

---

# Supplementary Material for Variable Selection with Rigorous Uncertainty Quantification using Deep Bayesian Neural Networks

---

## Contents

<b>A</b>	<b>Additional Background</b>	<b>2</b>
A.1	Rectified Linear Unit (ReLU) Neural Network in Matrix Notation . . . . .	2
A.2	ReLU Network in the reproducing kernel Hilbert space (RKHS) . . . . .	3
A.3	Bayesian Learning of Deep ReLU Neural Network . . . . .	3
<b>B</b>	<b>Proofs for Posterior Consistency</b>	<b>4</b>
B.1	Proof for Theorem 1 . . . . .	4
<b>C</b>	<b>Proofs for Bernstein-von Mises (BvM) Phenomenon</b>	<b>6</b>
C.1	Background: Semi-parametric BvM Theorem for Smooth Functionals . . . . .	6
C.2	Preliminary I: Notations and Basic Setup . . . . .	7
C.3	Preliminary II: Proof Strategy and Preliminary Theorems . . . . .	8
C.4	Proof for Theorem 2 . . . . .	9
C.5	Proof for Theorem 3 . . . . .	11
<b>D</b>	<b>Additional Proofs</b>	<b>13</b>
D.1	Proof for Theorem C.2 . . . . .	13
D.2	Proof for Theorem C.3 . . . . .	17
<b>E</b>	<b>Lemmas and Propositions</b>	<b>19</b>

## A Additional Background

### A.1 ReLU Neural Network in Matrix Notation

Given a sample  $\{y_i, \mathbf{x}_i\}_{i=1}^n$ , the neural network predictions  $\mathbf{f}_{n \times 1} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]^\top$  can be represented as a  $n \times 1$  vector in terms of its weight matrices. This matrix representation is mathematically convenient and provides important insight into how the data flow through the layers to interact with different components of the model. Interestingly, under the ReLU activation function,  $\mathbf{f}_{n \times 1}$  can be written as a linear product of model weight matrices despite the fact that  $f(\mathbf{x})$  itself is a nonlinear model. To keep the notation simple, in this section we will omit the bias term  $b_0$ . However we note that it can always be added back to the model by augmenting the feature matrix  $\Phi_{n \times K}$  (defined below) with a column vector of 1's.

We first consider the one-layer ReLU network  $f(\mathbf{x}) = \beta^\top \sigma(\mathcal{W}_1 \mathbf{x})$ . For a single hidden unit, the ReLU activation output is  $\sigma(\mathcal{W}_{1,k} \mathbf{x}) = I(\mathcal{W}_{1,k} \mathbf{x} > 0) * \mathcal{W}_{1,k} \mathbf{x}$ , and the ReLU output of the entire  $K$ -unit hidden layer is a  $K \times 1$  vector  $\sigma(\mathcal{W}_1 \mathbf{x}) = S_{1,\mathbf{x}} \mathcal{W}_1 \mathbf{x}$ , where  $S_{1,\mathbf{x}}$  is a  $K \times K$  diagonal matrix with indicator functions  $I(\mathcal{W}_{1,k} \mathbf{x} > 0)$  on the diagonal. As a result, the one-layer ReLU network evaluated at a single observation  $\mathbf{x}_i$  is:

$$f(\mathbf{x}_i) = \beta^\top S_{1,\mathbf{x}_i} \mathcal{W}_1 \mathbf{x}_i.$$

To express  $\mathbf{f}_{n \times 1} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]^\top$ , we define data matrix  $\mathbf{X}$ , weight matrix  $\mathbf{W}_1$  and activation matrix  $\mathbf{S}_1$  such that:

$$\mathbf{f}_{n \times 1} = \Phi_{\mathcal{W}} \beta = (\mathbf{X} \mathbf{W}_1 \mathbf{S}_1) \beta, \quad (1)$$

where we have defined  $\Phi_{\mathcal{W}, n \times K} = \mathbf{X} \mathbf{W}_1 \mathbf{S}_1$  the matrix of hidden features, and

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & 0 & \dots & 0 \\ 0 & \mathbf{x}_2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & 0 & \mathbf{x}_n \end{bmatrix}_{n \times nP}, \quad \mathbf{W}_1 = \mathcal{W}_1 \otimes \mathbf{I}_n = \begin{bmatrix} \mathcal{W}_1 & 0 & \dots & 0 \\ 0 & \mathcal{W}_1 & \dots & 0 \\ \dots & & & \\ 0 & 0 & 0 & \mathcal{W}_1 \end{bmatrix}_{nP \times nK}, \quad \mathbf{S}_1 = \begin{bmatrix} S_{1,\mathbf{x}_1} \\ S_{1,\mathbf{x}_2} \\ \vdots \\ \dots \\ S_{1,\mathbf{x}_n} \end{bmatrix}_{nK \times K}.$$

To express  $\mathbf{f}_{n \times 1}$  for a deep ReLU network  $f(\mathbf{x}) = \beta^\top (\sigma \mathcal{W}_L \dots (\sigma \mathcal{W}_1 \mathbf{x}))$ , we notice a  $L$ -layer ReLU neural network evaluated at a single observation is expressed as:

$$f(\mathbf{x}) = \beta^\top \left( \prod_{l=L}^1 S_{l,\mathbf{x}} \mathcal{W}_l \right) \mathbf{x}$$

using the short-hand notation  $\prod_{l=L}^1 M_l = M_L M_{L-1} \dots M_2 M_1$ . Here  $S_{l,\mathbf{x}}$  is a  $K \times K$  diagonal matrix with indicator functions  $I(\mathcal{W}_{l,k} \sigma \mathcal{W}_{l-1} \dots \sigma \mathcal{W}_1 \mathbf{x} > 0)$  on the diagonal. As a result:

$$\mathbf{f}_{n \times 1} = \Phi_{\mathcal{W}} \beta = \left( \mathbf{X} \left( \prod_{l=1}^L \mathbf{W}_l \mathbf{S}_l \right) \right) \beta, \quad (2)$$

where  $\Phi_{\mathcal{W}, n \times K} = \mathbf{X} \left( \prod_{l=1}^L \mathbf{W}_l \mathbf{S}_l \right)$  is the matrix of hidden features at the output layer, and we have denoted  $\mathbf{W}_l = \mathcal{W}_l \otimes \mathbf{I}_n$  and  $\{\mathbf{S}_l\}_{l=1}^{L-1}, \mathbf{S}_L$  as:

$$\mathbf{S}_l = \begin{bmatrix} S_{l,\mathbf{x}_1} & 0 & \dots & 0 \\ 0 & S_{l,\mathbf{x}_2} & \dots & 0 \\ \dots & & & \\ 0 & 0 & 0 & S_{l,\mathbf{x}_n} \end{bmatrix}_{nK \times nK}, \quad \mathbf{S}_L = \begin{bmatrix} S_{L,\mathbf{x}_1} \\ S_{L,\mathbf{x}_2} \\ \vdots \\ \dots \\ S_{L,\mathbf{x}_n} \end{bmatrix}_{nK \times K}.$$

It is worth noting that the gradient function of a ReLU network  $\partial_p \mathbf{f}_{n \times 1} = \left[ \frac{\partial}{\partial x_p} f(\mathbf{x}_1), \dots, \frac{\partial}{\partial x_p} f(\mathbf{x}_n) \right]^\top$  can also be expressed as a linear product of matrices. Specifically, since ReLU network is expressed as  $\mathbf{f} = \Phi \beta$  with  $\Phi = \mathbf{X} \left( \prod_{l=1}^L \mathbf{W}_l \mathbf{S}_l \right)$ , its derivative can be written as:

$$\partial_p \mathbf{f}_{n \times 1} = (\partial_p \Phi) \beta = \left( \mathbf{W}_{1,p} \mathbf{S}_1 \left( \prod_{l=2}^L \mathbf{W}_l \mathbf{S}_l \right) \right) \beta$$

where  $\mathbf{W}_{1,p}$  is a  $n \times nK$  matrix such that  $\mathbf{W}_{1,p} = \mathbf{I}_n \otimes \mathcal{W}_{1,p}$ . Correspondingly, the empirical norm of the gradient function can be written as:

$$\psi_p(f) = \left\| \frac{\partial}{\partial x_p} f \right\|_n^2 = \frac{1}{n} \left\| \partial_p \mathbf{f}_{n \times 1} \right\|_2^2 = \frac{1}{n} \beta^\top (\partial_p \Phi)^\top (\partial_p \Phi) \beta,$$

which is of a quadratic form with respect to a  $K \times K$  ‘‘gradient kernel matrix’’  $(\partial_p \Phi)^\top (\partial_p \Phi)$ .

## A.2 ReLU Network in the RKHS

It is important to notice that the function space  $\mathcal{F}(L, K, B)$  corresponds to a reproducing kernel Hilbert space (RKHS) [2, 3]. Specifically, notice that if we focus on the output layer, then every  $f \in \mathcal{F}(L, K, B)$  can be written as

$$f(\mathbf{x}) = \phi_{\mathcal{W}}(\mathbf{x})^\top \beta = \sum_{k=1}^K \beta_k \phi_k(\mathbf{x}|\mathcal{W}),$$

where  $\phi_k(\mathbf{x}|\mathcal{W}) = \mathcal{W}_{k,L} \left[ \circ_{l=1}^{L-1} (\sigma \mathcal{W}_l) \circ \mathbf{x} \right]$  are the  $K$  ‘‘basis functions’’ at the output layer that are formed by the hidden neurons.

Consequently,  $\mathcal{F}(L, K, B)$  corresponds to a RKHS  $\mathcal{H}$  equipped with positive definite kernel  $k_{\mathcal{W}}(\mathbf{x}, \mathbf{x}') = \phi_{\mathcal{W}}(\mathbf{x})^\top \phi_{\mathcal{W}}(\mathbf{x}') = \sum_{k=1}^K \phi_k(\mathbf{x}|\mathcal{W}) \phi_k(\mathbf{x}'|\mathcal{W})$ , and the hidden weights  $\mathcal{W}$  can be regarded as the hyper-parameters for this RKHS kernel [3]. For two functions in this RKHS  $f(\mathbf{x}) = \phi_{\mathcal{W}}(\mathbf{x})^\top \beta_f$  and  $g(\mathbf{x}) = \phi_{\mathcal{W}}(\mathbf{x})^\top \beta_g$ , the inner product in  $\mathcal{H}$  is defined as:

$$\langle f, g \rangle_{\mathcal{H}} = \beta_f^\top \beta_g,$$

i.e., we can interpret the output-layer weights  $\beta$  as a ‘‘representer’’ of  $f$  in  $\mathcal{H}$ . It is important to verify that the above definition of  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$  indeed gives rise to the reproducing property:

$$\langle f, k(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = \beta_f^\top \phi_{\mathcal{W}}(\mathbf{x}) = f(\mathbf{x}),$$

which can be seen easily by noticing that  $k_{\mathcal{W}}(\mathbf{x}, \cdot) = \phi_{\mathcal{W}}(\mathbf{x})^\top \phi_{\mathcal{W}}(\cdot)$  is an element in  $\mathcal{H}$  with its ‘‘ $\beta$ ’’ being  $\phi_{\mathcal{W}}(\mathbf{x})$ . Finally, the quadratic norm in this space is defined by  $\|f\|_{\mathcal{H}}^2 = \sum_{k=1}^K \beta_{f,w}^2 = \|\beta_f\|_2^2$  [2].

To compute the kernel function  $k_{\mathcal{W}}$  and the corresponding kernel matrix  $\mathbf{K}_{\mathcal{W}}$  for a ReLU network, we recall that, for the single-layer ReLU network  $f(\mathbf{x}) = \beta \sigma(\mathcal{W}_1 \mathbf{x})$ , the corresponding basis function is a  $K \times 1$  vector  $\phi_{\mathcal{W}}(\mathbf{x})_{K \times 1} = \sigma(\mathcal{W}_1 \mathbf{x}) = S_{1,\mathbf{x}} \mathcal{W}_1 \mathbf{x}$ . so we can write  $k_{\mathcal{W}}$  in terms of  $\phi_{\mathcal{W}}$  as:

$$k_{\mathcal{W}}(\mathbf{x}, \mathbf{x}') = \phi_{\mathcal{W}}(\mathbf{x})^\top \phi_{\mathcal{W}}(\mathbf{x}') = \mathbf{x}^\top \mathcal{W}_1^\top S_{1,\mathbf{x}} S_{1,\mathbf{x}'} \mathcal{W}_1 \mathbf{x}'.$$

Similarly,  $k_{\mathcal{W}}$  for a  $L$ -layer ReLU network is:

$$k_{\mathcal{W}}(\mathbf{x}, \mathbf{x}') = \phi_{\mathcal{W}}(\mathbf{x})^\top \phi_{\mathcal{W}}(\mathbf{x}') = \mathbf{x}^\top \left( \prod_{l=L}^1 S_{l,\mathbf{x}} \mathcal{W}_l \right)^\top \left( \prod_{l=L}^1 S_{l,\mathbf{x}'} \mathcal{W}_l \right) \mathbf{x}'.$$

To express the kernel matrix  $\mathbf{K}_{\mathcal{W}, n \times n}$ , recall the matrix of basis functions  $\Phi_{\mathcal{W}, n \times K} = [\phi_{\mathcal{W}}(\mathbf{x}_1), \dots, \phi_{\mathcal{W}}(\mathbf{x}_n)]^\top$  can be expressed compactly as  $\Phi_{\mathcal{W}} = \mathbf{X} \mathbf{W}_1 \mathbf{S}_1$  for a single-layer network, and  $\Phi_{n \times K} = \mathbf{X} \left( \prod_{l=1}^L \mathbf{W}_l \mathbf{S}_l \right)$  for a deep ReLU network. So the kernel matrices  $\mathbf{K}_{\mathcal{W}}$  for a single-layer and a deep ReLU network are, respectively:

$$\mathbf{K}_{\mathcal{W}} = \Phi_{\mathcal{W}} \Phi_{\mathcal{W}}^\top = \mathbf{X} \mathbf{W}_1 \mathbf{S}_1 \mathbf{S}_1^\top \mathbf{W}_1^\top \mathbf{X}^\top, \quad \mathbf{K}_{\mathcal{W}} = \Phi_{\mathcal{W}} \Phi_{\mathcal{W}}^\top = \mathbf{X} \left( \prod_{l=1}^L \mathbf{W}_l \mathbf{S}_l \right) \left( \prod_{l=1}^L \mathbf{W}_l \mathbf{S}_l \right)^\top \mathbf{X}^\top.$$

## A.3 Bayesian Learning of Deep ReLU Neural Network

It is well-known that for some common choices of  $\Pi(\beta)$ ,  $\Pi(f)$  corresponds to a (conditional) Gaussian process (GP) [15]. Specifically, by placing independent and identically distributed (i.i.d.) Gaussian prior  $N(0, \frac{1}{K})$  on  $\beta$  and  $N(0, \sigma_{b_0}^2)$  on  $b_0$ , the neural network model  $f(\cdot) = \phi_{\mathcal{W}}(\cdot)^\top \beta$  is equal

in distribution to Gaussian process with kernel function  $k_{\mathcal{W}}(\mathbf{x}, \mathbf{x}') = \frac{1}{K} \phi_{\mathcal{W}}(\mathbf{x})^\top \phi_{\mathcal{W}}(\mathbf{x}') + \sigma_{b_0}^2$ , i.e.,  $\Pi(f|\mathcal{W}) = GP(f|0, k_{\mathcal{W}})$ .

As a result, under the conditional Gaussian process (GP) representation, the prior distribution for  $f$  can be written as:

$$\Pi(f, \mathcal{W}) = \Pi(f|\mathcal{W})\Pi(\mathcal{W}) = GP(f|0, k_{\mathcal{W}})\Pi(\mathcal{W}), \quad (3)$$

The conditional GP representation in (3) is important for analyzing the asymptotic behavior of the Bayesian neural network. It suggests that, if the behavior of the conditional posterior  $\Pi_n(f|\mathcal{W})$  does not change drastically under  $\Pi_n(\mathcal{W})$ , then the asymptotic behavior of  $\Pi_n(f, \mathcal{W})$  is analogous to that of a Gaussian process, whose theoretical properties are well-understood in the literature [21, 7]. In Section D.1 - D.2, we take advantage of this representation to show an BvM phenomenon (i.e. asymptotic normality) for the posterior distribution of variable importance for a wide range of choices for  $\Pi(\mathcal{W})$ .

## B Proofs for Posterior Consistency

### B.1 Proof for Theorem 1

*Proof.* Denote  $A_n = \{f : \|f - f_0\|_n^2 > M_n \epsilon_n\}$  and  $B_n = \{f : |\psi_p(f) - \Psi_p(f_0)| > M_n \epsilon_n\}$ , then showing the statement in Theorem 1 is equivalent to showing  $\Pi_n(B_n) \rightarrow 0$ .

Specifically, we assume below three facts hold:

**Fact 1**  $|\psi_p(f) - \psi_p(f_0)| \leq \left\| \frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0 \right\|_n^2$

**Fact 2**  $\sup_{p \in \{1, \dots, P\}} \left\| \frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0 \right\|_n^2 \leq C * \|f - f_0\|_n^2$  for some constant  $C$ .

**Fact 3**  $\sup_{p \in \{1, \dots, P\}} |\psi_p(f_0) - \Psi_p(f_0)| \lesssim \|f - f_0\|_n^2$ .

Because if above facts hold, we then have

$$\begin{aligned} \sup_{p \in \{1, \dots, P\}} |\psi_p(f) - \Psi_p(f_0)| &\leq \sup_{p \in \{1, \dots, P\}} |\psi_p(f) - \psi_p(f_0)| + \sup_{p \in \{1, \dots, P\}} |\psi_p(f_0) - \Psi_p(f_0)| \\ &\leq \sup_{p \in \{1, \dots, P\}} \left\| \frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0 \right\|_n^2 + \sup_{p \in \{1, \dots, P\}} |\psi_p(f_0) - \Psi_p(f_0)| \\ &\leq C * \|f - f_0\|_n^2 + \sup_{p \in \{1, \dots, P\}} |\psi_p(f_0) - \Psi_p(f_0)| \\ &\lesssim \|f - f_0\|_n^2, \end{aligned}$$

it then follows that:

$$E_0 \Pi_n \left( \sup_{p \in \{1, \dots, P\}} |\psi_p(f) - \Psi_p(f_0)| \geq M_n \epsilon_n \right) \lesssim E_0 \Pi_n \left( \|f - f_0\|_n^2 \geq M_n' \epsilon_n \right) \rightarrow 0.$$

We now show Facts 1-3 are true:

- **Fact 1** follows simply from the triangular inequality:

$$\begin{aligned} |\psi_p(f) - \psi_p(f_0)| &= \left| \left\| \frac{\partial}{\partial x_p} f \right\|_n^2 - \left\| \frac{\partial}{\partial x_p} f_0 \right\|_n^2 \right| \\ &= \max \left\{ \left\| \frac{\partial}{\partial x_p} f \right\|_n^2 - \left\| \frac{\partial}{\partial x_p} f_0 \right\|_n^2, \left\| \frac{\partial}{\partial x_p} f_0 \right\|_n^2 - \left\| \frac{\partial}{\partial x_p} f \right\|_n^2 \right\} \leq \left\| \frac{\partial}{\partial x_p} f_0 - \frac{\partial}{\partial x_p} f \right\|_n^2. \end{aligned}$$

- **Fact 2.** First establish some notation. Given data  $\{\mathbf{x}_i, y_i\}_{i=1}^n$ , denote  $\mathbf{f}$  and  $\mathbf{f}_0$  the  $n \times 1$  vectors with their elements being  $f(\mathbf{x}_i)$ ,  $f_0(\mathbf{x}_i)$ , respectively. We then have  $\|f - f_0\|_n^2 = \frac{1}{n} \|\mathbf{f} - \mathbf{f}_0\|_2^2$  where  $\|\cdot\|_2$  is the matrix 2-norm. Furthermore, since  $f, f_0 \in \mathcal{F}(L, W, B)$ , there exists sets of weight matrices  $\{\mathbf{W}_l, \mathbf{S}_l\}_{l=1}^L$ ,  $\{\mathbf{W}_{0,l}, \mathbf{S}_{0,l}\}_{l=1}^L$  and output weights  $\beta, \beta_0$  such that  $\mathbf{f} = \mathbf{X} \mathbf{W}_1 \mathbf{S}_1 (\prod_{l=2}^L \mathbf{W}_l \mathbf{S}_l) \beta$

and  $\mathbf{f}_0 = \mathbf{X}\mathbf{W}_{0,1}\mathbf{S}_{0,1}(\prod_{l=2}^L \mathbf{W}_{0,l}\mathbf{S}_{0,l})\beta_0$ . To keep the notation simple, we write for  $\mathbf{f}$  its the input weights as  $\mathbf{W}$ , and the product of weight matrices after the input layer as  $\mathbf{D} = \mathbf{S}_1(\prod_{l=2}^L \mathbf{W}_l\mathbf{S}_l)\beta$ , such that  $\mathbf{f}$  and  $\mathbf{f}_0$  can be written as:

$$\mathbf{f} = \mathbf{X}\mathbf{W}\mathbf{D}, \quad \mathbf{f}_0 = \mathbf{X}\mathbf{W}_0\mathbf{D}_0$$

where recall  $\mathbf{X}$  is a  $n \times nP$  block diagonal matrix with  $1 \times P$  vectors  $\mathbf{x}_i$ 's on the diagonal. Furthermore, by the definition of gradient functions for ReLU network, we can write the  $n \times 1$  vectors of gradient functions as:

$$\partial_p \mathbf{f} = \mathbf{W}_p \mathbf{D}, \quad \partial_p \mathbf{f}_0 = \mathbf{W}_{0,p} \mathbf{D}_0,$$

where  $\mathbf{W}_p = \mathbf{I}_n \otimes \mathbf{w}_p$ ,  $\mathbf{W}_{0,p} = \mathbf{I}_n \otimes \mathbf{w}_{0,p}$  are  $n \times nK$  block diagonal matrices. Notice that  $\|\frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0\|_n^2 = \frac{1}{n} \|\partial_p \mathbf{f} - \partial_p \mathbf{f}_0\|_2^2$ , also define  $\mathbb{X} = \mathbf{X}^\top \mathbf{X} + c\mathbf{I}$  for a small positive constant  $c$ .

Consequently:

$$\begin{aligned} \sup_{p \in \{1, \dots, P\}} \left\| \frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0 \right\|_n^2 &\leq \sum_{p=1}^P \left\| \frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0 \right\|_n^2 \\ &= \frac{1}{n} \sum_{p=1}^P \|\mathbf{W}_p \mathbf{D} - \mathbf{W}_{0,p} \mathbf{D}_0\|_2^2 = \frac{1}{n} \|\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0\|_2^2. \end{aligned}$$

Also notice that:

$$\begin{aligned} \|\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0\|_2 &= \|\mathbb{X}^{-1} \mathbb{X}(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2 = \|\mathbb{X}^{-1}(\mathbf{X}^\top \mathbf{X} + c\mathbf{I})(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2 \\ &= \|\mathbb{X}^{-1}(\mathbf{X}^\top \mathbf{X})(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0) + \mathbb{X}^{-1}(c\mathbf{I})(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2 \\ &\leq \|\mathbb{X}^{-1}(\mathbf{X}^\top \mathbf{X})(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2 + \|\mathbb{X}^{-1}(c\mathbf{I})(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2, \quad (4) \end{aligned}$$

where in the second term of the above expression, we have  $\mathbb{X}^{-1}(c\mathbf{I}) = c * (\mathbf{X}^\top \mathbf{X} + c\mathbf{I})^{-1} \prec \mathbf{I}$ , this is because the eigenvalues of  $c * (\mathbf{X}^\top \mathbf{X} + c\mathbf{I})^{-1}$  are always smaller than 1. As a result, it is always true that  $\|\mathbb{X}^{-1}(c\mathbf{I})(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2 < \|\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0\|_2$ . Therefore there exists a positive constant  $\rho_c < 1$  that upper bounds  $\|\mathbb{X}^{-1}(c\mathbf{I})\|_2$ , such that

$$\|\mathbb{X}^{-1}(c\mathbf{I})(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2 = \rho_c * \|\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0\|_2.$$

Plugging the above expression into (4), move the second term on the right-hand side to the left, square both sides and finally multiply both sides by  $\frac{1}{n}$ , we have:

$$\begin{aligned} \frac{1}{n} (1 - \rho_c)^2 \|\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0\|_2^2 &\leq \frac{1}{n} \|\mathbb{X}^{-1} \mathbf{X}^\top \mathbf{X}(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2^2 \\ &\leq \frac{1}{n} \|\mathbb{X}^{-1} \mathbf{X}^\top\|_2^2 \|\mathbf{X}(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2^2 \\ &= \frac{1}{n} \|(\mathbf{X}^\top \mathbf{X} + c\mathbf{I})^{-1} \mathbf{X}^\top\|_2^2 \|\mathbf{X}(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)\|_2^2 \\ &= \|(\mathbf{X}^\top \mathbf{X} + c\mathbf{I})^{-1} \mathbf{X}^\top\|_2^2 \|f - f_0\|_n^2. \quad (5) \end{aligned}$$

Notice that in (5), the matrix inside the first term  $\|(\mathbf{X}^\top \mathbf{X} + c\mathbf{I})^{-1} \mathbf{X}^\top\|_2^2$  is the "projection matrix" for the coefficient of a Ridge regression estimator (i.e.,  $\hat{\beta} = (\mathbf{X}^\top \mathbf{X} + c\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$ ), and the 2-norm of this term is upper bounded by its largest singular value, i.e.,

$$\|(\mathbf{X}^\top \mathbf{X} + c\mathbf{I})^{-1} \mathbf{X}^\top\|_2 \leq \max_{\lambda > 0} \left( \frac{\lambda}{\lambda^2 + c} \right) \leq \frac{1}{2\sqrt{c}},$$

where the last inequality is obtained by noting that  $\lambda = \sqrt{c}$  is the solution to the above maximization problem. Using this fact, we can write (5) as:

$$\frac{1}{n} \|\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0\|_2^2 \leq \frac{1}{4c(1 - \rho_c)^2} \|f - f_0\|_n.$$

Denoting  $C = \frac{1}{4c(1 - \rho_c)^2}$  and noting that it is a constant, we have shown that

$$\sup_{p \in \{1, \dots, P\}} \left\| \frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0 \right\|_n^2 \leq \frac{1}{n} \|\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0\|_2^2 \leq C * \|f - f_0\|_n^2,$$

which is the statement of Fact 2.

- **Fact 3** follows from standard Bernstein-type concentration inequality (see, e.g. Lemma 18 of [18]). Specifically, for  $|\frac{\partial}{\partial x_p} f_0(\mathbf{x})|^2$  a random variable with respect to probability measure  $P(\mathbf{x})$  that is bounded by  $L$ . Given  $n$  i.i.d. samples  $\{|\frac{\partial}{\partial x_p} f_0(x_i)|^2\}_{i=1}^n$ , recall that  $\hat{\psi}(f_0) = \frac{1}{n} \sum_{i=1}^n |\frac{\partial}{\partial x_p} f_0(x_i)|^2$  and  $\psi(f_0) = E(\frac{\partial}{\partial x_p} f_0)$ , then with probability  $1 - \eta$ :

$$|\hat{\psi}(f_0) - \psi(f_0)| \leq n^{-\frac{1}{2}} * (2\sqrt{2} * L * \log(2/\eta)),$$

that is,  $|\hat{\psi}(f_0) - \psi(f_0)| \rightarrow 0$  at rate of  $O(n^{-\frac{1}{2}})$ . Notice that  $O(n^{-\frac{1}{2}})$  is the optimal parametric rate that cannot be surpassed by the convergence speed of the ReLU networks (recall the typical convergence rate is  $\epsilon_n \asymp n^{-\frac{\beta}{2\beta+\delta}} * \log(n)^\gamma$  for some  $\delta > 0$  and  $\gamma > 1$ ). Therefore we have:

$$\sup_{p \in \{1, \dots, P\}} |\psi_p(f_0) - \Psi_p(f_0)| \lesssim \|f - f_0\|_n^2.$$

□

## C Proofs for BvM Phenomenon

### C.1 Background: Semi-parametric BvM Theorem for Smooth Functionals

In this section, we provide background on a general semi-parametric BvM theorem for smooth nonlinear functionals [8]. In nonparametric regression, the regression function  $f \in \mathcal{F}$  is infinite-dimensional and the asymptotic distribution of  $f$  in this case is in general difficult to characterize [10]. However, in practical applications, we are mostly interested in a finite-dimensional parameter  $\psi : \mathcal{F} \rightarrow \mathbb{R}^d$  whose asymptotic distribution is easier to reason with. For example, a cumulative distribution function at a fixed point  $F(x_0) = \int \mathbb{I}(x < x_0) f(x) dx$  [17].

To this end, a series of work by [4, 17, 8] provided general sufficient conditions for BvM theorem for smooth functionals under general models. These results show that, if the functional of interest  $\psi$  and the model log likelihood  $l_n$  both satisfy certain smoothness conditions, then the marginal posterior of  $\psi(f)$  concentrates at the rate  $O(n^{-1/2})$ , and furthermore, the marginal posterior distribution of  $\sqrt{n}(\psi(f) - \hat{\psi})$  converges weakly to a  $N(0, V_0)$  under the data-generation distribution  $P_0$ , where  $\hat{\psi}$  is an efficient estimator of  $\psi(f_0)$ . Such properties have the implication that it allow the construction of credible regions for which have correct asymptotic frequentist coverage [7].

The main conditions for BvM theorem for smooth functionals are as below:

1. Locally Asymptotic Normal (LAN) Expansion of Likelihood Function  $l_n(f)$ :

$$l_n(f) - l_n(f_0) = -\frac{n}{2} \|f - f_0\|_n^2 + \sqrt{n} W_n(f - f_0). \quad (6)$$

2. Smoothness Expansion of Functional  $\psi(f)$ :

$$\psi(f) - \psi(f_0) = \langle \psi_1, f - f_0 \rangle_n + \frac{1}{2} \langle \psi_2(f - f_0), f - f_0 \rangle_n + o_p(\sqrt{n}). \quad (7)$$

3. Relation between  $l_n(f)$  and  $\psi(f)$ :

For a posterior distribution  $\Pi_n(f) = \Pi(f | \{y_i, \mathbf{x}_i\}_{i=1}^n)$  that concentrates around  $f_0$  at rate  $\epsilon_n$ , i.e.  $\Pi_n(f : \|f - f_0\|_n \leq \epsilon_n) = 1 + o_p(1)$ , define  $A_n$  as the sequence of sets that receive majority of probability mass from  $\Pi_n$ , i.e.

$$\Pi_n(A_n) = \Pi_n(f \in A_n : \|f - f_0\|_n \leq \epsilon_n) = 1 + o_p(1).$$

Assume there exists  $w_n \in \mathcal{F}$  such that  $W_n$  adopts a decomposition

$$W_n(f) = \langle w_n, f \rangle_n + \Delta_n(f),$$

where  $w_n$  is the "representor" of  $W_n$  such that  $\langle w_n, f \rangle_n$  retains majority of information from  $W_n(f)$ , and  $\Delta_n(f)$  is the corresponding residual term. It is required that both of these terms are sufficiently regular in the sense that they satisfy below two conditions:

$$\langle w_n, \psi_2(\psi_1) \rangle_n + \epsilon_n \|w_n\|_n = o_p(\sqrt{n}), \quad (8)$$

$$\sup_{f \in A_n} |\Delta_n(\psi_2(f - f_0))| = o_p(1). \quad (9)$$

Then under some mild additional conditions, BvM is valid in below sense:

**Theorem C.1** (Semiparametric BvM Theorem). *Let  $W_n$ ,  $w_n$  and  $\psi$ ,  $\psi_1$ ,  $\psi_2$  as defined above. Furthermore, denote*

$$f_t = f - \frac{t}{\sqrt{n}} \left( \psi_1 + \frac{1}{2} \psi_2 (f - f_0) \right) - \frac{t}{2n} \psi_2 (w_n)$$

and

$$\hat{\psi} = \psi(f_0) + \frac{W_n(\psi_1)}{\sqrt{n}} + \frac{1}{2} \frac{\langle w_n, \psi_2(w_n) \rangle_n}{n}, \quad V_{0,n} = \left\| \psi_1 - \frac{1}{2} \frac{\psi_2(w_n)}{\sqrt{n}} \right\|_n^2$$

Then the moment generating function (MGF) of  $\sqrt{n}(\psi(f) - \hat{\psi})$  under posterior distribution  $\Pi_n$  evaluated at the set  $A_n$  such that  $\Pi_n(A_n) = 1$  can be written as:

$$E_n(e^{t\sqrt{n}(\psi(f) - \hat{\psi})} | A_n) = e^{o_p(1) + t^2 V_{0,n}/2} * \mathcal{I}_n, \quad \text{where } \mathcal{I}_n = \frac{\int_{A_n} e^{l_n(f_t) - l_n(f_0)} d\Pi(f | \mathcal{W})}{\int_{A_n} e^{l_n(f) - l_n(f_0)} d\Pi(f | \mathcal{W})}$$

Moreover, if  $V_{0,n} = \|\psi_1\|_n^2 + o_p(1)$  and  $\mathcal{I}_n = o_p(1)$ , then the posterior distribution  $\sqrt{n}(\psi(f) - \hat{\psi}_p)$  is asymptotically normal with mean zero and variance  $\|\psi_1\|_n^2$ , i.e.

$$\Pi_n \left( \sqrt{n}(\psi(f) - \hat{\psi}_p) \right) \rightsquigarrow N(0, \|\psi_1\|_n^2) \quad (10)$$

*Proof.* [8], Theorem 2.1. □

Although the original theorem is stated under the scalar case, the generalization to multivariate case is possible, i.e., one need to generalize  $V_0$  to the corresponding matrix form following the definition of  $\psi_1$  [8].

## C.2 Preliminary I: Notations and Basic Setup

In this section, we set up the basic notations for showing semi-parametric BvM theorem for a general smooth functional  $\psi_p(f)$  in a nonparametric regression model. We will first verify the model likelihood  $l_n(f) = -\frac{1}{2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2$  and the functional  $\psi(f) = \langle H_p(f), f \rangle_n$  satisfy the three conditions for Theorem C.1, and by doing so, identify the expression for the technical terms  $W_n$ ,  $w_n$ ,  $\psi_1$ ,  $\psi_2$  that are relevant for deriving the asymptotic distribution of  $\sqrt{n}(\psi(f) - \hat{\psi})$ .

First verify the LAN condition (6) for model likelihood  $l_n(f)$  and derive expression for  $W_n$ . Under independent Gaussian assumption, the likelihood for nonparametric regression adopts the LAN expansion:

$$l_n(f) - l_n(f_0) = -\frac{n}{2} \|f - f_0\|_n^2 + \sqrt{n} W_n(f - f_0)$$

where  $\|f - f_0\|_n^2 = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2$ , and  $W_n$  is:

$$W_n(f) = \langle \sqrt{n} \epsilon, f \rangle_n = \frac{1}{n} \sum_{i=1}^n \sqrt{n} \epsilon_i * f(\mathbf{x}_i) \quad (11)$$

Now verify the rest of the two conditions, we consider two cases: the univariate case where  $\psi_p(f) = \left\| \frac{\partial}{\partial x_p} f \right\|_n^2$ , to be used by the univariate BvM Theorem 2, and the multivariate case  $\psi(f)_{P \times 1} = \left[ \left\| \frac{\partial}{\partial x_1} f \right\|_n^2, \dots, \left\| \frac{\partial}{\partial x_P} f \right\|_n^2 \right]^\top$ , to be used by the multivariate BvM Theorem 3..

### Univariate Case

Now verify (7) the smoothness condition for functional of interest  $\psi^c(f)$  and derive expressions for  $\psi_1$ ,  $\psi_2$ . The centered quadratic norm of gradient  $\psi_p^c(f) = \langle H_p(f), f \rangle_n - E(\langle H_p \omega, \omega \rangle_n)$  adopts the smoothness expansion:

$$\psi_p^c(f) - \psi_p^c(f_0) = \langle \psi_1, f - f_0 \rangle_n + \frac{1}{2} \langle \psi_2(f - f_0), f - f_0 \rangle_n + o_p(\sqrt{n}),$$

in which  $\psi_1, \psi_2$  take the form:

$$\psi_1 = 2H_p(f_0), \quad \psi_2(f) = 2H_p(f),$$

where  $H_p = D_p^\top D_p$  for  $D_p : f \rightarrow \frac{\partial}{\partial x_p} f$  the differentiation operator and  $D_p^\top$  the adjoint of  $D_p$ . Given data  $\{\mathbf{x}_i, y_i\}_{i=1}^n$ , recall the definition of  $\Phi$ ,  $\partial_p \Phi$  (Section A) and denote  $\Phi_{K \times n}^+$  the generalized inverse of  $\Phi$ , the operator  $D_p$  can be evaluated in matrix form as  $\mathbf{D}_p = \partial_p \Phi \Phi^+$ , and  $D_p(f)$  can be evaluated as  $\mathbf{D}_p \mathbf{f} = (\partial_p \Phi \Phi^+) \Phi \beta = \partial_p \Phi \beta$  for  $f \in \mathcal{F}(L, K, B)$ . Correspondingly, the operator  $H_p$  adopts matrix representation  $\mathbf{H}_p = \mathbf{D}_p^\top \mathbf{D}_p = (\Phi^+)^\top \partial_p \Phi^\top \partial_p \Phi \Phi^+$ , such that  $\langle H_p(f), f \rangle = (\mathbf{H}_p \Phi \beta)^\top \Phi \beta = (\partial_p \Phi \beta)^\top (\partial_p \Phi \beta)$ .

Finally, for the decomposition  $W_n(f) = \langle \omega, f \rangle_n + \Delta_n(f)$ , we will define  $\omega = P_{\mathcal{F}}(\epsilon)$  the projection of  $\epsilon$  to  $\mathcal{F}$ , and  $\Delta_n(f) = \langle P_{\mathcal{F}}^\perp(\epsilon), f \rangle$ . Given observations  $\{\mathbf{x}_i, y_i\}_{i=1}^n$ , the projection operator  $P_{\mathcal{F}}^\perp$  can be evaluated by computing the projection matrix  $\mathbf{P}_{\mathcal{F}} = \mathbf{U} \mathbf{U}^\top$  and  $\mathbf{P}_{\mathcal{F}}^\perp = \mathbf{I} - \mathbf{U} \mathbf{U}^\top$  for  $\Phi_{n \times K} = \mathbf{U}_{n \times K} \mathbf{D}_{K \times K} \mathbf{V}_{K \times K}^\top$ . By noticing that  $\mathbf{P}_{\mathcal{F}}$  is a rank  $K$  matrix, it is then easy to see that the two conditions (8) and (9) are satisfied since  $\|\omega\|_n = O(K) \lesssim O(\sqrt{n})$  and  $P_{\mathcal{F}}^\perp(\epsilon)$  is orthogonal to  $\psi_2(f - f_0) \in \mathcal{F}_{\mathcal{W}}$ .

As an aside, we note that due to the existence of the bias term at the output layer, the actual feature matrix is  $\Phi_1 = [\mathbf{1}, \Phi]$ . However, this does not impact the expression of  $\mathbf{D}_p$  or  $\mathbf{H}_p$ , since  $\mathbf{D}_p = \partial_p \Phi_1 \Phi_1^+ = [\mathbf{0}, \partial_p \Phi][(\mathbf{1}^+)^\top, (\Phi^+)^\top]^\top = \partial_p \Phi \Phi^+$  where  $\mathbf{1}_{n \times 1}^+$  is a vector that is orthogonal to  $\mathbf{1}_{n \times 1}$  and  $\Phi_{n \times K}^+$ .

### Multivariate Case

Now we consider the multivariate case for  $\psi(f) = [ \|\frac{\partial}{\partial x_1} f\|_n^2, \dots, \|\frac{\partial}{\partial x_P} f\|_n^2 ]^\top$  by changing  $\psi_1$  and  $\psi_2$  to their multivariate counterparts. Specifically, define the operator  $H$ :

$$H(f) = [H_1(f), \dots, H_P(f)]^\top,$$

where  $H_p = D_p^\top D_p$  for  $D_p : f \rightarrow \frac{\partial}{\partial x_p} f$ , so that the functional of interest  $\psi^c(f)$  can still be expressed as  $\psi^c(f) = \langle H(f), f \rangle_n - E(\langle H(\omega), \omega \rangle_n)$ .

Now verify (7) the smoothness condition for functional of interest  $\psi(f)$  and derive expressions for  $\psi_1, \psi_2$ . The centered quadratic norm of gradient  $\psi^c(f)$  adopts the smoothness expansion:

$$\psi(f) - \psi(f_0) = \langle \psi_1, f - f_0 \rangle_n + \frac{1}{2} \langle \psi_2(f - f_0), f - f_0 \rangle_n + o_p(\sqrt{n})$$

in which  $\psi_1$  and  $\psi_2$  are two  $P \times 1$  vectors that take the form:

$$\psi_1 = 2H(f_0), \quad \psi_2(f) = 2H(f)$$

Given data  $\{\mathbf{x}_i, y_i\}_{i=1}^n$ , recall that the operator  $D_p$  can be evaluated in matrix form as  $\mathbf{D}_p = \partial_p \Phi \Phi^+$  and  $H_p$  can be evaluated as  $\mathbf{H}_p = \mathbf{D}_p^\top \mathbf{D}_p = \Phi^{+\top} \partial_p \Phi^\top \partial_p \Phi \Phi^+$ . Correspondingly,  $H(f)$  adopts matrix representation  $\mathbf{H} \mathbf{f}_{n \times P} = [(\mathbf{H}_1 \mathbf{f})_{n \times 1}, \dots, (\mathbf{H}_P \mathbf{f})_{n \times 1}] = [\Phi^{+\top} \partial_1 \Phi^\top \partial_1 \Phi \beta, \dots, \Phi^{+\top} \partial_P \Phi^\top \partial_P \Phi \beta]$ , such that  $\langle H_p(f), f \rangle_n$  is evaluated as  $\psi_{P \times 1} = (\mathbf{H} \mathbf{f}_{n \times P})^\top \mathbf{f}_{n \times 1} = [ \|\partial_1 \Phi \beta\|_2^2, \dots, \|\partial_P \Phi \beta\|_2^2 ]^\top$ .

### C.3 Preliminary II: Proof Strategy and Preliminary Theorems

Recall that under a deep ReLU neural network, the prior distribution adopts a conditional Gaussian process representation (Section A.3):

$$\Pi(f, \mathcal{W}) = \Pi(f|\mathcal{W})\Pi(\mathcal{W}) = GP(f|0, k_{\mathcal{W}})\Pi(\mathcal{W}).$$

This decomposition suggests that a neural network model can be treated as a Gaussian process with an adaptive kernel function  $k_{\mathcal{W}}$ , whose hyperparameters  $\mathcal{W}$  follows a prior distribution  $\Pi(\mathcal{W})$ .

Consequently, we use a two-step strategy to show BvM phenomenon for ReLU network:



- **Step 1**, fix hidden weight  $\mathcal{W}$  and show BvM phenomenon hold for  $GP(f|0, k_{\mathcal{W}})$ . This essentially corresponding to performing Bayesian inference for a *randomized neural network* whose hidden weights are sampled *a priori* from certain fixed distribution [16]. Then
- **Step 2**, we show that such BvM phenomenon for  $\Pi_n(f|\mathcal{W})$  still holds under the posterior distribution of hidden weights  $\mathcal{W} \sim \Pi_n(\mathcal{W})$ .

Theorem C.2 establishes **Step 1**. Notice that in the fixed- $\mathcal{W}$  case,  $f \in \mathcal{F}$  follows an exact GP with effective model dimension  $K$  (i.e. the rank of the kernel matrix), for whom the BvM phenomenon are known to hold under suitable regularity conditions [6, 7]. Therefore it is expected that BvM to hold for randomized neural network  $f \in \mathcal{F}_{\mathcal{W}}$ , provided  $K$  does not grow too fast with respect to  $n$  (i.e. Assumption 1) and the functional  $\psi_p(f)$  is sufficiently smooth (i.e. satisfying (7)), which is true for  $\psi_p(f) = \|\frac{\partial}{\partial x_p} f\|_n^2$ :

**Theorem C.2** (Bernstein-von Mises (BvM) for  $\psi_p^c$ , Fixed Hidden Weights). *For  $f \in \mathcal{F}_{\mathcal{W}}(L, W, S, B)$  a deep ReLU network with hidden weight fixed to  $\mathcal{W}$ , denoting  $f_{0, \mathcal{W}}$  the projection of  $f_0$  to  $\mathcal{F}_{\mathcal{W}}$ , and assume the posterior distribution  $\Pi_n(f|\mathcal{W})$  contracts around  $f_{0, \mathcal{W}}$  at rate  $\epsilon_n$ . Denote  $D_{\mathcal{W}, p} : f \rightarrow \frac{\partial}{\partial x_p} f$  the differentiation operator in  $\mathcal{F}_{\mathcal{W}}$ , and  $H_{\mathcal{W}, p} = D_{\mathcal{W}, p}^\top D_{\mathcal{W}, p}$  the corresponding self-adjoint operator. For  $\omega_{\mathcal{W}} = Proj_{\mathcal{F}_{\mathcal{W}}}(\epsilon)$  the projection of  $\epsilon$  to  $\mathcal{F}_{\mathcal{W}}$ , define:*

$$\hat{\psi}_{\mathcal{W}, p} = \|D_{\mathcal{W}, p}(f_0 + \omega_{\mathcal{W}})\|_n^2 = \psi_{\mathcal{W}, p}(f_{0, \mathcal{W}}) + 2\langle H_{\mathcal{W}, p} f_{0, \mathcal{W}}, \omega_{\mathcal{W}} \rangle_n + \langle H_{\mathcal{W}, p} \omega_{\mathcal{W}}, \omega_{\mathcal{W}} \rangle_n, \quad (12)$$

Define  $\hat{\psi}_{\mathcal{W}, p}^c = \hat{\psi}_{\mathcal{W}, p} - \hat{\eta}_{\mathcal{W}, n}$  where  $\hat{\eta}_{\mathcal{W}, n} = \text{tr}(\hat{H}_{\mathcal{W}, p})/n$ . Then  $\hat{\psi}_{\mathcal{W}, p}^c$  is an unbiased estimator of  $\psi_{\mathcal{W}, p}(f_0)$ , and the posterior distribution for  $\psi_{\mathcal{W}, p}^c(f)$  is asymptotically normal surrounding  $\hat{\psi}_{\mathcal{W}, p}^c$ , i.e.

$$\Pi\left(\sqrt{n}(\psi_{\mathcal{W}, p}^c(f) - \hat{\psi}_{\mathcal{W}, p}^c) \mid \{\mathbf{x}_i, y_i\}_{i=1}^n, \mathcal{W}\right) \rightsquigarrow N(0, 4\|H_{\mathcal{W}, p} f_{0, \mathcal{W}}\|_n^2), \quad (13)$$

The proof is delayed to Section D.1. It should be stressed that both operators  $D_{\mathcal{W}, p}$  and  $H_{\mathcal{W}, p}$  are defined strictly with respect to  $\mathcal{F}_{\mathcal{W}}$ , such that given data, the operator  $D_{\mathcal{W}, p}$  is estimated in matrix form as  $\hat{D}_{\mathcal{W}, p} = \partial_p \Phi_{\mathcal{W}} \Phi_{\mathcal{W}}^+$ , and  $H_{\mathcal{W}, p}$  is evaluated as  $\hat{H}_{\mathcal{W}, p} = (\Phi_{\mathcal{W}}^+)^{\top} \partial_p \Phi_{\mathcal{W}} \partial_p \Phi_{\mathcal{W}} \Phi_{\mathcal{W}}^+$ . In comparison, the original  $D_p$  and  $H_p$  defined Section C.2 are with respect to the optimal solution  $f_0 \in \mathcal{F}$ .

#### C.4 Proof for Theorem 2

We now prove Theorem 2, which establishes **Step 2** of the proof strategy outlined in Section C.3.

Our goal is to show that the BvM phenomenon in Theorem C.2 still holds under the *adaptive* case (i.e.  $\mathcal{W}$  is not fixed but follows the posterior distribution  $\Pi_n(\mathcal{W})$ ), and furthermore, the whole posterior distribution of  $\sqrt{n}(\psi_p^c(f) - \hat{\psi}_p^c(f))$  converges to  $N(0, 4\|H_p f_0\|_n^2)$  where  $H_p$  is defined with respect to the optimal solution  $f_0 \in \mathcal{F}$ .

*Proof.* Our goal is to show:

$$\Pi\left(\sqrt{n}(\psi_p^c(f) - \hat{\psi}_p^c) \mid \{\mathbf{x}_i, y_i\}_{i=1}^n\right) \rightsquigarrow N(0, 4\|H_p f_0\|_n^2).$$

First notice that by Theorem C.2, the asymptotic distribution of the marginal posterior distribution can be represented as a mixture of Gaussian:

$$\begin{aligned} \Pi_n\left(\sqrt{n}(\psi_p^c(f) - \hat{\psi}_p^c) \leq z\right) &= \int_{\mathcal{W}} \Pi_n\left(\sqrt{n}(\psi_p^c(f) - \hat{\psi}_p^c) \leq z \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) \\ &= \int_{\mathcal{W}} \Pi_n\left(\sqrt{n}(\psi_p^c(f) - \hat{\psi}_{\mathcal{W}, p}^c) \leq z + \sqrt{n}(\hat{\psi}_p^c - \hat{\psi}_{\mathcal{W}, p}^c) \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) \\ &= \int_{\mathcal{W}} \Phi\left((z + \sqrt{n}(\hat{\psi}_p^c - \hat{\psi}_{\mathcal{W}, p}^c)) / \sqrt{V_{\mathcal{W}, 0}} \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) \end{aligned} \quad (14)$$

where the last line follows from Theorem C.2, where  $V_{\mathcal{W},0} = 4\|H_{\mathcal{W},p}f_{\mathcal{W},0}\|_n^2$  and  $\Phi$  is the standard Gaussian cumulative distribution function (CDF).

Clearly, for BvM to hold in the case of (17), it is sufficient to show below two conditions [8]:

$$|V_{\mathcal{W},0} - V_0| = o_p(1), \quad \sqrt{n}|\hat{\psi}_p^c - \hat{\psi}_{\mathcal{W},p}^c| = o_p(1). \quad (15)$$

The first condition follows from the continuous mapping theorem for  $V(H_p f_0) = 2\|H_p f_0\|_n^2$ , along with the fact that :

$$\begin{aligned} \|H_{\mathcal{W},p}f_{\mathcal{W},0} - H_p f_0\|_n &\leq \|(H_{\mathcal{W},p} - H_p)f_0\|_n + \|H_{\mathcal{W},p}(f_{\mathcal{W},0} - f_0)\|_n \\ &= O(\|H_{\mathcal{W},p} - H_p\|_n) + O(\|f_{\mathcal{W},0} - f_0\|_n), \\ &= O\left(\frac{1}{\sqrt{n}}\|H_{\mathcal{W},p} - H_p\|_F\right) + o_p(1) \\ &= O\left(\frac{K}{\sqrt{n}}\right) + o_p(1) = o_p(1), \end{aligned}$$

where the first equality follows from the boundedness of  $\|f_0\|_\infty$  and  $\|H_{\mathcal{W},p}\|_\infty$  (by assumption in main article and also Proposition 3), the second equality follows from the definition of  $\|\cdot\|_n$  for matrix and the fact about posterior concentration of  $\|f - f_0\|_n^2$  in the statement of Theorem 2. The last line follows since  $\|\mathbf{H}_p\|_F = O(K)$  by Proposition 3 and the assumption that  $K = o_p(\sqrt{n})$  (Assumption 1 in the main article).

The second condition in (15) is the important *no-bias* condition which ensures that under  $\mathcal{W} \sim \Pi_n(\mathcal{W})$ , all the conditional posterior  $\psi_p^c|\mathcal{W}$  converges toward the same target  $\hat{\psi}_p^c$  [8, 19]. Recall that  $\hat{\psi}_p^c = \psi_p(f_0) + 2\langle H_p f_0, \omega \rangle_n + \langle H_p \omega, \omega \rangle_n - E(\langle H_p \omega, \omega \rangle_n)$ , then the second condition can be written as:

$$\begin{aligned} \sqrt{n}|\hat{\psi}_p^c - \hat{\psi}_{\mathcal{W},p}^c| &\leq \sqrt{n}|\psi_p(f_0) - \psi_{\mathcal{W},p}(f_{\mathcal{W},0})| + 2\sqrt{n}|\langle H_p f_0, \omega \rangle_n - \langle H_{\mathcal{W},p}f_{\mathcal{W},0}, \omega_{\mathcal{W}} \rangle_n| + \\ &\quad \sqrt{n}|\langle H_{\mathcal{W},p}\omega_{\mathcal{W}}, \omega_{\mathcal{W}} \rangle_n - \langle H_p \omega, \omega \rangle_n| + \sqrt{n}|E(\langle H_{\mathcal{W},p}\omega_{\mathcal{W}}, \omega_{\mathcal{W}} \rangle_n) - E(\langle H_p \omega, \omega \rangle_n)|, \end{aligned} \quad (16)$$

where all four terms are  $o_p(1)$  since they are all  $O_p(K/\sqrt{n})$  and the model dimension  $K$  is by assumption not too large (i.e.  $K = o(n^{1/2})$ ). We delay the detailed arguments to the end of the proof.

Consequently, since both conditions in (15) are satisfied, the expression in (17) converge uniquely to a normal distribution under the posterior distribution  $\Pi_n(\mathcal{W})$ , i.e.,

$$\begin{aligned} \Pi_n\left(\sqrt{n}(\psi_p^c(f) - \hat{\psi}_p^c) \leq z\right) &= \int_{\mathcal{W}} \Phi\left((z + \sqrt{n}(\hat{\psi}_p^c - \hat{\psi}_{\mathcal{W},p}^c))/\sqrt{V_{\mathcal{W},0}} \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) \\ &= \int_{\mathcal{W}} \Phi\left((z + o_p(1))/\sqrt{V_0 + o_p(1)} \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) \\ &= \Phi(z/\sqrt{V_0}), \quad \text{where} \quad V_0 = 4\|H_p f_0\|_n^2 \end{aligned} \quad (17)$$

which implies the statement of interest:

$$\Pi\left(\sqrt{n}(\psi_p^c(f) - \hat{\psi}_p^c) \mid \{\mathbf{x}_i, y_i\}_{i=1}^n\right) \rightsquigarrow N(0, 4\|H_p f_0\|_n^2).$$

We are only left to show that all four terms in the expression (16) are  $o_p(1)$ . Specifically, recall that  $H_p = D_p^\top D_p$  such that  $\langle H_p a, b \rangle_n = \langle D_p a, D_p b \rangle_n$  for any  $a, b \in \mathcal{F}$ , then:

- **First Term:** Recall  $\psi_p(f_0) = \langle H_p f_0, f_0 \rangle_n = \|D_p f_0\|_n^2$ , then the first term can be expressed as:

$$\begin{aligned}
\sqrt{n}|\psi_p(f_0) - \psi_{\mathcal{W},p}(f_{\mathcal{W},0})| &= \sqrt{n} \left| \|D_p f_0\|_n^2 - \|D_{\mathcal{W},p} f_{\mathcal{W},0}\|_n^2 \right| \\
&\leq \sqrt{n} \left( \|D_p\|_n^2 \|f_0\|_n^2 + \|D_{\mathcal{W},p}\|_n^2 \|f_{\mathcal{W},0}\|_n^2 \right) \\
&= \sqrt{n} \left( O_p(\|D_p\|_n^2) + O_p(\|D_{\mathcal{W},p}\|_n^2) \right) = \frac{1}{\sqrt{n}} \left( O_p(\|\mathbf{D}_p\|_F^2) + O_p(\|\mathbf{D}_{\mathcal{W},p}\|_F^2) \right) \\
&= O\left(\frac{K}{\sqrt{n}}\right) = o_p(1)
\end{aligned}$$

where on the third line, the first equality follows since  $f_0$  and  $f_{\mathcal{W},0}$  are both bounded, the second equality follows by the definition of the matrix Euclidean norm  $\|\mathbf{M}\|_n^2 = \frac{1}{n} \sum_{i,j} \mathbf{M}_{i,j}^2 = \frac{1}{n} \|\mathbf{M}\|_F^2$ . On the last line, the first equality follows by  $\|\mathbf{D}_{\mathcal{W},p}\|_F^2 = \text{tr}(\mathbf{H}_{\mathcal{W},p}) = O(K)$  due to Proposition 3, and the second equality follows by Assumption  $K = o_p(n^{1/2})$ .

- **Second Term:** Similarly, the second term can be expressed as:

$$\begin{aligned}
\sqrt{n}|\langle H_p f_0, \omega \rangle_n - \langle H_{\mathcal{W},p} f_{\mathcal{W},0}, \omega_{\mathcal{W}} \rangle_n| &= \sqrt{n} |\langle D_p f_0, D_p \omega \rangle_n - \langle D_{\mathcal{W},p} f_{\mathcal{W},0}, D_{\mathcal{W},p} \omega_{\mathcal{W}} \rangle_n| \\
&\leq \sqrt{n} \left( \|D_p\|_n^2 \|f_0\|_n \|\omega\|_n + \|D_{\mathcal{W},p}\|_n^2 \|f_{\mathcal{W},0}\|_n \|\omega_{\mathcal{W}}\|_n \right) \\
&= \sqrt{n} \left( O_p(\|D_p\|_n^2) + O_p(\|D_{\mathcal{W},p}\|_n^2) \right) \\
&= O_p\left(\frac{K}{\sqrt{n}}\right) = o_p(1)
\end{aligned}$$

where the equality on the third line follows from the fact that  $f_{0,\mathcal{W}}$  is bounded and  $\omega = \text{Proj}_{\mathcal{F}}(\epsilon)$  is a random variable with bounded variance. The rest of the equalities follow similarly as those in the First term.

- **Third and Fourth Terms** are similar to the first term except for  $f_0$  is replaced by  $\omega$ . As a result:

$$\begin{aligned}
\sqrt{n}|\langle H_{\mathcal{W},p} \omega_{\mathcal{W}}, \omega_{\mathcal{W}} \rangle_n - \langle H_p \omega, \omega \rangle_n| &= \sqrt{n} \left| \|D_{\mathcal{W},p} \omega_{\mathcal{W}}\|_n^2 - \|D_p \omega\|_n^2 \right| \\
&= \sqrt{n} \left( O_p(\|D_p\|_n^2) + O_p(\|D_{\mathcal{W},p}\|_n^2) \right) \\
&= O_p\left(\frac{K}{\sqrt{n}}\right) = o_p(1) \\
\sqrt{n}|E(\langle H_{\mathcal{W},p} \omega_{\mathcal{W}}, \omega_{\mathcal{W}} \rangle_n) - E(\langle H_p \omega, \omega \rangle_n)| &= \sqrt{n} O_p \left( |\langle H_{\mathcal{W},p} \omega_{\mathcal{W}}, \omega_{\mathcal{W}} \rangle_n - \langle H_p \omega, \omega \rangle_n| \right) \\
&= O_p\left(\frac{K}{\sqrt{n}}\right) = o_p(1)
\end{aligned}$$

□

### C.5 Proof for Theorem 3

To prepare for the multivariate BvM theorem (Theorem 3), below theorem extends Theorem C.2 to the multivariate case for  $\psi(f) = [\|\frac{\partial}{\partial x_1} f\|_n^2, \dots, \|\frac{\partial}{\partial x_p} f\|_n^2]^\top$ .

**Theorem C.3** (Multivariate BvM for  $\psi^c$ , Fixed Hidden Weights). *For  $f \in \mathcal{F}_{\mathcal{W}}(L, W, S, B)$ , assuming the posterior distribution  $\Pi_n(f)$  contracts around  $f_{0,\mathcal{W}}$  at rate  $\epsilon_n$ . For  $\omega_{\mathcal{W}} = \text{Proj}_{\mathcal{F}_{\mathcal{W}}}(\epsilon)$ , denote  $\hat{\psi}_{\mathcal{W}}^c = [\hat{\psi}_{\mathcal{W},1}^c, \dots, \hat{\psi}_{\mathcal{W},P}^c]$  for  $\hat{\psi}_{\mathcal{W},p}^c$  as defined in Theorem C.2.*

*Then  $\hat{\psi}_{\mathcal{W}}^c$  is an unbiased and efficient estimator of  $\psi_{\mathcal{W}}(f_0)$ , and the posterior distribution for  $\psi_{\mathcal{W}}^c(f)$  asymptotically converge toward a multivariate normal distribution surrounding  $\hat{\psi}_{\mathcal{W}}^c$ , i.e.*

$$\Pi \left( \sqrt{n}(\psi_{\mathcal{W}}^c(f) - \hat{\psi}_{\mathcal{W}}^c) \middle| \{\mathbf{x}_i, y_i\}_{i=1}^n \right) \rightsquigarrow MVN(0, V_{0,\mathcal{W}}), \quad (18)$$

where  $V_{0,\mathcal{W}}$  is a  $P \times P$  matrix such that  $(V_{0,\mathcal{W}})_{p_1,p_2} = 4\langle H_{\mathcal{W},p_1} f_{0,\mathcal{W}}, H_{\mathcal{W},p_2} f_{0,\mathcal{W}} \rangle_n$ .

The proof is in Section D.2. We are now ready to prove Theorem 3.

*Proof.* Our goal is to show:

$$\Pi\left(\sqrt{n}(\psi^c(f) - \hat{\psi}^c) \mid \{\mathbf{x}_i, y_i\}_{i=1}^n\right) \rightsquigarrow N(0, V_0). \quad (19)$$

We show the convergence of above multivariate normal distribution by invoking the Cramér-Wold theorem [5], i.e. we show that for all  $t \in \mathbb{R}^P$ :

$$\Pi\left(\sqrt{n}(t^\top \psi^c(f) - t^\top \hat{\psi}^c) \mid \{\mathbf{x}_i, y_i\}_{i=1}^n\right) \rightsquigarrow N(0, t^\top V_0 t).$$

Similar to the scalar case, the asymptotic distribution of the marginal posterior distribution can be represented as a mixture of Gaussian:

$$\begin{aligned} \Pi_n\left(\sqrt{n}(t^\top \psi^c(f) - t^\top \hat{\psi}^c) \leq z\right) &= \int_{\mathcal{W}} \Pi_n\left(t^\top \sqrt{n}(\psi^c(f) - t^\top \hat{\psi}_p^c) \leq z \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) \\ &= \int_{\mathcal{W}} \Pi_n\left(\sqrt{n}(t^\top \psi^c(f) - t^\top \hat{\psi}_{\mathcal{W},p}^c) \leq z + \sqrt{n}(t^\top \hat{\psi}_p^c - t^\top \hat{\psi}_{\mathcal{W},p}^c) \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) \\ &= \int_{\mathcal{W}} \Phi\left((z + \sqrt{n}(t^\top \hat{\psi}_p^c - t^\top \hat{\psi}_{\mathcal{W},p}^c)) / \sqrt{t^\top V_{\mathcal{W},0} t} \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) + o_p(1) \end{aligned} \quad (20)$$

where the last line follows from Theorem C.3, where  $(V_{0,\mathcal{W}})_{p_1,p_2} = 4\langle H_{\mathcal{W},p_1} f_{0,\mathcal{W}}, H_{\mathcal{W},p_2} f_{0,\mathcal{W}} \rangle_n$  and  $\Phi$  is the standard Gaussian CDF.

Therefore it is sufficient to show below two conditions [8]:

$$t^\top (V_{\mathcal{W},0} - V_0)t = o_p(1), \quad \sqrt{n}|t^\top \hat{\psi}^c - t^\top \hat{\psi}_{\mathcal{W}}^c| = o_p(1). \quad (21)$$

However, both conditions follows from the coordinate-wise convergence  $\sqrt{n}|\hat{\psi}_p^c - \hat{\psi}_{\mathcal{W},p}^c| = o_p(1)$  and  $\|H_{\mathcal{W},p} f_{\mathcal{W},0} - H_p f_0\|_n = o_p(1)$ , which were established in the proof for Theorem 2.

Consequently, the expression in (20) converge uniquely to a normal distribution under the posterior distribution  $\Pi_n(\mathcal{W})$ , i.e.,

$$\begin{aligned} \Pi_n\left(\sqrt{n}(t^\top \psi^c(f) - t^\top \hat{\psi}^c) \leq z\right) &= \int_{\mathcal{W}} \Phi\left((z + \sqrt{n}(t^\top \hat{\psi}_p^c - t^\top \hat{\psi}_{\mathcal{W},p}^c)) / \sqrt{t^\top V_{\mathcal{W},0} t} \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) + o_p(1) \\ &= \int_{\mathcal{W}} \Phi\left((z + o_p(1)) / \sqrt{t^\top V_0 t + o_p(1)} \mid \mathcal{W}\right) d\Pi_n(\mathcal{W}) + o_p(1) \\ &= \Phi(z / \sqrt{t^\top V_0 t}) + o_p(1), \quad \text{where} \quad (V_0)_{p_1,p_2} = 4\langle H_{p_1} f_0, H_{p_2} f_0 \rangle_n \end{aligned}$$

which implies the statement of interest in (19). □

## D Additional Proofs

### D.1 Proof for Theorem C.2

*Proof.* Our goal is to show asymptotic normality

$$\Pi_n(\sqrt{n}(\psi_{\mathcal{W},p}^c(f) - \hat{\psi}_{\mathcal{W},p}^c)|\mathcal{W}) \rightsquigarrow N(0, \|\psi_1\|_n^2).$$

First derive the expressions for  $f_t$ ,  $\hat{\psi}_{\mathcal{W},p}$ ,  $V_{0,n}$  and  $\mathcal{I}_n$ . The expression for  $f_t$  is:

$$\begin{aligned} f_t &= f - \frac{t}{\sqrt{n}} \left( \psi_1 + \frac{1}{2} \psi_2(f - f_{0,\mathcal{W}}) \right) - \frac{t}{2n} \psi_2(\omega_{\mathcal{W}}) \\ &= \left( I - \frac{t}{\sqrt{n}} * H_p \right) (f) - \frac{t}{\sqrt{n}} * H_p(f_{0,\mathcal{W}} + \omega_{\mathcal{W}}). \end{aligned} \quad (22)$$

where  $I : f \rightarrow f$  is the identity operator. The expressions for  $\hat{\psi}_{\mathcal{W},p}$  and  $V_{0,n}$  are (recall  $\omega_{\mathcal{W}} = Proj_{\mathcal{F}_{\mathcal{W}}}(\epsilon)$ ):

$$\begin{aligned} \hat{\psi}_{\mathcal{W},p} &= \psi_{\mathcal{W},p}(f_{0,\mathcal{W}}) + \frac{W_n(\psi_1)}{\sqrt{n}} + \frac{1}{2} \frac{\langle \omega_{\mathcal{W}}, \psi_2(\omega_{\mathcal{W}}) \rangle_n}{n} \\ &= \psi_{\mathcal{W},p}(f_{0,\mathcal{W}}) + 2 \langle H_{\mathcal{W},p} f_{0,\mathcal{W}}, \omega_{\mathcal{W}} \rangle_n + \langle H_{\mathcal{W},p} \omega_{\mathcal{W}}, \omega_{\mathcal{W}} \rangle_n, \\ V_{0,n} &= \left\| \psi_1 - \frac{1}{2} \frac{\psi_2(\omega_{\mathcal{W}})}{\sqrt{n}} \right\|_n^2 = 4 * \left\| H_{\mathcal{W},p} f_{0,\mathcal{W}} + o_p(1) \right\|_n^2, \end{aligned}$$

and the expression for  $\mathcal{I}_n$  is:

$$\mathcal{I}_n = \frac{\int_{A_n} e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}{\int_{A_n} e^{l_n(f) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}.$$

It then follows from the Theorem C.1 that

$$E_n(e^{t\sqrt{n}(\psi_{\mathcal{W},p}^c(f) - \hat{\psi}_{\mathcal{W},p}^c)} | A_n, \mathcal{W}) = e^{o_p(1) + t^2 V_{0,n}/2} * \mathcal{I}_n. \quad (23)$$

Therefore to show asymptotic normality  $\Pi_n(\sqrt{n}(\psi_{\mathcal{W},p}^c(f) - \hat{\psi}_{\mathcal{W},p}^c)|\mathcal{W}) \rightsquigarrow N(0, \|\psi_1\|_n^2)$ , we only need to show  $\mathcal{I}_n \rightarrow 1$  for some suitable choice of  $A_n$  so that  $E_n(e^{t\sqrt{n}(\psi_{\mathcal{W},p}^c(f) - \hat{\psi}_{\mathcal{W},p}^c)} | A_n, \mathcal{W}) \rightarrow e^{t^2 V_0/2}$ . To this end, we consider the standard choice  $A_n = \left\{ f \mid \|f - f_{0,\mathcal{W}}\|_n^2 \leq M_n \epsilon_n \right\}$  for some sufficiently large  $M_n \rightarrow \infty$  and  $\epsilon_n$  the given rate of posterior convergence.

To analyze the asymptotic behavior of  $\mathcal{I}_n$ , analogous to the Theorem 3.1 of [8], we consider the asymptotically equivalent term  $\mathcal{I}'_n$ :

$$\mathcal{I}'_n = \frac{\int_{f \in A_n} e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}{\int e^{l_n(f) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}$$

where the denominator  $\int_{A_n}$  in  $\mathcal{I}_n$  is replaced by  $\int$  (i.e., integration over the entire space).  $\mathcal{I}'_n$  is asymptotically equivalent to  $\mathcal{I}_n$  since  $\Pi_n(A_n) = 1 + o_p(1)$ .

Notice that  $\mathcal{I}'_n$  can be written as:

$$\mathcal{I}'_n = \mathcal{J}_n * \mathcal{K}_n = \frac{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}{\int e^{l_n(f) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})} * \frac{\int_{f \in A_n} e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}$$

where we have denoted:

$$\mathcal{J}_n = \frac{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}{\int e^{l_n(f) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}, \quad \mathcal{K}_n = \frac{\int_{f \in A_n} e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}. \quad (24)$$

Therefore in order to show  $\mathcal{I}_n \rightarrow 1$ , it is sufficient to show  $\mathcal{J}_n \rightarrow 1$  and  $\mathcal{K}_n \rightarrow 1$ .

We show  $\mathcal{J}_n \rightarrow 1$  by first performing change of variables  $f \rightarrow f_t$  on the numerator, and then analyze the asymptotic behavior of the resulting expression. To this end, notice that the conditional prior

$\Pi(f|\mathcal{W})$  is a Gaussian process prior with kernel function  $k_{\mathcal{W}}(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^K \phi_k(\mathbf{x}|\mathcal{W})\phi_k(\mathbf{x}'|\mathcal{W}) + \sigma_{b_0}^2$ , and correspondingly, the kernel matrix evaluated at  $\{\mathbf{x}_i, y_i\}_{i=1}^n$  is  $\mathbf{K}_{\mathcal{W}} = \Phi_{\mathcal{W}}\Phi_{\mathcal{W}}^\top + \sigma_{b_0}^2\mathbf{I}$  for  $\Phi_{n \times K} = \mathbf{X}(\prod_{i=1}^L \mathbf{W}_i \mathbf{S}_i)$  (see Section A.1). Therefore given observations  $\{y_i, \mathbf{x}_i\}_{i=1}^n$ , the conditional prior distribution for  $\mathbf{f} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]^\top \in \mathbb{R}^n$  is a multivariate Gaussian with covariance matrix  $\mathbf{K}_{\mathcal{W}}$ :

$$d\Pi(\mathbf{f}|\mathcal{W}) \propto |\mathbf{K}_{\mathcal{W}}|^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{f}^\top \mathbf{K}_{\mathcal{W}}^{-1} \mathbf{f}}$$

where  $|\mathbf{K}_{\mathcal{W}}|$ ,  $\mathbf{K}_{\mathcal{W}}^{-1}$  are the determinant and inverse of  $\mathbf{K}_{\mathcal{W}}$ , respectively. To perform change of variable, we denote  $\mathbf{f}_t \in \mathbb{R}^n$  as  $f_t$  in (22) evaluated at  $\{y_i, \mathbf{x}_i\}_{i=1}^n$ , we can write  $\mathbf{f}_t = \mathbf{B}_{t,n}\mathbf{f} - \frac{t}{\sqrt{n}}\mathbf{a}_n$  such that

$$\begin{aligned} d\Pi(\mathbf{f}_t|\mathcal{W}) &\propto |\mathbf{K}_{n,\mathcal{W}}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{f} + \frac{t}{\sqrt{n}}\mathbf{a}_n)^\top \mathbf{K}_{n,\mathcal{W}}^{-1}(\mathbf{f} + \frac{t}{\sqrt{n}}\mathbf{a}_n)} \\ &= d\Pi(\mathbf{f}|\mathcal{W}) * \exp\left[-\frac{1}{2}\log\left(\frac{|\mathbf{K}_{n,\mathcal{W}}|}{|\mathbf{K}_{\mathcal{W}}|}\right) - \frac{1}{2}\mathbf{f}^\top (\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1})\mathbf{f} + \frac{t}{\sqrt{n}}\mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1}\mathbf{f} - \frac{t^2}{2n}\mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1}\mathbf{a}_n\right] \end{aligned}$$

where

$$\mathbf{K}_{n,\mathcal{W}} = \mathbf{B}_{t,n}\mathbf{K}_{\mathcal{W}}\mathbf{B}_{t,n} = \mathbf{K}_{\mathcal{W}} - \frac{t}{\sqrt{n}}\left[\mathbf{K}_{\mathcal{W}}\mathbf{H}_p + \mathbf{H}_p\mathbf{K}_{\mathcal{W}}\mathbf{B}_{t,n}\right] = \mathbf{K}_{\mathcal{W}} - \frac{t}{\sqrt{n}}d\mathbf{K}_{\mathcal{W}},$$

where  $\mathbf{B}_{t,n} = \mathbf{I} - \frac{t}{\sqrt{n}}\mathbf{H}_p$ ,  $\mathbf{a}_n = \mathbf{H}_p(\mathbf{f}_{0,\mathcal{W}} + \omega_{\mathcal{W}})$ , and we have denoted  $d\mathbf{K}_{\mathcal{W}} = \mathbf{K}_{\mathcal{W}}\mathbf{H}_p + \mathbf{H}_p\mathbf{K}_{\mathcal{W}}\mathbf{B}_{t,n}$ .

Consequently, the expression of  $\mathcal{J}_n$  after the change of variable becomes:

$$\mathcal{J}_n = \frac{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} * e^{-\mathcal{L}_{t,n}} d\Pi(f_t|\mathcal{W})}{\int e^{l_n(f) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})},$$

where

$$\mathcal{L}_{t,n} = \frac{1}{2}\log\left(\frac{|\mathbf{K}_{n,\mathcal{W}}|}{|\mathbf{K}_{\mathcal{W}}|}\right) + \frac{1}{2}\mathbf{f}^\top (\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1})\mathbf{f} - \frac{t}{\sqrt{n}}\mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1}\mathbf{f} + \frac{t^2}{2n}\mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1}\mathbf{a}_n. \quad (25)$$

Therefore in order to show  $\mathcal{J}_n \rightarrow 1$ , it is sufficient to show that all terms in  $\mathcal{L}_{t,n} = o_p(1)$  are asymptotically vanishing toward zero as  $n \rightarrow \infty$ . We defer the detailed arguments to the end of this proof. If this is true, then it holds that  $\mathcal{L}_{t,n} = o_p(1)$  and

$$\mathcal{J}_n = \frac{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}}) + o_p(1)} d\Pi(f_t|\mathcal{W})}{\int e^{l_n(f) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})} \rightarrow 1. \quad (26)$$

We show  $\mathcal{K}_n \rightarrow 1$  by noticing that  $\mathcal{L}_{t,n} = o_p(1)$  implies:

$$\mathcal{K}_n = \frac{\int_{f \in A_n} e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})}{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})} = \frac{\int_{f \in A_n} e^{l_n(f_t) - l_n(f_{0,\mathcal{W}}) + o_p(1)} d\Pi(f_t|\mathcal{W})}{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}}) + o_p(1)} d\Pi(f_t|\mathcal{W})}$$

Notice that in the above expression, the event  $f \in A_n$  is equivalent to  $f_t \in A_{n,t}$  for:

$$\begin{aligned} A_{n,t} &= \left\{f_t \mid \|f - f_0\|_n^2 \leq M_n \epsilon_n\right\} \\ &= \left\{f_t \mid \|f - f_0\|_n^2 \leq M_n \epsilon_n, \|f_t - f_0\|_n^2 \leq \|f_t - f\|_n^2 + \|f - f_0\|_n^2\right\} \\ &\supseteq \left\{f_t \mid \|f_t - f_0\|_n^2 \leq \|f_t - f\|_n^2 + M_n \epsilon_n\right\} \\ &= \left\{f_t \mid \|f_t - f_0\|_n^2 \leq O(n^{-1}) + M_n \epsilon_n\right\} \\ &= \left\{f_t \mid \|f_t - f_0\|_n^2 \leq M'_n \epsilon_n\right\} = A'_n \end{aligned}$$

where on the fourth line, the equality follows since  $\|f_t - f\|_n^2 = \frac{1}{n}\|\mathbf{f}_t - \mathbf{f}\|_2^2 = \frac{t}{\sqrt{n}}\|\mathbf{H}_p\mathbf{f} + \mathbf{a}_n\|_2^2/n \leq \frac{t}{\sqrt{n}}(\|\mathbf{H}_p\|_2^2\|\mathbf{f}\|_2^2/n + \|\mathbf{a}_n\|_2^2/n) = O(n^{-\frac{1}{2}})$  since  $\|\mathbf{H}_p\|_2^2 = O_p(1)$  by Proposition 3, and the facts

that  $\|\mathbf{f}\|_2^2/n = O_p(1)$  and  $\|\mathbf{a}_n\|_2^2/n = \|\mathbf{H}_p(\mathbf{f}_0, \mathcal{W} + \omega_{\mathcal{W}})\|^2/n = O_p(1)$  due to boundedness of  $\mathbf{f}_0$  and the fact that  $\omega_{\mathcal{W}}$  is a random variable with finite variance. Also, on the last line, the equality follows since  $n^{-1} \lesssim \epsilon_n$ , i.e. the learning rate  $\epsilon_n$  is expected to be slower than  $O(n^{-1})$ . As a result, we see that since the set  $A'_n = \left\{f \mid \|f - f_0\|_n \leq M'_n \epsilon_n\right\}$  is of the same form as  $A_n$ , we have that  $\Pi_n(A'_n) \rightarrow 1 + o_p(1)$  due to the posterior convergence of  $f$  to  $f_0$ . Consequently:

$$\mathcal{K}_n = \frac{\int_{f_t \in A_n, t} e^{l_n(f_t) - l_n(f_0, \mathcal{W}) + o_p(1)} d\Pi(f_t | \mathcal{W})}{\int e^{l_n(f_t) - l_n(f_0, \mathcal{W}) + o_p(1)} d\Pi(f_t | \mathcal{W})} \geq \frac{\int_{f_t \in A'_n} e^{l_n(f_t) - l_n(f_0, \mathcal{W}) + o_p(1)} d\Pi(f_t | \mathcal{W})}{\int e^{l_n(f_t) - l_n(f_0, \mathcal{W}) + o_p(1)} d\Pi(f_t | \mathcal{W})} \rightarrow 1 \quad (27)$$

Finally, since we have shown both  $\mathcal{J}_n \rightarrow 1$  and  $\mathcal{K}_n \rightarrow 1$ , it then follows by Theorem C.1 that the posterior distribution of  $\sqrt{n}(\psi_{\mathcal{W}, p}^c(f) - \hat{\psi}_{\mathcal{W}}^c)$  converge toward a normal law with mean zero and variance  $\|\psi_1\|_n^2 = 4\|H_{\mathcal{W}, p} f_0, \mathcal{W}\|_n^2$ , i.e. we have shown the statement of the interest:

$$\Pi\left(\sqrt{n}(\psi_{\mathcal{W}, p}^c(f) - \hat{\psi}_{\mathcal{W}}^c) \mid \{\mathbf{x}_i, y_i\}_{i=1}^n, \mathcal{W}\right) \rightsquigarrow N(0, 4\|H_{\mathcal{W}, p} f_0, \mathcal{W}\|_n^2). \quad (28)$$

We are only left to show that  $\mathcal{L}_{t, n} = o_p(1)$ , i.e. all the four terms in (25) are asymptotically vanishing toward zero. We achieve this by analyzing the asymptotic behavior of the four terms one by one:

- First term:  $\log\left(\frac{|\mathbf{K}_{n, \mathcal{W}}|}{|\mathbf{K}_{\mathcal{W}}|}\right) = \log|\mathbf{K}_{n, \mathcal{W}}| - \log|\mathbf{K}_{\mathcal{W}}|$

Performing Taylor Expansion on  $\log|\mathbf{K}_{n, \mathcal{W}}|$ :<sup>1 2</sup>

$$\begin{aligned} \log|\mathbf{K}_{n, \mathcal{W}}| &= \log\left|\mathbf{K}_{\mathcal{W}} - \frac{t}{\sqrt{n}}d\mathbf{K}_{\mathcal{W}}\right| \\ &= \log|\mathbf{K}_{\mathcal{W}}| - \frac{t}{\sqrt{n}}\text{tr}(\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}}) + \frac{t^2}{n}O_p(\text{tr}(\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}})^2). \end{aligned} \quad (29)$$

where

$$\begin{aligned} \frac{t}{\sqrt{n}}\text{tr}(\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}}) &= \frac{t}{\sqrt{n}}\text{tr}\left(\mathbf{K}_{\mathcal{W}}^{-1}(\mathbf{K}_{\mathcal{W}}\mathbf{H}_{\mathcal{W}, p} + \mathbf{H}_{\mathcal{W}, p}\mathbf{K}_{\mathcal{W}} - \frac{t}{\sqrt{n}}\mathbf{H}_{\mathcal{W}, p}\mathbf{K}_{\mathcal{W}}\mathbf{H}_{\mathcal{W}, p})\right) \\ &= \frac{2t}{\sqrt{n}}\text{tr}(\mathbf{H}_{\mathcal{W}, p}) - \frac{t^2}{n}\text{tr}(\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{H}_{\mathcal{W}, p}\mathbf{K}_{\mathcal{W}}\mathbf{H}_{\mathcal{W}, p}) \\ &\leq \frac{2t}{\sqrt{n}}\text{tr}(\mathbf{H}_{\mathcal{W}, p}) \\ &= \frac{t}{\sqrt{n}}O_p(K). \end{aligned} \quad (30)$$

In the above expression, the inequality follows from the fact that  $\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{H}_{\mathcal{W}, p}\mathbf{K}_{\mathcal{W}}\mathbf{H}_{\mathcal{W}, p}$  is positive semi-definite, and the last line follows from Proposition 3, i.e.,  $\text{tr}(\mathbf{H}_{\mathcal{W}, p}) = O_p(K)$ .

By combing (29) and (30):

$$\log|\mathbf{K}_{n, \mathcal{W}}| - \log|\mathbf{K}_{\mathcal{W}}| = -\frac{t}{\sqrt{n}}\text{tr}(\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}}) + \frac{t^2}{n}O_p(\text{tr}(\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}})^2) = t * O_p\left(\frac{K}{\sqrt{n}}\right) = o_p(1). \quad (31)$$

where recall  $O_p\left(\frac{K}{\sqrt{n}}\right) = o_p(1)$  since  $K = o_p(n^{1/2})$ .

<sup>1</sup>For the log pseudo-determinant function  $f(\mathbf{X}) = \log|\mathbf{X}|$ , we can compute its gradient function as  $\nabla_{\mathbf{X}}f(\mathbf{X}) = \mathbf{X}^+$  Hessian  $H_{\mathbf{X}}f(\mathbf{X}) = \mathbf{X}^+ \otimes \mathbf{X}^+$ .

<sup>2</sup>The second-order term in Taylor expansion is  $\text{vec}(d\mathbf{K}_{\mathcal{W}})^{\top}(\mathbf{K}_{\mathcal{W}}^{-1} \otimes \mathbf{K}_{\mathcal{W}}^{-1})\text{vec}(d\mathbf{K}_{\mathcal{W}})$ . Using property of Kronecker product  $(\mathbf{C}^{\top} \otimes \mathbf{A})\text{vec}(\mathbf{B}) = \text{vec}(\mathbf{A}\mathbf{B}\mathbf{C})$ , the second-order term can also be written as  $\text{vec}(d\mathbf{K}_{\mathcal{W}})^{\top}\text{vec}(\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}}\mathbf{K}_{\mathcal{W}}^{-1}) = \text{tr}(d\mathbf{K}_{\mathcal{W}}\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}}\mathbf{K}_{\mathcal{W}}^{-1}) = \text{tr}((\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}})^2) \leq \text{tr}(\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}})^2$ , where the last inequality follows from the fact that  $\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}}$  is positive semidefinite.

- Second term:  $\mathbf{f}^\top (\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1})\mathbf{f}$

Bound this term using the Cauchy-Schwarz inequality  $\mathbf{f}^\top (\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1})\mathbf{f} \leq \|\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1}\|_\infty \|\mathbf{f}\|_\infty^2$ . Notice that by applying the inversion formula  $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}$  [12], we have  $\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1} = \frac{t}{\sqrt{n}}(\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}})\mathbf{K}_{n,\mathcal{W}}^{-1}$ , and further manipulation reveals that:

$$\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}}\mathbf{K}_{n,\mathcal{W}}^{-1} = \mathbf{K}_{\mathcal{W}}^{-1}(\mathbf{K}_{\mathcal{W}}\mathbf{H}_{\mathcal{W},p} + \mathbf{H}_{\mathcal{W},p}\mathbf{K}_{\mathcal{W}}\mathbf{B}_{t,n})(\mathbf{B}_{t,n}^+\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{B}_{t,n}^+) = \mathbf{H}_{\mathcal{W},p}\mathbf{B}_{t,n}^+\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{B}_{t,n}^+ + \mathbf{K}_{\mathcal{W}}^{-1}\mathbf{H}_{\mathcal{W},p}\mathbf{B}_{t,n}^+$$

Then, by triangular inequality, we can bound  $\|\mathbf{K}_{n,\mathcal{W}}^{-1}\|_\infty$  as:

$$\begin{aligned} \|\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1}\|_\infty &= \frac{t}{\sqrt{n}} \|\mathbf{K}_{\mathcal{W}}^{-1}d\mathbf{K}_{\mathcal{W}}\mathbf{K}_{n,\mathcal{W}}^{-1}\|_\infty \\ &\leq \frac{t}{\sqrt{n}} \left( \|\mathbf{H}_{\mathcal{W},p}\mathbf{B}_{t,n}^+\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{B}_{t,n}^+\|_\infty + \|\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{H}_{\mathcal{W},p}\mathbf{B}_{t,n}^+\|_\infty \right) \\ &= \frac{t}{\sqrt{n}} \left( \|\mathbf{B}_{t,n}^+\mathbf{H}_{\mathcal{W},p}\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{B}_{t,n}^+\|_\infty + \|\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{H}_{\mathcal{W},p}\mathbf{B}_{t,n}^+\|_\infty \right) \\ &\leq \frac{t}{\sqrt{n}} \left( \|\mathbf{H}_{\mathcal{W},p}\mathbf{K}_{\mathcal{W}}^{-1}\|_\infty \|\mathbf{B}_{t,n}^+\|_\infty^2 + \|\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{H}_{\mathcal{W},p}\|_\infty \|\mathbf{B}_{t,n}^+\|_\infty \right) \\ &= \frac{t}{\sqrt{n}} \left( \|\mathbf{H}_{\mathcal{W},p}\mathbf{K}_{\mathcal{W}}^{-1}\|_\infty + \|\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{H}_{\mathcal{W},p}\|_\infty \right) = \frac{2t}{\sqrt{n}} \|\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{H}_{\mathcal{W},p}\|_\infty \\ &\leq \frac{2t}{\sqrt{n}} \|\mathbf{K}_{\mathcal{W}}^{-1}\|_\infty * \|\mathbf{H}_{\mathcal{W},p}\|_\infty = t * O_p(n^{-1/2}), \end{aligned} \quad (32)$$

where the second equality follows since  $\mathbf{H}_{\mathcal{W},p}\mathbf{B}_{t,n}^+ = \mathbf{B}_{t,n}^+\mathbf{H}_{\mathcal{W},p}$  due to the fact that  $\mathbf{B}_{t,n}^+ = \mathbf{H}_{\mathcal{W},p}^+(\mathbf{H}_{\mathcal{W},p}^+ - \frac{t}{\sqrt{n}}\mathbf{I})^+ = (\mathbf{H}_{\mathcal{W},p}^+ - \frac{t}{\sqrt{n}}\mathbf{I})^+\mathbf{H}_{\mathcal{W},p}^+$ . The third equality follows since  $\|\mathbf{B}_{t,n}^+\|_\infty \leq \|\mathbf{B}_{t,n}^+\|_2 = \|(\mathbf{I} - \frac{t}{\sqrt{n}}\mathbf{H}_{\mathcal{W},p})^+\|_2 = 1$  due to the fact that  $\mathbf{H}_{\mathcal{W},p}$  is not full-column rank, and the last line follows by the facts that  $\|\mathbf{K}_{\mathcal{W}}^{-1}\|_\infty \leq \|\mathbf{K}_{\mathcal{W}}^{-1}\|_2 = O_p(1)$  due to  $\mathbf{K}_{\mathcal{W}} = \Phi_{\mathcal{W}}\Phi_{\mathcal{W}}^\top + \sigma_{b_0}^2\mathbf{I}$  and  $\|\mathbf{H}_{\mathcal{W},p}\|_\infty \leq \|\mathbf{H}_{\mathcal{W},p}\|_2 = o_p(1)$  due to Proposition 3.

Finally, we know  $\|\mathbf{f}\|_\infty^2 = C_f^2 = o_p(1)$  due to assumption that  $\|f_0\|_\infty < C_f$ . As a result, the term  $\mathbf{f}^\top (\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1})\mathbf{f}$  can be bounded as:

$$\mathbf{f}^\top (\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1})\mathbf{f} \leq \|\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1}\|_\infty \|\mathbf{f}\|_\infty^2 = t * O_p(n^{-\frac{1}{2}}) = o_p(1). \quad (33)$$

- Third and fourth term:  $\frac{t}{\sqrt{n}}\mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1}\mathbf{f}$  and  $\frac{t^2}{2n}\mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1}\mathbf{a}_n^\top$ .

First notice that:

$$\|\mathbf{K}_{n,\mathcal{W}}^{-1}\|_\infty = \|\mathbf{B}_{t,n}^+\mathbf{K}_{\mathcal{W}}^{-1}\mathbf{B}_{t,n}^+\|_\infty \leq \|\mathbf{K}_{\mathcal{W}}^{-1}\|_\infty \|\mathbf{B}_{t,n}^+\|_\infty^2 = \|\mathbf{K}_{\mathcal{W}}^{-1}\|_\infty = O_p(1)$$

where the second equality follows since  $\|\mathbf{B}_{t,n}^+\|_\infty \leq \|\mathbf{B}_{t,n}^+\|_2 = \|(\mathbf{I} - \frac{t}{\sqrt{n}}\mathbf{H}_{\mathcal{W},p})^+\|_2 = 1$  due to the fact that  $\mathbf{H}_{\mathcal{W},p}$  is not full-column rank. The final equality follows from the fact that  $\|\mathbf{K}_{\mathcal{W}}^{-1}\|_\infty \leq \|\mathbf{K}_{\mathcal{W}}^{-1}\|_2 = \sigma_{b_0}^{-2} = O_p(1)$  due to  $\mathbf{K}_{\mathcal{W}} = \Phi_{\mathcal{W}}\Phi_{\mathcal{W}}^\top + \sigma_{b_0}^2\mathbf{I}$ .

Also notice that

$$\begin{aligned} \|\mathbf{a}_n\|_\infty &= \|\mathbf{H}_{\mathcal{W},p}(\mathbf{f}_{0,\mathcal{W}} + \omega_{\mathcal{W}})\|_\infty \leq \|\mathbf{H}_{\mathcal{W},p}\|_\infty \|\mathbf{f}_{0,\mathcal{W}}\|_\infty + \|\omega_{\mathcal{W}}\|_\infty \\ &\leq \|\mathbf{H}_{\mathcal{W},p}\|_2 \|\mathbf{f}_{0,\mathcal{W}}\|_\infty + \|\omega_{\mathcal{W}}\|_2 = O_p(1) \end{aligned}$$

since  $\|\omega_{\mathcal{W}}\|_2 = O_p(1)$ ,  $\|\mathbf{f}_{0,\mathcal{W}}\|_\infty = C_f = O_p(1)$  and  $\|\mathbf{H}_{\mathcal{W},p}\|_2 = O_p(1)$  by Proposition 3. Also

$$\|\mathbf{f}\|_\infty = \|\mathbf{X}(\prod_{l=1}^L \mathbf{W}_l \mathbf{S}_l)\boldsymbol{\beta}\|_\infty \leq \|\mathbf{X}\|_\infty * \prod_{l=1}^L (\|\mathbf{W}_l\|_\infty \|\mathbf{S}_l\|_\infty) * \|\boldsymbol{\beta}\|_\infty = O_p(1)$$

since  $\|\mathbf{W}_l\|_\infty \leq B \leq 1$ ,  $\|\boldsymbol{\beta}\|_\infty \leq B \leq 1$ , and the fact that elements in  $\mathbf{X}$  and  $\mathbf{S}_l$  are bounded between  $(0, 1)$ .



Therefore we can bound  $\frac{t}{\sqrt{n}} \mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1} \mathbf{f}$  as:

$$\frac{t}{\sqrt{n}} \mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1} \mathbf{f} \leq \frac{t}{\sqrt{n}} \|\mathbf{a}_n\|_\infty \|\mathbf{K}_{n,\mathcal{W}}^{-1}\|_\infty \|\mathbf{f}\|_\infty \leq t * O_p(n^{-\frac{1}{2}}) = o_p(1). \quad (34)$$

Similarly, we can bound the fourth term  $\frac{t^2}{2n} \mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1} \mathbf{a}_n^\top$  as:

$$\frac{t^2}{2n} \mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1} \mathbf{a}_n^\top \leq \frac{t^2}{2n} \|\mathbf{a}_n\|_\infty^2 \|\mathbf{K}_{n,\mathcal{W}}^{-1}\|_\infty \leq t^2 * O_p(n^{-1}) = o_p(1). \quad (35)$$

Consequently, by combining the asymptotic bounds for the four terms (31), (33), (34) and (35), we have that

$$\mathcal{J}_n = \frac{\int e^{l_n(f_t) - l_n(f_0, \mathcal{W})} * e^{-\mathcal{L}_{t,n}} d\Pi(f_t | \mathcal{W})}{\int e^{l_n(f) - l_n(f_0, \mathcal{W})} d\Pi(f | \mathcal{W})} = e^{-\frac{t}{\sqrt{n}} o_p(K)} \rightarrow 1,$$

where the last equality follows since  $K/\sqrt{n} = o_p(1)$  due to Assumption 1. □

## D.2 Proof for Theorem C.3

*Proof.* Our goal is to show the asymptotic normality

$$\Pi_n(\sqrt{n}(\psi_{\mathcal{W}}^c(f) - \hat{\psi}_{\mathcal{W}}^c) | \mathcal{W}) \rightsquigarrow MVN(0, V_{0,\mathcal{W}}).$$

First derive the expressions for  $f_t$ ,  $\hat{\psi}_{\mathcal{W}}$ ,  $V_{0,n}$  and  $\mathcal{I}_n$ . For  $t = [t_1, \dots, t_P] \in \mathbb{R}^P$ , the expression for  $f_t$  is:

$$\begin{aligned} f_t &= f - \frac{t^\top}{\sqrt{n}} \left( \psi_1 + \frac{1}{2} \psi_2(f - f_0, \mathcal{W}) \right) - \frac{t^\top}{2n} \psi_2(\omega_{\mathcal{W}}) \\ &= \left( I - \frac{t^\top}{\sqrt{n}} * H_{\mathcal{W}} \right) (f) - \frac{t^\top}{\sqrt{n}} * H_{\mathcal{W}}(f_0, \mathcal{W} + \omega_{\mathcal{W}}). \end{aligned} \quad (36)$$

The expressions for  $\hat{\psi}_{\mathcal{W}}$  and  $t^\top V_{0,n} t$  are:

$$\begin{aligned} \hat{\psi}_{\mathcal{W}} &= \hat{\psi}(f_0, \mathcal{W}) + \frac{W_n(\psi_1)}{\sqrt{n}} + \frac{1}{2} \frac{\langle \omega_{\mathcal{W}}, \psi_2(\omega_{\mathcal{W}}) \rangle_n}{n} \\ &= \hat{\psi}(f_0, \mathcal{W}) + 2 \langle H_{\mathcal{W}} f_0, \mathcal{W}, \omega_{\mathcal{W}} \rangle_n + \langle H_{\mathcal{W}} \omega_{\mathcal{W}}, \omega_{\mathcal{W}} \rangle_n, \\ t^\top V_{0,n} t &= \left\| t^\top \left( \psi_1 - \frac{1}{2} \frac{\psi_2(\omega_{\mathcal{W}})}{\sqrt{n}} \right) \right\|_n^2 = 4 * \left\| t^\top H_{\mathcal{W}} f_0, \mathcal{W} + o_p(1) \right\|_n^2, \end{aligned}$$

and the expression for  $\mathcal{I}_n$  is:

$$\mathcal{I}_n = \frac{\int_{A_n} e^{l_n(f_t) - l_n(f_0, \mathcal{W})} d\Pi(f | \mathcal{W})}{\int_{A_n} e^{l_n(f) - l_n(f_0, \mathcal{W})} d\Pi(f | \mathcal{W})}.$$

It then follows from the Theorem C.1 that

$$E_n(e^{t^\top \sqrt{n}(\psi_{\mathcal{W}}^c(f) - \hat{\psi}_{\mathcal{W}}^c)} | A_n, \mathcal{W}) = e^{o_p(1) + t^\top V_{0,n} t / 2} * \mathcal{I}_n.$$

Therefore to show asymptotic normality, we only need to show  $\mathcal{I}_n \rightarrow 1$  for some suitable choice of  $A_n$  so that  $E_n(e^{t^\top \sqrt{n}(\hat{\psi}_{\mathcal{W}}^c(f) - \hat{\psi}_{\mathcal{W}}^c)} | A_n, \mathcal{W}) \rightarrow e^{t^\top V_{0,n} t / 2}$ .

To this end, we consider the standard choice  $A_n = \left\{ f \mid \|f - f_0, \mathcal{W}\| \leq M_n \epsilon_n \right\}$  for some sufficiently large  $M_n \rightarrow \infty$  and  $\epsilon_n$  the given rate of posterior convergence. Similar to the proof for Theorem C.1, it is sufficient to show the terms  $\mathcal{J}_n \rightarrow 1$  and  $\mathcal{K}_n \rightarrow 1$ , where

$$\mathcal{J}_n = \frac{\int e^{l_n(f_t) - l_n(f_0, \mathcal{W})} d\Pi(f | \mathcal{W})}{\int e^{l_n(f) - l_n(f_0, \mathcal{W})} d\Pi(f | \mathcal{W})}, \quad \mathcal{K}_n = \frac{\int_{f \in A_n} e^{l_n(f_t) - l_n(f_0, \mathcal{W})} d\Pi(f | \mathcal{W})}{\int e^{l_n(f_t) - l_n(f_0, \mathcal{W})} d\Pi(f | \mathcal{W})}.$$

The argument for  $\mathcal{K}_n \rightarrow 1$  is essentially the same as those in the proof for Theorem C.2 (i.e., (27)). We show  $\mathcal{J}_n \rightarrow 1$  by performing change of variables  $f \rightarrow f_t$  on the numerator, and then analyze the asymptotic behavior of the resulting expression. Similar to the proof for Theorem C.1, by noticing the conditional prior  $\Pi(f|\mathcal{W})$  is a Gaussian process prior with kernel matrix  $\mathbf{K}_{\mathcal{W}} = \Phi_{\mathcal{W}}\Phi_{\mathcal{W}}^\top + \sigma_{b_0}^2 \mathbf{I}$  for  $\Phi_{n \times K} = \mathbf{X}(\prod_{l=1}^L \mathbf{W}_l \mathbf{S}_l)$ , we see that given observations  $\{y_i, \mathbf{x}_i\}_{i=1}^n$ , the conditional prior distribution for  $\mathbf{f} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]^\top \in \mathbb{R}^n$  is a multivariate Gaussian with covariance matrix  $\mathbf{K}_{\mathcal{W}}$ :

$$d\Pi(\mathbf{f}|\mathcal{W}) \propto |\mathbf{K}_{\mathcal{W}}|^{-\frac{1}{2}} e^{-\frac{1}{2} \mathbf{f}^\top \mathbf{K}_{\mathcal{W}}^{-1} \mathbf{f}}.$$

To perform change of variable, we denote  $\mathbf{f}_t \in \mathbb{R}^n$  as  $f_t$  in (36) evaluated at  $\{y_i, \mathbf{x}_i\}_{i=1}^n$ , we can write  $\mathbf{f}_t = \mathbf{B}_{t,n} \mathbf{f} - \frac{1}{\sqrt{n}} \mathbf{a}_n$  such that

$$\begin{aligned} d\Pi(\mathbf{f}_t|\mathcal{W}) &\propto |\mathbf{K}_{n,\mathcal{W}}|^{-\frac{1}{2}} e^{-\frac{1}{2} (\mathbf{f} + \frac{1}{\sqrt{n}} \mathbf{a}_n)^\top \mathbf{K}_{n,\mathcal{W}}^{-1} (\mathbf{f} + \frac{1}{\sqrt{n}} \mathbf{a}_n)} \\ &= d\Pi(\mathbf{f}|\mathcal{W}) * \exp \left[ -\frac{1}{2} \log \left( \frac{|\mathbf{K}_{n,\mathcal{W}}|}{|\mathbf{K}_{\mathcal{W}}|} \right) - \frac{1}{2} \mathbf{f}^\top (\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1}) \mathbf{f} + \frac{1}{\sqrt{n}} \mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1} \mathbf{f} - \frac{1}{2n} \mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1} \mathbf{a}_n \right] \end{aligned}$$

where  $\mathbf{B}_{t,n} = \mathbf{I} - \frac{1}{\sqrt{n}} \mathbf{H}_t$ ,  $\mathbf{a}_n = \mathbf{H}_t (\mathbf{f}_{0,\mathcal{W}} + \omega_{\mathcal{W}})$  for  $\mathbf{H}_t = \sum_{p=1}^P t_p \mathbf{H}_p$ , and

$$\mathbf{K}_{n,\mathcal{W}} = \mathbf{B}_{t,n} \mathbf{K}_{\mathcal{W}} \mathbf{B}_{t,n} = \mathbf{K}_{\mathcal{W}} - \frac{1}{\sqrt{n}} \left[ \mathbf{K}_{\mathcal{W}} \mathbf{H}_t + \mathbf{H}_t \mathbf{K}_{\mathcal{W}} \mathbf{B}_{t,n} \right] = \mathbf{K}_{\mathcal{W}} - \frac{1}{\sqrt{n}} d\mathbf{K}_{\mathcal{W}},$$

where we have denoted  $d\mathbf{K}_{\mathcal{W}} = \mathbf{K}_{\mathcal{W}} \mathbf{H}_t + \mathbf{H}_t \mathbf{K}_{\mathcal{W}} \mathbf{B}_{t,n}$ .

Consequently, the expression of  $\mathcal{J}_n$  after the change of variable becomes:

$$\mathcal{J}_n = \frac{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}})} * e^{-\mathcal{L}_{t,n}} d\Pi(f_t|\mathcal{W})}{\int e^{l_n(f) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})},$$

where

$$\mathcal{L}_{t,n} = \frac{1}{2} \log \left( \frac{|\mathbf{K}_{n,\mathcal{W}}|}{|\mathbf{K}_{\mathcal{W}}|} \right) + \frac{1}{2} \mathbf{f}^\top (\mathbf{K}_{n,\mathcal{W}}^{-1} - \mathbf{K}_{\mathcal{W}}^{-1}) \mathbf{f} - \frac{1}{\sqrt{n}} \mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1} \mathbf{f} + \frac{t^2}{2n} \mathbf{a}_n^\top \mathbf{K}_{n,\mathcal{W}}^{-1} \mathbf{a}_n. \quad (37)$$

Therefore in order to show  $\mathcal{J}_n \rightarrow 1$ , it is sufficient to show that all terms in  $\mathcal{L}_{t,n}$  are asymptotically vanishing toward zero as  $n \rightarrow \infty$ . Now only left to show that all the four terms in (37) are asymptotically vanishing toward zero. However, notice that the only difference between (37) and its counterpart (25) in the proof of Theorem C.2 is that  $\mathbf{H}_p$  is replaced by  $\mathbf{H}_t = \sum_{p=1}^P t_p \mathbf{H}_p$ . Furthermore, for fixed  $t$ , the asymptotic behavior of  $\mathbf{H}_p$  and  $\mathbf{H}_t$  are similar in the sense that  $\text{tr}(\mathbf{H}_t) = O_p(K)$  and  $\|\mathbf{H}_t\|_2 = O_p(1)$  since  $P = O(1)$ . Therefore we can follow exactly the same arguments as those in the proof for Theorem C.2 to show all the four terms asymptotically vanishing towards zero. Consequently, we have that  $\mathcal{J}_n = \frac{\int e^{l_n(f_t) - l_n(f_{0,\mathcal{W}}) + o_p(1)} d\Pi(f_t|\mathcal{W})}{\int e^{l_n(f) - l_n(f_{0,\mathcal{W}})} d\Pi(f|\mathcal{W})} \rightarrow 1..$

It then follows by the semi-parametric BvM theorem (Theorem C.1) that the posterior distribution of  $\sqrt{n}(\psi_{\mathcal{W}}^c(f) - \hat{\psi}_{\mathcal{W}}^c)$  converge toward a normal law with mean zero and variance  $V_0$ , i.e.

$$\Pi \left( \sqrt{n}(\psi_{\mathcal{W}}^c(f) - \hat{\psi}_{\mathcal{W}}^c) \middle| \{ \mathbf{x}_i, y_i \}_{i=1}^n \right) \rightsquigarrow N(0, V_0), \quad (38)$$

where  $V_0$  is the  $P \times P$  matrix that takes value  $t^\top V_{0,n} t = 4 * \|t^\top H_{\mathcal{W}} f_{0,\mathcal{W}}\|_n^2$  for any  $t \in \mathbb{R}^P$ . Specifically, the matrix  $V_0$  whose  $(p_1, p_2)^{th}$  entry defined as  $4 \langle H_{p_1} f_{0,\mathcal{W}}, H_{p_2} f_{0,\mathcal{W}} \rangle_n$  will satisfy this choice.  $\square$

## E Lemmas and Propositions

We first show the posterior concentration rate for a deep Bayesian neural network (BNN) with moderate level of sparsity (Proposition 1). This is an assumption that can hold even for a BNN without an explicit sparse-inducing prior. For example, for a convolutional neural network using separable convolutional layers, its convolutional kernels are in fact banded Toeplitz matrices with the higher off-diagonal entries set to zero, whose number of parameters is only proportional to the number of the output channels [9, 20]. For this case, we show that for a network with sparsity level (i.e., the number of unique non-zero weights)  $O(L * K)$ , the model can achieve an optimal, parametric rate of  $O(n^{-1/2})$  up to a logarithm factor.

For completeness, in Proposition 2, we also study the case where no sparsity is assumed. This corresponds to the case of a naive densely-connected network without any type of regularization. We show that in this case, for the model to achieve a optimal rate of  $O(n^{-1/2})$ , it in fact needs to be narrower (i.e.,  $K = O(n^{1/4})$  rather than  $O(n^{1/2})$ ), hence restricting the space of true functions it can reliably approximate in the finite data. The difference between these two cases highlights the importance of choosing a parameter-efficient architecture to ensure the effectiveness of variable selection in the finite sample. Finally, notice that the BvM theorem still applies in this second case, since the Assumption 1 still holds.

**Proposition 1** (Posterior Concentration for  $f_0 \in \mathcal{F}$ ). *For the space of ReLU network  $\mathcal{F} = \mathcal{F}(L, K, B)$ . Assuming*

- *The model architecture satisfies:*

$$L = O(\log(N)), \quad K = O(N), \quad S = O(N \log(N)).$$

where  $S = \sum_{l=1}^L \|\mathcal{W}_l\|_0$  is the number of non-zero parameters in the model, and  $N \in \mathbb{N}$  is a function of sample size  $n$  such that  $\log(N) \geq \sqrt{\log(n)}$ .

- *The prior distribution  $\Pi(\mathcal{W})$  is an i.i.d. product of Gaussian distributions.*

Then, for  $f_0 \in \mathcal{F}$ , the posterior distribution  $\Pi_n(f) = \Pi(f | \{\mathbf{x}_i, y_i\}_{i=1}^n)$  contracts toward  $f_0$  in a rate of at least  $\epsilon_n = O((N/n) * \log(N)^3)$ , i.e., for any  $M_n \rightarrow \infty$

$$E_0 \Pi_n(\|f - f_0\|_n^2 \geq M_n \epsilon_n) \rightarrow 0$$

In particular, if  $N = o_p(n^{1/2})$  (i.e. Assumption 1), we then have  $\epsilon_n = O(n^{-1/2} * \log(n)^3)$ .

*Proof.* We show posterior consistency by checking if the stated convergence rate  $\epsilon$  satisfy the classic posterior convergence theorem (i.e., Theorem 1 of [11]). Since we consider a well-specified case where  $f_0 \in \mathcal{F}$ , one of the theorem's three conditions regarding expressiveness of the prior model trivially holds (i.e.,  $\Pi(\mathcal{F}/\mathcal{F}_n) \leq e^{-n\epsilon_n(C+4)}$  for  $\mathcal{F}$  the true function space and  $\mathcal{F}_n$  the model space). However, even in this well-specified case, it is still important to understand if neural network can achieve the desired convergence rate  $\epsilon_n$  in the face of the prior stochasticity in its large number of hidden weight parameters. Therefore, we would like to check the below two conditions about the neural network prior distribution:

$$\log N(\epsilon_n, \mathcal{F}, \|\cdot\|_\infty) \leq Cn\epsilon_n \tag{39}$$

$$\Pi\left(\|f - f_0\|_\infty^2 \leq \epsilon_n\right) \geq e^{-Cn\epsilon_n} \tag{40}$$

where  $C > 1$  is an universal constant. The first condition (39) ensures the model size, which is measured by the covering number of  $\mathcal{F}$  (i.e., the minimal number of  $\|\cdot\|_\infty$ -balls with radius  $\delta$  that covers  $\mathcal{F}$ ) is not too large. The second condition (40) ensures the model's prior distribution places sufficient prior mass around the target function  $f_0$ . For the rest of this proof, we show the statement of interest by showing these two conditions.

- **Entropy Condition:**  $\log N(\epsilon_n, \mathcal{F}, \|\cdot\|_n) \leq n\epsilon_n$

We show this condition by invoking the classic result on the covering number of neural network approximation spaces (Lemma 2 [1]). Specifically, by setting  $\delta = \epsilon_n$  in Lemma 2, we have

$$\begin{aligned}
\log N(\epsilon_n, \mathcal{F}, \|\cdot\|_\infty) &\leq S \log(\epsilon_n^{-1} L K^{2L}) \\
&\lesssim S \log(\epsilon_n^{-1}) + S L \log(K) \\
&\lesssim N \log(N) \log\left(\frac{n}{N \log(N)^3}\right) + N \log(N)^3 \\
&\lesssim N \log(N) (\log(n) + \log(N)^2) \\
&\lesssim N \log(N)^3 = n \epsilon_n
\end{aligned}$$

where the third line follows by the definition of  $\epsilon_n$ , and the last line follows from the assumption that  $\log(N) \geq \sqrt{\log(n)}$  and the definition of  $\epsilon_n$ .

- **Prior Mass Condition:**  $\Pi\left(\|f - f_0\|_n \leq \epsilon_n\right) \geq e^{-Cn\epsilon_n}$ .

We show this using the classic concentration inequality for the centered Gaussian measure  $\Pi$  (Lemma 1). Denote  $\mathcal{W} = \{\mathcal{W}_i\}_{i=1}^L \cup \{\beta\}$  the set of all parameters for a neural network  $f$ . By the proof of Lemma 2 below, we can bound the difference between a model  $f$  and the true function  $f_0$  as  $|f - f_0| \leq L K^L \|\mathcal{W} - \mathcal{W}_0\|_\infty$ , therefore

$$\left\{f \mid \|f - f_0\|_\infty^2 \leq \epsilon_n\right\} \subset \left\{\mathcal{W} \mid \|\mathcal{W} - \mathcal{W}_0\|_\infty^2 \leq \frac{\epsilon_n}{(L K^L)^2}\right\}.$$

Then by Lemma 1:

$$\begin{aligned}
\Pi\left(\|f - f_0\|_\infty^2 \leq \epsilon_n\right) &\geq \Pi\left(\|\mathcal{W} - \mathcal{W}_0\|_\infty^2 \leq \frac{\epsilon_n}{(L K^L)^2}\right) \\
&\geq \exp\left(-\frac{\|\mathcal{W}_0\|_\Pi^2}{2}\right) \Pi\left(\|\mathcal{W}\|_\infty^2 < \frac{\epsilon_n}{(L K^L)^2}\right) \\
&\geq \exp\left(-\frac{\|\mathcal{W}_0\|_\Pi^2}{2}\right) \exp\left(-\frac{\epsilon_n}{(L K^L)^2}\right) \\
&\geq \exp(-Cn\epsilon_n)
\end{aligned}$$

for some constant  $C > 1$ . In the above expression, the second inequality follows from Lemma 1, and the third inequality follows from the Borell's inequality [22].

**Proposition 2** (Posterior Concentration for  $f_0 \in \mathcal{F}$ , No Sparsity). *For the space of ReLU network  $\mathcal{F} = \mathcal{F}(L, K, B)$ . Assuming*

- *The model architecture satisfies:*

$$L = O(\log(N)), \quad K = O(\sqrt{N}).$$

*Since we assume no sparsity, we set  $S = O(L * K^2) = O(N \log(N))$ .*

- *The prior distribution  $\Pi(\mathcal{W})$  is an i.i.d. product of Gaussian distributions.*

*Then, for  $f_0 \in \mathcal{F}$ , the posterior distribution  $\Pi_n(f) = \Pi(f \mid \{\mathbf{x}_i, y_i\}_{i=1}^n)$  contracts toward  $f_0$  in a rate of at least  $\epsilon_n = O((N^2/n) * \log(N)^3)$ , i.e., for any  $M_n \rightarrow \infty$*

$$E_0 \Pi_n(\|f - f_0\|_n^2 \geq M_n \epsilon_n) \rightarrow 0$$

*In particular, if  $N = o_p(n^{1/2})$  (i.e. Assumption 1), we then have  $\epsilon_n = O(n^{-1/2} * \log(n)^3)$ .*

*Proof.* The proof is similar to Proposition 1. The only difference is  $K$  is changed from  $O(N)$  to  $O(\sqrt{N})$  (and the order of the sparsity in fact stays the same). Specifically, we re-check the entropy condition and the prior mass condition below:

- **Entropy Condition:**  $\log N(\epsilon_n, \mathcal{F}, \|\cdot\|_n) \leq n\epsilon_n$

By setting  $\delta = \epsilon_n$  in Lemma 2, we have

$$\begin{aligned}
\log N(\epsilon_n, \mathcal{F}, \|\cdot\|_\infty) &\leq S \log(\epsilon_n^{-1} LK^{2L}) \\
&\lesssim S \log(\epsilon_n^{-1}) + SL \log(K) \\
&\lesssim N \log(N) \log\left(\frac{n}{N \log(N)^3}\right) + N^2 \log(N)^3 \\
&\lesssim N \log(N) (\log(n) + \log(N)^2) \\
&\lesssim N \log(N)^3 = n\epsilon_n
\end{aligned}$$

- **Prior Mass Condition:**  $\Pi\left(\|f - f_0\|_n \leq \epsilon_n\right) \geq e^{-Cn\epsilon_n}$ .

Similar to proof for Proposition 1, by the proof of Lemma 2, we can bound the difference between a model  $f$  and the true function  $f_0$  as  $|f - f_0| \leq LK^L \|\mathcal{W} - \mathcal{W}_0\|_\infty$ , therefore

$$\left\{f \mid \|f - f_0\|_\infty^2 \leq \epsilon_n\right\} \subset \left\{\mathcal{W} \mid \|\mathcal{W} - \mathcal{W}_0\|_\infty^2 \leq \frac{\epsilon_n}{(LK^L)^2}\right\}.$$

Then by Lemma 1:

$$\begin{aligned}
\Pi\left(\|f - f_0\|_\infty^2 \leq \epsilon_n\right) &\geq \Pi\left(\|\mathcal{W} - \mathcal{W}_0\|_\infty^2 \leq \frac{\epsilon_n}{(LK^L)^2}\right) \\
&\geq \exp\left(-\frac{\|\mathcal{W}_0\|_\Pi^2}{2}\right) \Pi\left(\|\mathcal{W}\|_\infty^2 < \frac{\epsilon_n}{(LK^L)^2}\right) \\
&\geq \exp\left(-\frac{\|\mathcal{W}_0\|_\Pi^2}{2}\right) \exp\left(-\frac{\epsilon_n}{(LK^L)^2}\right) \\
&\geq \exp(-Cn\epsilon_n)
\end{aligned}$$

for some constant  $C > 1$ .

□

**Proposition 3.** Recall that a ReLU network  $f \in \mathcal{F}(L, K, B)$  adopts basis function representation  $f(\mathbf{x}) = \sum_{k=1}^K \beta_k \phi_k(\mathbf{x}) = \phi(\mathbf{x})^\top \beta$  at the output layer. Denote  $\Phi$  and  $\partial_p \Phi$  the  $n \times K$  matrices of  $\phi$  and  $\frac{\partial}{\partial x_p} \phi$  evaluated at observations  $\{\mathbf{x}_i, y_i\}_{i=1}^n$ , such that  $\Phi_{n \times K} = \mathbf{X} \mathbf{W}_1 \mathbf{S}_1 (\prod_{l=2}^L \mathbf{W}_l \mathbf{S}_l)$  and  $\partial_p \Phi_{n \times K} = \mathbf{W}_{1,p} \mathbf{S}_1 (\prod_{l=1}^L \mathbf{W}_l \mathbf{S}_l)$ . Denote

$$\mathbf{H}_p = (\Phi^+)^{\top} \partial_p \Phi^{\top} \partial_p \Phi \Phi^+$$

where  $\Phi_{K \times n}^+$  is the generalized inverse of  $\Phi_{n \times K}$ . Then:

$$\|\mathbf{H}_p\|_2 = O_p(1), \quad \text{tr}(\mathbf{H}_p) = O_p(K),$$

*Proof.* Notice that  $\|\mathbf{H}_p\|_2 = O_p(1)$  implies  $\text{tr}(\mathbf{H}_p) = O_p(K)$  since

$$\text{tr}(\mathbf{H}_p) \leq \text{rank}(\mathbf{H}_p) * \|\mathbf{H}_p\|_2 = O_p(\text{rank}(\mathbf{H}_p)) = O_p(K).$$

Therefore we only need to show  $\|\mathbf{H}_p\|_2 = O_p(1)$ .

Show  $\|\mathbf{H}_p\|_2 = O_p(1)$  by showing that it is upper bounded by certain constant that does not depend on  $n$ . Recall that  $\mathbf{H}_p = (\Phi^+)^{\top} (\partial_p \Phi^{\top} \partial_p \Phi) \Phi^+$ , by the fact that 2-norm is invariant under cyclic permutations, we then have

$$\begin{aligned} \|\mathbf{H}_p\|_2 &= \|(\partial_p \Phi \Phi^+)^{\top} (\partial_p \Phi \Phi^+)\|_2 = \|\partial_p \Phi (\Phi^+ \Phi^{+\top}) \partial_p \Phi^{\top}\|_2 = \|\partial_p \Phi (\Phi^{\top} \Phi)^{-1} \partial_p \Phi^{\top}\|_2 \\ &= \|(\Phi^{\top} \Phi)^{-1} (\partial_p \Phi^{\top} \partial_p \Phi)\|_2 \end{aligned}$$

Denote  $\mathbf{M} = \mathbf{S}_1 (\prod_{l=2}^L \mathbf{W}_l \mathbf{S}_l)$ , then we have  $\Phi = \mathbf{X} \mathbf{W}_1 \mathbf{M}$  and  $\partial_p \Phi = \mathbf{W}_{1,p} \mathbf{M}$ . Denote  $\mathbf{X}_p = \text{diag}(x_{1,p}, x_{2,p}, \dots, x_{n,p})$  and notice that  $\mathbf{X} \mathbf{W}_1 = \sum_p \mathbf{X}_p \mathbf{W}_{1,p}$ , so we see that  $\Phi$  is related to  $\partial_p \Phi$  through the expression  $\Phi = \sum_p \mathbf{X}_p \partial_p \Phi$ , and

$$\Phi^{\top} \Phi = \left( \sum_p \mathbf{X}_p \partial_p \Phi \right)^{\top} \left( \sum_p \mathbf{X}_p \partial_p \Phi \right) = \sum_{p_1, p_2} \partial_{p_1} \Phi^{\top} \mathbf{X}_{p_1} \mathbf{X}_{p_2} \partial_{p_2} \Phi.$$

Using the inversion formula  $(\mathbf{A} + \mathbf{B})^{-1} = (\mathbf{I} + \mathbf{A}^{-1} \mathbf{B})^{-1} \mathbf{A}^{-1}$  [12], we can write  $(\Phi^{\top} \Phi)^{-1}$  as:

$$\begin{aligned} (\Phi^{\top} \Phi)^{-1} &= \left( \partial_p \Phi^{\top} \mathbf{X}_p^2 \partial_p \Phi + \sum_{p_1, p_2 \neq p} \partial_{p_1} \Phi^{\top} \mathbf{X}_{p_1} \mathbf{X}_{p_2} \partial_{p_2} \Phi \right)^{-1} \\ &= \left( \mathbf{I} + (\partial_p \Phi^{\top} \mathbf{X}_p^2 \partial_p \Phi)^{-1} \sum_{p_1, p_2 \neq p} \partial_{p_1} \Phi^{\top} \mathbf{X}_{p_1} \mathbf{X}_{p_2} \partial_{p_2} \Phi \right)^{-1} (\partial_p \Phi^{\top} \mathbf{X}_p^2 \partial_p \Phi)^{-1}. \end{aligned}$$

Consequently, we can bound  $\|\mathbf{H}_p\|_2$  as:

$$\begin{aligned} \|\mathbf{H}_p\|_2 &= \|(\Phi^{\top} \Phi)^{-1} (\partial_p \Phi^{\top} \partial_p \Phi)\|_2 \\ &= \left\| \left( \mathbf{I} + (\partial_p \Phi^{\top} \mathbf{X}_p^2 \partial_p \Phi)^{-1} \sum_{p_1, p_2 \neq p} \partial_{p_1} \Phi^{\top} \mathbf{X}_{p_1} \mathbf{X}_{p_2} \partial_{p_2} \Phi \right)^{-1} (\partial_p \Phi^{\top} \mathbf{X}_p^2 \partial_p \Phi)^{-1} (\partial_p \Phi^{\top} \partial_p \Phi) \right\|_2 \\ &\leq \left\| \left( \mathbf{I} + (\partial_p \Phi^{\top} \mathbf{X}_p^2 \partial_p \Phi)^{-1} \sum_{p_1, p_2 \neq p} \partial_{p_1} \Phi^{\top} \mathbf{X}_{p_1} \mathbf{X}_{p_2} \partial_{p_2} \Phi \right)^{-1} \right\|_2 * \left\| (\partial_p \Phi^{\top} \mathbf{X}_p^2 \partial_p \Phi)^{-1} (\partial_p \Phi^{\top} \partial_p \Phi) \right\|_2 \end{aligned}$$

Notice in the last line of the above expression, the first term can be bounded as:

$$\left\| \left( \mathbf{I} + (\partial_p \Phi^{\top} \mathbf{X}_p^2 \partial_p \Phi)^{-1} \sum_{p_1, p_2 \neq p} \partial_{p_1} \Phi^{\top} \mathbf{X}_{p_1} \mathbf{X}_{p_2} \partial_{p_2} \Phi \right)^{-1} \right\|_2 \leq 1 \quad (41)$$

This is because  $\lambda_{\min} \left( \mathbf{I} + (\partial_p \Phi^{\top} \mathbf{X}_p^2 \partial_p \Phi)^{-1} \sum_{p_1, p_2 \neq p} \partial_{p_1} \Phi^{\top} \mathbf{X}_{p_1} \mathbf{X}_{p_2} \partial_{p_2} \Phi \right) \geq 1$  due to the fact that  $\partial_{p_1} \Phi^{\top} \mathbf{X}_{p_1} \mathbf{X}_{p_2} \partial_{p_2} \Phi$  is full column rank for all  $p_1, p_2$ , which implies that  $\lambda_{\max}$  of (41) is no larger than 1.

The second term can be bounded as:

$$\begin{aligned}
\|(\partial_p \Phi^\top \mathbf{X}_p^2 \partial_p \Phi)^{-1} (\partial_p \Phi^\top \partial_p \Phi)\|_2 &= \|(\partial_p \Phi^\top \mathbf{X}_p^{-2} \partial_p \Phi^\top) (\partial_p \Phi^\top \partial_p \Phi)\|_2 \\
&= \|\partial_p \Phi^\top \mathbf{X}_p^{-2} \partial_p \Phi\|_2 \\
&= \|\mathbf{X}_p^{-2} \partial_p \Phi \partial_p \Phi^\top\|_2 \\
&\leq \|\mathbf{X}_p^{-2}\|_2 \|\partial_p \Phi \partial_p \Phi^\top\|_2 \\
&= \|\mathbf{X}_p^{-1}\|_2^2, \tag{42}
\end{aligned}$$

where we have denoted  $\Phi^+ = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top$  for  $\Phi = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ . In above expression, the third equality follows from the fact that the eigenvalue of the product of square matrices is invariant under cyclic permutation of the product order [14].

Combining (41) and (42), we have:

$$\begin{aligned}
\|\mathbf{H}_p\|_2 &= \|(\Phi^\top \Phi)^{-1} (\partial_p \Phi^\top \partial_p \Phi)\|_2 \\
&\leq \left\| \left( \mathbf{I} + (\partial_p \Phi^\top \mathbf{X}_p^2 \partial_p \Phi)^{-1} \sum_{p_1, p_2 \neq p} \partial_{p_1} \Phi^\top \mathbf{X}_{p_1} \mathbf{X}_{p_2} \partial_{p_2} \Phi \right)^{-1} \right\|_2 * \left\| (\partial_p \Phi^\top \mathbf{X}_p^2 \partial_p \Phi)^{-1} (\partial_p \Phi^\top \partial_p \Phi) \right\|_2 \\
&\leq \|\mathbf{X}_p^{-1}\|_2^2
\end{aligned}$$

Recall that since  $\mathbf{X}_p = \text{diag}(x_{1,p}, x_{2,p}, \dots, x_{n,p})$  is a diagonal matrix, we have  $\|\mathbf{X}_p^{-1}\|_2 = \max(1/x_{i,p}) = 1/\min(x_{i,p}) \leq 1/c_x$ , i.e. bounded by a constant that does not depend on  $n$ . Therefore  $\|\mathbf{H}_p\|_2 = O_p(1)$ .  $\square$

**Lemma 1** (Gaussian Shift-ball Inequality). *Let  $\Pi(f)$  be a centered Gaussian measure and  $H_\Pi$  the Hilbert space induced by  $\Pi$  with norm  $\|\cdot\|_\Pi$ . For  $f_0 \in H_\Pi$ , it holds that*

$$\Pi(\|f - f_0\|_\infty \leq \delta) \geq \exp\left(-\frac{\|f_0\|_\Pi^2}{2}\right)\Pi(\|f\|_\infty < \delta)$$

*Proof.* [13], Theorem 2

**Lemma 2** (Covering Number for  $\mathcal{F}(L, K, B)$ ). *For the space of ReLU network  $\mathcal{F} = \mathcal{F}(L, K, B)$  as defined in (2), define  $S = \sum_{l=1}^L \|\mathcal{W}_l\|_0$ , then the covering number of  $\mathcal{F}$  as  $N(\delta, \mathcal{F}, \|\cdot\|_\infty)$ , i.e., the minimal number of  $\|\cdot\|_\infty$ -balls with radius  $\delta$  that covers  $\mathcal{F}$ , can be bounded as:*

$$\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \leq S \log(\delta^{-1} L K^{2L})$$

*Proof.* The proof is an adaptation of the classic entropy number result for neural networks [1] to our current setting.

Consider two networks  $f, f' \in \mathcal{F}(L, K, B)$  where  $f(\mathbf{x}) = \beta\sigma\mathcal{W}_L \dots \sigma\mathcal{W}_1\mathbf{x}$  and  $f'(\mathbf{x}) = \beta'\sigma\mathcal{W}'_L \dots \sigma\mathcal{W}'_1\mathbf{x}$ . Note that  $\|\beta - \beta'\|_\infty \leq KB$  and  $\|\mathcal{W}_l - \mathcal{W}'_l\|_\infty \leq KB$ . Also denote  $A_l(f) = \sigma\mathcal{W}_{l-1} \dots \sigma\mathcal{W}_1\mathbf{x}$  the hidden activation before the  $l^{\text{th}}$  layer, and  $B_l(f) = \beta\sigma\mathcal{W}_L \sigma \dots \mathcal{W}_l \sigma$  the hidden mapping at and after the  $l^{\text{th}}$  layer. Then for bounded input  $\|\mathbf{x}\|_\infty \leq 1$ :

$$\|A_l(f)\|_\infty \leq \|\mathcal{W}_{l-1} A_{l-1}(f)\|_\infty \leq KB \|A_{l-1}(f)\|_\infty \leq (KB)^{l-1},$$

and similarly  $\|B_l(f)\|_\infty \leq (KB)^{L-l+1}$ . Then it holds that:

$$\begin{aligned} |f(\mathbf{x}) - f'(\mathbf{x})| &\leq \left| \sum_{l=1}^L [B_{l+1}(f')\mathcal{W}_l A_l(f)(x) - B_{l+1}(f')\mathcal{W}'_l A_l(f)(x)] \right| \\ &\leq \sum_{l=1}^L (KB)^{L-l} \|\mathcal{W}_l A_l(f)(x) - \mathcal{W}'_l A_l(f)(x)\|_\infty \\ &\leq \sum_{l=1}^L (KB)^{L-l} * \|\mathcal{W}_l - \mathcal{W}'_l\|_\infty * (KB)^{l-1} \leq L * (KB)^{L-l} * (KB) * (KB)^{l-1} \\ &= L * (KB)^L \end{aligned}$$

Thus for  $\|\cdot\|_\infty$  balls with radius  $\delta$ , the covering number for a single sparsity configuration is bounded by  $(\frac{\delta}{L*(KB)^L})^{-S}$ . Now, since the number of possible sparsity patterns is bounded by  $\binom{K^L}{S} \leq K^{LS}$ , then the covering number of the whole space is bounded as  $K^{LS} * (\frac{\delta}{L*(KB)^L})^{-S} = (\frac{\delta}{LK^{2L}B^L})^{-S}$ .

Taking log on both sides and recall we have  $B \leq 1$  due to assumption in (2), then:

$$\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \leq S \log(\delta^{-1} L K^{2L} B^L) \leq S \log(\delta^{-1} L K^{2L})$$

□



## References

- [1] M. Anthony. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1 edition edition, Aug. 2009.
- [2] F. Bach. Breaking the Curse of Dimensionality with Convex Neural Networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.
- [3] A. Berlines and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, Boston, 2004 edition edition, Dec. 2003.
- [4] P. J. Bickel and B. J. K. Kleijn. The semiparametric Bernstein von Mises theorem. *The Annals of Statistics*, 40(1):206–237, Feb. 2012.
- [5] P. Billingsley. *Probability and Measure*. Wiley, Hoboken, N.J, anniversary edition edition, Feb. 2012.
- [6] I. Castillo. A semiparametric Bernstein von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152(1):53–99, Feb. 2012.
- [7] I. Castillo and R. Nickl. Nonparametric Bernstein von Mises theorems in Gaussian white noise. *The Annals of Statistics*, 41(4):1999–2028, Aug. 2013.
- [8] I. Castillo and J. Rousseau. A Bernstein von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(6):2353–2383, Dec. 2015.
- [9] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, July 2017. ISSN: 1063-6919.
- [10] D. Freedman. Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27(4):1119–1141, Aug. 1999.
- [11] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, Apr. 2000.
- [12] H. V. Henderson and S. R. Searle. On Deriving the Inverse of a Sum of Matrices. *SIAM Review*, 23(1):53–60, 1981.
- [13] J. Kuelbs, W. V. Li, and W. Linde. The Gaussian measure of shifted balls. *Probability Theory and Related Fields*, 98(2):143–162, June 1994.
- [14] J. K. Merikoski and R. Kumar. Inequalities for spreads of matrix sums and products. In *Applied Mathematics E-Notes*, 4:150–159, 2004.
- [15] R. M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer-Verlag, New York, 1996.
- [16] A. Rahimi and B. Recht. Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1313–1320. Curran Associates, Inc., 2009.
- [17] V. Rivoirard and J. Rousseau. Bernstein von Mises theorem for linear functionals of the density. *The Annals of Statistics*, 40(3):1489–1523, June 2012.
- [18] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri. Nonparametric Sparsity and Regularization. *Journal of Machine Learning Research*, 14:1665–1714, 2013.
- [19] J. Rousseau. On the Frequentist Properties of Bayesian Nonparametric Methods. *Annual Review of Statistics and Its Application*, 3(1):211–231, June 2016.
- [20] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897, Aug. 2020. Publisher: Institute of Mathematical Statistics.

- [21] A. W. van der Vaart and J. H. van Zanten. Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.
- [22] A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *arXiv:0805.3252 [math, stat]*, pages 200–222, 2008. arXiv: 0805.3252.