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
\[ d \leq \|\phi(X)\|_k/\|X\|_k \leq D. \] (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:
\[ \|\phi(X)\|_k > \|d_1 u_+\|_k + \|\phi(u_-)\|_k + d_1 u_- \geq d_1 \|u\|_k. \]
It yields asymptotic equivalence as in eq.(5) of the main paper.

Proposition 3.1 proof.

Proof. Let $X_i \sim \text{subW}(\theta)$ 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
\[ \mathbb{P}(|X_i| \geq x) \leq \exp \left( -x^{1/\sigma}/K \right) \] 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 $\theta$. 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
\[ \|X\|_k \approx \sqrt{k}. \]

Proof. The moments of central normal absolute random variable $|X|$ are equal to
\[ \mathbb{E}[|X|^k] = \int_{\mathbb{R}} |x|^k \mathbb{P}(x) \, dx \]
\[ = 2 \int_{0}^{\infty} x^k \mathbb{P}(x) \, dx \]
\[ = \frac{1}{\sqrt{\pi}} \sigma^{k+2/2} \Gamma \left( \frac{k+1}{2} \right). \] (2)
By the Stirling approximation of the Gamma function:
\[ \Gamma(z) \approx \sqrt{2\pi} \left( \frac{z}{e} \right)^z \left( 1 + O \left( \frac{1}{z} \right) \right). \] (3)
Substituting (3) into the central normal absolute moment (2), we obtain
\[ \mathbb{E}[|X|^k] = \frac{\sigma^{k+2/2}}{\sqrt{\pi}} \sqrt{\frac{4\pi}{k+1}} \left( \frac{k+1}{2e} \right)^k \left( 1 + O \left( \frac{1}{k} \right) \right) \]
\[ = 2\sigma^{k} \left( \frac{k+1}{e} \right)^k \left( 1 + O \left( \frac{1}{k} \right) \right). \]
Then we have the following asymptotic equivalence
\[ \|X\|_k \approx \frac{\sigma}{\sqrt{\varepsilon}} \sqrt{k+1} \left( 1 + O \left( \frac{1}{k^2} \right) \right)^{1/k} \]
\[ = \frac{\sqrt{k+1}}{\sqrt{\varepsilon}} \left( 1 + O \left( \frac{1}{k^2} \right) \right) \]
\[ = \frac{\sqrt{\varepsilon}}{\sqrt{c_k}} \sqrt{k+1}. \]

Here the coefficient \( c_k \) denotes
\[ c_k = \frac{1}{\sqrt{\varepsilon}} \left( 1 + O \left( \frac{1}{k^2} \right) \right) \rightarrow 1, \]
with \( k \rightarrow \infty \). Thus, asymptotic equivalence holds
\[ \|X\|_k \approx \sqrt{k+1} \approx \sqrt{k}. \]

\( \Box \)

Lemma 0.2 (Multiplication moments). Let \( W \) and \( X \) be independent random variables such that \( W \sim N(0, \sigma^2) \) and for some \( p > 0 \) it holds
\[ \|X\|_k \approx k^p. \] (4)

Let \( W_i \) be independent copies of \( W \), and \( X_i \) be copies of \( X \), \( i = 1, \ldots, H \) with non-negative covariance between moments of copies
\[ Cov \left[ X_i^s, X_j^t \right] \geq 0, \quad \text{for } i \neq j, s, t \in \mathbb{N}. \] (5)
Then we have the following asymptotic equivalence
\[ \left\| \sum_{i=1}^{H} W_iX_i \right\|_k \approx k^{p+1/2}. \] (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
\[ \|WX\|_k = \left( \mathbb{E}[|WX|^k] \right)^{1/k} \approx \left( \mathbb{E}[|W|^k] \mathbb{E}[|X|^k] \right)^{1/k} \]
\[ = \|W\|_k \|X\|_k. \] (7)
Since the random variable \( W \) follows Gaussian distribution, then Lemma 0.1 implies
\[ \|W\|_k \approx \sqrt{k}. \] (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
\[ \left\| \sum_{i=1}^{n-1} W_iX_i \right\|_k \approx k^{p+1/2}. \] (9)
Then, according to the covariance assumption (5), for \( H = n \) we get
\[ \left\| \sum_{i=1}^{n} W_iX_i \right\|_k \approx \left\| \sum_{i=1}^{n-1} W_iX_i + W_nX_n \right\|_k \]
\[ \geq \sum_{j=0}^{k} C_j^n \left\| \sum_{i=1}^{n-1} W_iX_i \right\|_j \left\| W_nX_n \right\|_{k-j}. \] (10)

Using the equivalence definition (Def. 3.2), from the induction assumption (9) for all \( j = 0, \ldots, k \) there exists absolute constant \( d_1 > 0 \) such that
\[ \left\| \sum_{i=1}^{n-1} W_iX_i \right\|_j \approx \left( d_1 j^{p+1/2} \right)^j. \] (12)
Recalling previous equivalence results in the base case, there exists constant \( m_2 > 0 \) such that
\[ \left\| W_nX_n \right\|_{k-j} \geq \left( d_2 (k-j)^{p+1/2} \right)^{k-j}. \] (13)
Substitute obtained bounds (12) and (13) into equation (10) with denoted \( d = \min\{d_1, d_2\} \), obtain
\[ \left\| \sum_{i=1}^{n} W_iX_i \right\|_k \approx d^k \sum_{j=0}^{k} C_j^n \left[ j^j (k-j)^{k-j} \right]^{p+1/2} \]
\[ = d^k k^{(p+1/2)} \sum_{j=0}^{k} C_j^n \left[ \left( \frac{j}{k} \right)^j \left( 1 - \frac{j}{k} \right)^{k-j} \right]^{p+1/2}. \] (14)

Notice the lower bound of the following expression
\[ \sum_{j=0}^{k} C_j^n \left[ \left( \frac{j}{k} \right)^j \left( 1 - \frac{j}{k} \right)^{k-j} \right]^{p+1/2} \]
\[ \geq \sum_{j=0}^{k} \left[ \left( \frac{j}{k} \right)^j \left( 1 - \frac{j}{k} \right)^{k-j} \right]^{p+1/2} \geq 2. \] (15)
Substituting found lower bound (15) into (14), get
\[ \left\| \sum_{i=1}^{n} W_iX_i \right\|_k \approx 2 d^k k^{(p+1/2)} > d^k k^{(p+1/2)}. \] (16)
Now prove the upper bound. For random variables \( Y \) and \( Z \) the Holder’s inequality holds
\[
\|YZ\|_1 = \mathbb{E}[|YZ|] \leq \left( \mathbb{E}[|Y|^p] \mathbb{E}[|Z|^q] \right)^{1/p} = \|YZ\|_2 \|Y\|_2.
\]

Holder’s inequality leads to the inequality for \( L^k \) norm
\[
\|YX\|_k^k \leq \|Y\|_2^k \|Z\|_2^k.
\] (17)

Obtain the upper bound of \( \|\sum_{i=1}^n W_iX_i\|_k^k \) from the norm property (17) for the random variables \( Y = (n-1)W_iX_i\) and \( Z = (W_nX_n)\)
\[
\left\| \sum_{i=1}^n W_iX_i \right\|_k^k \leq \left\| \sum_{i=1}^{n-1} W_iX_i + W_nX_n \right\|_k^k \leq \sum_{j=0}^{n-1} C_k^j \left\| \sum_{i=1}^{n-1} W_iX_i \right\|_{2j} \left\| W_nX_n \right\|_{2(k-j)}^{k-j}.
\] (19)

From the induction assumption (9) for all \( j = 0, \ldots, k \) there exists absolute constant \( D_1 > 0 \) such that
\[
\left\| \sum_{i=1}^{n-1} W_iX_i \right\|_{2j} \leq \left(D_1 (2j)^{p+1/2}\right)^j.
\] (20)

Recalling previous equivalence results in the base case, there exists constant \( D_2 > 0 \) such that
\[
\left\| W_nX_n \right\|_{2(k-j)}^{k-j} \leq \left(D_2 (2(k-j))^{p+1/2}\right)^{k-j}.
\] (21)

Substitute obtained bounds (20) and (21) into equation (18) with denoted \( D = \max\{D_1, D_2\} \), obtain
\[
\left\| \sum_{i=1}^n W_iX_i \right\|_k^k \leq D^k \sum_{j=0}^{k-1} C_k^j \left(2j\right) (2(k-j))^{k-j}^p+1/2.
\]

Find an upper bound for \( \left(1 - \frac{2}{k}\right)^{k-j} \left(\frac{2}{k}\right)^{k-j} \). Since expressions \( 1 - \frac{2}{k} \) and \( \frac{2}{k} \) are less than 1, then
\[
\left(1 - \frac{2}{k}\right)^{k-j} \left(\frac{2}{k}\right)^{k-j} < 1 \text{ holds for all natural numbers } k > 0.
\]

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
\[
\left\| \sum_{i=1}^n W_iX_i \right\|_k^k \leq 2^k D^k (2k)^{p+1/2}.
\] (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
\[
d'^{k^{p+1/2}} \leq \left\| \sum_{i=1}^n W_iX_i \right\|_k \leq D'^k k^{p+1/2},
\]