

---

# Posterior Concentration for Sparse Deep Learning

---

Nicholas G. Polson and Veronika Ročková  
 Booth School of Business  
 University of Chicago  
 Chicago, IL 60637

## 1 Supplemental Materials

### 1.1 Proof of Theorem 6.1

We prove the theorem by verifying Condition (11) and (12), setting  $\mathcal{F}_n = \mathcal{F}(L^*, \mathbf{p}^*, s^*)$ . First, we need to verify the entropy condition and show that

$$\sup_{\varepsilon > \varepsilon_n} \log \mathcal{E} \left( \frac{\varepsilon}{36}, \{f_{\mathbf{B}}^{DL} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*) : \|f - f_0\|_n < \varepsilon\}, \|\cdot\|_n \right) \leq n \varepsilon_n^2. \quad (1)$$

We can upper-bound the local entropy (1) with the global metric entropy. In addition, because

$$\{f_{\mathbf{B}}^{DL} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*) : \|f\|_\infty \leq \varepsilon\} \subset \{f_{\mathbf{B}}^{DL} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*) : \|f\|_n \leq \varepsilon\},$$

we can upper-bound (1) with

$$\begin{aligned} \log \mathcal{E} \left( \frac{\varepsilon_n}{36}, f_{\mathbf{B}}^{DL} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*), \|\cdot\|_\infty \right) &\leq (s^* + 1) \log \left( \frac{72}{\varepsilon_n} (L^* + 1) (12pN + 1)^{2(L^*+2)} \right) \\ &\lesssim n^{p/(2\alpha+p)} \log(n) \log \left( n / \log^\delta(n) \right) \lesssim n^{p/(2\alpha+p)} \log^2(n) \lesssim n \varepsilon_n^2 \end{aligned}$$

for  $\delta > 1$ , where we used Lemma 10 of Schmidt-Hieber (2017) and the fact that  $s^* \lesssim n^{p/(2\alpha+p)}$  and  $N \asymp n^{p/(2\alpha+p)} / \log(n)$ . This verifies the entropy Condition (11).

Next, we want to show that the prior concentrates enough mass around the truth in the sense that

$$\Pi(f_{\mathbf{B}}^{DL} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*) : \|f_{\mathbf{B}}^{DL} - f_0\|_n \leq \varepsilon_n) \geq e^{-dn \varepsilon_n^2} \quad (2)$$

for some  $d > 2$ . Choosing  $N^* = C_N \lfloor n^{p/(2\alpha+p)} / \log(n) \rfloor$  in Lemma 5.1, there exists a neural network  $\hat{f}_{\hat{\mathbf{B}}} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*)$  consisting of  $\mathbf{p}^*$  nodes aligned in  $L^* \lesssim \log(n)$  layers and indexed by  $\|\hat{\mathbf{B}}\|_0 = s^* \lesssim n^{p/(2\alpha+p)} \log(n)$  nonzero parameters such that

$$\|\hat{f}_{\hat{\mathbf{B}}} - f_0\|_n \leq C_\infty n^{-\alpha/(2\alpha+p)} \log^{\delta\alpha/p}(n) \lesssim \varepsilon_n/2,$$

where the last inequality follows from  $\alpha < p$ , absorbing  $C_\infty$  in the concentration rate. The approximation  $\hat{f}_{\hat{\mathbf{B}}}$  sits on a network architecture characterized by a specific pattern  $\hat{\gamma}$  of nonzero links among  $\hat{\mathbf{B}}$ , i.e.  $\hat{W}_l$  and  $\hat{a}_l$  for  $1 \leq l \leq L + 1$ . We denote by  $\mathcal{F}(\hat{\gamma}, L^*, \mathbf{p}^*, s^*) \subset \mathcal{F}(L^*, \mathbf{p}^*, s^*)$  all the functions supported on this particular architecture. These functions differ only in the size of the  $s^*$  nonzero coefficients among  $\hat{\mathbf{B}}$ , denoted by  $\beta \in \mathbb{R}^{s^*}$ . With  $\hat{\beta}$ , we denote the  $s^*$ -vector associated with the nonzero elements in  $\hat{\mathbf{B}}$ .

Note that there are  $\binom{T}{s^*} \leq (12pN)^{(L^*+1)s^*}$  combinations to pick  $s^*$  the nonzero coefficients and each one, according to prior (9), has an equal prior probability of occurrence  $\frac{1}{\binom{T}{s^*}}$ .

To continue, we note (from the triangle inequality) that

$$\{f_{\mathbf{B}}^{DL} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*) : \|f_{\mathbf{B}}^{DL} - f_0\|_n \leq \varepsilon_n\} \supset \{f_{\mathbf{B}}^{DL} \in \mathcal{F}(\hat{\gamma}) : \|f_{\mathbf{B}}^{DL} - \hat{f}_{\hat{\mathbf{B}}}\|_\infty \leq \varepsilon_n/2\}.$$

Next, we denote with  $\{\beta \in \mathbb{R}^{s^*} : \|\beta\|_\infty \leq 1 \text{ and } \|\beta - \widehat{\beta}\|_\infty \leq \varepsilon_n\}$  the set of coefficients that are at most  $\varepsilon$ -away from the best approximating coefficients  $\beta$  of the neural network  $\widehat{f}_{\widehat{B}} \in \mathcal{F}(\widehat{\gamma}, L^*, \mathbf{p}^*, s^*)$ . From the proof of Lemma 10 of Schmidt-Hieber (2017), it follows that

$$\left\{ f_{\mathbf{B}}^{DL} \in \mathcal{F}(\widehat{\gamma}) : \|f_{\mathbf{B}}^{DL} - \widehat{f}_{\widehat{B}}\|_\infty \leq \frac{\varepsilon_n}{2} \right\} \supset \left\{ \beta \in \mathbb{R}^{s^*} : \|\beta\|_\infty \leq 1 \text{ and } \|\beta - \widehat{\beta}\|_\infty \leq \frac{\varepsilon_n}{2V(L^* + 1)} \right\},$$

where  $V = \prod_{l=0}^{L^*+1} (p_l^* + 1)$ . Now we have all the pieces needed to find a lower bound to the probability in (2). We can write, for some suitably large  $C > 0$ ,

$$\begin{aligned} \Pi(f_{\mathbf{B}}^{DL} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*) : \|f_{\mathbf{B}}^{DL} - f_0\|_n \leq \varepsilon_n) &> \frac{\Pi(f_{\mathbf{B}}^{DL} \in \mathcal{F}(\widehat{\gamma}, L^*, \mathbf{p}^*, s^*) : \|f_{\mathbf{B}} - \widehat{f}_{\widehat{B}}\|_\infty \leq \varepsilon_n/2)}{\binom{T}{s^*}} \\ &> e^{-(L^*+1)s^* \log(12pN^*)} \Pi\left(\beta \in \mathbb{R}^{s^*} : \|\beta\|_\infty \leq 1 \text{ and } \|\beta - \widehat{\beta}\|_\infty \leq \frac{\varepsilon_n}{2V(L^* + 1)}\right). \end{aligned}$$

To continue to lower-bound the expression above, we note that

$$e^{-(L^*+1)s^* \log(12pN^*)} > e^{-C \log^2(n) n^{p/(2\alpha+p)}}$$

for some  $C > 0$ . Under the uniform prior distribution on a cube  $[-1, 1]^{s^*}$  we can write

$$\begin{aligned} \Pi\left(\beta \in \mathbb{R}^{s^*} : \|\beta\|_\infty \leq 1 \text{ and } \|\beta - \widehat{\beta}\|_\infty \leq \frac{\varepsilon_n}{2V(L^* + 1)}\right) &= \left(\frac{\varepsilon_n}{2V(L^* + 1)}\right)^{s^*} \\ &\geq e^{-s^*(L^*+2) \log(12pn/\log^\delta(n))} \geq e^{-D n^{p/(2\alpha+p)} \log^2(n)} \end{aligned}$$

for some  $D > 0$ . We can combine this bound with the preceding expressions to conclude that  $e^{-(C+D) n^{p/(2\alpha+p)} \log^2(n)} \geq e^{-dn\varepsilon_n^2}$  for  $\delta > 1$  and  $d > C + D$ . This concludes the proof of (17).

## 1.2 Proof of Theorem 6.2

First we show that the sieve  $\mathcal{F}_n$  defined in (20) is still reasonably small in the sense that the log covering number can be upper-bounded by a constant multiple of  $n^{p/(2\alpha+p)} \log^{2\delta}(n)$ . It follows from the proof of Theorem 6.1 that the global metric entropy satisfies

$$\begin{aligned} \mathcal{E}\left(\frac{\varepsilon_n}{36}, \mathcal{F}_n, \|\cdot\|_n\right) &\leq \sum_{N=1}^{N_n} \sum_{s=0}^{s_n} e^{(s+1) \log\left(\frac{72}{\varepsilon_n}(L^*+2)(12pN+1)^{2(L^*+2)}\right)} \\ &\lesssim N_n s_n e^{C(L^*+1)(s_n+1) \log(pN_n L^*/\varepsilon_n)} \end{aligned}$$

for some  $C > 0$  and thereby

$$\log \mathcal{E}\left(\frac{\varepsilon_n}{36}, \mathcal{F}_n, \|\cdot\|_n\right) \lesssim \log N_n + \log s_n + n\varepsilon_n^2 \lesssim n\varepsilon_n^2.$$

This verifies Condition (11).

Next, we need to show that the prior charges the sieve in the sense that  $\Pi[\mathcal{F}_n^c] = o(e^{(d+2)n\varepsilon_n^2})$  for some  $d > 2$  (determined below). We have

$$\Pi[\mathcal{F}_n^c] < \Pi(N > N_n) + \Pi(s > s_n).$$

We apply the Chernoff bound to find that

$$\Pi(N > N_n) < e^{-t(N_n+1)} \mathbb{E} e^{tN} \propto e^{-t(N_n+1)} \left(e^{e^t} - 1\right) \quad (3)$$

for any  $t > 0$ . With our choice  $N_n = \lfloor \widetilde{C}_N n^{p/(2\alpha+p)} \log^{2\delta-1} n \rfloor$  and with  $t = \log N_n$  we obtain

$$\Pi(N > N_n) e^{(d+2)n\varepsilon_n^2} \lesssim e^{-(N_n+1) \log N_n + \lambda N_n + (d+2)n\varepsilon_n^2} \rightarrow 0$$

for a large enough constant  $\widetilde{C}_N$ . Next, we find that

$$\Pi(s > s_n) e^{(d+2)n\varepsilon_n^2} \lesssim e^{-C_s(\lfloor L^* N_n \rfloor + 1) + (d+2)n\varepsilon_n^2} \rightarrow 0$$

for some suitably large  $\tilde{C}_N > 0$ . This verifies Condition (13).

Finally, we verify the prior concentration Condition (12). For  $N^* < N_n$  and  $s^* < s_n$  we know from the proof of Theorem 6.1 that

$$\Pi(f_{\mathbf{B}}^{DL} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*) : \|f_{\mathbf{B}}^{DL} - f_0\|_n \leq \varepsilon_n) \geq e^{-D_1 n \varepsilon_n^2}$$

for some  $D_1 > 2$ . Our priors put enough mass at the “right choices”  $(N^*, s^*)$  in the sense that  $\pi(N^*) \gtrsim e^{-N_n \log(N_n/\lambda)} \gtrsim e^{-D n \varepsilon_n^2}$  and  $\pi(s^*) \gtrsim e^{-D n \varepsilon_n^2}$  for some suitable  $D > 0$ . Then we can write

$$\begin{aligned} & \Pi(f_{\mathbf{B}}^{DL} \in \mathcal{F}_n : \|f_{\mathbf{B}}^{DL} - f_0\|_n \leq \varepsilon_n) \\ & \geq \pi(N^*)\pi(s^*)\Pi(f_{\mathbf{B}}^{DL} \in \mathcal{F}(L^*, \mathbf{p}^*, s^*) : \|f_{\mathbf{B}}^{DL} - f_0\|_n \leq \varepsilon_n) \geq e^{-(2D+D_1) n \varepsilon_n^2}. \end{aligned}$$

With these considerations, we conclude the proof of Theorem 6.2.