the gamma [Scardapane et al., 2017, Wang et al., 2017], the inverse gamma [Ober and Aitchison, 2021], and the improper Jeffrey distributions [Louizos et al., 2017]. The review paper [Fortuin, 2021] also discusses several examples in the context of a Bayesian framework for neural networks.

Bayesian priors are also related to non-Bayesian estimators based on regularised empirical risk minimisation, where the estimator can be interpreted as a maximum a posteriori estimator under these priors. A typical example is the group lasso penalty on the weights of a neural network, used in a number of articles [Murray and Chiang, 2015, Scardapane et al., 2017, Wang et al., 2017, Ochiai et al., 2017], which can be interpreted as a negative log-prior on the weights when $\lambda_{p_l,j}^{(l)}$ follows a gamma distribution.

Finally, random weights of the form in Equation (1) have been used to initialise weights in stochastic gradient descent algorithms, departing from the standard iid Gaussian initialisation commonly used for training deep neural networks [Glorot and Bengio, 2010]. Blier et al. [2019] use per-node random learning rates in stochastic gradient descent. This is equivalent to using the prior in Equation (1) at initialisation, and then learning $V_{jk}^{(l+1)}$ while keeping the variances fixed after initialisation. A similar approach was considered by Wolinski et al. [2020], using deterministic variances.

References

Léonard Blier, Pierre Wolinski, and Yann Ollivier. Learning with random learning rates. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'19), pages 449–464, 2019.

D. Bracale, S. Favaro, S. Fortini, and S. Peluchetti. Large-width functional asymptotics for deep Gaussian neural networks. arXiv preprint arXiv:2102.10307, 2021.

R. Der and D. Lee. Beyond Gaussian processes: On the distributions of infinite networks. In Proceedings of the 20th Conference on Neural Information Processing Systems (NeurIPS'06), pages 275–282, 2006.

S. Favaro, S. Fortini, and S. Peluchetti. Stable behaviour of infinitely wide deep neural networks. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS'20), pages 1137–1146, 2020.

V. Fortuin. Priors in Bayesian deep learning: A review. arXiv preprint arXiv:2105.06868, 2021.

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

S. Ghosh, J. Yao, and F. Doshi-Velez. Structured variational learning of Bayesian neural networks with horseshoe priors. In Proceedings of the 35th International Conference on Machine Learning (ICML'18), pages 1744–1753, 2018.

S. Ghosh, J. Yao, and F. Doshi-Velez. Model selection in Bayesian neural networks via horseshoe priors. Journal of Machine Learning Research, 20(182):1–46, 2019.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS'10), pages 249–256, 2010.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV'15), pages 1026–1034, 2015.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS'18), pages 8571–8580, 2018.

S. Jantre, S. Bhattacharya, and T. Maiti. Layer adaptive node selection in Bayesian neural networks: Statistical guarantees and implementation details. arXiv preprint arXiv:2108.11000, 2021.

P. Jung, H. Lee, J. Lee, and H. Yang. α-stable convergence of heavy-tailed infinitely-wide neural networks. Advances in Applied Probability, 2023.

H. Lee, F. Ayed, P. Jung, J. Lee, H. Yang, and F. Caron. Deep neural networks with dependent weights: Gaussian process mixture limit, heavy tails, sparsity and compressibility. arXiv preprint arXiv:2205.08187, 2022.

J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. In Proceedings of the 6th International Conference on Learning Representations (ICLR'18), 2018.

C. Louizos, K. Ullrich, and M. Welling. Bayesian compression for deep learning. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS'17), pages 3288–3298, 2017.

C. H. Martin and M. W. Mahoney. Traditional and heavy-tailed self regularization in neural network models. arXiv preprint arXiv:1901.08276, 2019.

A. G de G Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR'18), 2018.

K. Murray and D. Chiang. Auto-sizing neural networks: With applications to n-gram language models. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15), pages 908–916, 2015.

R. M. Neal. Bayesian Learning for Neural Networks. PhD thesis, Department of Computer Science, University of Toronto, 1995.

Radford M. Neal. Priors for infinite networks. In Bayesian Learning for Neural Networks, pages 29–53. Springer New York, 1996.

S. W. Ober and L. Aitchison. Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In Proceedings of the 38th International Conference on Machine Learning (ICML'21), pages 8248–8259, 2021.

T. Ochiai, S. Matsuda, H. Watanabe, and S. Katagiri. Automatic node selection for deep neural networks using group lasso regularization. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), pages 5485–5489. IEEE, 2017.

A.-L. Popkes, H. Overweg, A. Ercole, Y. Li, J. M. Hernández-Lobato, Y. Zaykov, and C. Zhang. Interpretable outcome prediction with sparse Bayesian neural networks in intensive care. arXiv preprint arXiv:1905.02599, 2019.

S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini. Group sparse regularization for deep neural networks. Neurocomputing, 241:81–89, 2017.

J. Wang, C. Xu, X. Yang, and J. M. Zurada. A novel pruning algorithm for smoothing feedforward neural networks based on group lasso method. IEEE Transactions on Neural Networks and Learning Systems, 29(5):2012–2024, 2017.

Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? In Proceedings of the 37th International Conference on Machine Learning (ICML'20), pages 10248–10259, 2020.

P. Wolinski, G. Charpiat, and Y. Ollivier. Asymmetrical scaling layers for stable network pruning. 2020.

Greg Yang. Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS'19), pages 9947–9960, 2019.