Dependence between Bayesian neural network units

Mariia Vladimirova* **Julyan Arbel** **Stéphane Girard**
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Abstract

The connection between Bayesian neural networks and Gaussian processes gained a lot of attention in the last few years, with the flagship result that hidden units converge to a Gaussian process limit when the layers width tends to infinity. Underpinning this result is the fact that hidden units become independent in the infinite-width limit. Our aim is to shed some light on hidden units dependence properties in practical finite-width Bayesian neural networks. In addition to theoretical results, we assess empirically the depth and width impacts on hidden units dependence properties.

1 Introduction

Pre-activations and post-activations of layer ℓ in Bayesian neural networks are respectively defined as

$$\boldsymbol{g}^{(\ell)} = \boldsymbol{W}^{(\ell)^\mathsf{T}} \boldsymbol{h}^{(\ell-1)}, \quad \boldsymbol{h}^{(\ell)} = \phi(\boldsymbol{g}^{(\ell)}), \tag{1}$$

where $\boldsymbol{W}^{(\ell)} \in \mathbb{R}^{H_{\ell-1}} \times \mathbb{R}^{H_\ell}$ are weights that follow some prior distribution, ϕ is a nonlinear function called activation function, $\boldsymbol{g}^{(\ell)} \in \mathbb{R}^{H_\ell}$ is a vector of pre-activations, and $\boldsymbol{h}^{(\ell)} \in \mathbb{R}^{H_\ell}$ is a vector of post-activations. For $\ell = 0$, $\boldsymbol{h}^{(0)}$ is an input vector of deterministic numerical object features. For $\ell > 0$, H_ℓ is the width of layer ℓ. When we talk about both $\boldsymbol{g}^{(\ell)}$ or $\boldsymbol{h}^{(\ell)}$ or when we do not need to specify if we consider pre-activations or post-activations, we refer to units of layer ℓ. The distributions induced on units are priors in functional space, or induced priors, also called prior predictives in the literature.

Induced priors in Bayesian neural networks with Gaussian weights become Gaussian processes when the number of hidden units per layer tends to infinity (Neal, 1996; Matthews et al., 2018; Lee et al., 2018; Garriga-Alonso et al., 2019). Stable distributions also lead to stable processes which are generalizations of Gaussian ones (Favaro et al., 2020). Tightening hidden units closer to the Gaussian process can be considered as reducing the induced dependence between units. Since it is not the case for finite-width neural networks, dealing with the induced dependence is one of the problems in describing the prior predictive.

In this note, we focus on dependence properties that help in better characterizing hidden unit priors. We study dependence properties between hidden units in Bayesian neural networks and establish analytically, in Section 2, and empirically, as illustrated on Figure 1, positive and negative dependence induced by weight priors.

2 Dependence properties

We start by showing that hidden units of the same layer are uncorrelated for uncorrelated weights. This theorem refines the non-negative covariance theorem from Vladimirova et al. (2019), the proof is deferred to Appendix A.

*Corresponding author: mariia.vladimirova@inria.fr.

Bayesian Deep Learning workshop, NeurIPS 2021.

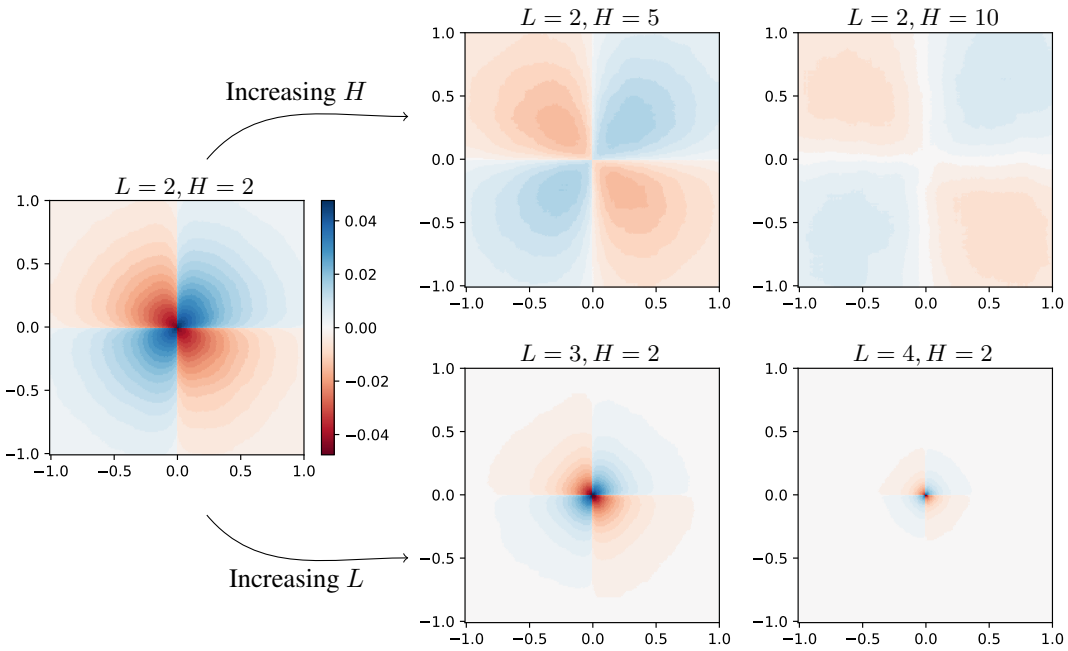

Figure 1: Influence of neural network width H (first row) and depth L (second row) on the units dependence measured through $\Delta^{(L)}(z_1, z_2)$, defined in Equation (2).

Theorem 2.1 (Covariance between hidden units)**.** *Consider a Bayesian neural network as described in Equation (1) with ReLU activation function. Assume that weights $w^{(\ell)}$ are centered and independent from units $h^{(\ell-1)}$. If weights are uncorrelated, then any pre-activations of the same layer ℓ are uncorrelated.*

Let $g^{(\ell)}$ and $\tilde{g}^{(\ell)}$ be two distinct pre-nonlinearities of layer ℓ, and define

$$\Delta^{(\ell)}(z_1, z_2) := \mathbb{P}(g^{(\ell)} \geq z_1, \tilde{g}^{(\ell)} \geq z_2) - \mathbb{P}(g^{(\ell)} \geq z_1)\mathbb{P}(\tilde{g}^{(\ell)} \geq z_2). \tag{2}$$

The following theorem represents how the sign of $\Delta^{(\ell)}(z_1, z_2)$ depends on signs of z_1 and z_2. Usually the weights in Bayesian neural networks are assumed to be independent (Neal, 1996; Matthews et al., 2018; Lee et al., 2018; Garriga-Alonso et al., 2019). However, some works (Garriga-Alonso and van der Wilk, 2021; Fortuin et al., 2021) proposed correlated priors for convolutional neural networks since trained weights are empirically strongly correlated. They showed that these correlated priors can improve overall performance. Our results take into account Bayesian neural networks with possibly dependent priors. More precisely, for ℓ-th layer pre-activations $g_j = \sum_{i=1}^{H_{\ell-1}} W_{ij} h_i$, $j \in \{1, \ldots, H_\ell\}$, weights $W_{i_1 j}$ and $W_{i_2 j}$ can be dependent for distinct $i_1, i_2 \in \{1, \ldots, H_{\ell-1}\}$, while W_{ij_1} and W_{ij_2} are independent for any distinct $j_1, j_2 \in \{1, \ldots, H_\ell\}$. By applying Lemma A.2 and A.3 from Appendix, we have the relationship between Δ and values of z_1, z_2:

Theorem 2.2 (Hidden units dependence)**.** *Consider a Bayesian neural network as described in Equation (1) with some activation function ϕ. Let elements of weight vector $\boldsymbol{W}^{(\ell)}$ follow some zero-center elliptical (possibly different and possibly dependent) distributions, and weight vectors be independent for distinct units of the following layer. If $\ell = 1$, then $\Delta^{(\ell)}(z_1, z_2) = 0$ for all z_1, z_2. If $\ell \geq 2$, then $\Delta^{(\ell)}(z_1, z_2) \geq 0$ if $z_1 z_2 \geq 0$, and $\Delta^{(\ell)}(z_1, z_2) \leq 0$ otherwise. If (and only if) the activation function ϕ satisfies $\mathbb{P}\left(\phi(\boldsymbol{g}^{(\ell)}) = 0\right) = 0$ for any $\boldsymbol{g}^{(\ell)}$, then $\Delta^{(\ell)}(0, z_2) = 0$ and $\Delta^{(\ell)}(z_1, 0) = 0$ for any z_1, z_2.*

The case of $\mathbb{P}\left(\phi(\boldsymbol{g}^{(\ell)}) = 0\right) = 0$ corresponds to post-activations without critical mass at zero. They can be obtained after applying activation functions such as identity, sigmoid, ELU, and others, but not ReLU.

Remark 2.1. *Due to the ellipticity and zero-centering of distributions, the statement of Theorem 2.2 is also true for $\tilde{\Delta}^{(\ell)}(z_1, z_2) := \mathbb{P}(g^{(\ell)} \leq z_1, \tilde{g}^{(\ell)} \leq z_2) - \mathbb{P}(g^{(\ell)} \leq z_1)\mathbb{P}(\tilde{g}^{(\ell)} \leq z_2)$.*

2.1 Corollaries

Dependence measures and properties are interrelated. Widely used measures such as Kendall's tau and Spearman's rho (Nelsen, 2007), take into account the concordance. Based on Theorem 2.2, we establish that for hidden units these coefficients are equal to zero.

Corollary 2.1. *In Bayesian neural networks under assumptions of Theorem 2.2, Kendall's tau and Spearman's rho computed for hidden units are equal to zero.*

The following dependence condition of hidden units is defined by Vladimirova et al. (2021) in order to establish the Weibull-tail property of hidden units. Random variables $X_1, \dots, X_N$ satisfy the *positive dependence (PD) condition* if the following inequalities hold for all $z \in \mathbb{R}$ and some constant $C > 0$:

$$\mathbb{P}\left(X_1 \geq 0, \dots, X_{N-1} \geq 0 | X_N \geq z\right) \geq C \quad \text{(right tail)},$$
$$\mathbb{P}\left(X_1 \leq 0, \dots, X_{N-1} \leq 0 | X_N \leq z\right) \geq C \quad \text{(left tail)}.$$

The proof of Theorem 2.2 can be adapted to prove the following property for hidden units, originally proved in Vladimirova et al. (2021).

Corollary 2.2 (Vladimirova et al., 2021)**.** *Let $X_1, \dots, X_N$ be some possibly dependent random variables and $W_1, \dots, W_N$ be symmetric, mutually independent and independent from $X_1, \dots, X_N$, then random variables $X_1 W_1, \dots, X_N W_N$ satisfy the PD condition.*

3 Experiments

We have built neural networks of $L = 2, 3, 4$ hidden layers, with $H = 2, 5, 10$ hidden units on each layer. We used a fixed input $\mathbf{x}$ of size 10^4, which can be thought of as an image of dimension 100×100. This input was sampled once for all with standard Gaussian entries. In order to obtain samples from the prior distribution of the neural network units, we have sampled the weights from independent centered Gaussians from which units were obtained by forward evaluation with the ReLU non-linearity. This process was iterated $n = 10^5$ times. We propagated the priors and calculated values of $\Delta^{(L)}$, defined in Equation (2), for z_1, z_2 on a grid $(-1.0, 1.0) \times (-1.0, 1.0)$. The results are illustrated on Figure 1. All subplots are appeared to be divided into four quadrants of negative and positive values, confirming Theorem 2.2: $\Delta^{(L)}$ is positive when z_1 and z_2 are of the same sign, and $\Delta^{(L)}$ is negative otherwise.

The increase of the number of hidden units H leads to less dependence between hidden units as the obtained values of $\Delta^{(L)}$ are smaller. Moreover, the center of the plot takes values closer to zero than the corners. The $\Delta^{(L)}$ values are more spread out and less peaked. The increase of the depth L leads to the opposite result when the corners become closer to zero than the center while the $\Delta^{(L)}$ values become more peaked around zero.

4 Discussion

We described analytically and empirically the dependence between hidden units in Bayesian neural networks. We proved that Kendall's tau and Spearman's rho are equal to zero. These results help to understand better the influence of changing the width and depth in Bayesian neural networks.

Representation learning. Aitchison (2020) studied the prior over representations in finite and infinite Bayesian neural networks. The narrower, deeper networks offer more flexibility because the covariance of the outputs gradually disappears as network size increases. The results are obtained by considering the variability in the top-layer kernel induced by the prior over a finite neural network. Our empirical results show that such deep narrow neural networks keep hidden units highly dependent in the center. Therefore, there might be a connection between the prior over representations and highly-peaked dependence between units.

Width-depth trade-off. From a deep Gaussian process perspective, Pleiss and Cunningham (2021) argue that width becomes harmful to model fit and performance as the posterior becomes less data-dependent with width. Empirically, there is a sweet spot in width for convolutional neural networks, depending on the dataset. The increase of width beyond this sweet spot degrades the

performance. The tail analysis demonstrates that width and depth have opposite effects: depth accentuates a model's non-Gaussianity, while width makes models increasingly Gaussian. Indeed, it was proved that Bayesian neural network units are heavier-tailed with depth (Vladimirova et al., 2019; Zavatone-Veth and Pehlevan, 2021; Noci et al., 2021; Vladimirova et al., 2021). So the increase of width might make the resulting units distributions more Gaussian in the center.

References

Aitchison, L. (2020). Why bigger is not always better: on finite and infinite neural networks. In *International Conference on Machine Learning*, pages 156–164.

Cambanis, S., Huang, S., and Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385.

Favaro, S., Fortini, S., and Stefano, P. (2020). Stable behaviour of infinitely wide deep neural networks. In *International Conference on Artificial Intelligence and Statistics*.

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

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. (2019). Deep convolutional networks as shallow Gaussian processes. In *International Conference on Learning Representations*.

Garriga-Alonso, A. and van der Wilk, M. (2021). Correlated weights in infinite limits of deep convolutional neural networks. *arXiv preprint arXiv:2101.04097*.

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

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

Neal, R. M. (1996). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Noci, L., Bachmann, G., Roth, K., Nowozin, S., and Hofmann, T. (2021). Precise characterization of the prior predictive distribution of deep ReLU networks. *arXiv preprint arXiv:2106.06615*.

Pleiss, G. and Cunningham, J. P. (2021). The limitations of large width in neural networks: A deep gaussian process perspective. *arXiv preprint arXiv:2106.06529*.

Vladimirova, M., Arbel, J., and Girard, S. (2021). Bayesian neural network unit priors and generalized Weibull-tail property. In *Asian Conference on Machine Learning*.

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

Zavatone-Veth, J. A. and Pehlevan, C. (2021). Exact priors of finite neural networks. *arXiv preprint arXiv:2104.11734*.

A Bayesian neural network properties

A.1 Covariance

Further, we provide the proof of the following theorem that refines the non-negative covariance theorem from Vladimirova et al. (2019).

Theorem 2.1. *Consider a Bayesian neural network as described in Equation (1) with ReLU activation function. Assume that weights $w^{(\ell)}$ are centered and independent from units $h^{(\ell-1)}$. If weights are uncorrelated, then any pre-activations of the same layer ℓ are uncorrelated.*

Proof. Consider first hidden layer distinct pre-activations $g^{(1)} = \boldsymbol{W}^{(1)^\mathsf{T}} \boldsymbol{h}^{(0)}$ and $\tilde{g}^{(1)} = \tilde{\boldsymbol{W}}^{(1)^\mathsf{T}} \boldsymbol{h}^{(0)}$ as described in Equation (1). Since $h^{(0)}$ is a deterministic vector, the covariance between pre-activations is of the same sign as the covariance between the weights:

$$\text{Cov}\left[\boldsymbol{W}^{(1)^\mathsf{T}} \boldsymbol{h}^{(0)}, \tilde{\boldsymbol{W}}^{(1)^\mathsf{T}} \boldsymbol{h}^{(0)}\right] = \sum_{i=1}^{H_1} \sum_{j=1}^{H_1} \left(\mathbb{E}\left[W_i^{(1)} \tilde{W}_j^{(1)}\right] - \mathbb{E}\left[W_i^{(1)}\right] \mathbb{E}\left[\tilde{W}_j^{(1)}\right]\right) h_i^{(0)} h_j^{(0)}.$$

If the weights are uncorrelated, then the units are uncorrelated, therefore, $\text{Cov}\left[g^{(1)}, \tilde{g}^{(1)}\right] = 0$.

Consider the case where $\ell \geq 2$. Let $\boldsymbol{X} \in \mathbb{R}^{H_{\ell-1}}$ be outputs of hidden layer $\ell - 1$, $\boldsymbol{W} \in \mathbb{R}^{H_{\ell-1}}$ be weights that follow some prior distribution, $\tilde{\boldsymbol{W}} \in \mathbb{R}^{H_{\ell-1}}$ be an independent copy of $\boldsymbol{W}$. Two distinct units of layer ℓ can be written as $g^{(\ell)} = \boldsymbol{W}^\mathsf{T} \boldsymbol{X}$ and $\tilde{g}^{(\ell)} = \tilde{\boldsymbol{W}}^\mathsf{T} \boldsymbol{X}$. Then, the covariance between pre-activations $g^{(\ell)}$ and $\tilde{g}^{(\ell)}$ can be expressed as

$$\text{Cov}\left[g^{(\ell)}, \tilde{g}^{(\ell)}\right] = \sum_{i=1}^{H_{\ell-1}} \sum_{j=1}^{H_{\ell-1}} \left(\mathbb{E}[W_i \tilde{W}_j] \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[W_i\right] \mathbb{E}[\tilde{W}_j] \mathbb{E}\left[X_i\right] \mathbb{E}\left[X_j\right]\right).$$

Since the weights are uncorrelated, we have

$$\text{Cov}\left[g^{(\ell)}, \tilde{g}^{(\ell)}\right] = \sum_{i=1}^{H_{\ell-1}} \sum_{j=1}^{H_{\ell-1}} \mathbb{E}\left[W_i\right] \mathbb{E}[\tilde{W}_j] \left(\mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right] \mathbb{E}\left[X_j\right]\right).$$

If $\mathbb{E}[W_i] = 0$ for all $i = 1, \ldots, H_{\ell-1}$, then $\text{Cov}\left[g^{(\ell)}, \tilde{g}^{(\ell)}\right] = 0$.

□

A.2 Dependence

We provide an auxiliary lemma that we will further use for the dependence theorem proof.

Lemma A.1. *Let Y be a random variable on $\mathbb{R}$ and $\xi_1, \xi_2 : \mathbb{R} \to \mathbb{R}$ be monotonic functions. Then $\text{Cov}(\xi_1(Y), \xi_2(Y)) \geq 0$ if ξ_1 and ξ_2 have the same monotonicity (are both non-increasing or both non-decreasing), and $\text{Cov}(\xi_1(Y), \xi_2(Y)) \leq 0$ otherwise.*

Proof. Let Y_1 be an independent copy of Y. Let us consider the following expectation:

$$\mathbb{E}\left[\left(\xi_1(Y) - \xi_1(Y_1)\right)\left(\xi_2(Y) - \xi_2(Y_1)\right)\right] = \\ \mathbb{E}\left[\xi_1(Y)\xi_2(Y)\right] - \mathbb{E}\left[\xi_1(Y)\xi_2(Y_1)\right] - \mathbb{E}\left[\xi_1(Y_1)\xi_2(Y)\right] + \mathbb{E}\left[\xi_1(Y_1)\xi_2(Y_1)\right].$$

The independence of Y and Y_1 yields $\mathbb{E}\left[\xi_1(Y)\xi_2(Y_1)\right] = \mathbb{E}\left[\xi_1(Y)\right] \mathbb{E}\left[\xi_2(Y_1)\right]$. Since Y and Y_1 are identically distributed, then we get

$$\mathbb{E}\left[\left(\xi_1(Y) - \xi_1(Y_1)\right)\left(\xi_2(Y) - \xi_2(Y_1)\right)\right] = 2\text{Cov}\left[\xi_1(Y), \xi_2(Y)\right].$$

If ξ_1 and ξ_2 are both increasing or both decreasing, then, for all $x, y \in \mathbb{R}$,

$$(\xi_1(x) - \xi_1(y))(\xi_2(x) - \xi_2(y)) \geq 0.$$

Otherwise, for all $x, y \in \mathbb{R}$, we have

$$(\xi_1(x) - \xi_1(y))(\xi_2(x) - \xi_2(y)) \leq 0.$$

Taking the expectation leads to the conclusion. □

Lemma A.2. *Consider a Bayesian neural network as described in Equation (1) with some activation function. Let elements of weight vector $\boldsymbol{W}^{(\ell)}$ follow some zero-center elliptical (possibly different and possibly dependent) distributions, and weight vectors be independent for distinct units of the following layer. If $\ell = 1$, then $\Delta^{(\ell)}(z_1, z_2) = 0$ for all z_1, z_2. If $\ell \geq 2$, then $\Delta^{(\ell)}(z_1, z_2) \geq 0$ if $z_1 z_2 \geq 0$, and $\Delta^{(\ell)}(z_1, z_2) \leq 0$ otherwise.*

Proof. The case where $\ell = 1$ trivially holds as pre-activations are independent for independent weights.

Consider the case where $\ell \geq 2$. Let $\boldsymbol{X} \in \mathbb{R}^{H_{\ell-1}}$ be outputs of hidden layer $\ell - 1$, $\boldsymbol{W} \in \mathbb{R}^{H_{\ell-1}}$ be weights that follow some prior distribution, and $\tilde{\boldsymbol{W}} \in \mathbb{R}^{H_{\ell-1}}$ be an independent copy of $\boldsymbol{W}$. Since $g^{(\ell)}$ and $\tilde{g}^{(\ell)}$ are two distinct units of layer ℓ, they can be written as $g^{(\ell)} = \boldsymbol{W}^\intercal \boldsymbol{X}$ and $\tilde{g}^{(\ell)} = \tilde{\boldsymbol{W}}^\intercal \boldsymbol{X}$, then

$$\begin{aligned}
\mathbb{P}\left(g^{(\ell)} \geq z_1, \tilde{g}^{(\ell)} \geq z_2\right) &= \mathbb{P}\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1, \tilde{\boldsymbol{W}}^\intercal \boldsymbol{X} \geq z_2\right) \\
&= \mathbb{E}\left[\mathbb{I}\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1, \tilde{\boldsymbol{W}}^\intercal \boldsymbol{X} \geq z_2\right)\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_W\left[\mathbb{I}\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1, \tilde{\boldsymbol{W}}^\intercal \boldsymbol{X} \geq z_2\right)\right] \middle| \boldsymbol{X}\right] \\
&= \mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1, \tilde{\boldsymbol{W}}^\intercal \boldsymbol{X} \geq z_2 \middle| \boldsymbol{X}\right)\right].
\end{aligned}$$

Since the weights $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}}$ of different hidden units are independent, pre-activations are independent conditionally on $\boldsymbol{X}$. Therefore, we can express the conditional joint probability as a product of conditional probabilities:

$$\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1, \tilde{\boldsymbol{W}}^\intercal \boldsymbol{X} \geq z_2 \middle| \boldsymbol{X}\right) = \mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1 \middle| \boldsymbol{X}\right) \mathbb{P}_W\left(\tilde{\boldsymbol{W}}^\intercal \boldsymbol{X} \geq z_2 \middle| \boldsymbol{X}\right).$$

Weights $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}}$ are identically distributed, so the conditional probabilities differ only by the lower bound values z_1 and z_2. Therefore, we get

$$\mathbb{P}\left(g^{(\ell)} \geq z_1, \tilde{g}^{(\ell)} \geq z_2\right) = \mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1 \middle| \boldsymbol{X}\right) \mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_2 \middle| \boldsymbol{X}\right)\right]. \tag{3}$$

Now consider the product of probabilities

$$\begin{aligned}
\mathbb{P}\left(g^{(\ell)} \geq z_1\right) \mathbb{P}\left(\tilde{g}^{(\ell)} \geq z_2\right) &= \mathbb{P}\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1\right) \mathbb{P}\left(\tilde{\boldsymbol{W}}^\intercal \boldsymbol{X} \geq z_2\right) \\
&= \mathbb{P}\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1\right) \mathbb{P}\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_2\right) \\
&= \mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1 \middle| \boldsymbol{X}\right)\right] \mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_2 \middle| \boldsymbol{X}\right)\right]. \qquad (4)
\end{aligned}$$

Then, by combining Equations (3) and (4), at the ℓ-th layer we get

$$\Delta(z_1, z_2) = \mathrm{Cov}\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_1 \middle| \boldsymbol{X}\right), \mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z_2 \middle| \boldsymbol{X}\right)\right].$$

Since $\boldsymbol{W}$ follows a centered elliptical distribution, then for some positive-definite matrix Σ and some scalar function ψ the density function has the form $f(\boldsymbol{w}) = \psi(\boldsymbol{w}^\intercal \Sigma^{-1} \boldsymbol{w})$ (Cambanis et al., 1981).

Consider the case when $z \neq 0$. From ellipticity we have

$$\begin{aligned}
\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z \middle| \boldsymbol{X}\right) &= \int \mathbb{I}\left[\boldsymbol{w}^\intercal \boldsymbol{X} \geq z\right] \psi\left(\boldsymbol{w}^\intercal \Sigma^{-1} \boldsymbol{w}\right) \mathrm{d}\boldsymbol{w} \\
&= \int \mathbb{I}\left[\boldsymbol{w}^\intercal \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_\Sigma} \geq \frac{z}{\|\boldsymbol{X}\|_\Sigma}\right] \psi\left(\boldsymbol{w}^\intercal \Sigma^{-1} \boldsymbol{w}\right) \mathrm{d}\boldsymbol{w}.
\end{aligned}$$

Introduce the change of variables $\boldsymbol{v} = Q_{\boldsymbol{X}}^\intercal \Sigma^{-1/2} \boldsymbol{w}$ for some rotation (orthogonal) matrix $Q_{\boldsymbol{X}}$ (which satisfies $Q_{\boldsymbol{X}}^{-1} = Q_{\boldsymbol{X}}^\intercal$) such that $Q_{\boldsymbol{X}}^{-1} \Sigma^{1/2} \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_\Sigma}$ equals the first basis vector $\boldsymbol{e}_1$. Since $\det(Q_{\boldsymbol{X}}) = 1$ is independent of $\boldsymbol{X}$, this shows that

$$\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z \middle| \boldsymbol{X}\right) = \int \mathbb{I}\left[\boldsymbol{v}^\intercal \boldsymbol{e}_1 \geq \frac{z}{\|\boldsymbol{X}\|_\Sigma}\right] \psi\left(\boldsymbol{v}^\intercal \boldsymbol{v}\right) \det(\Sigma^{1/2}) \mathrm{d}\boldsymbol{v},$$

thus establishing that function $\boldsymbol{X} \mapsto \mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z \middle| \boldsymbol{X}\right)$ is actually a function of $Y = \|\boldsymbol{X}\|_\Sigma$, a one-dimensional random variable, i.e. for $Y > 0$ and for some function ξ_z, $\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z \middle| \boldsymbol{X}\right) = \xi_z(Y)$.

Determine $\xi_z(0) = \mathbb{I}[z \leq 0]$. Then,

$$\Delta^{(\ell)}(z_1, z_2) = \mathrm{Cov}\left[\xi_{z_1}(Y), \xi_{z_2}(Y)\right].$$

If $z > 0$, then $Y \to \xi_z(Y)$ is non-decreasing as $\mathbb{I}[z \leq 0] = 0 \leq \Psi\left(\frac{z}{Y}\right)$. Similarly, if $z < 0$, then $Y \to \xi_z(Y)$ is non-increasing.

Therefore, ξ_{z_1} and ξ_{z_2} have the same monotonicity if z_1 and z_2 are of the same sign. According to Lemma A.1, in this case $\Delta(z_1, z_2) = \mathrm{Cov}(\xi_{z_1}(Y), \xi_{z_2}(Y)) \geq 0$. If z_1 and z_2 are of different signs, then $\Delta(z_1, z_2) = \mathrm{Cov}(\xi_{z_1}(Y), \xi_{z_2}(Y)) \leq 0$.

If $z = 0$, then $\xi_0(Y) \leq 1$ for $Y > 0$ and $\xi_0(0) = 1$. Thus, since at the smallest value the function has the maximum, $\xi_0(Y)$ is non-increasing, and Lemma A.1 can also be applied to the case when z_1 or z_2 is zero.

□

Lemma A.3. *Consider a Bayesian neural network as described in Equation (1) with some activation function ϕ. Let elements of weight vector $\boldsymbol{W}^{(\ell)}$ follow some zero-center elliptical (possibly different and possibly dependent) distributions, and weight vectors be independent for distinct units of the following layer. The activation function satisfies $\mathbb{P}\left(\phi(\boldsymbol{g}^{(\ell)}) = 0\right) = 0$ at layer ℓ iff $\Delta^{(\ell)}(0, z) = 0$ for any z.*

Proof. With previous notations, we set $z_1 = 0$ and $z_2 = z$. Note that we could invert the roles of z_1 and z_2 without loss of generality.

Let $\mathbb{P}(\boldsymbol{X} = 0) = p$, then $\mathbb{P}(\boldsymbol{X} \neq 0) = 1 - p$. Notice that $\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z \middle| \boldsymbol{X} = 0\right) = \mathbb{I}[z \leq 0]$, and, in particular, if $z = 0$, $\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq 0 \middle| \boldsymbol{X} = 0\right) = 1$. Moreover, $\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq 0 \middle| \boldsymbol{X} \neq 0\right) = 1/2$ due to ellipticity.

Therefore, for the case when $z = 0$ we have

$$\begin{aligned}
\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq 0 \middle| \boldsymbol{X}\right)\right] &= \mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq 0 \middle| \boldsymbol{X} = 0\right)\right] p \\
&\quad + \mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq 0 \middle| \boldsymbol{X} \neq 0\right)\right] (1 - p) \\
&= p + \frac{1 - p}{2} = \frac{p + 1}{2},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_X\left[\mathbb{P}_W^2\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq 0 \middle| \boldsymbol{X}\right)\right] &= \mathbb{E}_X\left[\mathbb{P}_W^2\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq 0 \middle| \boldsymbol{X} = 0\right)\right] p \\
&\quad + \mathbb{E}_X\left[\mathbb{P}_W^2\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq 0 \middle| \boldsymbol{X} \neq 0\right)\right] (1 - p) \\
&= p + \frac{1 - p}{4} = \frac{3p + 1}{4}.
\end{aligned}$$

Thus, we get

$$\Delta^{(\ell)}(0, 0) = \frac{3p + 1}{4} - \frac{(p + 1)^2}{4} = \frac{p(1 - p)}{4} \geq 0.$$

We see that $\Delta^{(\ell)}(0, 0) = 0$ iff $p = \mathbb{P}(\boldsymbol{X} = 0) = 0$ or $1 - p = \mathbb{P}(\boldsymbol{X} \neq 0) = 0$.

Now let us consider more general case, where $z \neq 0$:

$$\begin{aligned}
\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z \middle| \boldsymbol{X}\right)\right] &= \mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z \middle| \boldsymbol{X} = 0\right)\right] p \\
&\quad + \mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z \middle| \boldsymbol{X} \neq 0\right)\right] (1 - p) \\
&= p\,\mathbb{I}[z \leq 0] + (1 - p)\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\intercal \boldsymbol{X} \geq z \middle| \boldsymbol{X} \neq 0\right)\right],
\end{aligned}$$

$$
\begin{aligned}
&\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq 0\middle|\boldsymbol{X}\right)\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}\right)\right]\\
&\quad=\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq 0\middle|\boldsymbol{X}=0\right)\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}=0\right)\right]p\\
&\quad+\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq 0\middle|\boldsymbol{X}\neq 0\right)\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}\neq 0\right)\right](1-p)\\
&\quad=p\,\mathbb{I}[z\leq 0]+\frac{1-p}{2}\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}\neq 0\right)\right].
\end{aligned}
$$

Further,

$$
\begin{aligned}
\Delta^{(\ell)}(0,z)&=p\,\mathbb{I}[z\leq 0]+\frac{1-p}{2}\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}\neq 0\right)\right]\\
&\quad-\frac{p+1}{2}\left(p\,\mathbb{I}[z\leq 0]+(1-p)\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}\neq 0\right)\right]\right)\\
&=\frac{p(1-p)}{2}\,\mathbb{I}[z\leq 0]-\frac{p(1-p)}{2}\,\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}\neq 0\right)\right].
\end{aligned}
$$

If $z>0$, then $\Delta^{(\ell)}(0,z)=-\frac{p(1-p)}{2}\,\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}\neq 0\right)\right]\leq 0$.

If $z<0$, then $\Delta^{(\ell)}(0,z)=\frac{p(1-p)}{2}\left(1-\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}\neq 0\right)\right]\right)\geq 0$, as $\mathbb{E}_X\left[\mathbb{P}_W\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{X}\geq z\middle|\boldsymbol{X}\right)\right]\leq 1$.

Notice that $\Delta^{(\ell)}(0,z)=0$ iff $p=0$ or $1-p=0$ for any z.

The case when $p=1$ means that $\boldsymbol{X}=\varphi\left(\boldsymbol{g}^{(\ell)}\right)=0$ for any $\boldsymbol{g}^{(\ell)}$. It cannot be the case for an activation function, thus, we get the statement of the lemma. □

A.3 Corollaries

Corollary 2.1 requires a generalization of Theorem 2.2 to sums and differences of pre-activations. Let Δ_2^+ and Δ_2^- be defined as

$$
\Delta_2^+(z_1,z_2):=\mathbb{P}(g_1^{(\ell)}+g_2^{(\ell)}\geq z_1,\tilde{g}_1^{(\ell)}+\tilde{g}_2^{(\ell)}\geq z_2)-\mathbb{P}(g_1^{(\ell)}+g_2^{(\ell)}\geq z_1)\mathbb{P}(\tilde{g}_2^{(\ell)}+\tilde{g}_2^{(\ell)}\geq z_2), \tag{5}
$$

$$
\Delta_2^-(z_1,z_2):=\mathbb{P}(g_1^{(\ell)}-g_2^{(\ell)}\geq z_1,\tilde{g}_1^{(\ell)}-\tilde{g}_2^{(\ell)}\geq z_2)-\mathbb{P}(g_1^{(\ell)}-g_2^{(\ell)}\geq z_1)\mathbb{P}(\tilde{g}_2^{(\ell)}-\tilde{g}_2^{(\ell)}\geq z_2), \tag{6}
$$

where $g_1^{(\ell)}$, $g_2^{(\ell)}$ are independent copies of hidden unit $g^{(\ell)}$, and $\tilde{g}_1^{(\ell)}$, $\tilde{g}_2^{(\ell)}$ are independent copies of hidden unit $\tilde{g}^{(\ell)}$.

Theorem A.1. *Under the assumptions of Theorem 2.2, the same result holds for Δ_2^+ and Δ_2^-.*

Proof. We say $g^{(\ell)}=\boldsymbol{W}^\mathsf{T}\boldsymbol{X}$, where $\boldsymbol{X}\in\mathbb{R}^{H_{\ell-1}}$ be outputs of hidden layer $\ell-1$, $\boldsymbol{W}\in\mathbb{R}^{H_{\ell-1}}$ be weights that follow some prior distribution, independent of $\boldsymbol{X}$. Similarly, $\tilde{g}^{(\ell)}=\tilde{\boldsymbol{W}}^\mathsf{T}\boldsymbol{X}$, where $\tilde{\boldsymbol{W}}\in\mathbb{R}^{H_{\ell-1}}$ be independent copy of $\boldsymbol{W}$. We can express the joint probability in Δ_2^+ as

$$
\mathbb{P}\left(g_1^{(\ell)}+g_2^{(\ell)}\geq z_1,\tilde{g}_1^{(\ell)}+\tilde{g}_2^{(\ell)}\geq z_2\right)=\mathbb{P}\left(\boldsymbol{W}_1^\mathsf{T}\boldsymbol{X}_1+\boldsymbol{W}_2^\mathsf{T}\boldsymbol{X}_2\geq z_1,\tilde{\boldsymbol{W}}_1^\mathsf{T}\boldsymbol{X}_1+\tilde{\boldsymbol{W}}_2^\mathsf{T}\boldsymbol{X}_2\geq z_2\right).
$$

Following the proof of Theorem 2.2, we have

$$
\begin{aligned}
&\mathbb{P}\left(g_1^{(\ell)}+g_2^{(\ell)}\geq z_1,\tilde{g}_1^{(\ell)}+\tilde{g}_2^{(\ell)}\geq z_2\right)\\
&\qquad=\mathbb{P}\left(\boldsymbol{W}_1^\mathsf{T}\boldsymbol{x}_1+\boldsymbol{W}_2^\mathsf{T}\boldsymbol{x}_2\geq z_1,\tilde{\boldsymbol{W}}_1^\mathsf{T}\boldsymbol{x}_1+\tilde{\boldsymbol{W}}_2^\mathsf{T}\boldsymbol{x}_2\geq z_2\middle|\boldsymbol{X}_1=\boldsymbol{x}_1,\boldsymbol{X}_2=\boldsymbol{x}_2\right).
\end{aligned}
$$

Let us denote $\boldsymbol{W}_0=[\boldsymbol{W}_1,\boldsymbol{W}_2]\in\mathbb{R}^{2H_{\ell-1}}$, $\tilde{\boldsymbol{W}}_0=[\tilde{\boldsymbol{W}}_1,\tilde{\boldsymbol{W}}_2]\in\mathbb{R}^{2H_{\ell-1}}$, $\boldsymbol{X}_0=[\boldsymbol{X}_1,\boldsymbol{X}_2]\in\mathbb{R}^{2H_{\ell-1}}$, and $\boldsymbol{x}_0=[\boldsymbol{x}_1,\boldsymbol{x}_2]\in\mathbb{R}^{2H_{\ell-1}}$. We obtain $\boldsymbol{W}_0$ and $\tilde{\boldsymbol{W}}_0$ are vectors of elliptical distributions independent of $\boldsymbol{X}_0$. Now, we can rewrite

$$
\begin{aligned}
\mathbb{P}\left(g_1^{(\ell)}+g_2^{(\ell)}\geq z_1,\tilde{g}_1^{(\ell)}+\tilde{g}_2^{(\ell)}\geq z_2\right)&=\mathbb{P}\left(\boldsymbol{W}_0^\mathsf{T}\boldsymbol{x}_0\geq z_1,\tilde{\boldsymbol{W}}_0^\mathsf{T}\boldsymbol{x}_0\geq z_2\middle|\boldsymbol{X}_0=\boldsymbol{x}_0\right)\\
&=\mathbb{P}\left(\boldsymbol{W}_0^\mathsf{T}\boldsymbol{x}_0\geq z_1\middle|\boldsymbol{X}_0=\boldsymbol{x}_0\right)\mathbb{P}\left(\tilde{\boldsymbol{W}}_0^\mathsf{T}\boldsymbol{x}_0\geq z_2\middle|\boldsymbol{X}_0=\boldsymbol{x}_0\right).
\end{aligned}
$$

The same way we get an equation for a product of probabilities

$$\mathbb{P}\left(g_1^{(\ell)} + g_2^{(\ell)} \geq z_1\right) \mathbb{P}\left(\tilde{g}_1^{(\ell)} + \tilde{g}_2^{(\ell)} \geq z_2\right) = \mathbb{P}\left(\boldsymbol{W}_0^{\intercal}\boldsymbol{x}_0 \geq z_1\right) \mathbb{P}\left(\tilde{\boldsymbol{W}}_0^{\intercal}\boldsymbol{x}_0 \geq z_2\right).$$

The rest of the proof is exactly the same as in Theorem 2.2.

Notice that if $\boldsymbol{W}$ is elliptical, then $-\boldsymbol{W}$ is elliptical. Then, for the case of Δ_2^-, we denote $\boldsymbol{W}_0 = [\boldsymbol{W}_1, -\boldsymbol{W}_2]$ and $\tilde{\boldsymbol{W}}_0 = [\tilde{\boldsymbol{W}}_1, -\tilde{\boldsymbol{W}}_2]$, which are also elliptical vectors independent of $\boldsymbol{X}_0$. Similarly as for Δ_2^+, we obtain the statement for Δ_2^-. □

Corollary 2.1. *In Bayesian neural networks under assumptions of Theorem 2.2, Kendall's tau and Spearman's rho computed for hidden units are equal to zero.*

Proof. Consider random variables (X, Y) with some joint distribution. Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed random copies of (X, Y). From Nelsen (2007), Kendall's tau τ can be expressed as

$$\tau = \tau_{X,Y} = \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right].$$

Let X and Y be different hidden units from Bayesian neural networks satisfying the assumptions in the statement.

Notice that $\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_2)Y > 0\right] = \mathbb{P}\left[X_1 - X_2 > 0, Y_1 - Y_2 > 0\right] + \mathbb{P}\left[X_1 - X_2 < 0, Y_1 - Y_2 < 0\right]$. From Theorem A.1, we have $\Delta(0, 0) = 0$, so $\mathbb{P}\left[X_1 - X_2 > 0, Y_1 - Y_2 > 0\right] = \mathbb{P}\left[X_1 - X_2 > 0\right]\mathbb{P}\left[Y_1 - Y_2 > 0\right]$. Since X_1 and X_2 are independent copies of X, $\mathbb{P}\left[X_1 > X_2\right] = 1/2$ Similarly, combining Theorem A.1 with Remark 2.1, $\mathbb{P}\left[X_1 - X_2 < 0, Y_1 - Y_2 < 0\right] = \mathbb{P}\left[X_1 - X_2 < 0\right]\mathbb{P}\left[Y_1 - Y_2 < 0\right]$ and $\mathbb{P}\left[X_1 < X_2\right] = 1/2$. Therefore, $\tau = 0$.

Spearman's rho ρ is defined as

$$\rho = \rho_{X,Y} = 3\left(\mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] - \mathbb{P}\left[(X_1 - X_2)(Y_1 - Y_3) < 0\right]\right),$$

where (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) are independent and identically distributed random copies of (X, Y) (Nelsen, 2007). The proof for ρ is identical. □