Understanding Priors in Bayesian Neural Networks at the Unit Level — Supplementary Material —

Mariia Vladimirova [1 2] **Jakob Verbeek** [1] **Pablo Mesejo** [3] **Julyan Arbel** [1]

In this supplementary material we provide a number of additional technical results that are referred to in the main paper.

Lemma 3.1 proof.

Proof. According to asymptotic equivalence definition there must exist positive constants d and D such that for all $k \in \mathbb{N}$ it holds

$$d \leq \|\phi(X)\|_k / \|X\|_k \leq D. \tag{1}$$

The extended envelope property upper bound and the triangle inequality for norms imply the right-hand side of (1), since

$$\|\phi(X)\|_k \leq \|c_2 + d_2|u|\|_k \leq c_2 + d_2\|u\|_k.$$

Assume that $|\phi(u)| \geq c_1 + d_1|u|$ for $u \in \mathbb{R}_+$. Consider the lower bound of the nonlinearity moments

$$\|\phi(X)\|_k \geq \|d_1 u_+\|_k + c_1 + \|\phi(u_-)\|_k,$$

where $\{u_- : u \in \mathbb{R}_-\}$ and $\{u_+ : u \in \mathbb{R}_+\}$. For negative u_- there are constants $c_1 \geq 0$ and $d_1 > 0$ such that $c_1 - d_1 u > |\phi(u)|$, or $c_1 > |\phi(u)| + d_1 u$:

$$\|\phi(X)\|_k > \|d_1 u_+\|_k + \||\phi(u_-)| + d_1 u_-\|_k \geq d_1\|u\|_k.$$

It yields asymptotic equivalence as in eq.(5) of the main paper. □

Proposition 3.1 proof.

Proof. Let $X_i \sim \mathbf{subW}(\theta)$ for $1 \leq i \leq N$ be units from one region where pooling operation is applied. Using Definition 3.3, for all $x \geq 0$ and some constant $K > 0$ we have

$$\mathbb{P}(|X_i| \geq x) \leq \exp\left(-x^{1/\theta}/K\right) \text{ for all } i.$$

Max pooling operation takes the maximum element in the region. Since X_i, $1 \leq i \leq N$ are the elements in one region, we want to check if the tail of $\max_{1\leq i \leq N} X_i$ obeys sub-Weibull property with optimal tail parameter is equal to θ. Since max pooling operation can be decomposed into linear and ReLU operations, which does not harm the distribution tail (Lemma 3.1), it leads to the proposition statement first part.

Summation and division by a constant does not influence the distribution tail, yielding the proposition result regarding the averaging operation.

□

Lemma 0.1 (Gaussian moments)**.** *Let X be a normal random variable such that $X \sim \mathcal{N}(0, \sigma^2)$, then the following asymptotic equivalence holds*

$$\|X\|_k \asymp \sqrt{k}.$$

Proof. The moments of central normal absolute random variable $|X|$ are equal to

$$\begin{aligned}\mathbb{E}[|X|^k] &= \int_{\mathbb{R}} |x|^k \, p(x)\, dx \\ &= 2\int_0^\infty x^k \, p(x)\, dx \\ &= \frac{1}{\sqrt{\pi}} \sigma^k 2^{k/2} \Gamma\left(\frac{k+1}{2}\right). \end{aligned} \tag{2}$$

By the Stirling approximation of the Gamma function:

$$\Gamma(z) = \sqrt{\frac{2\pi}{z}} \left(\frac{z}{\mathrm{e}}\right)^z \left(1 + O\left(\frac{1}{z}\right)\right). \tag{3}$$

Substituting (3) into the central normal absolute moment (2), we obtain

$$\begin{aligned}\mathbb{E}[|X|^k] &= \frac{\sigma^k 2^{k/2}}{\sqrt{\pi}} \sqrt{\frac{4\pi}{k+1}} \left(\frac{k+1}{2\mathrm{e}}\right)^{\frac{k+1}{2}} \left(1 + O\left(\frac{1}{k}\right)\right) \\ &= \frac{2\sigma^k}{\sqrt{2\mathrm{e}}} \left(\frac{k+1}{\mathrm{e}}\right)^{k/2} \left(1 + O\left(\frac{1}{k}\right)\right). \end{aligned}$$

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France [2]Moscow Institute of Physics and Technology, 141701 Dolgoprudny, Russia [3]Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071 Granada, Spain. Correspondence to: Mariia Vladimirova <mariia.vladimirova@inria.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

Then the roots of absolute moments can be written in the form of

$$\begin{aligned}\|X\|_k &= \frac{\sigma}{\mathrm{e}^{1/(2k)}}\sqrt{\frac{k+1}{\mathrm{e}}}\left(1+O\left(\frac{1}{k}\right)\right)^{1/k} \\ &= \frac{\sigma}{\sqrt{\mathrm{e}}}\frac{\sqrt{k+1}}{\mathrm{e}^{1/(2k)}}\left(1+O\left(\frac{1}{k^2}\right)\right) \\ &= \frac{\sigma}{\sqrt{\mathrm{e}}}c_k\sqrt{k+1}.\end{aligned}$$

Here the coefficient c_k denotes

$$c_k = \frac{1}{\mathrm{e}^{1/(2k)}}\left(1+O\left(\frac{1}{k^2}\right)\right) \to 1,$$

with $k \to \infty$. Thus, asymptotic equivalence holds

$$\|X\|_k \asymp \sqrt{k+1} \asymp \sqrt{k}.$$

□

Lemma 0.2 (Multiplication moments)**.** *Let W and X be independent random variables such that $W \sim \mathcal{N}(0,\sigma^2)$ and for some $p > 0$ it holds*

$$\|X\|_k \asymp k^p. \tag{4}$$

Let W_i be independent copies of W, and X_i be copies of X, $i = 1,\dots,H$ with non-negative covariance between moments of copies

$$Cov\left[X_i^s, X_j^t\right] \geq 0, \quad \text{for } i \neq j,\ s,t \in \mathbb{N}. \tag{5}$$

Then we have the following asymptotic equivalence

$$\Big\|\sum_{i=1}^{H} W_i X_i\Big\|_k \asymp k^{p+1/2}. \tag{6}$$

Proof. Let us proof the statement, using mathematical induction.

Base case: show that the statement is true for $H = 1$. For independent variables W and X, we have

$$\begin{aligned}\|WX\|_k &= \left(\mathbb{E}[|WX|^k]\right)^{1/k} = \left(\mathbb{E}[|W|^k]\mathbb{E}[|X|^k]\right)^{1/k} \\ &= \|W\|_k\|X\|_k.\end{aligned} \tag{7}$$

Since the random variable W follows Gaussian distribution, then Lemma 0.1 implies

$$\|W\|_k \asymp \sqrt{k}. \tag{8}$$

Substituting assumption (4) and weight norm asymptotic equivalence (8) into (7) leads to the desired asymptotic equivalence (6) in case of $H = 1$.

Inductive step: show that if for $H = n-1$ the statement holds, then for $H = n$ it also holds.

Suppose for $H = n-1$ we have

$$\Big\|\sum_{i=1}^{n-1} W_i X_i\Big\|_k \asymp k^{p+1/2}. \tag{9}$$

Then, according to the covariance assumption (5), for $H = n$ we get

$$\Big\|\sum_{i=1}^{n} W_i X_i\Big\|_k^k = \Big\|\sum_{i=1}^{n-1} W_i X_i + W_n X_n\Big\|_k^k \tag{10}$$

$$\geq \sum_{j=0}^{k} C_k^j \Big\|\sum_{i=1}^{n-1} W_i X_i\Big\|_j^j \Big\|W_n X_n\Big\|_{k-j}^{k-j}. \tag{11}$$

Using the equivalence definition (Def. 3.2), from the induction assumption (9) for all $j = 0,\dots,k$ there exists absolute constant $d_1 > 0$ such that

$$\Big\|\sum_{i=1}^{n-1} W_i X_i\Big\|_j^j \geq \left(d_1\, j^{p+1/2}\right)^j. \tag{12}$$

Recalling previous equivalence results in the base case, there exists constant $m_2 > 0$ such that

$$\Big\|W_n X_n\Big\|_{k-j}^{k-j} \geq \left(d_2\,(k-j)^{p+1/2}\right)^{k-j}. \tag{13}$$

Substitute obtained bounds (12) and (13) into equation (10) with denoted $d = \min\{d_1, d_2\}$, obtain

$$\begin{aligned}\Big\|\sum_{i=1}^{n} W_i X_i\Big\|_k^k &\geq d^k \sum_{j=0}^{k} C_k^j \left[j^j\,(k-j)^{k-j}\right]^{p+1/2} \\ &= d^k\, k^{k(p+1/2)} \sum_{j=0}^{k} C_k^j \left[\left(\frac{j}{k}\right)^j \left(1-\frac{j}{k}\right)^{k-j}\right]^{p+1/2}.\end{aligned} \tag{14}$$

Notice the lower bound of the following expression

$$\begin{aligned}&\sum_{j=0}^{k} C_k^j \left[\left(\frac{j}{k}\right)^j \left(1-\frac{j}{k}\right)^{k-j}\right]^{p+1/2} \\ &\geq \sum_{j=0}^{k} \left[\left(\frac{j}{k}\right)^j \left(1-\frac{j}{k}\right)^{k-j}\right]^{p+1/2} \geq 2.\end{aligned} \tag{15}$$

Substituting found lower bound (15) into (14), get

$$\Big\|\sum_{i=1}^{n} W_i X_i\Big\|_k^k \geq 2\,d^k\,k^{k(p+1/2)} > d^k\,k^{k(p+1/2)}. \tag{16}$$

Now prove the upper bound. For random variables Y and Z the Holder's inequality holds

$$\begin{aligned}\|YZ\|_1 = \mathbb{E}\left[|YZ|\right] &\le \left(\mathbb{E}\left[|Y|^2\right]\mathbb{E}\left[|Z|^2\right]\right)^{1/2} \\ &= \|YZ\|_2\|YZ\|_2.\end{aligned}$$

Holder's inequality leads to the inequality for L^k norm

$$\|YX\|_k^k \le \|Y\|_{2k}^k \|Z\|_{2k}^k. \tag{17}$$

Obtain the upper bound of $\|\sum_{i=1}^n W_i X_i\|_k^k$ from the norm property (17) for the random variables $Y = \left(\sum_{i=1}^{n-1} W_i X_i\right)^{k-j}$ and $Z = (W_n X_n)^j$

$$\Big\|\sum_{i=1}^{n} W_i X_i\Big\|_k^k = \Big\|\sum_{i=1}^{n-1} W_i X_i + W_n X_n\Big\|_k^k \tag{18}$$

$$\le \sum_{j=0}^{k} C_k^j \Big\|\sum_{i=1}^{n-1} W_i X_i\Big\|_{2j}^j \Big\|W_n X_n\Big\|_{2(k-j)}^{k-j}. \tag{19}$$

From the induction assumption (9) for all $j = 0, \dots, k$ there exists absolute constant $D_1 > 0$ such that

$$\Big\|\sum_{i=1}^{n-1} W_i X_i\Big\|_{2j}^j \le \Big(D_1\,(2j)^{p+1/2}\Big)^j. \tag{20}$$

Recalling previous equivalence results in the base case, there exists constant $D_2 > 0$ such that

$$\Big\|W_n X_n\Big\|_{2(k-j)}^{k-j} \le \Big(D_2\,\big(2(k-j)\big)^{p+1/2}\Big)^{k-j}. \tag{21}$$

Substitute obtained bounds (20) and (21) into equation (18) with denoted $D = \max\{D_1, D_2\}$, obtain

$$\Big\|\sum_{i=1}^{n} W_i X_i\Big\|_k^k \le D^k \sum_{j=0}^{k} C_k^j \Big[(2j)^j \big(2(k-j)\big)^{k-j}\Big]^{p+1/2}.$$

Find an upper bound for $\Big[\left(1-\frac{j}{k}\right)^{k-j}\left(\frac{j}{k}\right)^j\Big]^{p+1/2}$. Since expressions $\left(1-\frac{j}{k}\right)$ and $\left(\frac{j}{k}\right)$ are less than 1, then $\Big[\left(1-\frac{j}{k}\right)^{k-j}\left(\frac{j}{k}\right)^j\Big]^{p+1/2} < 1$ holds for all natural numbers $p > 0$. For the sum of binomial coefficients it holds the inequality $\sum_{j=0}^k C_k^j < 2^k$. So the final upper bound is

$$\Big\|\sum_{i=1}^{n} W_i X_i\Big\|_k^k \le 2^k\, D^k\, (2k)^{k(p+1/2)}. \tag{22}$$

Hence, taking the k-th root of (16) and (22), we have upper and lower bounds which imply the equivalence for $H = n$ and the truth of inductive step

$$d' k^{p+1/2} \le \Big\|\sum_{i=1}^{n} W_i X_i\Big\|_k \le D' k^{p+1/2},$$

where $d' = d$ and $D' = 2^{p+3/2} D$. Since both the base case and the inductive step have been performed, by mathematical induction the equivalence holds for all $H \in \mathbb{N}$

$$\Big\|\sum_{i=1}^{H} W_i X_i\Big\|_k \asymp k^{p+1/2}.$$

□