Uncertainty Quantification for Sparse Deep Learning

Yuexi Wang and Veronika Ročková

Booth School of Business, University of Chicago

6 Supplemental Material

6.1 Rudiments

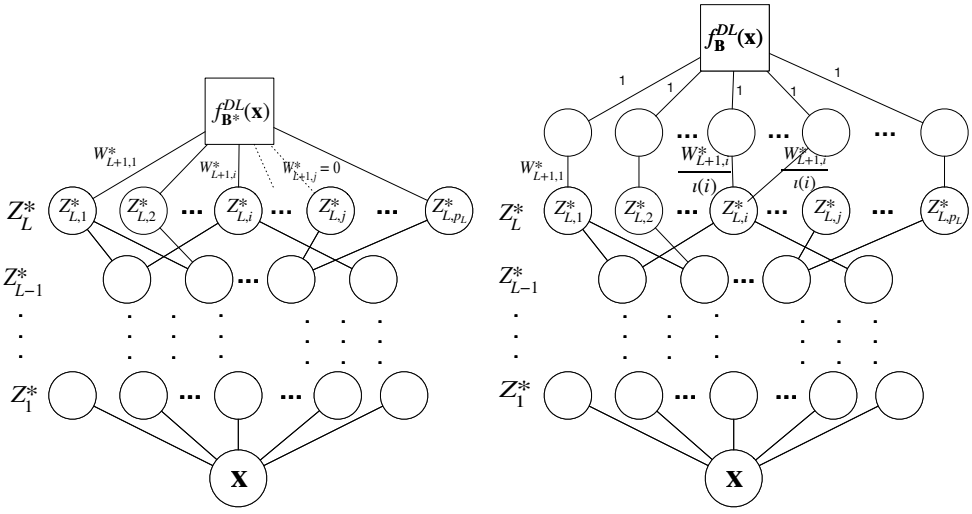
With the prior measure $\Pi(\cdot)$ on $\mathcal{F}(L, \mathbf{p}, s)$, given observed data $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)'$, inference about f_0 is carried out via the posterior distribution

$$\Pi(A|\boldsymbol{Y}^{(n)}, \{\boldsymbol{x}_i\}_{i=1}^n) = \frac{\int_A \prod_{i=1}^n \Pi_f(Y_i|\boldsymbol{x}_i) d\Pi(f)}{\int \prod_{i=1}^n \Pi_f(Y_i|\boldsymbol{x}_i) d\Pi(f)}, \forall A \in \mathcal{B}$$

where $\mathcal{B}$ is a σ-field on $\mathcal{F}(L, \mathbf{p}, s)$ and where $\Pi_f(Y_i|\boldsymbol{x}_i)$ is the likelihood function for the output Y_i under f.

6.2 Posterior Concentration Rate

First, we show that the posterior concentrates at the optimal (near-minimax) rate. We modify the result in Polson and Rockova (2018) to our prior which differs in two aspects: (1) the top layer is fully connected, (2) the top layer coefficients are assigned a Gaussian prior. First, we show that our fully-connected top layer networks can approximate f_0 as well as the networks considered in Polson and Rockova (2018) (i.e. with a sparse top layer). The following Lemma demonstrates how one can construct a fully connected top layer network from any network considered in PR18 so that their outputs are the same. A graphical illustration of this construction can be found in Figure 3.



(a) $f_{\boldsymbol{B}}^{DL}(\boldsymbol{x})$ with sparse top layer (b) $f_{\boldsymbol{B}^*}^{DL}(\boldsymbol{x})$ with fully-connected top layer

Figure 3: Network Construction

Lemma 6.1. *Assume a sparse network $f_{\boldsymbol{B}^*}^{DL} \in \tilde{\mathcal{F}}(L, \mathbf{p}^*, s^*)$ of the form (6) in PR18 with a sparsity pattern γ, where $\tilde{\mathcal{F}}(L, \mathbf{p}^*, s^*)$ is defined in Section 4 of PR18. With $\mathbf{p}^* = (p, p_1^*, \ldots, p_L^*, 1) \in \mathbb{N}^{L+2}$ and $|\gamma| = s^*$, there exists at least one network $f_{\boldsymbol{B}}^{DL} \in \mathcal{F}(L+1, \mathbf{p}, s)$ with $\mathbf{p} = (p, p_1^*, \ldots, p_L^*, p_L^*, 1) \in \mathbb{N}^{L+3}$ and $|\gamma| = s \leq s^* + 2p_L^*$ such that $f_{B^*}^{DL}(\boldsymbol{x}) = f_{\boldsymbol{B}}^{DL}(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^p$.*

Proof. We construct one function $f_{\boldsymbol{B}}^{DL}$ that satisfies the stated conditions. We denote $\boldsymbol{B} = \{(W_l, b_l) : 1 \leq l \leq L+2\}$ such that $\mathbf{p} = (p, p_1^*, \ldots, p_L^*, p_L^*, 1) \in \mathbb{N}^{L+3}$ and choose the same deep coefficients $\{W_l, b_l\} = \{W_l^*, b_l^*\}$ for each $1 \leq l \leq L$. The parameters of the top layer are set as $W_{L+2} = 1'_{p_L^*}$ and $b_{L+2} = b_{L+1}^*$. Choosing the matrix W_{L+1} in a way such that $W'_{L+1} 1_{p_L^*} = W_{L+1}^{*\prime}$ we obtain

$$f_{\boldsymbol{B}}^{DL}(\boldsymbol{x}) = W_{L+2} Z_{L+1} + b_{L+2} = W_{L+2} W_{L+1} Z_L^* + b_{L+1}^* = W_{L+1}^* Z_L^* + b_{L+1}^* = f_{\boldsymbol{B}*}^{DL}(\boldsymbol{x}).$$

The procedure we use to generate W_{L+1} from W_{L+1}^* can be found in Algorithm 1.

Algorithm 1 Network Construction of $\mathcal{F}(L+1, \mathbf{p}, s)$ from $\tilde{\mathcal{F}}(L, \mathbf{p}^*, s^*)$

We assume $W_{L+1,1}^* \neq 0$
Initialize $\{W_l, b_l\}_{l=1}^L = \{W_l^*, b_l^*\}_{l=1}^L, W_{L+1} = 0_{p_L \times p_L}, b_{L+1} = 0, W_{L+2} = \mathbb{I}'_{p_L}, b_{L+2} = b_{L+1}^*$
function $h(j)$ ▷ the index of last connected node (up to j) in layer L+1 of $f_{\boldsymbol{B}^*}^{DL}$
  $h(j) := \max\{k \leq j : W_{L+1,k} \neq 0\}$
function $\iota(j)$ ▷ #nodes in layer L+1 in f_B^{DL} that will be connected to $Z_{L,h(j)}$
  $\iota(j) := \sum_{i=1}^{p_L} \mathbb{I}(h(i) = h(j))$
procedure GENERATE W_{L+1} FROM W_{L+1}^*
  for each $j = 1 : p_L$ **do**
    if $h(j) = j$ **then** ▷ when $Z_{L,j}$ previously connected in $f_{\boldsymbol{B}^*}^{DL}$
      $W_{L+1,i,i} = \frac{W_{L+1,j}^*}{\iota(j)}$ ▷ connect $Z_{L,j}$ to $Z_{L+1,j}$ with the averaged weights
    else ▷ when $Z_{L,j}$ previously unconnected in $f_{\boldsymbol{B}^*}^{DL}$
      $W_{L+1,j,h(j)} = \frac{W_{L+1,h(j)}^*}{\iota(j)}$ ▷ connect $Z_{L,h(j)}$ to $Z_{L+1,j}$ with the averaged weights

It turns out that the sparsity of this extended network satisfies

$$s = s^* + \|W_{L+2}\|_0 + \|W_{L+1}\|_0 - \|W_{L+1}^*\|_0 = s^* + 2p_L^* - \|W_{L+1}^*\|_0 \leq s^* + 2p_L^*. \qquad \square$$

With the construction from Lemma 7.1, our network class could achieve at least the same approximation error as the one in Schmidt-Hieber (2017). To recover the posterior concentration rate results in Theorem 6.1 in PR18, we impose the following conditions on (L, s, N)

$$\begin{cases} L^* \propto \log(n) \\ s^* \lesssim n^{p/(2\alpha+p)} \\ N^* \propto n^{p/(2\alpha+p)}/\log(n) \end{cases} \Rightarrow \begin{cases} L = L^* + 1 \propto \log(n) \\ s \leq s^* + 2p_L^* = s^* + 24pN^* \lesssim n^{p/(2\alpha+p)} + n^{p/(2\alpha+p)} \frac{p}{\log(n)} \lesssim n^{p/(2\alpha+p)} \\ N = N^* \propto n^{p/(2\alpha+p)}/\log(n) \end{cases}$$

The assumptions on the network structure (depth, width and sparsity) maintain very similar for our new prior.

We formally state the posterior concentration result for our prior below.

Theorem 6.1. *Assume $f_0 \in \mathcal{H}_p^\alpha$ where $p = O(1)$ as $n \to \infty$, $\alpha < p$ and $\|f_0\|_\infty \leq F$. Let L, s be as in (14), and $\mathbf{p} = (p, 12pN, \ldots, 12pN, 1)' \in \mathbb{N}^{L+2}$, where $N = C_N \lfloor n^{p/(2\alpha+p)}/\log(n) \rfloor$ for some $C_N > 0$. Under the priors from Section 2.1, the posterior distribution concentrates at the rate $\epsilon_n = n^{-\alpha/(2\alpha+p)} \log^\delta(n)$ for some $\delta > 1$ in the sense that*

$$\Pi(f_{\boldsymbol{B}}^{DL} \in \mathcal{F}(L, p, s) : \|f - f_0\|_n > M_n \epsilon_n \mid Y^{(n)}) \to 0$$

in $\mathbb{P}_0^n$ probability as $n \to \infty$ for any $M_n \to \infty$.

Proof. The statement can be proved as in Rockova and Polson (2018) by verifying the following three conditions (adopted from Ghosal and Van Der Vaart (2007))

$$\sup_{\epsilon>\epsilon_n} \log \mathcal{E}\left(\frac{\epsilon}{36}; A_{\epsilon,1} \cap \mathcal{F}_n; \|\cdot\|_n\right) \leq n\epsilon_n^2 \tag{24}$$

$$\Pi(A_{\epsilon_n,1}) \geq e^{-dn\epsilon_n^2} \tag{25}$$

$$\Pi(\mathcal{F}\backslash\mathcal{F}_n) \leq e^{-(d+2)n\epsilon_n^2} \quad \text{for some } d > 2. \tag{26}$$

We define $\mathcal{F}_n$, for some $C_n = Cn^{p/(2\alpha+p)}\log^{2\delta}(n)$ and $C > 0$, as

$$\mathcal{F}_n = \{f_{\boldsymbol{B}}^{DL} \in \mathcal{F}(L, \mathbf{p}, s) : \|W_{L+1}\|_2^2 + b_{L+1}^2 \leq C_n\}.$$

Here $\mathcal{F}_n \subset \mathcal{F}(L, \mathbf{p}, s)$ is an approximating space (a sieve) consisting of functions whose top layer weights are contained in a ball of radius $\sqrt{C_n}$ in $\mathbb{R}^{p_L+1}$. We show that this sieve contains most of the prior mass as required in (26) for $C > 0$ large enough. Indeed, because $p = \mathcal{O}(1)$ and

$$p_L + 1 = 12\,p\,N + 1 \asymp n^{p/(2\alpha+p)}/\log(n)$$

we have

$$\begin{aligned}\Pi(\mathcal{F}\backslash\mathcal{F}_n) &= \mathbb{P}\left(\|W_{L+1}\|_2^2 + b_{L+1}^2 > C_n\right) \\ &= \mathbb{P}(\chi^2_{p_L+1} > C_n) = \mathbb{P}(e^{\frac{1}{4}\chi^2_{p_L+1}} > e^{\frac{C_n}{4}}) \leq e^{-\frac{C_n}{4}} 2^{(p_L+1)/2} \to 0.\end{aligned}$$

Next, we want to verify the entropy condition (24). Because

$$\{f_{\boldsymbol{B}}^{DL} \in \mathcal{F}_n : \|f\|_\infty \leq \epsilon\} \subset \{f_{\boldsymbol{B}}^{DL} \in \mathcal{F}_n : \|f\|_n \leq \epsilon\}$$

we have

$$\begin{aligned}\sup_{\epsilon>\epsilon_n} \log \mathcal{E}\left(\frac{\epsilon}{36}; f \in \mathcal{F}_n; \|\cdot\|_\infty\right) &\lesssim \log \left\{ \underbrace{\left(\frac{2}{\frac{\epsilon_n/36}{V(L+1)}}\right)^{s-(p_L+1)}}_{(I)} \underbrace{\left(\frac{\sqrt{C_n}}{\frac{\epsilon_n/36}{V(L+1)}}\right)^{p_L+1}}_{(II)} \right\} \\ &\lesssim (s+1)\log\left(\frac{72}{\epsilon_n}(L+1)(12pN+1)^{2(L+1)}\right) + (p_L+1)\log(n^{p/(2\alpha+p)}\log^{2\delta}(n)) \\ &\lesssim n^{p/(2\alpha+p)}\log(n)\log\left(n/\log^\delta(n)\right) + n^{p/(2\alpha+p)}/\log(n)\log\left(n\log(n)\right) \\ &\lesssim n^{p/(2\alpha+p)}\log^2(n) \lesssim n\epsilon_n^2\end{aligned}$$

for some $\delta > 1$, where

$$V = \prod_{l=0}^{L+1}(P_l + 1) \tag{27}$$

and using the fact that $s \lesssim n^{p/(2\alpha+p)}$ and $L \asymp \log(n)$.

The covering number $\mathcal{E}(\frac{\epsilon}{36}; f \in \mathcal{F}_n; \|\cdot\|_\infty)$ consists of two parts. The part (I) stands for the covering number for the deep architecture, while the part (II) is the covering number for the top layer. The calculations of the covering numbers are derived from Lemma 12 of Schmidt-Hieber (2017) which shows

$$\left\|f_{\boldsymbol{B}}^{DL} - f_{\boldsymbol{B}^*}^{DL}\right\|_\infty \leq \|\boldsymbol{B} - \boldsymbol{B}^*\|_\infty V(L+1)$$

with V defined as in (27). To make sure $\left\|f_{\boldsymbol{B}}^{DL} - f_{\boldsymbol{B}^*}^{DL}\right\|_\infty \leq \frac{\epsilon_n}{36}$, we want $\|\boldsymbol{B} - \boldsymbol{B}^*\|_\infty \leq \frac{\epsilon_n/36}{2V(L+1)}$. Since all deep parameters are bounded in absolute value by one, we can discretize the unit cube $[-1,1]^{s-p_L-1}$ with a grid of a diameter $\frac{\epsilon_n/36}{2V(L+1)}$ and obtain the covering number in part (I). For the top layer, the weights and the

bias term are contained inside a (p_L+1)-dimensional ball with a radius $\sqrt{C_n}$. Part(II) for $\|\cdot\|_\infty$ is bounded by the $\frac{\epsilon_n/36}{2V(L+1)}$-covering number of a Euclidean ball of radius $\sqrt{C_n}$ in (p_L+1)-dimensional space (Edmunds and Triebel, 2008).

Last, we need to show that the prior concentrates enough mass around the truth in the sense of (25). From Lemma 7.1 and Lemma 5.1 in PR18, we know that there exists a neural network $\hat{f}_{\hat{\boldsymbol{B}}} \in \mathcal{F}_n(L, \mathbf{p}, s)$, such that

$$\left\|\hat{f}_{\hat{\boldsymbol{B}}} - f_0\right\|_n \leq \epsilon/2.$$

We denote the connectivity pattern of $\hat{f}_{\hat{\boldsymbol{B}}}$ as $\hat{\gamma}$ (with $\hat{s} = |\hat{\gamma}|$) and the corresponding set of coefficients as $\hat{\boldsymbol{B}}$. Following the same arguments as in PR18, we have

$$\{f_{\boldsymbol{B}}^{DL} \in \mathcal{F}_n(L, \mathbf{p}, s) : \left\|f_{\boldsymbol{B}}^{DL} - f_0\right\|_n \leq \epsilon_n\} \supset \{f_{\boldsymbol{B}}^{DL} \in \mathcal{F}_n(L, \mathbf{p}, \hat{\gamma}) : \left\|f_{\boldsymbol{B}}^{DL} - f_0\right\|_n \leq \epsilon_n/2\}.$$

We now denote with $\boldsymbol{\beta} \in \mathbb{R}^T$ and $\hat{\boldsymbol{\beta}} \in \mathbb{R}^T$ the vectorized nonzero coefficients in $\boldsymbol{B}$ and $\hat{\boldsymbol{B}}$ that have the sparsity pattern $\hat{\gamma}$. We use $\gamma(\boldsymbol{\beta})$ to pin down the sparsity pattern of $\boldsymbol{\beta}$. Using Lemma 12 of Schmidt-Hieber (2017) we have

$$\left\{f_{\boldsymbol{B}}^{DL} \in \mathcal{F}_n(L, \mathbf{p}, \hat{\gamma}) : \left\|f_{\boldsymbol{B}}^{DL} - f_0\right\|_n \leq \epsilon_n/2\right\} \supset \left\{\boldsymbol{\beta} \in \mathbb{R}^T : \gamma(\beta) = \hat{\gamma} \text{ and } \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_\infty \leq \frac{\epsilon_n}{2V(L+1)}\right\}. \tag{28}$$

Altogether, we can write

$$\begin{aligned}
&\Pi(f_{\boldsymbol{B}}^{DL} \in \mathcal{F}_n(L, \mathbf{p}, \hat{s}) : \left\|f_{\boldsymbol{B}}^{DL} - f_0\right\|_n \leq \epsilon_n) > \frac{\Pi(f_{\boldsymbol{B}}^{DL} \in \mathcal{F}_n(L, \mathbf{p}, \hat{\gamma}) : \left\|f_{\boldsymbol{B}}^{DL} - f_0\right\|_n \leq \epsilon_n/2)}{\binom{T-p_L-1}{\hat{s}-p_L-1}} \\
&> \frac{1}{\binom{T-p_L-1}{\hat{s}-p_L-1}} \Pi\left(\boldsymbol{\beta} \in \mathbb{R}^T : \gamma(\beta) = \hat{\gamma} \text{ and } \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_\infty \leq \frac{\epsilon_n}{2V(L+1)}\right).
\end{aligned}$$

We note that with $\hat{s} \asymp n^{p/(2\alpha+p)}, L \asymp \log(n)$ and $N \asymp n^{p/(2\alpha+p)}/\log(n)$

$$\frac{1}{\binom{T-p_L-1}{\hat{s}-p_L-1}} \geq e^{-(L+1)\hat{s}\log(12pN)} > e^{-D_1 \log^2(n) n^{p/(2\alpha+p)}}$$

for some $D_1 > 0$. In addition, under the uniform prior on the deep coefficients and the standard normal prior on the top layer, we can write

$$\begin{aligned}
&\Pi\left(\boldsymbol{\beta} \in \mathbb{R}^T : \gamma(\beta) = \hat{\gamma} \text{ and } \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_\infty \leq \frac{\epsilon_n}{2V(L+1)}\right) \geq \left(\frac{\epsilon_n}{2V(L+1)}\right)^{\hat{s}-p_L-1} \prod_{j>T-p_L-1} \Pi\left(\left|\beta_j - \hat{\beta}_j\right| \leq \frac{\epsilon_n}{2V(L+1)}\right) \\
&= \left(\frac{\epsilon_n}{2V(L+1)}\right)^{\hat{s}-p_L-1} \prod_{j>T-p_L-1} \int_{-\frac{\epsilon_n}{2V(L+1)}}^{\frac{\epsilon_n}{2V(L+1)}} d\Pi(\beta_j - \hat{\beta}_j).
\end{aligned} \tag{29}$$

where the last $T-p_L-1$ coefficients in $\boldsymbol{\beta}$ are the top layer weights and bias as shown in (9).

We want to recenter the normal distribution at 0 rather than $\hat{\beta}_j$ by using the following inequality

$$\frac{dN(\hat{\beta}_j, 1)}{dN(0, \frac{1}{2})} = e^{-\frac{1}{2}(\beta_j - \hat{\beta}_j)^2 + \beta_j^2} = e^{\frac{1}{2}(\beta_j + \hat{\beta}_j)^2 - \hat{\beta}_j^2} \geq e^{-\hat{\beta}_j^2}.$$

Then we can continue with the lower bound for (29) as follows

$$\begin{aligned}
(29) &\geq \left(\frac{\epsilon_n}{2V(L+1)}\right)^{\hat{s}-p_L-1} e^{-\sum_{j>T-p_L-1} \hat{\beta}_j^2} \left(\int_{-\frac{\epsilon_n}{2V(L+1)}}^{\frac{\epsilon_n}{2V(L+1)}} dN\left(0, \frac{1}{2}\right)\right)^{p_L+1} \\
&\geq \left(\frac{\epsilon_n}{2V(L+1)}\right)^{\hat{s}-p_L-1} e^{-C_n} \left(e^{-(\frac{\epsilon_n}{2V(L+1)})^2} \frac{\epsilon_n}{\sqrt{\pi}V(L+1)}\right)^{p_L+1} \\
&\geq \left(\frac{2}{\sqrt{2\pi}}\right)^{p_L+1} \left(\frac{\epsilon_n}{2V(L+1)}\right)^{\hat{s}} e^{-C_n} e^{-\frac{(p_L+1)\epsilon_n}{4(12pN+1)^{(L+1)}(L+1)}} \geq e^{-D_2 n^{p/(2\alpha+p)} \log^2(n)}
\end{aligned}$$

for some $D_2 > 0$ and recall that $C_n = Cn^{p/(2\alpha+p)} \log^{2\delta}(n)$.Thus we can combine the bounds and conclude that $e^{-(D_1+D_2)n^{p/(2\alpha+p)}\log^2(n)} \geq e^{-dn\epsilon_n^2}$ for some $\delta > 1$ and $d > D_1 + D_2$. The proof is now complete.

□

It is worth noting that the same concentration rate still holds if we use $N(0,1)$ prior on *all* parameters. We could define

$$\mathcal{F}_n = \{\|\beta\|_2^2 \leq C_n\}.$$

The prior mass condition in (26) is

$$\Pi(\mathcal{F}\backslash\mathcal{F}_n) = \mathbb{P}(\chi_s^2 > C_n) \leq e^{-C_1 n^{p/(2\alpha+p)}\log^{2\delta}(n)}.$$

The entropy condition in (24) is

$$\begin{aligned}
\sup_{\epsilon>\epsilon_n} \log \mathcal{E}(\frac{\epsilon}{36}, f \in \mathcal{F}_n; \|\cdot\|_\infty) &\lesssim \log\left\{\left(\frac{\sqrt{C_n}}{\frac{\epsilon_n/36}{V(L+1)}}\right)^s\right\} \\
&\lesssim (s+1)\log\left(\frac{72}{\epsilon_n}(L+1)(12pN+1)^{2(L+1)}\right) + s\log(Cn^{p/(2\alpha+p)}\log^{2\delta}(n)) \\
&\lesssim n^{p/(2\alpha+p)}\log(n)\log\left(n/\log^\delta(n)\right) + n^{p/(2\alpha+p)}\log(n\log(n)) \\
&\lesssim n\epsilon_n^2
\end{aligned}$$

for some $\delta > 1$, using the fact that $s \lesssim n^{p/(2\alpha+p)}$ and $L \asymp \log(n)$.

The prior concentration condition in (25) can be proved by changing (29) into

$$\begin{aligned}
&\Pi(\beta \in \mathbb{R}^T : \gamma(\beta) = \hat{\gamma}, \sum_j \beta_j^2 \leq C_n \text{ and } \left\|\beta - \hat{\beta}\right\|_\infty \leq \frac{\epsilon_n}{2V(L+1)}) \\
&\geq e^{-\sum_j \hat{\beta}_j^2}\left(\int_{-\frac{\epsilon_n}{2V(L+1)}}^{\frac{\epsilon_n}{2V(L+1)}} dN(0,\frac{1}{2})\right)^{\hat{s}} \\
&\geq e^{-C_n}\left(e^{-(\frac{\epsilon_n}{2V(L+1)})^2}\frac{\epsilon_n}{\sqrt{\pi}V(L+1)}\right)^{\hat{s}} \\
&\geq e^{-C_n}\left(\frac{\epsilon_n}{\sqrt{\pi}V(L+1)}\right)^{\hat{s}} e^{-\frac{\hat{s}\epsilon_n}{4(12pN+1)^{(L+1)}(L+1)}} \geq e^{-Dn^{p/(2\alpha+p)}\log^2(n)}.
\end{aligned}$$

□

Theorem 6.2. *(adaptive priors) Assume $f_0 \in \mathcal{H}_p^\alpha$, where $p = O(1)$ as $n \to \infty$, $\alpha < p$, and $\|f_0\|_\infty \leq F$. Let $L \asymp \log(n)$ and assume priors for N and s as in (20) and (21). Assume the prior of f as given by (7) and (8). Then the posterior distribution concentrates at the rate $\xi_n = n^{-\alpha/(2\alpha+p)}\log^\delta(n)$ for $\delta > 1$ in the sense that*

$$\Pi(f \in \mathcal{F}(L) : \|f - f_0\|_L > M_n\xi_n \mid \boldsymbol{Y}^{(n)}) \to 0$$

in $\mathbb{P}_0^n$ probability as $n \to \infty$ for any $M_n \to \infty$.

The proof for Theorem 7.2 follows the same techniques used in Theorem 6.2 of PR18. And this adaptive results also hold for networks with standard normal priors on all weights.

6.3 Preparations for Main Theorems

The general framework for first-order approximation of functionals is as follows

Theorem 6.3. *(Castillo and Rousseau, 2015) Consider the model $\mathbb{P}_0^n$, a real-valued functional $f \to \Psi(f)$ and $\langle\cdot,\cdot\rangle_L, \Psi_0^{(1)}, W_n$, as defined above. Suppose that (16) is satisfied, and denote*

$$\hat{\Psi} = \Psi(f_0) + \frac{W_n(\Psi_0^{(1)})}{\sqrt{n}}, \qquad V_0 = \left\|\Psi_0^{(1)}\right\|_L^2.$$

Let Π be a prior distribution on f. Let A_n be any measurable set such that

$$\Pi(A_n \mid \boldsymbol{Y}^{(n)}) = 1 + o_P(1), \quad as\ n \to \infty.$$

Then for any real t with f_t as

$$f_t = f - \frac{t\Psi_0^{(1)}}{\sqrt{n}},$$

we could write

$$\mathbb{E}^{\Pi}[e^{t\sqrt{n}(\Psi(f)-\hat{\Psi})} \mid \boldsymbol{Y}^{(n)}, A_n] = e^{o_P(1)+t^2V_0/2}\frac{\int_{A_n} e^{\ell_n(f_t)-\ell_n(f_0)d\Pi(f)}}{\int_{A_n} e^{\ell_n(f)-\ell_n(f_0)d\Pi(f)}}.$$

Moreover, if

$$\frac{\int_{A_n} e^{\ell_n(f_t)-\ell_n(f_0)d\Pi(f)}}{\int_{A_n} e^{\ell_n(f)-\ell_n(f_0)d\Pi(f)}} = 1 + o_P(1), \forall t \in \mathbb{R} \tag{30}$$

is satisfied, then the posterior distribution of $\sqrt{n}(\Psi(f)-\hat{\Psi})$ is asymptotically normal and mean-zero, with variance V_0.

Proof. Set $R_n(\cdot,\cdot) = 0, \Psi_0^{(2)} = 0, \mu_n = 0$ in Theorem 2.1 of Castillo and Rousseau (2015). □

Projection of Functions The intuition of our projection conditional on (γ, Z) is to maintain the same partitions for the shifted function in (17) and perform the *change of measure* locally. We first give the notation for Z^L, which are the nodes in the top layer. Let $Z_{Lj}, j = 1, \ldots, p_L$ denote the j^{th} node in L^{th} layer, which can be written as a sum of local linear functions, respectively:

$$Z_{Lj}(\boldsymbol{x}) = \sum_{k=1}^{K_L} \mathbb{I}(\boldsymbol{x} \in \Omega_k^j)\{\tilde{\beta}_k^{j'}\boldsymbol{x} + \tilde{\alpha}_k^j\}$$

here the partitions $\{\Omega_k^j\}_{k=1}^{K_L}$ and coefficients $\{\tilde{\beta}_k^j, \tilde{\alpha}_k^j\}_{k=1}^{K_L}$ are determined by $\{W_l, b_l\}_{l=1}^L$.

For simplicity of notation, we denote $W_{L+1} = (w_1, \ldots, w_{p_L})'$. Then the output can be written as:

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{j=1}^{p_L} w_j Z_{Lj}(\boldsymbol{x}) + b_{L+1} \\ &= \sum_{k_1=1}^{K_L} \cdots \sum_{k_{p_L}=1}^{K_L} \mathbb{I}\left(\boldsymbol{x} \in \bigcap_{j=1}^{p_L} \Omega_{k_j}^j\right)\left\{\left(\sum_{j=1}^{p_L} w_j \tilde{\beta}_{k_j}^{j'}\right)\boldsymbol{x} + \left(\sum_{j=1}^{p_L} w_j \tilde{\alpha}_{k_j}^j + b_{L+1}\right)\right\}. \end{aligned}$$

We denote the projection of function $a(\boldsymbol{x})$ conditional on $\{W_l, b_l\}_{l=1}^L$ with $a_{[Z]}^{\gamma}$, since conditional on $\{W_l, b_l\}_{l=1}^L$ is equivalent to conditional on (γ, Z):

$$\begin{aligned} (W^a, b^a) &= \arg\min_{W_{L+1}, b_{L+1} \in \mathcal{F}_n(L, \mathbf{p}, \gamma, Z)} \|WZ_L(\boldsymbol{x}) + b - a(\boldsymbol{x})\|_L, \\ a_{[Z]}^{\gamma}(\boldsymbol{x}) &= W^a Z_L(\boldsymbol{x}) + b^a. \end{aligned}$$

The projection $a_{[Z]}^{\gamma}$ can also be viewed as the best approximation to a conditional on (γ, Z).

Similarly, we denote projection of f_0 onto $\{W_l, b_l\}_{l=1}^L$ as $f_{0[Z]}^{\gamma}$:

$$(W^0, b^0) = \arg\min_{W_{L+1}, b_{L+1} \in \mathcal{F}_n(L, \mathbf{p}, \gamma, Z)} \|WZ_L(\boldsymbol{x}) + b - f_0(\boldsymbol{x})\|_L, \tag{31}$$

$$f_{0[Z]}^{\gamma}(\boldsymbol{x}) = W^0 Z_L(\boldsymbol{x}) + b^0. \tag{32}$$

Note that $f \in \{WZ_L(\boldsymbol{x}) + b : W \in \mathbb{R}^{p_L}, b \in \mathbb{R}\}$, so naturally we have $\left\|f_{0[Z]}^{\gamma} - f_0\right\|_L \leq \|f - f_0\|_L$.

6.4 Proof of Theorem 3.1

We will perform the analysis locally on the sets $A_n \equiv A_n^{M_n}$ from (15) for some $M_n \to \infty$. We use the fact that convergence of Laplace transforms for all t in probability implies convergence in distribution in probability (Castillo and Rousseau, 2015). The posterior decomposes into a mixture of laws with weights $\Pi(\gamma \mid \boldsymbol{Y}^{(n)})$, where γ is the vector encoding the connectivity pattern with prior in (10). We denote with $I_{n,\gamma} = \mathbb{E}^{\Pi}[e^{t\sqrt{n}(\Psi(f)-\hat{\Psi})} \mid \boldsymbol{Y}^{(n)}, A_n, \gamma]$ and write

$$I_n := \mathbb{E}^{\Pi}[e^{t\sqrt{n}(\Psi(f)-\hat{\Psi})} \mid \boldsymbol{Y}^{(n)}, A_n] = \sum_{\gamma \in \mathcal{V}^{\mathbf{p},s}} \Pi(\gamma \mid \boldsymbol{Y}^{(n)}, A_n) I_{n,\gamma}.$$

Next, we want to show that on the event A_n and uniformly for all $\gamma \in \mathcal{V}^{\mathbf{p},s}$

$$I_{n,\gamma} = e^{o_P(1)+t^2V_0/2}(1+o(1)) \quad \text{as } n \to \infty$$

so that $I_n = e^{o_P(1)+t^2V_0/2}(1+o(1))$.

We choose γ such that $\mathcal{F}(L,\mathbf{p},\gamma) \cap A_n \neq \emptyset$ and for $f \in \mathcal{F}(L,\mathbf{p},\gamma) \cap A_n$ we expand the linear functional as $\Psi(f) - \Psi(f_0) = \langle a, f - f_0 \rangle_L$ which yields

$$\begin{aligned} \Psi_0^{(1)} &= a, \\ r(f, f_0) &= 0. \end{aligned}$$

The remainder condition (16) is thus trivially satisfied. To verify the second condition (17), we choose the shifted function f_t as

$$f_t = f - \frac{ta}{\sqrt{n}}.$$

Due to the fact that our class of neural networks has a top linear layer, the function f_t shares the same deep connectivity structure as f where only the top layer intercepts b_{L+1}^t have been shifted. The *change of measure* thus only influences b_{L+1} where $b_{L+1}^t = b_{L+1} - \frac{ta}{\sqrt{n}}$. Next, we can write

$$I_{n,\gamma} = e^{\frac{t^2}{2}\|a\|_L^2} \times \frac{\int_{A_n} e^{\ell_n(f_t)-\ell_n(f_0)} d\Pi(f \mid \gamma)}{\int_{A_n} e^{\ell_n(f)-\ell_n(f_0)} d\Pi(f \mid \gamma)} \tag{33}$$

$$= e^{\frac{t^2}{2}\|a\|_L^2} \times \frac{\int_{f_t + \frac{ta}{\sqrt{n}} \in A_n} e^{\ell_n(f_t)-\ell_n(f_0)} d\Pi(f_t \mid \gamma) \frac{d\Pi(f|\gamma)}{d\Pi(f_t|\gamma)}}{\int_{A_n} e^{\ell_n(f)-\ell_n(f_0)} d\Pi(f \mid \gamma)}. \tag{34}$$

Next, we show that the ratio above converges to 1 as $n \to \infty$. We have

$$\frac{d\Pi(f \mid \gamma)}{d\Pi(f_t \mid \gamma)} = \frac{d\Pi(\{W_l, b_l\}_{l=1}^L, W_{L+1}, b_{L+1} \mid \gamma)}{d\Pi(\{W_l, b_l\}_{l=1}^L, W_{L+1}, b_{L+1}^t \mid \gamma)} = \frac{d\Pi(\{W_l, b_l\}_{l=1}^L \mid \gamma) d\Pi(W_{L+1}) d\Pi(b_{L+1})}{d\Pi(\{W_l, b_l\}_{l=1}^L \mid \gamma) \Pi(W_{L+1}) \Pi(b_{L+1}^t)} = \frac{d\Pi(b_{L+1})}{d\Pi(b_{L+1}^t)}$$

$$\frac{d\Pi(b_{L+1})}{d\Pi(b_{L+1}^t)} = \frac{\phi(b_{L+1})}{\phi(b_{L+1} - \frac{ta}{\sqrt{n}})} = \exp\left\{-\frac{1}{2}\left[b_{L+1}^2 - (b_{L+1} - \frac{ta}{\sqrt{n}})^2\right]\right\} = \exp\left(-\frac{atb_{L+1}}{\sqrt{n}} + \frac{t^2a^2}{2n}\right)$$

Next, we note (from the definition of the sieve $\mathcal{F}_n$ and C_n in the proof of Theorem 7.1)

$$\frac{|b_{L+1}|}{\sqrt{n}} \leq \frac{\sqrt{C_n}}{\sqrt{n}} \lesssim n^{-\frac{\alpha}{2\alpha+p}\log^{\delta}(n)}.$$

Going back to (33), we now have for some $c > 0$

$$e^{-c\, n^{-\frac{\alpha}{2\alpha+p}}\log^{\delta}(n) + \frac{t^2a^2}{2n} + \frac{t^2}{2}\|a\|_L^2} \times \frac{\Pi\left(f + \frac{ta}{\sqrt{n}} \in A_n \mid \boldsymbol{Y}^{(n)}, \gamma\right)}{\Pi\left(f \in A_n \mid \boldsymbol{Y}^{(n)}, \gamma\right)} \leq I_{n,\gamma}$$

$$\leq e^{c\, n^{-\frac{\alpha}{2\alpha+p}}\log^{\delta}(n) + \frac{t^2a^2}{2n} + \frac{t^2}{2}\|a\|_L^2} \times \frac{\Pi\left(f + \frac{ta}{\sqrt{n}} \in A_n \mid \boldsymbol{Y}^{(n)}, \gamma\right)}{\Pi\left(f \in A_n \mid \boldsymbol{Y}^{(n)}, \gamma\right)}. \tag{35}$$

Next, from

$$\|f-f_0\|_L - \left\|\frac{ta}{\sqrt{n}}\right\|_L \le \left\|f+\frac{ta}{\sqrt{n}}-f_0\right\|_L \le \|f-f_0\|_L + \left\|\frac{ta}{\sqrt{n}}\right\|_L$$

it is clear that

$$\left\{f:\|f-f_0\|_L \le M_n\xi_n - \left\|\frac{ta}{\sqrt{n}}\right\|_L\right\} \subset \left\{f:\left\|f+\frac{ta}{\sqrt{n}}-f_0\right\|_L \le M_n\xi_n\right\} \subset \left\{f:\|f-f_0\|_L \le M_n\xi_n + \left\|\frac{ta}{\sqrt{n}}\right\|_L\right\}$$

This yields

$$\Pi\left(f:\|f-f_0\|_L \le \xi_n - \left\|\frac{ta}{\sqrt{n}}\right\|_L \mid \boldsymbol{Y}^{(n)},\gamma\right) \le \Pi\left(f: f+\frac{ta}{\sqrt{n}} \in A_n \mid \boldsymbol{Y}^{(n)},\gamma\right) \le \Pi\left(f:\|f-f_0\|_L \le \xi_n + \left\|\frac{ta}{\sqrt{n}}\right\|_L \mid \boldsymbol{Y}^{(n)},\gamma\right).$$

Since the concentration rate is slower than $1/\sqrt{n}$, i.e. $\xi_n = n^{-\alpha/(2\alpha+p)}\log^{\delta}(n) \gtrsim n^{-1/2}$, we have $\Pi(f+\frac{ta}{\sqrt{n}} \in A_n) \to \Pi(f \in A_n)$, as $n\to\infty$. From the sandwich inequality (35), we have $I_{n,\gamma} \to e^{\frac{t^2\|a\|_L^2}{2}}$ for any $t\in\mathbb{R}$ as $n\to\infty$.

6.5 Proof of Theorem 3.2

Similar to the linear functional case, the posterior decomposes into a mixture of laws with weights $\Pi(\gamma \mid \boldsymbol{Y}^{(n)})$, where γ is the vector encoding the connectivity pattern with a prior in (10). We can write

$$I_n := \mathbb{E}^{\Pi}[e^{t\sqrt{n}(\Psi(f)-\hat{\Psi})} \mid \boldsymbol{Y}^{(n)}, A_n] = \sum_{\gamma\in\mathcal{V}^{\mathbf{p}},s} \Pi(\gamma \mid \boldsymbol{Y}^{(n)}, A_n) I_{n,\gamma} \tag{36}$$

where

$$I_{n,\gamma} := \mathbb{E}^{\Pi}[e^{t\sqrt{n}(\Psi(f)-\hat{\Psi})} \mid \boldsymbol{Y}^{(n)}, A_n, \gamma].$$

We further decompose each $I_{n,\gamma}$ by conditioning on the deep weights $\{W_l,b_l\}_{l=1}^L$. We can write

$$\begin{aligned}\Pi(\{W_l,b_l\}_{l=1}^{L+1} \mid \boldsymbol{Y}^{(n)}, A_n,\gamma) &= \Pi(W_{L+1},b_{L+1} \mid \{W_l,b_l\}_{l=1}^L, \boldsymbol{Y}^{(n)}, A_n,\gamma)\Pi(\{W_l,b_l\}_{l=1}^L \mid \boldsymbol{Y}^{(n)}, A_n,\gamma)\\ &= \Pi(W_{L+1},b_{L+1} \mid \boldsymbol{Y}^{(n)}, A_n,\gamma,Z)\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n,\gamma),\end{aligned}$$

since $Z=\{Z_l\}_{l=1}^L$ is fully determined by $\{W_l,b_l\}_{l=1}^L$ and we can thereby replace conditioning on $\{W_l,b_l\}_{l=1}^L$ by conditioning on Z. We can further dissect $I_{n,\gamma}$ by conditioning on Z

$$I_{n,\gamma} = \int I_{n,\gamma}^Z d\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n,\gamma), \quad \text{where} \quad I_{n,\gamma}^Z := \int e^{t\sqrt{n}(\Psi(f)-\hat{\Psi})} d\Pi(W_{L+1},b_{L+1} \mid \boldsymbol{Y}^{(n)}, A_n,\gamma,Z).$$

In the rest of the proof, we show that $I_{n,\gamma}^Z \to \exp(-t^2V_0/2)$ uniformly for all γ and Z such that $f\in A_n$. This can be done in two steps. First, we show that conditional on $(\boldsymbol{Y}^{(n)}, A_n,\gamma,Z)$, $\Psi(f)$ asymptotically centers at a local (γ,Z)-dependent centering point $\hat{\Psi}_Z^\gamma$ with a local (γ,Z)-dependent variance V_Z^γ (both defined later). In the second step, we show that the local centering points $\hat{\Psi}_Z^\gamma$ are close to the global centering point $\hat{\Psi}$ and that the local variances V_Z^γ converge to V_0 uniformly for all γ and Z such that $f\in A_n$.

We define the (γ,Z)-dependent local centering point and variance as

$$\hat{\Psi}_Z^\gamma = \Psi(f_0) + \frac{W_n(2f_{0[Z]}^\gamma)}{\sqrt{n}} \quad \text{and} \quad V_Z^\gamma = 4\left\|f_{0[Z]}^\gamma\right\|_L^2, \tag{37}$$

where $f_{0[Z]}^\gamma$ is the $\|\cdot\|_L$ projection of f_0 on the set of deep learning networks f with a connectivity pattern γ and hidden nodes Z defined in (32).

For any $f \in \mathcal{F}(L, \mathbf{p}, \gamma)$, the squared L^2-norm functional can be expanded as

$$\begin{aligned}\Psi(f) - \Psi(f_0) &= 2\langle f_0, f - f_0\rangle_L + \|f - f_0\|_L^2 \\ &= 2\langle f^\gamma_{0[Z]}, f - f_0\rangle_L + \|f - f_0\|_L^2 + 2\langle f_0 - f^\gamma_{0[Z]}, f - f_0\rangle_L.\end{aligned}$$

Note that $\left\|f^\gamma_{0[Z]} - f_0\right\|_L \leq \|f - f_0\|_L$ for any f which has a connectivity pattern γ and hidden nodes Z.

This expansion yields the first-order and remainder terms

$$\begin{aligned}\Psi_0^{(1)} &= 2f^\gamma_{0[Z]}, \\ r(f, f_0) &= \|f - f_0\|_L^2 + 2\langle f_0 - f^\gamma_{0[Z]}, f - f_0\rangle_L.\end{aligned}$$

To ensure asymptotical normality of $\Psi(f)$, we first need to ensure the local shape condition in (16). Assuming that the smoothness α satisfies

$$\alpha > p/2 \tag{38}$$

we have for $f \in A_n$ with a connectivity γ and hidden nodes Z

$$\begin{aligned}r(f, f_0) &= \|f - f_0\|_L^2 + 2\langle f_0 - f^\gamma_{0[Z]}, f - f_0\rangle_L \\ &\leq 2\|f - f_0\|_L^2 + \left\|f_0 - f^\gamma_{0[Z]}\right\|_L^2 \\ &\leq 3\|f - f_0\|_L^2 \lesssim \xi_n^2 = n^{-\frac{2\alpha}{2\alpha+p}} \log^{2\delta} = o\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Next, to verify the second sufficient condition (17) we define the shifted function f_t as

$$f_t = f - \frac{2t f^\gamma_{0[Z]}}{\sqrt{n}}.$$

Then we use the local centering point $\hat{\Psi}^\gamma_Z$ in (37) to define

$$\begin{aligned}\tilde{I}^Z_{n,\gamma} :=& \mathbb{E}^\Pi[e^{t\sqrt{n}(\Psi(f) - \hat{\Psi}^\gamma_Z)} \mid \boldsymbol{Y}^{(n)}, A_n, \gamma, Z] \\ =& e^{2t^2\left\|f^\gamma_{0[Z]}\right\|_L^2} \times \frac{\int_{A_n} e^{\ell_n(f_t) - \ell_n(f_0)} d\Pi(f \mid \gamma, Z)}{\int_{A_n} e^{\ell_n(f) - \ell_n(f_0)} d\Pi(f \mid \gamma, Z)} \\ =& e^{2t^2\left\|f^\gamma_{0[Z]}\right\|_L^2} \times \frac{\int_{f_t + \frac{2t f^\gamma_{0[Z]}}{\sqrt{n}} \in A_n} e^{\ell_n(f_t) - \ell_n(f_0)} d\Pi(f_t \mid \gamma, Z) \frac{d\Pi(f|\gamma,Z)}{d\Pi(f_t|\gamma,Z)}}{\int_{A_n} e^{\ell_n(f) - \ell_n(f_0)} d\Pi(f \mid \gamma, Z)}\end{aligned} \tag{39}$$

For simplicity of notation, we first denote $\zeta = (W_{L+1}, b_{L+1})' \in \mathbb{R}^{p_L+1}$ and $\zeta^t = (W^t_{L+1}, b^t_{L+1})' \in \mathbb{R}^{p_L+1}$ and $\Delta = (W^0, b^0)'$ as defined in (31). Then we can simply write $\zeta^t = \zeta - \frac{2t}{\sqrt{n}}\Delta$.

Since all parameters are a-priori independent and there is no sparsity structure placed on $\{W_{L+1}, b_{L+1}\}$, the prior ratio $\frac{d\Pi(f|\gamma,Z)}{d\Pi(f_t|\gamma,Z)}$ can be calculated as

$$\begin{aligned}\frac{d\Pi(f \mid \gamma, Z)}{d\Pi(f_t \mid \gamma, Z)} &= \frac{d\Pi(W_{L+1})}{d\Pi(W^t_{L+1})} \frac{d\Pi(b_{L+1})}{d\Pi(b^t_{L+1})} = \frac{d\Pi(\zeta)}{d\Pi(\zeta^t)} \\ &= \prod_{i=1}^{p_{L+1}} \exp\left\{-\frac{1}{2}\left[\zeta^2 - (\zeta_i - \frac{2t}{\sqrt{n}}\Delta_i)^2\right]\right\} \\ &= \exp\left\{\sum_{i=1}^{p_{L+1}}\left[-\zeta_i \frac{\Delta_i t}{\sqrt{n}} + \frac{2t^2\Delta_i^2}{n}\right]\right\}.\end{aligned}$$

Similar to our previous proof, we have under the assumption $\alpha > p/2$

$$\left|\sum_{i=1}^{p_{L+1}} \zeta_i \frac{\Delta_i t}{\sqrt{n}}\right| \leq \frac{t}{\sqrt{n}} \|\zeta\|_2 \|\Delta\|_2 \lesssim \frac{C_n}{\sqrt{n}} = o(1), \tag{40}$$

where we used the fact that both f and $f^{\gamma}_{0[Z]}$ are contained in A_n and thereby have their top coefficients contained in a ball of radius $\sqrt{C_n}$ (recall the definition of C_n in the proof of Theorem 7.1).

Now, using the fact that

$$\|f - f_0\|_L - 2\left\|\frac{t f^{\gamma}_{0[Z]}}{\sqrt{n}}\right\|_L \leq \left\|f + \frac{2t f^{\gamma}_{0[Z]}}{\sqrt{n}} - f_0\right\|_L \leq \|f - f_0\|_L + 2\left\|\frac{t f^{\gamma}_{0[Z]}}{\sqrt{n}}\right\|_L$$

we have

$$\begin{aligned}
&\Pi\left(f : \|f - f_0\|_L \leq \xi_n - 2\left\|\frac{t f^{\gamma}_{0[Z]}}{\sqrt{n}}\right\|_L \mid \boldsymbol{Y}^{(n)}, \gamma, Z\right) \\
&\qquad \leq \Pi\left(f + \frac{2t f^{\gamma}_{0[Z]}}{\sqrt{n}} \in A_n \mid \boldsymbol{Y}^{(n)}, \gamma, Z\right) \leq \Pi\left(f : \|f - f_0\|_L \leq \xi_n + 2\left\|\frac{t f^{\gamma}_{0[Z]}}{\sqrt{n}}\right\|_L \mid \boldsymbol{Y}^{(n)}, \gamma, Z\right).
\end{aligned}$$

Again, since the concentration rate is slower than $1/\sqrt{n}$, i.e. $\xi_n = n^{-\alpha/(2\alpha+p)} \log^{\delta}(n) \gtrsim n^{-1/2}$, we have

$$\frac{\Pi(f + \frac{2t f^{\gamma}_{0[Z]}}{\sqrt{n}} \in A_n \mid \boldsymbol{Y}^{(n)}, \gamma, Z)}{\Pi(A_n \mid \boldsymbol{Y}^{(n)}, \gamma, Z)} \to 1, \, \forall t \in \mathbb{R}. \tag{41}$$

Hence, with (38), (40) and (41), one concludes $\tilde{I}^{Z}_{n,\gamma} \to e^{2t^2 \left\|f^{\gamma}_{0[Z]}\right\|^2_L}$ as $n \to \infty$ using a similar sandwich inequality in (35). In other words, we have

$$\tilde{I}^{Z}_{n,\gamma} = e^{t^2 V^{\gamma}_Z/2}(1 + o(1)). \tag{42}$$

Recall the definition of a local centering point $\hat{\Psi}^{\gamma}_Z$ and a local variance V^{γ}_Z in (37). Then we can write

$$\begin{aligned}
I^{Z}_{n,\gamma} &= \mathbb{E}^{\Pi}[e^{t\sqrt{n}(\Psi(f) - \hat{\Psi})} \mid \boldsymbol{Y}^{(n)}, A_n, \gamma, Z] \\
&= \mathbb{E}^{\Pi}[e^{t\sqrt{n}[(\Psi(f) - \hat{\Psi}^{\gamma}_Z) + (\hat{\Psi}^{\gamma}_Z - \hat{\Psi})]} \mid \boldsymbol{Y}^{(n)}, A_n, \gamma, Z] \\
&= \tilde{I}^{Z}_{n,\gamma} \times e^{t\sqrt{n}(\hat{\Psi}^{\gamma}_Z - \hat{\Psi})} \\
&= (1 + o(1)) e^{t^2 V^{\gamma}_Z/2 + t\sqrt{n}(\hat{\Psi}^{\gamma}_Z - \hat{\Psi})} \\
&= (1 + o(1)) e^{t^2 V_0/2 + t^2(V^{\gamma}_Z - V_0)/2 + t\sqrt{n}(\hat{\Psi}^{\gamma}_Z - \hat{\Psi})}.
\end{aligned}$$

The proof will be complete once we show the following condition uniformly for all γ such that $f \in A_n$

$$\begin{aligned}
I_{n,\gamma} &= \int I^{Z}_{n,\gamma} d\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n, \gamma) \\
&= (1 + o(1)) e^{t^2 V_0/2} \int e^{t^2(V^{\gamma}_Z - V_0)/2 + t\sqrt{n}(\hat{\Psi}^{\gamma}_Z - \hat{\Psi})} d\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n, \gamma) \to e^{t^2 V_0/2}, \text{ as } n \to \infty.
\end{aligned}$$

This is equivalent to showing

$$\int e^{t^2(V^{\gamma}_Z - V_0)/2 + t\sqrt{n}(\hat{\Psi}^{\gamma}_Z - \hat{\Psi})} d\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n, \gamma) = 1 + o_P(1). \tag{43}$$

Since we work conditionally on the set A_n, we have $\|f_{0[Z]}^{\gamma} - f_0\|_L \lesssim \xi_n$ and thereby

$$\begin{aligned}\sqrt{n}(\hat{\Psi} - \hat{\Psi}_Z^{\gamma}) &= W_n(f_{0[Z]}^{\gamma} - f_0) = o_P(1),\\ |V_z^{\gamma} - V| &= 4\left|\left\|f_{0[Z]}^{\gamma}\right\|_L^2 - \|f_0\|_L^2\right|\\ &\lesssim 2\|f_0\|_L\left\|f_{0[Z]}^{\gamma} - f_0\right\|_L + \left\|f_{0[Z]}^{\gamma} - f_0\right\|_L^2\\ &\lesssim \left\|f_{0[Z]}^{\gamma} - f_0\right\|_L \le \xi_n\end{aligned}$$

under the assumption that $\|f_0\|_L \le F$.

Using the smoothness assumption (38), we have $\xi_n^2 = o\left(\frac{1}{\sqrt{n}}\right)$. We can bound the integral in (43) using the uniform bounds on $\sqrt{n}(\hat{\Psi} - \hat{\Psi}_Z^{\gamma})$ and $|V_z^{\gamma} - V|$ as

$$\begin{aligned}(43) &= \int e^{t^2\xi_n/2 + t\times o_P(1)} d\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n, \gamma)\\ &= e^{t^2\xi_n/2 + t\times o_P(1)} = e^{o_P(1)} = 1 + o_P(1).\end{aligned}$$

Putting the pieces together, we write I_n from (36) as

$$I_n = \sum_{\gamma\in\mathcal{V}^{\mathbf{p},s}} \Pi(\gamma \mid \boldsymbol{Y}^{(n)}, A_n) I_{n,\gamma} = \sum_{\gamma\in\mathcal{V}^{\mathbf{p},\gamma}} \Pi(\gamma \mid \boldsymbol{Y}^{(n)}, A_n) e^{t^2V_0/2}(1 + o_P(1)) = e^{t^2V_0/2}(1 + o_P(1))$$

which completes the proof. □

6.6 Proof of Theorem 4.1

For our proof for Theorem 4.1, the analysis is locally conducted on the set

$$A_n^M = \{f \in \mathcal{F}(L) : \|f - f_0\|_L \le M_n\xi_n\} \tag{44}$$

with $\xi_n = n^{-\alpha/(2\alpha+p)}\log^{\delta}(n)$ for some $M > 0$ and $\delta > 0$. And from the results in Theorem 7.2, we know $\Pi(A_n^M \mid Y^{(n)}) = 1 + o_p(1)$ for any $M_n \to \infty$.

Conditioning on A_n in (44), the posterior consists of a mixture of laws conditional on N, s and γ

$$\begin{aligned}I_n &= \mathbb{E}^{\Pi}[e^{t\sqrt{n}(\Psi(f) - \hat{\Psi})} \mid \boldsymbol{Y}^{(n)}, A_n]\\ &= \sum_{N=1}^{\infty}\Pi(N \mid \boldsymbol{Y}^{(n)}, A_n)\sum_{s=1}^{T}\Pi(s \mid \boldsymbol{Y}^{(n)}, A_n, N)\sum_{\gamma\in\mathcal{V}^{\mathbf{p},s}}\Pi(\gamma \mid \boldsymbol{Y}^{(n)}, A_n, N, s) I_{n,s,\gamma}\\ &= \sum_{N=1}^{N_n}\Pi(N \mid \boldsymbol{Y}^{(n)}, A_n)\sum_{s=1}^{s_n}\pi(s \mid \boldsymbol{Y}^{(n)}, A_n, N)\sum_{\gamma\in\mathcal{V}^{\mathbf{p},s}}\Pi(\gamma \mid \boldsymbol{Y}^{(n)}, A_n, N, s) I_{n,s,\gamma} + o_p(1)\end{aligned}$$

where we denote with

$$I_{n,s,\gamma} = \mathbb{E}^{\Pi}[e^{t\sqrt{n}(\Psi(f) - \hat{\Psi})} \mid \boldsymbol{Y}^{(n)}, A_n, N, s, \gamma].$$

The second equality follows from the fact that $\Pi(N > N_n \mid \boldsymbol{Y}^{(n)}) \to 0$ and $\Pi(s > s_n \mid \boldsymbol{Y}^{(n)}) \to 0$ in $\mathbb{P}_0^n$ probability as $n \to \infty$, using Corollary 6.1 of Polson and Rockova (2018). Thereby the set A_n eventually excludes all the deep learning mappings outside the sieve.

Linear functionals For $\Psi(f) = \langle a, f\rangle_L$, when $a(\cdot)$ is a constant function, following the same strategy as in the proof of Theorem 3.1, we have

$$I_{n,s,\gamma} = e^{t^2\|a\|_L^2/2}(1 + o(1))$$

and thereby the BvM holds.

Squared L^2-norm functionals For $\Psi(f) = \|f\|_2^2$, we use the same strategy as in the proof of Theorem 3.2. For $\alpha \in (\frac{p}{2}, p)$, we have

$$\left\| f_{0[Z]}^{N,s,\gamma} - f_0 \right\|_L^2 \leq \|f - f_0\|_L^2 = o\left(\frac{1}{\sqrt{n}}\right) \tag{45}$$

here $f_{0[Z]}^{N,s,\gamma}$ denotes the projection of f_0 onto deep learning networks with a fixed sparsity and hidden structure (γ, Z) where $|\gamma| = s$ and the width equals N (similarly as in (32)). The inequality (45) holds for all f with a deep structure determined by (γ, Z).

The following arguments are similar to the proof of Theorem 3.2 but will be conditional on N and s. Since

$$\Pi(\{W_l, b_l\}_{l=1}^{L+1} \mid \boldsymbol{Y}^{(n)}, A_n, N, s, \gamma) = \Pi(W_{L+1}, b_{L+1} \mid \boldsymbol{Y}^{(n)}, A_n, N, s, \gamma, Z) d\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n, N, s, \gamma)$$

we can rewrite $I_{n,s,\gamma}$ as

$$\begin{aligned} I_{n,s,\gamma} &= \int \left(\int e^{t\sqrt{n}(\Psi(f) - \hat{\Psi})} d\Pi(W_{L+1}, b_{L+1} \mid \boldsymbol{Y}^{(n)}, A_n, N, s, \gamma, Z) \right) d\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n, N, s, \gamma) \\ &= (1 + o(1)) e^{2t^2 \|f_0\|_L^2} \int e^{t^2 (V_Z^{N,s,\gamma} - V_0)/2 + t\sqrt{n}(\hat{\Psi}_Z^{N,s,\gamma} - \hat{\Psi})} d\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n, N, s, \gamma) \end{aligned}$$

where

$$\hat{\Psi}_Z^{N,s,\gamma} = \Psi(f_0) + \frac{1}{\sqrt{n}} W_n(2 f_{0[Z]}^{N,s,\gamma}), \quad V_Z^{N,s,\gamma} = 4 \left\| f_{0[Z]}^{N,s,\gamma} \right\|_L^2.$$

and the term $(1 + o(1))$ comes from similar considerations as in (42).

Now we need to show $I_{n,s,\gamma} \to e^{2t^2 \|f_0\|_L^2}$ for all N, s and γ in the local neighborhood A_n. In other words,

$$\sup_{N \leq N_n} \sup_{s \leq s_n} \sup_{\gamma \in \mathcal{V}^{\mathbf{p},s}} \int e^{t^2 (V_Z^{N,s,\gamma} - V_0)/2 + t\sqrt{n}(\hat{\Psi}_Z^{N,s,\gamma} - \hat{\Psi})} d\Pi(Z \mid \boldsymbol{Y}^{(n)}, A_n, N, s, \gamma) = o_P(1). \tag{46}$$

Then we can write for $\alpha > p/2$

$$\begin{aligned} \sqrt{n}(\hat{\Psi}^{N,s,\gamma} - \hat{\Psi}) &= W_n(f_{0[Z]}^{N,s,\gamma} - f_0) = o_P(1), \\ |V_{N,s,\gamma} - V_0| &= 4 \left| \left\| f_{0[Z]}^{N,s,\gamma} \right\|_L^2 - \|f_0\|_L^2 \right| \\ &\lesssim 2 \|f_0\|_L \left\| f_{0[Z]}^{N,s,\gamma} - f_0 \right\|_L + \left\| f_{0[Z]}^{N,s,\gamma} - f_0 \right\|_L^2 \\ &\lesssim \left\| f_{0[Z]}^{N,s,\gamma} - f_0 \right\|_L \leq \xi_n. \end{aligned}$$

With $\alpha > p/2$, (46) is satisfied. Aggregating the sum of $I_{N,s,\gamma}$ over N, s and γ, we have

$$\begin{aligned} I_n &= \sum_{N=1}^{N_n} \Pi(N \mid \boldsymbol{Y}^{(n)}, A_n) \sum_{s=1}^{s_n} \Pi(s \mid \boldsymbol{Y}^{(n)}, A_n, N) \sum_{\gamma \in \mathcal{V}^{\mathbf{p},s}} \Pi(\gamma \mid \boldsymbol{Y}^{(n)}, A_n, N, s) I_{n,s,\gamma} + o_P(1) \\ &= \sum_{N=1}^{N_n} \Pi(N \mid \boldsymbol{Y}^{(n)}, A_n) \sum_{s=1}^{s_n} \Pi(s \mid \boldsymbol{Y}^{(n)}, A_n, N) \sum_{\gamma \in \mathcal{V}^{\mathbf{p},s}} \Pi(\gamma \mid \boldsymbol{Y}^{(n)}, A_n, N, s)(1 + o(1)) e^{2t^2 \|f_0\|_L^2 + o_P(1)} + o_P(1). \end{aligned}$$

As a result, we have $I_n \to e^{2t^2 \|f_0\|_L^2}$ for all $t \in \mathbb{R}$ as $n \to \infty$, which concludes the proof for the L^2-norm functional case.