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 $\mathcal{F}_n = \mathcal{F}(L^*, p^*, s^*)$. First, we need to verify the entropy condition and show that

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

We can upper-bound the local entropy [\(1\)](#page-0-0) with the global metric entropy. In addition, because

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

we can upper-bound [\(1\)](#page-0-0) with

$$
\log \mathcal{E}\left(\frac{\varepsilon_n}{36}, f^{DL}_{\mathbf{B}} \in \mathcal{F}(L^\star, \mathbf{p}^\star, s^\star), \|\.\|_\infty\right) \le (s^\star + 1) \log \left(\frac{72}{\varepsilon_n} (L^\star + 1)(12pN + 1)^{2(L^\star + 2)}\right)
$$

$$
\lesssim n^{p/(2\alpha + p)} \log(n) \log \left(n / \log^{\delta}(n)\right) \lesssim n^{p/(2\alpha + p)} \log^2(n) \lesssim n\varepsilon_n^2
$$

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

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

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

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

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

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

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

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

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

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 ε -away from the best approximating coefficients β of the neural network $f\hat{B} \in \mathcal{F}(\hat{C}, L^* + L^*)$. Expecting the network of $\mathcal{F}(\hat{C}, L^* + L^*)$. $\mathcal{F}(\hat{\gamma}, L^{\star}, p^{\star}, s^{\star})$. From the proof of Lemma 10 of Schmidt-Hieber (2017), it follows that

$$
\begin{aligned} \Big\{f^{DL}_{\pmb{B}} \in \mathcal{F}(\widehat{\gamma}): \|f^{DL}_{\pmb{B}} - \widehat{f}\hat{\mathbf{B}}\|_{\infty} \leq \frac{\varepsilon_n}{2}\Big\} \supset \\ \Big\{\boldsymbol{\beta} \in \mathbb{R}^{s^\star}: \|\boldsymbol{\beta}\|_{\infty} \leq 1 \text{ and } \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_{\infty} \leq \frac{\varepsilon_n}{2V(L^\star+1)}\Big\}\,, \end{aligned}
$$

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 [\(2\)](#page-0-1). We can write, for some suitably large $C > 0$,

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

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

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

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

$$
\Pi\left(\boldsymbol{\beta} \in \mathbb{R}^{s^*} : \|\boldsymbol{\beta}\|_{\infty} \le 1 \text{ and } \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_{\infty} \le \frac{\varepsilon_n}{2V(L^*+1)}\right) = \left(\frac{\varepsilon_n}{2V(L^*+1)}\right)^{s^*}
$$

$$
\ge e^{-s^*(L^*+2)\log(12\,p\,n/\log^{\delta}(n))} \ge e^{-D\,n^{p/(2\alpha+p)}\log^2(n)}
$$

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

1.2 Proof of Theorem [6.2](#page-0-2)

First we show that the sieve \mathcal{F}_n defined in [\(20\)](#page-0-2) 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](#page-0-2) that the global metric entropy satisfies

$$
\mathcal{E}\left(\frac{\varepsilon_n}{36}, \mathcal{F}_n, \|\!.\|\!_n\right) \le \sum_{N=1}^{N_n} \sum_{s=0}^{s_n} e^{(s+1)\log\left(\frac{72}{\varepsilon_n}(L^*+2)(12pN+1)^{2(L^*+2)}\right)}
$$

$$
\lesssim N_n s_n e^{C\left(L^*+1\right)(s_n+1)\log\left(pN_nL^*/\varepsilon_n\right)}
$$

for some $C > 0$ and thereby

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

This verifies Condition [\(11\)](#page-0-2).

Next, we need to show that the prior charges the sieve in the sense that $\Pi[\mathcal{F}_n^c] = o(e^{(d+2)n\epsilon_n^2})$ 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

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

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

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

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

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

for some suitably large $\widetilde{C}_N > 0$. This verifies Condition [\(13\)](#page-0-2).

Finally, we verify the prior concentration Condition [\(12\)](#page-0-2). For $N^* < N_n$ and $s^* < s_n$ we know from the proof of Theorem [6.1](#page-0-2) that

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

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

$$
\Pi(f^{DL}_{\mathbf{B}} \in \mathcal{F}_n : ||f^{DL}_{\mathbf{B}} - f_0||_n \leq \varepsilon_n)
$$

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

With these considerations, we conclude the proof of Theorem [6.2.](#page-0-2)