Bayesian Analysis (0000) **00**, Number 0, pp. 1

Supplementary Material for "Adaptive Bayesian nonparametric regression using a kernel mixture of polynomials with application to partial linear model"

Fangzheng Xie[*] and Yanxun Xu[*,†]

A Proof of Theorem 1

We first list some auxiliary results that are needed to prove Theorem 1. The proof of these auxiliary results are deferred to Section C. First of all Lemma A.1 below guarantees that in order to prove theorem 1, it suffices to show that

$$\Pi\left(H^2\left(p_{f,\sigma}, p_0\right) > M\epsilon_n^2 \middle| \mathcal{D}_n\right) \to 0 \tag{1}$$

in $\mathbb{P}_0$-probability.

Lemma A.1. *Suppose $\mathcal{G} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ is a function class, where the design space $\mathcal{X}$ is the unit hypercube $[0,1]^p$. Then*

$$\|f - g\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2 + |\sigma_1 - \sigma_2|^2 \lesssim H^2(p_{f,\sigma_1}, p_{g,\sigma_2}) \lesssim \|f - g\|_{L_1(\mathbb{P}_{\mathbf{x}})} + |\sigma_1 - \sigma_2|^2,$$

and hence for all sufficiently small $\epsilon > 0$ and for some constant $C_1 > 0$,

$$\mathcal{N}(\epsilon, \mathcal{M}_{\underline{\sigma}}^{\overline{\sigma}}(\mathcal{G}), H) \leq \mathcal{N}(C_1\epsilon^2, \mathcal{G}, \|\cdot\|_{L_1(\mathbb{P}_{\mathbf{x}})}) \left[\left(\frac{\overline{\sigma} - \underline{\sigma}}{C_1\epsilon}\right) + \mathbf{1}(\underline{\sigma} = \overline{\sigma})\right],$$

where $\mathcal{M}_{\underline{\sigma}}^{\overline{\sigma}}(\mathcal{G}) = \{p_{f,\sigma} : f \in \mathcal{G}, \sigma \in [\underline{\sigma}, \overline{\sigma}]\}$.

Recall that we need to verify the prior concentration condition. To estimate the prior concentration $\Pi(p_{f,\sigma} \in B_{\mathrm{KL}}(p_0, \epsilon))$, we need to bound the Kullback-Leibler numbers $D_{\mathrm{KL}}(p_0||p_{f,\sigma})$ and $\mathbb{E}_0[\log(p_0/p_{f,\sigma})]^2$. Lemma A.2 below provides upper bounds for these two quantities in terms of $\|f - f_0\|_{L_2(\mathbb{P}_{\mathbf{x}})}$, and hence connects the Kullback-Leibler ball $B_{\mathrm{KL}}(p_0, \epsilon)$ with the $L_2(\mathbb{P}_{\mathbf{x}})$ neighborhood of f_0.

Lemma A.2. *Suppose $p_0(\mathbf{x}, y) = \phi_{\sigma_0}(y - f_0(\mathbf{x}))p_{\mathbf{x}}(\mathbf{x})$ and $p_{f,\sigma}(\mathbf{x}, y) = \phi_\sigma(y - f(\mathbf{x}))p_{\mathbf{x}}(\mathbf{x})$ are two joint densities on $\mathcal{X} \times \mathbb{R}$, where f and f_0 lie in some uniformly bounded function class with $\|f\|_\infty, \|f_0\|_\infty < A$, and $\mathcal{X} = [0,1]^p$. Assume that $\sigma_0, \sigma \in [\underline{\sigma}, \overline{\sigma}] \subset (0, \infty)$. Then*

$$\max\left\{D_{\mathrm{KL}}(p_0||p_{f,\sigma}), \mathbb{E}_0\left[\log\frac{p_0(\mathbf{x}, y)}{p_{f,\sigma}(\mathbf{x}, y)}\right]^2\right\} \lesssim |\sigma_0 - \sigma| + \|f - f_0\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2.$$

[*]Department of Applied Mathematics and Statistics, Johns Hopkins University
[†]Correspondence should be addressed to yanxun.xu@jhu.edu

 DOI: 0000

The following proposition explicitly estimates the desired quantity $\Pi(p_{f,\sigma} \in B_{\mathrm{KL}}(p_0, \epsilon))$.

Proposition A.1 (Prior concentration)**.** *Assume that f_0 is in the α-Hölder function class $\mathfrak{C}^{\alpha,B}(\mathcal{X})$with envelope B. Suppose Π is the prior constructed in Section 2.2. Then there exists some constant $C_2 > 0$ such that for all sufficiently small $\epsilon > 0$,*

$$\Pi\left(p_{f,\sigma} \in B_{\mathrm{KL}}(p_0, \epsilon)\right) \geq \exp\left[-C_2 \epsilon^{-p/\alpha}\left(\log\frac{1}{\epsilon}\right)^{\max(r_0,1)}\right].$$

We are now ready to prove the main Theorem 1.

Proof of Theorem 1. We first prove that (1) holds with $\epsilon_n = n^{-\alpha/(2\alpha+p)}(\log n)^{t/2}$. For any function class $\mathcal{G}$, denote $\mathcal{M}_{\underline{\sigma}}^{\overline{\sigma}}(\mathcal{G}) = \{p_{f,\sigma} : f \in \mathcal{G}, \sigma \in [\underline{\sigma}, \overline{\sigma}]\}$ to be the associated class of densities. Define: $\epsilon = [t - \{2\alpha\max(r_0,1)/(p+2\alpha) + \max(1-r_0,0)\}]/3$, $\delta = \{2\alpha\max(r_0,1)/(p+2\alpha) - r_0 + 2\epsilon\}/p$, and $\gamma = \alpha\max(r_0,1)(p+2\alpha) + \epsilon/2$. Then simple algebra shows that $t > p\delta + 1$, $p\delta + r_0 > 2\gamma$, and $2\gamma > \max(r_0,1) - p\gamma/\alpha$.

Let $K_n = \lceil n^{1/(2\alpha+p)}(\log n)^{\delta}\rceil$, $\underline{\epsilon}_n = n^{-\alpha/(p+2\alpha)}(\log n)^{\gamma}$, and $\mathcal{M}_n = \bigcup_{K=1}^{K_n}\mathcal{M}_{\underline{\sigma}}^{\overline{\sigma}}(\mathcal{F}_K)$. We see that $\epsilon_n > \underline{\epsilon}_n$ since by construction $t/2 = \alpha\max(r_0,1)/(p+2\alpha) + \max(1-r_0,0)/2 + 3\epsilon/2 > \alpha\max(r_0,1)/(p+2\alpha) + \epsilon/2 = \gamma$. By the construction, Lemma A.1, Proposition 1, and the fact that $\|f\|_{L_r(\mathbb{P}_{\mathbf{x}})} \leq \|f\|_\infty$ for any $r \geq 1$, we have

$$\begin{aligned}
&\exp(-n\epsilon_n^2)\sum_{K=1}^{K_n}\sqrt{\mathcal{N}(\epsilon_n, \mathcal{M}_{\underline{\sigma}}^{\overline{\sigma}}(\mathcal{F}_K), H)}\sqrt{\Pi(p_{f,\sigma} \in \mathcal{M}_{\underline{\sigma}}^{\overline{\sigma}}(\mathcal{F}_K))}\\
&\lesssim \exp\left(-n\epsilon_n^2 + \log\frac{1}{\epsilon_n}\right)\sum_{K=1}^{K_n}\exp\left\{K^p[(m+1)^p + p + 1]\left(\log\frac{1}{C_3\epsilon_n^2}\right)\right\}\\
&\leq \exp\left(-n\epsilon_n^2 + \log\frac{1}{\epsilon_n}\right)K_n\exp\left\{3K_n^p[(m+1)^p + p + 1]\left(\log\frac{1}{\epsilon_n}\right)\right\}\\
&\lesssim \exp(-n\epsilon_n^2)\exp\left\{4K_n^p[(m+1)^p + p + 2]\left(\log\frac{1}{\epsilon_n}\right)\right\}\\
&\lesssim \exp\left[-n^{p/(p+2\alpha)}(\log n)^t + C_1' n^{p/(p+2\alpha)}(\log n)^{p\delta+1}\right] \to 0
\end{aligned}$$

for some constant $C_1' > 0$, where we have used the fact $t > p\delta + 1$ in the last inequality. On the other hand, for sufficiently large n and some constants $b_1', B_1 > 0$, we argue that $\mathcal{M}_n$ capture sufficiently large prior mass. In fact, simple algebra yields

$$\begin{aligned}
\Pi\left(p_{f,\sigma} \in \mathcal{M}_n^c\right) &\leq B_1\exp\left[-b_1 K_n^p\left(\log K_n^p\right)^{r_0}\right]\\
&\leq \exp\left[-b_1' n^{p/(p+2\alpha)}(\log n)^{p\delta+r_0}\right]\\
&\leq \exp(-4n\underline{\epsilon}_n^2),
\end{aligned}$$

where the fact $p\delta + r_0 > 2\gamma$ is applied. Lastly, for the prior concentration, we have

$$\Pi(p_{f,\sigma} \in B_{\mathrm{KL}}(p_0, \underline{\epsilon}_n)) \geq \exp\left[-C_2\underline{\epsilon}_n^{-p/\alpha}\left(\log\frac{1}{\underline{\epsilon}_n}\right)^{\max(r_0,1)}\right]$$

$$\geq \exp\left[-C_2' n^{p/(p+2\alpha)} (\log n)^{\max(r_0,1)-p\gamma/\alpha}\right]$$
$$\geq \exp(-n\underline{\epsilon}_n^2)$$

for some constant $C_2' > 0$ by Proposition A.1, where we use the fact $2\gamma > \max(r_0, 1) - p\gamma/\alpha$ in the last inequality. Hence we conclude that $\Pi\left(H\left(p_{f,\sigma}, p_0\right) > M\epsilon_n \mid \mathcal{D}_n\right) \to 0$ in $\mathbb{P}_0$-probability for some constant $M > 0$. The proof is completed by applying Lemma A.1. □

B Additional Proofs for Section 3

Proof of Lemma 1. Suppose $f \in B_K^\star$. Define $\theta_{\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{s}:|\mathbf{s}|=0}^m \xi_{\mathbf{ks}}\psi_{\mathbf{ks}}(x)$, $\mathbf{k} \in [K]^p$, where $\psi_{\mathbf{ks}}(\mathbf{x})$'s are the kernel mixture of polynomial system, and $\widetilde{\theta}_{\mathbf{k}}(\mathbf{x})$ to be the Taylor polynomial of f_0 at $\boldsymbol{\mu}_{\mathbf{k}}^\star$:

$$\widetilde{\theta}_{\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil\alpha-1\rceil} \frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}_{\mathbf{k}}^\star)}{s_1!\dots s_p!}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}.$$

Notice that $\|(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}\|_\infty$ is bounded uniformly over $\boldsymbol{\mu}_{\mathbf{k}}^\star, \mathbf{s}, \mathbf{k}$. By the Taylor's expansion, for all $\mathbf{x} \in \mathcal{X}$ we have

$$\left|f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right| = \left|f_0(\mathbf{x}) - \sum_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil\alpha-1\rceil} \frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}_{\mathbf{k}}^\star)}{s_1!\dots s_p!}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}\right| \leq \tilde{C}_1 \|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^\star\|_\infty^\alpha$$

for some constant $\tilde{C}_1 > 0$. Since we assume that f_0 satisfies the α-Hölder condition globally over $\mathcal{X}$, the constant $\tilde{C}_1$ does not depend on $\boldsymbol{\mu}_{\mathbf{k}}^\star$. By the Cauchy-Schwarz inequality $(a + b)^2 \leq 2a^2 + 2b^2$, we write

$$\begin{aligned}\|f - f_0\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2 &\leq 2\mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x})\left(f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right)\right]^2 + 2\mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x})\left(\widetilde{\theta}_{\mathbf{k}}(\mathbf{x}) - \theta_{\mathbf{k}}(\mathbf{x})\right)\right]^2 \\ &= 2I_K + 2J_K.\end{aligned}$$

By the Jensen's inequality, for any $a > 0$, we proceed to derive

$$\begin{aligned}I_K &\leq \mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left(f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right)^2\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left(f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right)^2 \mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}\|_\infty > a)\right] \\ &\quad + \mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left(f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right)^2 \mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}\|_\infty \leq a)\right].\end{aligned}$$

Since $\|f_0 - \widehat{\theta}_{\mathbf{k}}\|_\infty \le A + B$ for all k, where the constant A is the uniform upper bound on $\{\|f\|_\infty : f \in \bigcup_{K=1}^{\infty} \mathcal{F}_K\}$, then we apply the Taylor approximation to obtain

$$\begin{aligned}
I_K &\lesssim \mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\mathbf{1}(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_\infty > a)\right] \\
&\quad + \mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^\star\|_\infty^{2\alpha}\mathbf{1}(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_\infty \le a)\right] \\
&\le \mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\mathbf{1}(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_\infty > a)\right] \\
&\quad + \mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_\infty + \|\boldsymbol{\mu}_{\mathbf{k}}-\boldsymbol{\mu}_{\mathbf{k}}^\star\|_\infty\right)^{2\alpha}\mathbf{1}(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_\infty \le a)\right] \\
&\le \mathbb{E}_{\mathbf{x}}\left[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\mathbf{1}(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_\infty > a)\right] + (a+h)^{2\alpha}.
\end{aligned}$$

Now pick $a = h$. Since $\varphi(\mathbf{x}) \le \mathbf{1}(\|\mathbf{x}\| \le 1)$, then $w_{\mathbf{k}}(\mathbf{x})\mathbf{1}(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_\infty > a) = 0$, and hence $\mathbb{E}_{\mathbf{x}}[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\mathbf{1}(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_\infty > a)] = 0$. It follows that $I_K \le \tilde{C}_1^2(a+h)^{2\alpha} \lesssim h^{2\alpha} \lesssim \epsilon^2$ when ϵ is sufficiently small. Similarly by Jensen's inequality and Cauchy's inequality $(a+b)^2 \le 2a^2 + 2b^2$ we write

$$\begin{aligned}
J_K &\le \mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[\sum_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil\alpha-1\rceil}\left(\xi_{\mathbf{ks}} - \frac{D^{\mathbf{s}}f_0(\boldsymbol{\mu}_{\mathbf{k}}^\star)}{s_1!\dots s_p!}\right)(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}} + \sum_{\mathbf{s}:|\mathbf{s}|=\lceil\alpha-1\rceil+1}^{m}\xi_{\mathbf{ks}}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}\right]^2\right\} \\
&\le 2\mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[\sum_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil\alpha-1\rceil}\left(\xi_{\mathbf{ks}} - \frac{D^{\mathbf{s}}f_0(\boldsymbol{\mu}_{\mathbf{k}}^\star)}{s_1!\dots s_p!}\right)(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}\right]^2\right\} \\
&\quad + 2\mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[\sum_{\mathbf{s}:|\mathbf{s}|=\lceil\alpha-1\rceil+1}^{m}\xi_{\mathbf{ks}}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}\right]^2\right\} \\
&\le 2\epsilon^2\mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[]\sum_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil\alpha-1\rceil}\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^\star\|_\infty^{|\mathbf{s}|}\right]^2\right\} \\
&\quad + 2B^2\mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[\sum_{\mathbf{s}:|\mathbf{s}|=\lceil\alpha-1\rceil+1}^{m}\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^\star\|_\infty^{|\mathbf{s}|}\right]^2\right\}.
\end{aligned}$$

The first term on the right-hand side is upper bounded by ϵ^2 up to a constant. Now we analyze the second term. Write

$$
\begin{aligned}
&B^2\mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[\sum_{\mathbf{s}:|\mathbf{s}|=\lceil\alpha-1\rceil+1}^{m}\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty}^{|\mathbf{s}|}\right]^2\right\}\\
&\quad\lesssim \mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[\sum_{\mathbf{s}:|\mathbf{s}|=\lceil\alpha-1\rceil+1}^{m}(h+\|\boldsymbol{\mu}_{\mathbf{k}}-\boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty})^{|\mathbf{s}|}\mathbf{1}(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_{\infty}\le h)\right]^2\right\}\\
&\quad\lesssim \mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[\sum_{\mathbf{s}:|\mathbf{s}|=\lceil\alpha-1\rceil+1}^{m}\left(h+\frac{1}{2K}\right)^{|\mathbf{s}|}\mathbf{1}(\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}\|_{\infty}\le h)\right]^2\right\}.
\end{aligned}
$$

Now that $h+1/2K\lesssim 1/K$, $h+1/2K\le 1$ for sufficiently large K, and $|s|\ge\alpha$, it follows that

$$
\begin{aligned}
&B^2\mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[\sum_{\mathbf{s}:|\mathbf{s}|=\lceil\alpha-1\rceil+1}^{m}\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty}^{|\mathbf{s}|}\right]^2\right\}\\
&\quad\lesssim \mathbb{E}_{\mathbf{x}}\left\{\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left[\sum_{\mathbf{s}:|\mathbf{s}|=\lceil\alpha-1\rceil+1}^{m}\left(\frac{1}{K}\right)^{|\mathbf{s}|}\right]^2\right\}\lesssim\left(\frac{1}{K}\right)^{2\alpha}\le\epsilon^2.
\end{aligned}
$$

We conclude that $J_K\lesssim\epsilon^2$, and hence $2I_K+2J_K\lesssim\epsilon^2$. To sum up, there exists a constant C_1, such that for sufficiently small $\epsilon>0$, $\|f-f_0\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2\le 2I_K+2J_K\le C_1\epsilon^2$. The proof is thus completed. $\square$

The following lemma that quantifies the local averaging behavior of the kernel mixture weights $(w_{\mathbf{k}}(\mathbf{x}))_{\mathbf{k}\in[K]^p}$ plays a fundamental role in estimating the metric entropies of $\mathcal{F}_K$. This lemma is also used in the proof of Proposition 1.

Lemma B.1. *Let $w_{j\mathbf{k}}(\mathbf{x})=\varphi_{h_j}(\mathbf{x}-\boldsymbol{\mu}_{j\mathbf{k}})/D_j(\mathbf{x})$, where $D_j(\mathbf{x})=\sum_{\mathbf{l}\in[K]^p}\varphi_{h_j}(\mathbf{x}-\boldsymbol{\mu}_{j\mathbf{l}})$, $\boldsymbol{\mu}_{j\mathbf{k}}\in\mathcal{X}_K(\mathbf{k})$, and $Kh_j\in[\underline{h},\overline{h}]$, $j=1,2$, $\mathbf{k}\in[K]^p$. Then*

$$
\|w_{1\mathbf{k}}(\mathbf{x})-w_{2\mathbf{k}}(\mathbf{x})\|_{\infty}\lesssim\left|\frac{1}{h_1}-\frac{1}{h_2}\right|+\frac{1}{h_2}\max_{\mathbf{l}\in[K]^p}\|\boldsymbol{\mu}_{1\mathbf{l}}-\boldsymbol{\mu}_{2\mathbf{l}}\|_{\infty}.
$$

Proof of Lemma B.1. Let $\tilde{\mathbf{v}}=\left[\underline{h}^{-1},\dots,\underline{h}^{-1}\right]^{\mathrm{T}}\in\mathbb{R}^p$, and $b=\varphi(\tilde{v})>0$. Suppose $\mathbf{x}\in\mathcal{X}$ is a fixed point. Then there exists a unique $\mathbf{k}_{\mathbf{x}}\in[K]^p$ such that $\mathbf{x}\in\mathcal{X}_K(\mathbf{k}_{\mathbf{x}})$. Since

$$
\left\|\frac{\mathbf{x}-\boldsymbol{\mu}_{j\mathbf{k}_{\mathbf{x}}}}{h_j}\right\|_{\infty}\le\frac{K}{\underline{h}}\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}_{\mathbf{x}}}^{\star}\|_{\infty}+\frac{K}{\underline{h}}\|\boldsymbol{\mu}_{\mathbf{k}_{\mathbf{x}}}^{\star}-\mu_{jk_x}\|_{\infty}\le\frac{1}{\underline{h}}=\|\tilde{v}\|_{\infty},
$$

it follows that

$$D_j(\mathbf{x}) = \sum_{\mathbf{l}\in[K]^p} \varphi_{h_j}(\mathbf{x}-\boldsymbol{\mu}_{j\mathbf{l}}) \geq \varphi_{h_j}(\mathbf{x}-\boldsymbol{\mu}_{j\mathbf{k}}) \geq \varphi(\tilde{\mathbf{v}}) = b,$$

since $\varphi(\mathbf{x})$ decreases as $\|\mathbf{x}\|_\infty$ increases. On the other hand, observe that for non-negative $(u_\mathbf{l}, v_\mathbf{l})_{\mathbf{l}\in[K]^p}$ with $\sum_{\mathbf{l}\in[K]^p} u_\mathbf{l} > 0$ and $v_\mathbf{k} > 0$ for some $\mathbf{k} \in [K]^p$,

$$\begin{aligned}\left|\frac{u_\mathbf{k}}{\sum_{\mathbf{l}\in[K]^p} u_\mathbf{l}} - \frac{v_\mathbf{k}}{\sum_{\mathbf{l}\in[K]^p} v_\mathbf{l}}\right| &\leq \frac{|u_\mathbf{k}-v_\mathbf{k}|}{\sum_{\mathbf{l}\in[K]^p} u_\mathbf{l}} + v_\mathbf{k}\left|\frac{1}{\sum_{\mathbf{l}\in[K]^p} u_\mathbf{l}} - \frac{1}{\sum_{\mathbf{l}\in[K]^p} v_\mathbf{l}}\right| \\ &\leq \frac{|u_\mathbf{k}-v_\mathbf{k}|}{\sum_{\mathbf{l}\in[K]^p} u_\mathbf{l}} + v_\mathbf{k}\sum_{\mathbf{l}\in[K]^p} \frac{|v_\mathbf{l}-u_\mathbf{l}|}{\sum_{\mathbf{l}\in[K]^p} u_\mathbf{l}\sum_{\mathbf{l}\in[K]^p} v_\mathbf{l}} \\ &\leq \frac{|u_\mathbf{k}-v_\mathbf{k}|}{\sum_{\mathbf{l}\in[K]^p} u_\mathbf{l}} + v_\mathbf{k}\sum_{\mathbf{l}\in[K]^p} \frac{|v_\mathbf{l}-u_\mathbf{l}|}{v_\mathbf{k}\sum_{\mathbf{l}\in[K]^p} u_\mathbf{l}} \lesssim \frac{\sum_{\mathbf{l}\in[K]^p}|u_\mathbf{l}-v_\mathbf{l}|}{\sum_{\mathbf{l}\in[K]^p}|u_\mathbf{l}|}.\end{aligned} \tag{2}$$

Suppose that $w_{1\mathbf{k}}(\mathbf{x}) > 0$ or $w_{2\mathbf{k}}(vx) > 0$. Without loss of generality we may assume that $w_{2\mathbf{k}}(\mathbf{x}) > 0$. It follows that

$$\begin{aligned}|w_{1\mathbf{k}}(\mathbf{x}) - w_{2\mathbf{k}}(\mathbf{x})| &\lesssim \frac{\sum_{\mathbf{l}\in[K]^p} |\varphi_{h_1}(\mathbf{x}-\boldsymbol{\mu}_{1\mathbf{l}}) - \varphi_{h_2}(\mathbf{x}-\boldsymbol{\mu}_{2\mathbf{l}})|}{\sum_{\mathbf{l}\in[K]^p} \varphi_{h_1}(\mathbf{x}-\boldsymbol{\mu}_{1\mathbf{l}})} \\ &\leq \frac{1}{b}\sum_{\mathbf{l}\in\mathcal{K}_1(\mathbf{x})\cup\mathcal{K}_2(\mathbf{x})} \left|\varphi\left(\frac{\mathbf{x}-\boldsymbol{\mu}_{1\mathbf{l}}}{h_1}\right) - \varphi\left(\frac{\mathbf{x}-\boldsymbol{\mu}_{2\mathbf{l}}}{h_2}\right)\right|\end{aligned}$$

where $\mathcal{K}_j(\mathbf{x}) = \{\mathbf{l}\in[K]^p : \|\mathbf{x}-\boldsymbol{\mu}_{j\mathbf{l}}\|_\infty \leq h_j\}$, $j = 1,2$. This is because when $\mathbf{l} \notin \mathcal{K}_1(\mathbf{x}) \cup \mathcal{K}_2(\mathbf{x})$, $\varphi_{h_1}(\mathbf{x}-\boldsymbol{\mu}_{1\mathbf{l}}) = \varphi_{h_2}(\mathbf{x}-\boldsymbol{\mu}_{2\mathbf{l}}) = 0$. Furthermore when $w_{1\mathbf{k}}(\mathbf{x}) = w_{2\mathbf{k}}(\mathbf{x}) = 0$, the above inequality also holds. Now let $\mathbf{k}_\mathbf{x} \in [K]^p$ to be the unique index such that $\mathbf{x} \in \mathcal{X}_K(\mathbf{k}_\mathbf{x})$. We claim that

$$\mathcal{K}_1(\mathbf{x}) \cup \mathcal{K}_2(\mathbf{x}) \subset \mathcal{K}(\mathbf{k}_\mathbf{x}) := \left\{\mathbf{l}\in[K]^p : \|\boldsymbol{\mu}^\star_{\mathbf{k}_\mathbf{x}} - \boldsymbol{\mu}^\star_\mathbf{l}\|_\infty \leq \frac{2\overline{h}}{K}\right\}.$$

In fact, if $l \notin \mathcal{K}(\mathbf{k}_\mathbf{x})$, then we have

$$\|\mathbf{x}-\boldsymbol{\mu}_{j\mathbf{l}}\|_\infty \geq \|\boldsymbol{\mu}^\star_{\mathbf{k}_\mathbf{x}} - \boldsymbol{\mu}^\star_\mathbf{l}\|_\infty - \|\mathbf{x}-\boldsymbol{\mu}^\star_{\mathbf{k}_\mathbf{x}}\|_\infty - \|\boldsymbol{\mu}^\star_\mathbf{l} - \boldsymbol{\mu}_{j\mathbf{l}}\|_\infty > h_j$$

by triangle inequality for $j = 1,2$, Namely, $l \notin \mathcal{K}_1(\mathbf{x}) \cup \mathcal{K}_2(\mathbf{x})$, $j = 1,2$. Hence $\mathcal{K}_1(\mathbf{x}) \cup \mathcal{K}_2(\mathbf{x}) \subset \mathcal{K}(\mathbf{k}_\mathbf{x})$. Since φ is continuous and is compactly supported, then by defining L_φ to be the Lipschitz constant of φ we proceed to compute

$$\begin{aligned}|w_{1\mathbf{k}}(\mathbf{x}) - w_{2\mathbf{k}}(\mathbf{x})| &\lesssim \sum_{\mathbf{l}'\in\mathcal{K}(\mathbf{k}_\mathbf{x})} \left|\varphi\left(\frac{\mathbf{x}-\boldsymbol{\mu}_{1\mathbf{l}'}}{h_1}\right) - \varphi\left(\frac{\mathbf{x}-\boldsymbol{\mu}_{2\mathbf{l}'}}{h_2}\right)\right| \\ &\leq \sum_{\mathbf{l}'\in\mathcal{K}(k_x)} L_\varphi \max_{\mathbf{l}\in[K]^p}\left\|\frac{\mathbf{x}-\boldsymbol{\mu}_{1\mathbf{l}}}{h_1} - \frac{\mathbf{x}-\boldsymbol{\mu}_{2\mathbf{l}}}{h_2}\right\|_\infty\end{aligned}$$

$$
\begin{aligned}
&\leq L_\varphi |\mathcal{K}(\mathbf{k_x})| \max_{\mathbf{l}\in[K]^p} \left[(\|\mathbf{x}\|_\infty + \|\boldsymbol{\mu}_{1\mathbf{l}}\|_\infty) \left| \frac{1}{h_1} - \frac{1}{h_2} \right| + \frac{1}{h_2} \|\boldsymbol{\mu}_{1\mathbf{l}} - \boldsymbol{\mu}_{2\mathbf{l}}\|_\infty \right] \\
&\lesssim \left| \frac{1}{h_1} - \frac{1}{h_2} \right| + \frac{1}{h_2} \max_{\mathbf{l}\in[K]^p} \|\boldsymbol{\mu}_{1\mathbf{l}} - \boldsymbol{\mu}_{2\mathbf{l}}\|_\infty,
\end{aligned}
$$

where we have used the fact that $|\mathcal{K}(\mathbf{k_x})| \leq \lceil 4\bar{h}+2 \rceil^p$ for all $\mathbf{x} \in \mathcal{X}$, $\|\mathbf{x}\|_\infty \leq 1$, and $\|\boldsymbol{\mu}_{1\mathbf{k}}\|_\infty \leq 1$. To see why $|\mathcal{K}(\mathbf{k_x})| \leq \lceil 4\bar{h}+2 \rceil^p$, notice $\mathcal{K}(\mathbf{k_x}) = \{\mathbf{l} \in [K]^p : \|\mathbf{k_x} - l\|_\infty \leq 2\bar{h}\}$, and hence has cardinality at most $\lceil 4\bar{h}+2 \rceil^p$. Now taking the supremum over $\mathbf{x} \in \mathcal{X}$ to the above display completes the proof. □

Proof of Proposition 1. Let

$$
f_j(\mathbf{x}) = \sum_{\mathbf{k}\in[K]^p} \sum_{\mathbf{s}:|\mathbf{s}|=0}^{m} \xi_{\mathbf{ks}}^{(j)} w_{j\mathbf{k}}(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}
$$

be in $\mathcal{F}_K$, where $w_{j\mathbf{k}}(\mathbf{x}) = \varphi_{h_j}(\mathbf{x} - \boldsymbol{\mu}_{j\mathbf{k}})/D_j(\mathbf{x})$ and $D_j(\mathbf{x}) = \sum_{\mathbf{l}\in[K]^p} \varphi_{h_j}(\mathbf{x} - \boldsymbol{\mu}_{j\mathbf{l}})$, $j = 1, 2$. Denote $\theta_{j\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{s}:|\mathbf{s}|=0}^m \xi_{\mathbf{ks}}^{(j)} w_{j\mathbf{k}}(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}$, $j = 1, 2$. We proceed to compute

$$
\begin{aligned}
\|f_1 - f_2\|_\infty &\leq \max_{\mathbf{x}\in\mathcal{X}} \left| \sum_{\mathbf{k}} w_{1\mathbf{k}}(\mathbf{x})(\theta_{1\mathbf{k}}(\mathbf{x}) - \theta_{2\mathbf{k}}(\mathbf{x})) \right| + \max_{\mathbf{x}\in\mathcal{X}} \left| \sum_{\mathbf{k}} (w_{1\mathbf{k}}(\mathbf{x}) - w_{2\mathbf{k}}(\mathbf{x}))\theta_{2\mathbf{k}}(\mathbf{x}) \right| \\
&\leq \max_{\mathbf{k}\in[K]^p} \max_{0\leq|\mathbf{s}|\leq m} \left| \xi_{\mathbf{ks}}^{(1)} - \xi_{\mathbf{ks}}^{(2)} \right| \max_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{k}} w_{1\mathbf{k}}(\mathbf{x}) \max_{\mathbf{l}\in[K]^p} \sum_{\mathbf{s}:|\mathbf{s}|=0}^{m} \|(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{l}}^\star)^{\mathbf{s}}\|_\infty \\
&\quad + \max_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{k}} |w_{1\mathbf{k}}(\mathbf{x}) - w_{2\mathbf{k}}(\mathbf{x})| \sum_{\mathbf{s}:|\mathbf{s}|=0}^{m} \left| \xi_{\mathbf{ks}}^{(2)} \right| \|(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}\|_\infty.
\end{aligned}
$$

Since $\max_{|\mathbf{s}|=0,1,\ldots,m} |\xi_{\mathbf{ks}}^{(2)}| \leq B$ for all $\mathbf{k}$ and $\|(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{l}}^\star)^{\mathbf{s}}\|_\infty$ is upper bounded by a universal constant, $\mathbf{l} \in [K]^p$, $|\mathbf{s}| = 0, 1, \ldots, m$, it follows that

$$
\begin{aligned}
\|f_1 - f_2\|_\infty &\lesssim \max_{\mathbf{k}\in[K]^p} \max_{0\leq|\mathbf{s}|\leq m} \left| \xi_{\mathbf{ks}}^{(1)} - \xi_{\mathbf{ks}}^{(2)} \right| + \max_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{k}} |w_{1\mathbf{k}}(\mathbf{x}) - w_{2\mathbf{k}}(\mathbf{x})| \\
&\leq \max_{\mathbf{k}\in[K]^p} \max_{0\leq|\mathbf{s}|\leq m} \left| \xi_{\mathbf{ks}}^{(1)} - \xi_{\mathbf{ks}}^{(2)} \right| \\
&\qquad + \max_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{k}} |w_{1\mathbf{k}}(\mathbf{x}) - w_{2\mathbf{k}}(\mathbf{x})| \left[\mathbf{1}(w_{1\mathbf{k}}(\mathbf{x}) > 0) + \mathbf{1}(w_{2\mathbf{k}}(\mathbf{x}) > 0) \right] \\
&\lesssim \max_{\mathbf{k}\in[K]^p} \max_{0\leq|\mathbf{s}|\leq m} \left| \xi_{\mathbf{ks}}^{(1)} - \xi_{\mathbf{ks}}^{(2)} \right| \\
&\qquad + \max_{\mathbf{l}\in[K]^p} \|w_{1\mathbf{k}}(\mathbf{x}) - w_{2\mathbf{k}}(\mathbf{x})\|_\infty \sum_{j=1}^{2} \left\| \sum_{\mathbf{k}\in[K]^p} \mathbf{1}(w_{j\mathbf{k}}(\mathbf{x}) > 0) \right\|_\infty .
\end{aligned}
$$

For any $\mathbf{x} \in \mathcal{X}$, there exists a unique $\mathbf{k_x} \in [K]^p$ such that $\mathbf{x} \in \mathcal{X}_K(\mathbf{k_x})$. Observe that $\sum_{\mathbf{k}\in[K]^p} \mathbf{1}(w_{j\mathbf{k}}(\mathbf{x}) > 0)$ is the same as the cardinality of the index set $\{\mathbf{l} \in [K]^p :$

$\|\mathbf{x}-\boldsymbol{\mu}_{j\mathbf{l}}\|_\infty < h\}$. We now argue that

$$\{\mathbf{l}\in[K]^p : \|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{l}}\|_\infty < h\} \subset \left\{\mathbf{l}\in[K]^p : \left\|\boldsymbol{\mu}^\star_{\mathbf{k}_\mathbf{x}}-\boldsymbol{\mu}^\star_{\mathbf{l}}\right\|_\infty \le \frac{4\overline{h}}{K}\right\},$$

where the cardinality of the right-hand side set of the last display is upper bounded by $\lceil 8\overline{h}+2\rceil^p$. Suppose $\mathbf{l}$ is in the complement of the right-hand side of the last display. Then the (reverse) triangle inequality yields

$$\begin{aligned}\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{l}}\|_\infty &\ge \|\boldsymbol{\mu}^\star_{\mathbf{k}_\mathbf{x}}-\boldsymbol{\mu}^\star_{\mathbf{l}}\|_\infty - \|\mathbf{x}-\boldsymbol{\mu}^\star_{\mathbf{k}_\mathbf{x}}\|_\infty - \|\boldsymbol{\mu}^\star_{\mathbf{l}}-\boldsymbol{\mu}_{\mathbf{l}}\|_\infty \\ &\ge 4\overline{h}/K - 1/(2K) - \underline{h}/K > h,\end{aligned}$$

finishing the argument for the claim that $\sum_{\mathbf{k}\in[K]^p}\mathbf{1}(w_{j\mathbf{k}}(\mathbf{x})>0)$ can be upper bounded by a constant only depending on $\overline{h}$.

Therefore, we obtain

$$\begin{aligned}&\|f_1-f_2\|_\infty \\ &\lesssim \max_{\mathbf{k}\in[K]^p}\|w_{1\mathbf{k}}(\mathbf{x})-w_{2\mathbf{k}}(\mathbf{x})\|_\infty + \max_{\mathbf{k}\in[K]^p}\max_{0\le|\mathbf{s}|\le m}\left|\xi^{(1)}_{\mathbf{ks}}-\xi^{(2)}_{\mathbf{ks}}\right| \\ &\lesssim K\max_{\mathbf{k}\in[K]^p}\|\boldsymbol{\mu}_{1\mathbf{k}}-\boldsymbol{\mu}_{2\mathbf{k}}\|_\infty + K^2|h_1-h_2| + \max_{\mathbf{k}\in[K]^p,0\le|\mathbf{s}|\le m}\left|\xi^{(1)}_{\mathbf{ks}}-\xi^{(2)}_{\mathbf{ks}}\right|, \qquad (3)\end{aligned}$$

where we have applied Lemma B.1 in the last inequality.

We now construct the ϵ-net for $\mathcal{F}_K$. Let $\mathcal{E}_\xi(\epsilon)$ be an ϵ-net of $[-B,B]$, $\mathcal{E}_{\boldsymbol{\mu}_\mathbf{k}}(\epsilon)$ be an ϵ-net of $\mathcal{X}_K(\mathbf{k})$ with respect to $\|\cdot\|_\infty$, $\mathbf{k}\in[K]^p$, and $\mathcal{E}_h(\epsilon)$ be an ϵ-net of $\left[\underline{h}/K,\overline{h}/K\right]$. Then $|\mathcal{E}_\xi(\epsilon)|\le 3B/\epsilon$, $|\mathcal{E}_{\boldsymbol{\mu}_\mathbf{k}}(\epsilon)|\le(2/K\epsilon)^p$, and $|\mathcal{E}_h(\epsilon)|\le(2\overline{h}-2\underline{h})/(K\epsilon)$ for sufficiently small ϵ. We claim that

$$\left\{\sum_{\mathbf{k}\in[K]^p}\sum_{|\mathbf{s}|=0}^m \xi'_{\mathbf{ks}}\psi'_{\mathbf{ks}}(\mathbf{x}) : \xi'_{\mathbf{ks}}\in\mathcal{E}_\xi(\epsilon), \boldsymbol{\mu}'_\mathbf{k}\in\mathcal{E}_{\boldsymbol{\mu}_\mathbf{k}}\left(\frac{\epsilon}{K}\right), h'\in\mathcal{E}_h\left(\frac{\epsilon}{K^2}\right)\right\}$$

is an $\tilde{c}_1\epsilon$-net of $\mathcal{F}_K$ with $\|\cdot\|_{L_1(\mathbb{P}_\mathbf{x})}$ for some constant $\tilde{c}_1>0$, where

$$\psi'_{\mathbf{ks}}(x) = w'_\mathbf{k}(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}^\star_\mathbf{k})^\mathbf{s}, \quad w'_\mathbf{k}(\mathbf{x}) = \varphi_{h'}(\mathbf{x}-\boldsymbol{\mu}'_\mathbf{k})/D'(\mathbf{x}), \quad D'(x) = \sum_{\mathbf{l}\in[K]^p}\varphi_{h'}(\mathbf{x}-\boldsymbol{\mu}'_\mathbf{l}).$$

In fact, for all $f(\mathbf{x})\in\mathcal{F}_K$ of the form (5), there exist some $h'\in\mathcal{E}_h(\epsilon/K^2)$, $\xi'_{\mathbf{ks}}\in\mathcal{E}_{\xi_{\mathbf{ks}}}(\epsilon), \boldsymbol{\mu}'_\mathbf{k}\in\mathcal{E}_{\boldsymbol{\mu}_\mathbf{k}}(\epsilon/K)$ for each $\mathbf{k}$ and each $\mathbf{s}$, such that $|h-h'|<\epsilon/K^2$, $|\xi_{\mathbf{ks}}-\xi'_{\mathbf{ks}}|<\epsilon$, $\|\boldsymbol{\mu}_\mathbf{k}-\boldsymbol{\mu}'_\mathbf{k}\|<\epsilon/K$. Let $f'(\mathbf{x})=\sum_{\mathbf{k}\in[K]^p}\sum_{|\mathbf{s}|=0}^m\xi'_{\mathbf{ks}}w'_\mathbf{k}(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}^\star_\mathbf{k})^\mathbf{s}\in\mathcal{F}'_K$. It follows by (3) that $\|f-f'\|_\infty\le C_1\epsilon$. Hence

$$\begin{aligned}\mathcal{N}\left(C_1\epsilon,\mathcal{F}_K,\|\cdot\|_\infty\right) &\le |\mathcal{E}_\xi(\epsilon)|^{K^p(m+1)^p}\left|\mathcal{E}_h\left(\frac{\epsilon}{K^2}\right)\right|\prod_k\left|\mathcal{E}_{\mu_k}\left(\frac{\epsilon}{K}\right)\right| \\ &\le \left(\frac{3B}{\epsilon}\right)^{K^p(m+1)^p}\left[\frac{2(\overline{h}-\underline{h})K}{\epsilon}\right]\left[\left(\frac{2}{\epsilon}\right)^p\right]^{K^p}\end{aligned}$$

$$\leq \exp\left\{2K^p\left[(m+1)^p+p+1\right]\left(\log\frac{1}{\epsilon}\right)\right\}$$

when ϵ is sufficiently small. Taking logarithm to both sides of the last display completes the proof of the second inequality.

Now we prove the first inequality. Suppose $(f_j)_{j=1}^N$ forms an ϵ-net of $\mathcal{F}_K$ with respect to $\|\cdot\|_\infty$ such that $N=\mathcal{N}(\epsilon,\mathcal{F}_K,\|\cdot\|_\infty)$. Then define $l_j(\mathbf{x})=\max(f_j(\mathbf{x})-\epsilon,-A)$ and $u_j(\mathbf{x})=\min(f_j(\mathbf{x})+\epsilon,A)$, yielding the brackets $([u_j,l_j])_{j=1}^N$ such that $\mathcal{F}_K\subset\bigcup_{j=1}^N[l_j,u_j]$. Furthermore, $\|l_j-u_j\|_{L_r(\mathbb{P}_{\mathbf{x}})}\leq\|l_j-u_j\|_\infty\leq 2\epsilon$. The proof is completed by the fact that $\log N=\log\mathcal{N}(\epsilon,\mathcal{F}_K,\|\cdot\|_\infty)$. □

Proof of Theorem 2. Note that we take the closure of $\mathcal{X}_K(k)$ so that the sieve maximum likelihood estimator exists. Also, when σ_0 is unknown, the computation of $\widehat{f}_K$ is not affected since the maximum likelihood estimator of σ_0 is equivalent to the least-squared estimator under the assumption of Gaussian noises. Furthermore, we remark that the metric entropy bound for $\mathcal{F}_K$ in Proposition 1 also applies to $\mathcal{G}_K$.

We follow the notation $\mathcal{M}_{\underline{\sigma}}^{\overline{\sigma}}(\mathcal{G})=\{p_{f,\sigma}:f\in\mathcal{G},\sigma\in[\underline{\sigma},\overline{\sigma}]\}$ used in the proof of Theorem 1. Denote $\tilde{\epsilon}_n=(\log n/n)^{\alpha/(2\alpha+p)}$. We first give an upper bound for the bracketing integral $J_{[\cdot]}\left(\tilde{\epsilon}_n,\mathcal{M}_{\sigma_0}^{\sigma_0}(\mathcal{G}_{K_n}),H\right)$. For convenience denote $\gamma=\alpha/(p+2\alpha)$. By Lemma A.1 and Proposition 1, we have by simple algebra:

$$J_{[\cdot]}\left(\tilde{\epsilon}_n,\mathcal{M}_{\sigma_0}^{\sigma_0}(\mathcal{G}_{K_n}),H\right)\leq\int_0^{\tilde{\epsilon}_n}\sqrt{\log\mathcal{N}\left(2C_3\epsilon^2,\mathcal{G}_{K_n},\|\cdot\|_\infty\right)}\mathrm{d}\epsilon\lesssim\frac{n^{\frac{p}{2(p+2\alpha)}}}{(\log n)^{\frac{p\gamma}{2\alpha}}}\int_0^{\tilde{\epsilon}_n}\sqrt{\log\frac{1}{\epsilon}}\mathrm{d}\epsilon.$$

Observe the following fact $\lim_{x\to\infty}\int_x^\infty u^2\mathrm{e}^{-u^2}\mathrm{d}u/(x\mathrm{e}^{-x^2})=1/2<1$. Then change of variable $\epsilon\mapsto\sqrt{\log(1/\epsilon)}$ yields

$$\int_0^{\tilde{\epsilon}_n}\sqrt{\log\frac{1}{\epsilon}}\mathrm{d}\epsilon=2\int_{(\log 1/\tilde{\epsilon}_n)^{1/2}}^\infty u^2\mathrm{e}^{-u^2}\mathrm{d}u\lesssim\tilde{\epsilon}_n\sqrt{\log\frac{1}{\tilde{\epsilon}_n}}\lesssim\tilde{\epsilon}_n\sqrt{\log n}. \tag{4}$$

Hence $J_{[\cdot]}\left(\tilde{\epsilon}_n,\mathcal{M}_{\sigma_0}^{\sigma_0}(\mathcal{G}_{K_n}),H\right)\lesssim n^{p/(2p+4\alpha)}(\log n)^{1/2-p\gamma/(2\alpha)}\tilde{\epsilon}_n=\sqrt{n}\tilde{\epsilon}_n^2$. Now define

$$\begin{aligned}\widetilde{f}_n(x)&=\underset{f\in\mathcal{G}_{K_n}}{\arg\min}\,D_{\mathrm{KL}}(p_0\|p_{f,\sigma_0}),\\ \delta_n&=D_{\mathrm{KL}}\left(p_0\middle\|p_{\widetilde{f}_n,\sigma_0}\right),\\ \tau_n&=\mathbb{E}_0\left[\log\frac{p_0(\mathbf{x},y)}{p_{\widetilde{f}_n,\sigma_0}(\mathbf{x},y)}\right]^2.\end{aligned}$$

Direct computation yields that for any function f,

$$D_{\mathrm{KL}}\left(p_0\|p_{f,\sigma_0}\right)=\frac{1}{2\sigma_0^2}\|f-f_0\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2.$$

Since $K_n \geq \left\lceil \tilde{\epsilon}_n^{-1/\alpha} \right\rceil$, then there exists some $\widetilde{g}_n \in \mathcal{G}_{K_n} \cap B^\star_{K_n}$ such that

$$\delta_n = D_{\mathrm{KL}}\left(p_0 \| p_{f,\sigma_0}\right) \leq D_{\mathrm{KL}}\left(p_0 \| p_{\widetilde{g}_n,\sigma_0}\right) \lesssim \|f - f_0\|^2_{L_2(\mathbb{P}_{\mathbf{x}})} \lesssim \tilde{\epsilon}_n^2,$$

where $B^\star_K$ is defined in Lemma 1. Since

$$\tau_n \lesssim \|f - f_0\|^2_{L_2(\mathbb{P}_{\mathbf{x}})} \lesssim D_{\mathrm{KL}}\left(p_0 \middle\| p_{\widetilde{f}_n,\sigma_0}\right) = \delta_n \lesssim \tilde{\epsilon}_n^2$$

by Lemma A.2, it follows that $\max(\delta_n, \tau_n) \lesssim \tilde{\epsilon}_n^2$. Now we apply Theorem 4 (ii) in Wong and Shen (1995) to conclude that

$$\mathbb{P}_0(\|f_0 - \widehat{f}_{K_n}\|_{L_2(\mathbb{P}_{\mathbf{x}})} > M\tilde{\epsilon}_n) \leq \mathbb{P}_0(H(p_0, p_{\widehat{f}_{K_n}}) > M\tilde{\epsilon}_n) \to 0$$

for some constant $M > 0$. □

C Proofs for the Results in Appendix B

Proof of Lemma A.1. The keys of the proof are a basic inequality $x/(x+1) \leq 1 - \mathrm{e}^{-x} \leq x$ for $x > 0$ and the closed-form formula for Hellinger distance between Gaussians:

$$\begin{aligned} &\frac{1}{2}\int_{\mathbb{R}} \left[\sqrt{\phi_{\sigma_1}(y-\mu_1)} - \sqrt{\phi_{\sigma_2}(y-\mu_2)}\right]^2 \mathrm{d}y \\ &\quad = 1 - \left[1 - \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2}\right]^{\frac{1}{2}} \exp\left[-\frac{1}{4\sigma_1^2+4\sigma_2^2}(\mu_1-\mu_2)^2\right], \end{aligned}$$

where $\phi_\sigma(\cdot)$ is the density of $\mathrm{N}(0,\sigma^2)$. Now we derive

$$\begin{aligned} &H^2(p_{f,\sigma_1}, p_{g,\sigma_2}) \\ &\leq \int_{\mathcal{X}} \left\{1 - \left[1 - \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2}\right]^{\frac{1}{2}} \exp\left[-\frac{(f(\mathbf{x})-g(\mathbf{x}))^2}{8\underline{\sigma}^2}\right]\right\} p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\leq \int_{\mathcal{X}} \left\{1 - \left[1 - \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2}\right]^{\frac{1}{2}} + \left[1 - \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2}\right]^{\frac{1}{2}} \frac{(f(\mathbf{x})-g(\mathbf{x}))^2}{8\underline{\sigma}^2}\right\} p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\leq \int_{\mathcal{X}} \left\{1 - \left[1 - \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2}\right]^{\frac{1}{2}} + \frac{(f(\mathbf{x})-g(\mathbf{x}))^2}{8\underline{\sigma}^2}\right\} p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\leq \frac{(\sigma_1-\sigma_2)^2}{2\underline{\sigma}^2} + \frac{1}{8\underline{\sigma}^2}\int_{\mathcal{X}} (f(\mathbf{x})-g(\mathbf{x}))^2 p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\leq \frac{(\sigma_1-\sigma_2)^2}{2\underline{\sigma}^2} + \frac{A}{4\underline{\sigma}^2}\int_{\mathcal{X}} |f(\mathbf{x})-g(\mathbf{x})| p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\lesssim |\sigma_1-\sigma_2|^2 + \|f-g\|_{L_1(\mathbb{P}_{\mathbf{x}})}, \end{aligned}$$

and hence the right inequality of the first assertion holds. It follows that there exists some constant $C_3 > 0$, such that for any $f, g \in \mathcal{G}$, $\sigma_1, \sigma_2 \in [\underline{\sigma}, \overline{\sigma}]$, $2C_3 H^2\left(p_{f,\sigma_1}, p_{g,\sigma_2}\right) \leq$

$|\sigma_1 - \sigma_2| + \|f - g\|_{L_1(\mathbb{P}_{\mathbf{x}})}$, and hence, $\{p_{g,\sigma'} : \sigma' \in B(\sigma, C_3\epsilon), g \in B_{\|\cdot\|_{L_1(\mathbb{P}_{\mathbf{x}})}}(f, C_3\epsilon^2)\} \subset B_H(p_{f,\sigma}, \epsilon^2)$ for any $\sigma \in [\underline{\sigma}, \overline{\sigma}]$ and $f \in \mathcal{G}$. The second entropy inequality naturally follows. On the other hand,

$$\begin{aligned}
&H^2(p_{f,\sigma_1}, p_{g,\sigma_2}) \\
&\geq \int_{\mathcal{X}} \left\{ 1 - \left[1 - \frac{(\sigma_1 - \sigma_2)^2}{\sigma_1^2 + \sigma_2^2} \right]^{\frac{1}{2}} \exp\left[-\frac{(f(\mathbf{x}) - g(\mathbf{x}))^2}{8\overline{\sigma}^2} \right] \right\} p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\geq \int_{\mathcal{X}} \left\{ \frac{(f(\mathbf{x}) - g(\mathbf{x}))^2/(8\overline{\sigma}^2) + 1 - \left[1 - (\sigma_1 - \sigma_2)^2/(\sigma_1^2 + \sigma_2^2)\right]^{\frac{1}{2}}}{(f(\mathbf{x}) - g(\mathbf{x}))^2/(8\overline{\sigma}^2) + 1} \right\} p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\geq \int_{\mathcal{X}} \left[\frac{(f(\mathbf{x}) - g(\mathbf{x}))^2/(8\overline{\sigma}^2) + (\sigma_1 - \sigma_2)^2/(4\overline{\sigma}^2)}{(f(\mathbf{x}) - g(\mathbf{x}))^2/(8\overline{\sigma}^2) + 1} \right] p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\gtrsim \left[2|\sigma_1 - \sigma_2|^2 + \int_{\mathcal{X}} (f(\mathbf{x}) - g(\mathbf{x}))^2 p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \right] \\
&\gtrsim \|f - g\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2 + |\sigma_1 - \sigma_2|^2.
\end{aligned}$$

Therefore the left inequality of the first assertion holds. □

Proof of Lemma A.2. Directly compute

$$\log \frac{p_0(\mathbf{x}, y)}{p_{f,\sigma}(\mathbf{x}, y)} = \frac{1}{2\sigma^2}(y - f(\mathbf{x}))^2 - \frac{1}{2\sigma_0^2}(y - f_0(\mathbf{x}))^2 + \frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2}.$$

Therefore

$$D_{\mathrm{KL}}(p_0 || p_{f,\sigma}) \leq \frac{1}{2}\left(\frac{|\sigma_0^2 - \sigma^2|}{\underline{\sigma}^2} + \left|\log\frac{\sigma^2}{\sigma_0^2}\right| + \frac{1}{\underline{\sigma}^2}\|f - f_0\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2 \right).$$

Since $\sigma \mapsto \sigma^2, \sigma \mapsto \log\sigma^2$ are continuously differentiable, and σ, σ_0 lie in the compact interval $[\underline{\sigma}, \overline{\sigma}]$, it follows that they are Lipschitz-continuous, and hence $D_{\mathrm{KL}}(p_0||p_{f,\sigma}) \lesssim |\sigma_0 - \sigma| + \|f - f_0\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2$. For the second moment of the log-likelihood ratio, by the Cauchy-Schwarz inequality $(a + b)^2 \leq 2a^2 + 2b^2$ we have

$$\mathbb{E}_0\left[\log\frac{p_0(\mathbf{x}, y)}{p_{f,\sigma}(\mathbf{x}, y)}\right]^2 \leq 2\mathbb{E}_0\left[\frac{(y - f(\mathbf{x}))^2}{2\sigma^2} - \frac{(y - f_0(\mathbf{x}))^2}{2\sigma_0^2}\right]^2 + 2\left(\frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2}\right)^2.$$

Since

$$\begin{aligned}
&\mathbb{E}_0\left[\frac{(y - f(\mathbf{x}))^2}{2\sigma^2} - \frac{(y - f_0(\mathbf{x}))^2}{2\sigma_0^2}\right]^2 \\
&\quad= \mathbb{E}_{\mathbf{x}}\left[\frac{(f_0 - f)^4 + 6\sigma_0^2(f_0 - f)^2 - 2\sigma^2(f - f_0)^2}{4\sigma^4} + \frac{3(\sigma^2 - \sigma_0^2)^2}{4\sigma^4}\right] \\
&\quad\leq \mathbb{E}_{\mathbf{x}}\left[\frac{4A^2 + 6\sigma_0 + 2\overline{\sigma}^2}{4\underline{\sigma}^4}(f_0 - f)^2 + \frac{6\overline{\sigma}^3}{\underline{\sigma}^4}|\sigma - \sigma_0|\right],
\end{aligned}$$

we conclude that $\mathbb{E}_0[\log p_0(\mathbf{x}, y)/p_{f,\sigma}(\mathbf{x}, y)]^2 \lesssim |\sigma - \sigma_0| + \|f - f_0\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2$. □

Proof of Proposition A.1. By Lemma A.2 there exists some constant C_1', such that for sufficiently small $\epsilon > 0$, $B_{\mathrm{KL}}(p_0, \epsilon) \supset \{p_{f,\sigma} : \|f - f_0\|^2_{L_2(\mathbb{P}_{\mathbf{x}})} \le C_1'\epsilon^2, |\sigma - \sigma_0| \le C_1'\epsilon^2\}$. By Lemma 1, we can directly compute

$$
\begin{aligned}
\Pi\left(B_{\mathrm{KL}}(p_0,\epsilon)\right) &\ge \sum_{K \ge \epsilon^{-1/\alpha}} \pi_K(K) \prod_{\mathbf{k}\in[K]^p} \prod_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil \alpha-1 \rceil} \Pi\left(\left|\xi_{\mathbf{ks}} - \frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}_{\mathbf{k}}^\star)}{s_1!\ldots s_p!}\right| \le C_1''\epsilon \middle| K\right) \\
&\quad \times \Pi(\sigma : |\sigma - \sigma_0| \le C_1''\epsilon^2) \\
&= \sum_{K \ge \epsilon^{-1/\alpha}} \pi_K(K) \prod_{\mathbf{k}\in[K]^p} \prod_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil \alpha-1 \rceil} \int_{\max\left\{\frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}_{\mathbf{k}}^\star)}{s_1!\ldots s_p!} - C_1''\epsilon, -B\right\}}^{\min\left\{\frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}_{\mathbf{k}}^\star)}{s_1!\ldots s_p!} + C_1''\epsilon, B\right\}} \pi_\xi(\xi)\mathrm{d}\xi \\
&\quad \times \int_{\max(\underline{\sigma}, \sigma_0 - C_1''\epsilon)}^{\min(\overline{\sigma}, \sigma_0 + C_1''\epsilon)} \pi_\sigma(\sigma)\mathrm{d}\sigma \\
&\ge \sum_{K = \lceil \epsilon^{-1/\alpha} \rceil}^{\infty} \exp\left[-b_0 K^p \left(\log K^p\right)^{r_0} - \tilde{C}_2 (K\lceil \alpha - 1 \rceil + K)^p \left(\log \frac{1}{\epsilon}\right)\right] \\
&\ge \exp\left[-C_2 \epsilon^{-\frac{p}{\alpha}} \left(\log\frac{1}{\epsilon}\right)^{\max(r_0,1)}\right]
\end{aligned}
$$

for some constants $C_1'', \tilde{C}_2, C_2 > 0$. The proof is thus completed. □

D Remaining Proofs for Section 4

This section contains the proofs of Theorem 3 and Theorem 4 of the manuscript. In what follows we introduce several notations that will be extensively used in the following proofs. Given a function $f : \mathcal{X} \to \mathbb{R}$, we use $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(\mathbf{x}_i)$ to denote the empirical measure of f, and $\mathbb{G}_n f = n^{-1/2}\sum_{i=1}^n (f(\mathbf{x}_i) - \mathbb{E}f(\mathbf{x}_i))$ to denote the empirical process of f, given the independent and identically distributed data $(\mathbf{x}_i)_{i=1}^n$. For any $t > 2\alpha\max(r_0, 1)/(p + 2\alpha) - \max(1 - r_0, 0) > 0$, define

$$
\begin{cases}
\epsilon = \dfrac{t}{3} - \dfrac{2\alpha\max(r_0,1)}{3(p+2\alpha)} - \dfrac{\max(1-r_0,0)}{3}, \\
\delta = \dfrac{2\alpha\max(r_0,1)}{p(p+2\alpha)} - \dfrac{r_0 + 2\epsilon}{p}, \\
\gamma = \dfrac{\alpha\max(r_0,1)}{p+2\alpha} + \dfrac{\epsilon}{2}.
\end{cases} \tag{5}
$$

Then simple algebra shows

$$
t > p\delta + 1, \quad p\delta + r_0 > 2\gamma, \quad 2\gamma > \max(\gamma_0, 1) - \frac{p\gamma}{\alpha}. \tag{6}
$$

For any $\Theta \subset \mathbb{R}^q$ and any function class $\mathcal{G}$, denote $\mathcal{P}(\mathcal{G}, \Theta) = \{p_{\beta,\eta}(x,z,y) : \eta \in \mathcal{G}, \beta \in \Theta\}$. Let $\Theta_J = \{\boldsymbol{\beta} \in \mathbb{R}^q : \|\boldsymbol{\beta}\| \le J\}$. For any functions $f, g : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, define the ρ_r distance between f and g to be

$$\rho_r(f,g) = \left[\int_{\mathcal{X}}\int_{\mathcal{Z}} |f(\mathbf{x},\mathbf{z}) - g(\mathbf{x},\mathbf{z})|^r \, p_{(\mathbf{x},\mathbf{z})}(\mathbf{x},\mathbf{z}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z}\right]^{1/r}$$

for any $r \in [1, +\infty)$. The proof of Theorem 3 relies on the following Proposition concerning the contraction rate of density estimation with respect to the Hellinger topology.

Proposition D.1. *Assume that f_0 is in the α-Hölder function class $\mathfrak{C}^{\alpha,B}(\mathcal{X})$ with envelope B and the design space $\mathcal{X}$ is the unit hypercube $[0,1]^p$. Under the setup and prior specification in Section 4.1, there exists some constant $M > 0$ such that*

$$\Pi\left(H\left(p_{\boldsymbol{\beta},\eta}, p_0\right) > M\epsilon_n \mid \mathcal{D}_n\right) \to 0$$

in $\mathbb{P}_0$-probability, where $\epsilon_n = n^{-\alpha/(2\alpha+p)}(\log n)^{t/2}$, and $t > 2\alpha \max(r_0, 1)/(2\alpha + p) + \max(0, 1 - r_0)$.

Proof. Denote

$$B_{\mathrm{KL}}(p_0, \epsilon) = \left\{p_{\boldsymbol{\beta},\eta} : D_{\mathrm{KL}}(p_0 || p_{\boldsymbol{\beta},\eta}) < \epsilon^2, \mathbb{E}_0\left[\log \frac{p_0(\mathbf{x},\mathbf{z},y)}{p_{\boldsymbol{\beta},\eta}(\mathbf{x},\mathbf{z},y)}\right]\}^2 < \epsilon^2\right\}.$$

Let $K_n = \lceil n^{1/(2\alpha+p)}(\log n)^{\delta}\rceil$, $J_n = K_n^p$, $\mathcal{M}_n = \bigcup_{K=1}^{K_n} \mathcal{M}_{nK}$ where $\mathcal{M}_{nK} = \mathcal{P}\left(\mathcal{F}_K, \Theta_{J_n}\right)$, $\epsilon_n = n^{-\alpha/(2\alpha+p)}(\log n)^{t/2}$ and $\underline{\epsilon}_n = n^{-\alpha/(p+2\alpha)}\left(\log n\right)^{\gamma}$, where δ and γ are defined in (5). We complete the proof by verifying (12), (13), and (14), which are originally presented in Kruijer et al. (2010).

We first verify (13). Since $\mathcal{M}_n = \bigcup_{K=1}^{K_n} \mathcal{P}\left(\mathcal{F}_K, \Theta_{J_n}\right)$, then

$$\begin{aligned}\Pi(p_{\boldsymbol{\beta},\eta} \in \mathcal{M}_n^c) &\le \Pi_\eta(K > K_n) + \Pi_{\boldsymbol{\beta}}(\boldsymbol{\beta} \in \Theta_{J_n}^c) \\ &\le B_1 \exp\left[-b_1 K_n^p (\log K_n^p)^{r_0}\right] + \Pi\left(\|\boldsymbol{\beta}\|_2^2 \ge J_n^2\right) \\ &\le B_1 \exp\left[-b_1 K_n^p (\log K_n^p)^{r_0}\right] + \sqrt{2^q}\exp\left(-\frac{J_n^2}{4}\right)\end{aligned}$$

for some constants $b_1, B_1 > 0$, where we have used the Chernoff bound, the fact that $\|\boldsymbol{\beta}\|^2 \sim \chi^2(q)$, and $E_\Pi\left\{\exp\left(\|\boldsymbol{\beta}\|_2^2/4\right)\right\} = \sqrt{2^q}$ under the prior Π in the last inequality. Since for sufficiently large n, $K_n^p(\log K_n^p)^{r_0} \lesssim J_n^2$, it follows from simple algebra that

$$\Pi(p_{\boldsymbol{\beta},\eta} \in \mathcal{M}_n^c) \le \exp\left[-b_1'' n^{\frac{p}{p+2\alpha}} (\log n)^{p\delta + r_0}\right] \le \exp(-4n\underline{\epsilon}_n^2)$$

for some constant $b_1'' > 0$ when n is sufficiently large, where (6) is applied.

We next verify (14). Simple algebra leads to $H^2(p_{\boldsymbol{\beta}_1,\eta_1}, p_{\boldsymbol{\beta}_2,\eta_2}) \lesssim \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + \|\eta_1 - \eta_2\|_{L_1(\mathbb{P}_{\mathbf{x}})}$. Hence $\mathcal{N}\left(\epsilon, \mathcal{M}_{nK}, H\right) \le \mathcal{N}\left(C_4\epsilon^2, \mathcal{F}_K, \|\cdot\|_{L_1(\mathbb{P}_{\mathbf{x}})}\right) \times \mathcal{N}(C_4\epsilon, \Theta_{J_n}, \|\cdot\|)$ for some constant $C_4 > 0$. For the finite dimensional Θ_{J_n}, the ϵ-covering number can be easily upper bounded: $\mathcal{N}(C_4\epsilon, \Theta_{J_n}, \|\cdot\|) \lesssim (J_n/\epsilon)^q$. Hence by Lemma A.1 and Proposition 1, for sufficiently large K and sufficiently small ϵ_n,

$$\exp(-n\epsilon_n^2)\sum_{K=1}^{K_n}\sqrt{\mathcal{N}\left(\epsilon_n, \mathcal{M}_{nK}, H\right)}\sqrt{\Pi(p_{\beta,\eta} \in \mathcal{M}_{nK})}$$

$$
\begin{aligned}
&\lesssim \exp(-n\epsilon_n^2)\sqrt{\mathcal{N}(C_4\epsilon_n, \Theta_{J_n}, \|\cdot\|)}\sum_{K=1}^{K_n}\sqrt{\mathcal{N}\left(C_4\epsilon_n^2, \mathcal{F}_K, \|\cdot\|_{L_1(\mathbb{P}_{\mathbf{x}})}\right)}\\
&\lesssim \exp\left(-n\epsilon_n^2 + q\log\frac{J_n}{\epsilon_n}\right)\sum_{K=1}^{K_n}\exp\left\{K^p[(m+1)^p+p+1]\left(\log\frac{1}{C_4\epsilon_n^2}\right)\right\}\\
&\lesssim \exp(-n\epsilon_n^2 + C_4'\log n)\exp\left\{4K_n^p[(m+1)^p+p+1]\left(\log\frac{1}{\epsilon_n}\right)\right\}\\
&\lesssim \exp\left[-n^{p/(p+2\alpha)}(\log n)^t + C_4'' n^{p/(p+2\alpha)}(\log n)^{p\delta+1}\right] \to 0
\end{aligned}
$$

for some constants $C_4', C_4'' > 0$, where (6) is used in the last inequality.

Lastly, we verify (12). Since

$$
\rho_2^2(\boldsymbol{\beta}^{\mathrm{T}}z + \eta(x), \boldsymbol{\beta}_0^{\mathrm{T}}z + \eta_0(x)) \le \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2\left\|\mathbb{E}\mathbf{z}\mathbf{z}^{\mathrm{T}}\right\| + \|\eta-\eta_0\|_{L_2(\mathbb{P}_{\mathbf{x}})}^2,
$$

which can be derived by direct computation, then as a consequence of Lemma 1, we see that

$$
B_K^\star \times B_2(\boldsymbol{\beta}_0, \epsilon) \subset \left\{(\eta, \boldsymbol{\beta}) : \rho_2^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z} + \eta(\mathbf{x}), \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{z} + \eta_0(\mathbf{x})) < (C_1 + \|\mathbb{E}\mathbf{z}\mathbf{z}^{\mathrm{T}}\|)\epsilon^2\right\}.
$$

where $K = \lceil \epsilon^{-1/\alpha} \rceil$, and $B_K^\star$ is defined in Lemma 1. Furthermore by simple algebra $D_{\mathrm{KL}}(p_0 || p_{\boldsymbol{\beta},\eta}) = \rho_2^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}+\eta(\mathbf{x}), \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{z}+\eta_0(\mathbf{x}))/2$, and the second moment of the likelihood ratio can be upper bounded using the Cauchy-Schwarz inequality $(a+b)^2 \le 2a^2 + 2b^2$:

$$
\begin{aligned}
&\mathbb{E}_0\left[\log\frac{p_0(\mathbf{x},\mathbf{z},y)}{p_{\boldsymbol{\beta},\eta}(\mathbf{x},\mathbf{z},y)}\right]^2\\
&\quad = \frac{1}{4}\mathbb{E}\left\{\left[\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}_0-\boldsymbol{\beta}) + (\eta_0-\eta)(\mathbf{x})\right]^2\left[\left(\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}_0-\boldsymbol{\beta}) + (\eta_0-\eta)(\mathbf{x})\right)^2 + 4\right]\right\}\\
&\quad \le \frac{1}{4}\mathbb{E}\left\{\left[\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}_0-\boldsymbol{\beta}) + (\eta_0-\eta)(\mathbf{x})\right]^2\left[2\max_{\mathbf{z}\in\mathcal{Z}}\|z\|^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + 8A^2 + 4\right]\right\}\\
&\quad \lesssim (\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + 1)\rho_2^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z} + \eta(\mathbf{x}), \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{z} + \eta_0(\mathbf{x})).
\end{aligned}
$$

When $\boldsymbol{\beta} \in B_2(\boldsymbol{\beta}_0, \epsilon)$ for $\epsilon \le 1$, the second moment of the log-likelihood ratio is upper bounded by $\rho_2^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z} + \eta(\mathbf{x}), \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{z} + \eta_0(\mathbf{x}))$ up to a multiplicative constant. It follows that $\{p_{\boldsymbol{\beta},\eta} : (\eta, \boldsymbol{\beta}) \in B_K^\star \times B_2(\boldsymbol{\beta}_0, \epsilon)\} \subset B_{\mathrm{KL}}(p_0, C_5\epsilon)$ for some constant $C_5 > 0$ when $K = \lceil \epsilon^{-1/\alpha} \rceil$. For the prior mass of β in $B_2(\boldsymbol{\beta}_0, \epsilon)$, we have

$$
\Pi_{\boldsymbol{\beta}}(B_2(\boldsymbol{\beta}_0, \epsilon)) \ge \frac{\mathrm{Vol}(B_2(0,\epsilon))}{\sqrt{(2\pi)^q}}\exp\left(-\frac{\|\boldsymbol{\beta}_0\|+1}{2}\right) \ge \exp\left[-C_6\left(\log\frac{1}{\epsilon}\right)\right]
$$

for some constant $C_6 > 0$ when $\epsilon < 1$. Hence by Proposition A.1, for sufficiently small ϵ,

$$
\begin{aligned}
\Pi\left(p_{\boldsymbol{\beta},\eta} \in B_{\mathrm{KL}}(p_0, C_6\epsilon)\right) &\ge \Pi_\eta(\eta \in B_K^\star)\Pi_{\boldsymbol{\beta}}(\boldsymbol{\beta} \in B_2(\boldsymbol{\beta}_0, \epsilon))\\
&\ge \exp\left[-2C_2\epsilon^{-p/\alpha}\left(\log\frac{1}{\epsilon}\right)^{\max(r_0,1)}\right].
\end{aligned}
$$

Substituting $C_6\epsilon$ by $\underline{\epsilon}_n$, we obtain

$$\Pi\left(p_{\boldsymbol{\beta},\eta} \in B_{\mathrm{KL}}(p_0, \underline{\epsilon}_n)\right) \geq \exp\left[-C_7 \underline{\epsilon}_n^{-p/\alpha}\left(\log\frac{1}{\underline{\epsilon}_n}\right)^{\max(r_0,1)}\right] \geq \exp(-n\underline{\epsilon}_n^2)$$

for some constant $C_7 > 0$, where (6) is used in the last inequality. $\square$

Proof of Theorem 3. Since

$$\frac{(f-g)^2}{[8+(f-g)^2]} \leq 1 - \exp\left[-\frac{(f-g)^2}{8}\right] = H^2(\phi(y-f), \phi(y-g))$$

and $\left((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^{\mathrm{T}}\mathbf{z} + (\eta_1 - \eta_2)\right)^2 \leq 2\overline{B}^2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + 8A^2$, it follows that

$$\begin{aligned}
H^2(p_{\beta_1,\eta_1}, p_{\beta_2,\eta_2}) &= \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[\frac{1}{2}\int_{\mathbb{R}}\left[\sqrt{\phi(y - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_1 - \eta_1(\mathbf{x}))} - \sqrt{\phi(y - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2 - \eta_2(\mathbf{x}))}\right]^2 \mathrm{d}y\right] \\
&\geq \frac{\mathbb{E}_{\mathbf{x},\mathbf{z}}\left\{\left[(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^{\mathrm{T}}\mathbf{z} + (\eta_1(\mathbf{x}) - \eta_2(\mathbf{x}))\right]^2\right\}}{2\overline{B}^2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + 8 + 8A^2} \\
&\geq \frac{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 \lambda_{\min}\left(\mathbb{E}\mathbf{z}\mathbf{z}^{\mathrm{T}}\right) + \|\eta_1 - \eta_2\|^2_{L_2(\mathbb{P}_{\mathbf{x}})}}{2\overline{B}^2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + 8 + 8A^2} \\
&\gtrsim \frac{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + \|\eta_1 - \eta_2\|^2_{L_2(\mathbb{P}_{\mathbf{x}})}}{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + 1},
\end{aligned}$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a real symmetric matrix. Hence

$$\{(\boldsymbol{\beta}, \eta) : H(p_{\boldsymbol{\beta},\eta}, p_0) \leq M\epsilon_n\} \subset \left\{(\boldsymbol{\beta}, \eta) : \frac{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\eta - \eta_0\|^2_{L_2(\mathbb{P}_{\mathbf{x}})}}{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + 1} \leq M_1^2\epsilon_n^2\right\}$$

for some large constant $M' > 0$. Since for sufficiently large n, $M'^2\epsilon_n^2 < 1/2$, we see that

$$\begin{aligned}
&\left\{(\boldsymbol{\beta}, \eta) : \frac{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\eta - \eta_0\|^2_{L_2(\mathbb{P}_{\mathbf{x}})}}{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + 1} \leq M_1^2\epsilon_n^2\right\} \\
&\quad \subset \left\{(\boldsymbol{\beta}, \eta) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\eta - \eta_0\|^2_{L_2(\mathbb{P}_{\mathbf{x}})} \leq \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + M_1^2\epsilon_n^2\right\} \\
&\quad \subset \left\{(\boldsymbol{\beta}, \eta) : \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\eta - \eta_0\|^2_{L_2(\mathbb{P}_{\mathbf{x}})} \leq M_1^2\epsilon_n^2\right\}.
\end{aligned}$$

It follows that $\{(\boldsymbol{\beta}, \eta) : H(p_{\boldsymbol{\beta},\eta}, p_0) \leq M\epsilon_n\} \subset \{(\boldsymbol{\beta}, \eta) : \|\eta - \eta_0\|^2_{L_2(\mathbb{P}_{\mathbf{x}})} \leq M_1^2\epsilon_n^2\}$. Now applying Proposition D.1 completes the proof. $\square$

Proof of Theorem 4. The proof consists of verifying the conditions F.2, F.3, and F.4 of Theorem F.3, which was originally proved in Yang et al. (2015). Let $t > 2\alpha\max(r_0, 1)/(p+$

$2\alpha)+\max(1-r_0,0)$, $\epsilon_n = n^{-\alpha/(2\alpha+p)}(\log n)^{t/2}$, $\underline{\epsilon}_n = n^{-\alpha(p+2\alpha)}(\log n)^{\gamma}$, $K_n = \lceil n^{1/(2\alpha+p)}(\log n)^{\delta}\rceil$, and $J_n = K_n^p$, where γ and δ are given by (5).

We first verify condition F.2. Exploiting the proof of Proposition D.1, one finds that

$$\Pi\left(p_{\boldsymbol{\beta},\eta} \in B_{\mathrm{KL}}(p_0, \underline{\epsilon}_n)\right) \geq \exp(-n\underline{\epsilon}_n^2), \quad \Pi\left\{\left[\|\boldsymbol{\beta}\| \leq J_n, \eta \in \bigcup_{K=1}^{K_n} \mathcal{F}_K\right]^c\right\} \leq \exp(-4n\underline{\epsilon}_n^2),$$

where

$$B_{\mathrm{KL}}(p_0, \epsilon) = \left\{p_{\boldsymbol{\beta},\eta} : D_{\mathrm{KL}}(p_0 || p_{\boldsymbol{\beta},\eta}) < \epsilon^2, E\left[\log \frac{p_0(\mathbf{x},\mathbf{z},y)}{p_{\boldsymbol{\beta},\eta}(\mathbf{x},\mathbf{z},y)}\right]\right\}$$

is the Kullback-Leibler ball. Then Lemma 1 in Ghosal et al. (2007) (also see Lemma F.1) yields

$$\mathbb{E}_0\left[\Pi\left(\|\boldsymbol{\beta}\| \leq J_n, \eta \in \bigcup_{K=1}^{K_n} \mathcal{F}_K \middle| \mathcal{D}_n\right)\right] \leq \mathbb{E}_0\left[\Pi\left(\eta \in \bigcup_{K=1}^{K_n} \mathcal{F}_K \middle| \mathcal{D}_n\right)\right] \to 1.$$

By Theorem 3, $\Pi\left(\eta : \|\eta-\eta_0\|_{L_2(\mathbb{P}_{\mathbf{x}})} \leq M\epsilon_n \middle| \mathcal{D}_n\right) = 1 - o_{\mathbb{P}_0}(1)$ for some constant $M > 0$. Let $\widehat{\mathcal{F}}_n = \left\{\eta : \|\eta-\eta_0\|_{\mathbb{L}_2(\mathbb{P}_{\mathbf{x}})} \leq M\epsilon_n\right\} \cap \bigcup_{K=1}^{K_n} \mathcal{F}_K$. It follows that $\Pi(\eta \in \widehat{\mathcal{F}}_n \mid \mathcal{D}_n) = 1 - o_{\mathbb{P}_0}(1)$. Now we consider bounding $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|$. By the proof of Theorem 3, there exists some constant $M_1 > 0$ such that for sufficiently large n,

$$\Pi\left(\frac{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + \|\eta-\eta_0\|^2_{\mathbb{L}_2(\mathbb{P}_{\mathbf{x}})}}{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2+1} \leq M_1^2\epsilon_n^2 \middle| \mathcal{D}_n\right) = 1 - o_{\mathbb{P}_0}(1).$$

Observing that for sufficiently large n with $1 - M_1^2\epsilon_n^2 \geq 1/4$,

$$\left\{(\boldsymbol{\beta},\eta) : \frac{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + \|\eta-\eta_0\|^2_{\mathbb{L}_2(\mathbb{P}_{\mathbf{x}})}}{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2+1} \leq M_1^2\epsilon_n^2\right\} \subset \left\{(\boldsymbol{\beta},\eta) : \frac{1}{4}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 \leq M_1^2\epsilon_n^2\right\},$$

we conclude that $\Pi\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \leq M\epsilon_n\right) = 1 - o_{\mathbb{P}_0}(1)$ by replacing $2M_1$ by M. Therefore

$$\Pi\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \leq M\epsilon_n, \eta \in \widehat{\mathcal{F}}_n \middle| \mathcal{D}_n\right) = 1 - o_{\mathbb{P}_0}(1).$$

We now turn to condition F.4. Since the semiparametric bias $\boldsymbol{\Delta}\eta_{\boldsymbol{\beta}} = \mathbf{0}$ for all $\boldsymbol{\beta}$ as the least favorable submodel coincides with η_0 for all $\boldsymbol{\beta} \in \mathbb{R}^q$, condition F.4 automatically holds with $\widetilde{G}_n = 0$.

Now we are left with the verification of condition F.3. Since the least favorable curve $\eta^*_{\boldsymbol{\beta}}$ coincides with η_0 for all $\boldsymbol{\beta}$, the least favorable submodel is given by $\{p_{\boldsymbol{\beta},\eta_0} : \boldsymbol{\beta} \in \mathbb{R}^q\}$. The score function and Fisher information of $p_{\boldsymbol{\beta},\eta_0}$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ are

$$\boldsymbol{\ell}_0(\mathbf{x},\mathbf{z},y) = \mathbf{z}\left[y - \eta_0(\mathbf{x}) - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0\right], \quad \mathbf{I}_0 = \mathbb{E}_0\left[\ell_0(\mathbf{x},\mathbf{z},y)\ell_0(\mathbf{x},\mathbf{z},y)^{\mathrm{T}}\right] = \mathbb{E}\mathbf{z}\mathbf{z}^{\mathrm{T}}.$$

Let $\boldsymbol{\Delta\beta}_n = \boldsymbol{\beta}_n - \boldsymbol{\beta}_0$. Then direct computation yields

$$n\mathbb{P}_n \log \frac{p_{\boldsymbol{\beta}_n,\eta+\boldsymbol{\Delta}\eta_{\boldsymbol{\beta}_n}}}{p_{\boldsymbol{\beta}_0,\eta}} = n\boldsymbol{\Delta\beta}_n^{\mathrm{T}}\mathbb{P}_n\boldsymbol{\ell}_0 - \frac{1}{2}n\boldsymbol{\Delta\beta}_n^{\mathrm{T}}\mathbf{I}_0\boldsymbol{\Delta\beta}_n + \sqrt{n}\boldsymbol{\Delta\beta}_n^{\mathrm{T}}\mathbb{G}_n\mathbf{z}(\eta_0-\eta)$$

$$
\begin{aligned}
&-\frac{1}{2}\sqrt{n}\boldsymbol{\Delta\beta}_n^{\mathrm{T}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{z}\mathbf{z}^{\mathrm{T}}-\mathbb{E}\mathbf{z}\mathbf{z}^{\mathrm{T}}\right)\right]\boldsymbol{\Delta\beta}_n\\
&= n(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbb{P}_n\boldsymbol{\ell}_0-\frac{1}{2}n(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{I}_0(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)\\
&\quad+\sqrt{n}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbb{G}_n\mathbf{z}(\eta_0-\eta)+O_{\mathbb{P}_0}\left(\sqrt{n}\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_0\|^2\right).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\sup_{\eta\in\widehat{\mathcal{F}}_n}\left|n\mathbb{P}_n\log\frac{p_{\boldsymbol{\beta}_n,\eta+\boldsymbol{\Delta}\eta_{\boldsymbol{\beta}_n}}}{p_{\boldsymbol{\beta}_0,\eta}}-n(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbb{P}_n\ell_0+\frac{1}{2}n(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\mathrm{T}}I_0(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)\right|\\
&\quad\le\sup_{\eta\in\widehat{\mathcal{F}}_n}\left|\sqrt{n}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbb{G}_n\mathbf{z}(\eta_0-\eta)\right|+O_{\mathbb{P}_0}\left(\sqrt{n}\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_0\|^2\right)\\
&\quad\le\sqrt{n}\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_0\|\sum_{j=1}^{q}\sup_{\eta\in\widehat{\mathcal{F}}_n}|\mathbb{G}_n z_j(\eta_0-\eta)|+O_{\mathbb{P}_0}\left(\sqrt{n}\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_0\|^2\right),
\end{aligned}
$$

where $\mathbf{z}=[z_1,\dots,z_q]\in\mathbb{R}^q$. We now bound the supremum on the right-hand side of the last display. This requires the use of the maximum inequality for empirical process. Define $\mathcal{H}_{nj}=\{z_j(\eta_0-\eta)(\mathbf{x}):\eta\in\widehat{\mathcal{F}}_n\}$ for $j=1,\dots,q$. It follows that $\|z_j(\eta_0-\eta_1)(\mathbf{x})-z_j(\eta_0-\eta_2)(\mathbf{x})\|_{L_2(\mathbb{P}_{(\mathbf{x},\mathbf{z})})}\lesssim\|\eta_1-\eta_2\|_{L_2(\mathbb{P}_{\mathbf{x}})}$ whenever $\eta_1,\eta_2\in\widehat{\mathcal{F}}_n$. Namely, we obtain the following metric entropy relation:

$$
\begin{aligned}
\log\mathcal{N}_{[\cdot]}\left(\epsilon,\mathcal{H}_{nj},\|\cdot\|_{L_2(\mathbb{P}_{(x,z)})}\right)&\le\log\mathcal{N}_{[\cdot]}\left(\widehat{C}\epsilon,\widehat{\mathcal{F}}_n,\|\cdot\|_{L_2(\mathbb{P}_{\mathbf{x}})}\right)\\
&\le\log\mathcal{N}_{[\cdot]}\left(\widehat{C}\epsilon,\bigcup_{K=1}^{K_n}\mathcal{F}_K,\|\cdot\|_{L_2(\mathbb{P}_{\mathbf{x}})}\right)\\
&\le\log\left[\sum_{K=1}^{K_n}\mathcal{N}_{[\cdot]}\left(\widehat{C}\epsilon,\mathcal{F}_K,\|\cdot\|_{L_2(\mathbb{P}_{\mathbf{x}})}\right)\right]
\end{aligned}
$$

for some constant $\widehat{C}>0$. By Proposition 1 we obtain for sufficiently large n:

$$
\sum_{K=1}^{K_n}\mathcal{N}_{[\cdot]}\left(\widehat{C}\epsilon,\mathcal{F}_K,\|\cdot\|_{L_2(\mathbb{P}_{\mathbf{x}})}\right)\le\exp\left\{5K_n^p[(m+1)^p+p+1]\left(\log\frac{1}{\epsilon}\right)\right\}.
$$

Therefore we estimate the bracketing integral

$$
J_{[\cdot]}\left(\epsilon_n,\mathcal{H}_{nj},\|\cdot\|_{L_2(\mathbb{P}_{(\mathbf{x},\mathbf{z})})}\right)\lesssim K_n^{\frac{p}{2}}\int_0^{\epsilon_n}\left(\log\frac{1}{\epsilon}\right)^{\frac{1}{2}}\mathrm{d}\epsilon\lesssim K_n^{\frac{p}{2}}\epsilon_n\left(\log n\right)^{\frac{1}{2}},
$$

here (4) is applied. Applying (6) yields

$$
J_{[\cdot]}\left(\epsilon_n,\mathcal{H}_{nj},\|\cdot\|_{L_2(\mathbb{P}_{(x,z)})}\right)\le n^{-\frac{\alpha-p/2}{2\alpha+p}}\left(\log n\right)^t=\sqrt{n}\epsilon_n^2.
$$

Now we apply Lemma 19.36 in van der Vaart (2000) (see Theorem F.4). Since $\mathbb{E}[z_j(\eta_0-\eta)(\mathbf{x})]^2 = \|\eta-\eta_0\|^2_{L_2(\mathbb{P}_{\mathbf{x}})}\left(\mathbb{E}z_j^2\right) \leq (\overline{B}M)^2\epsilon_n^2$ whenever $\eta\in\widehat{\mathcal{F}}_n$, we obtain

$$\mathbb{E}\left(\sup_{\eta\in\widehat{\mathcal{F}}_n}|\mathbb{G}_n z_j(\eta_0-\eta)|\right) = E\|\mathbb{G}_n\|_{\mathcal{H}_{nj}} \lesssim J_{[\cdot]}\left(\epsilon_n,\mathcal{H}_{nj},\|\cdot\|_{L_2(\mathbb{P}_{(x,z)})}\right) \lesssim \sqrt{n}\epsilon_n^2.$$

Markov's inequality thus implies that $\sup_{\eta\in\widehat{\mathcal{F}}_n}|\mathbb{G}_n z_j(\eta_0-\eta)| = O_{\mathbb{P}_0}(\sqrt{n}\epsilon_n^2)$. Together with the previous results, we conclude that condition F.3 is satisfied with $G_n(s) = n\epsilon_n^2 s + \sqrt{n}s^2$.

To sum up, we have verified that conditions F.1, F.2, F.3, and F.4 in Theorem F.3 are satisfied with $(G_n+\widetilde{G}_n)\left(n^{-1/2}\log n\right) = \sqrt{n}\epsilon_n^2\log n + n^{-1/2}(\log n)^2 = o(1)$ due to the assumption that $\alpha > p/2$. The proof is thus completed. □

E Proof of Theorem 5

Define

$$\mathcal{F}^{B_n} = \left\{f(\mathbf{x}) = \sum_{\mathbf{k}\in[K]^p}\sum_{\mathbf{s}:|\mathbf{s}|=0}^{m}\xi_{\mathbf{ks}}\psi_{\mathbf{ks}}(\mathbf{x}) : \max_{\mathbf{k}\in[K_n]^p,|\mathbf{s}|=0,\cdots,m}|\xi_{\mathbf{ks}}|\leq B_n\right\}$$

Let $\Pi_{[\underline{\sigma},\overline{\sigma}]}(\cdot) = \Pi_\sigma(\cdot\mid\underline{\sigma}\leq\sigma\leq\overline{\sigma})$, $\Pi_f^n(\cdot\mid\sigma)$ be the prior introduced above on f given σ, $\Pi^n(\cdot,\cdot) = \Pi_f^n(\cdot\mid\sigma)\times\Pi_\sigma^n(\cdot)$, and $\Pi^n_{[\underline{\sigma},\overline{\sigma}]}(\cdot,\cdot) = \Pi_f^n(\cdot\mid\sigma)\times\Pi_{[\underline{\sigma},\overline{\sigma}]}(\cdot)$.

We first provide the following approximation result, which is analogous to Lemma 1 but is with respect to the empirical L_2-distance. In this section we use $\mathbb{E}_n f = n^{-1}\sum_{i=1}^n f(\mathbf{x}_i)$ to denote the expectation with respect to the empirical distribution of $(\mathbf{x}_i)_{i=1}^n$.

Lemma E.1 (Approximation lemma, fixed-design). *Assume that f_0 is in the α-Hölder function class $\mathfrak{C}^{\alpha,B}(\mathcal{X})$ with envelope B and the design space $\mathcal{X}$ is the unit hypercube $[0,1]^p$. Let f be of the form (8). Then there exists some constant C_1 such that*

$$\begin{aligned}B_K^\star := &\left\{f : \max_{|\mathbf{s}|=0,1,\ldots,\lceil\alpha-1\rceil}\left|\xi_{\mathbf{ks}} - \frac{D^{\mathbf{s}}f_0(\boldsymbol{\mu}_k^\star)}{s_1!\ldots s_p!}\right|\leq\epsilon, \max_{\lceil\alpha-1\rceil<|\mathbf{s}|\leq m}|\xi_{\mathbf{ks}}|\leq B, \mathbf{k}\in[K_n]^p\right\}\\ &\subset\left\{f : \|f-f_0\|^2_{L_2(\mathbb{P}_n)} < C_1\epsilon^2\right\}.\end{aligned}$$

for sufficiently small ϵ when $K\geq\epsilon^{-1/\alpha}$.

Proof. Suppose $f\in B_K^\star$. Define $\theta_{\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{s}:|\mathbf{s}|=0}^m\xi_{\mathbf{ks}}\psi_{\mathbf{ks}}(\mathbf{x})$, $\mathbf{k}\in[K_n]^p$, where $\psi_{\mathbf{ks}}(\mathbf{x})$'s are given by (8), and $\widetilde{\theta}_{\mathbf{k}}(\mathbf{x})$ to be the Taylor polynomial of f_0 at $\boldsymbol{\mu}_{\mathbf{k}}^\star$:

$$\widetilde{\theta}_{\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil\alpha-1\rceil}(s_1!\ldots s_p!)^{-1}D^{\mathbf{s}}f_0(\boldsymbol{\mu}_{\mathbf{k}}^\star)(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^\star)^{\mathbf{s}}.$$

Notice that $\|(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{k}}^{\star})^{\mathbf{s}}\|_{\infty}$ is bounded uniformly over $\boldsymbol{\mu}_{\mathbf{k}}^{\star}, \mathbf{s}, \mathbf{k}$. By the Taylor's expansion, for all $\mathbf{x} \in \mathcal{X}$ we have

$$\left|f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right| = \left|f_0(\mathbf{x}) - \sum_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil \alpha-1 \rceil} \frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}_{\mathbf{k}}^{\star})}{s_1! \dots s_p!}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star})^{\mathbf{s}}\right| \leq \tilde{C}_1 \|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty}^{\alpha}$$

for some constant $\tilde{C}_1 > 0$. Since we assume that f_0 satisfies the α-Hölder condition globally over $\mathcal{X}$, the constant $\tilde{C}_1$ does not depend on $\mu_k^{\star}$. Denote $\mathbb{E}_n$ to be the expectation with respect to the empirical distribution $\mathbb{P}_n$ on $(\mathbf{x}_i)_{i=1}^n$. By the Cauchy-Schwarz inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we write

$$\begin{aligned}
\|f - f_0\|_{L_2(\mathbb{P}_n)}^2 &\leq 2\mathbb{E}_n\left[\sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x})\left(f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right)\right]^2 + 2\mathbb{E}_n\left[\sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x})\left(\widetilde{\theta}_{\mathbf{k}}(\mathbf{x}) - \theta_{\mathbf{k}}(\mathbf{x})\right)\right]^2 \\
&= 2I_K + 2J_K.
\end{aligned}$$

By the Jensen's inequality, for any $a > 0$, we proceed to derive

$$\begin{aligned}
I_K &\leq \mathbb{E}_n\left[\sum_{\mathbf{k}\in[K_n]^p} w_{\mathbf{k}}(\mathbf{x})\left(f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right)^2\right] \\
&= \mathbb{E}_n\left[\sum_{\mathbf{k}\in[K_n]^p} w_{\mathbf{k}}(\mathbf{x})\left(f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right)^2 \mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty} > a)\right] \\
&\quad + \mathbb{E}_n\left[\sum_{\mathbf{k}\in[K]^p} w_{\mathbf{k}}(\mathbf{x})\left(f_0(\mathbf{x}) - \widetilde{\theta}_{\mathbf{k}}(\mathbf{x})\right)^2 \mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty} \leq a)\right].
\end{aligned}$$

Since $\|f_0 - \widetilde{\theta}_{\mathbf{k}}\|_{\infty}$ is uniformly bounded for all $\mathbf{k}$, then we apply the Taylor approximation to obtain

$$\begin{aligned}
I_K &\lesssim \mathbb{E}_n\left[\sum_{\mathbf{k}\in[K_n]^p} w_k(x)\mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty} > a)\right] \\
&\quad + \mathbb{E}_n\left[\sum_{\mathbf{k}\in[K_n]^p} w_{\mathbf{k}}(\mathbf{x})\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty}^{2\alpha}\mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty} \leq a)\right] \\
&\leq \mathbb{E}_n\left[\sum_{\mathbf{k}\in[K_n]^p} w_{\mathbf{k}}(\mathbf{x})\mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty} > a)\right] + \frac{1}{K^{2\alpha}}.
\end{aligned}$$

Now pick $a = h$. Since $\varphi(\mathbf{x}) \leq \mathbf{1}(\|\mathbf{x}\| \leq 1)$, then $w_{\mathbf{k}}(\mathbf{x})\mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_k^{\star}\|_{\infty} > a) = 0$, and hence

$$\mathbb{E}_n\left[\sum_{\mathbf{k}\in[K_n]^p} w_{\mathbf{k}}(\mathbf{x})\mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty} > a)\right] = 0.$$

It follows that $I_K \lesssim \epsilon^2$ when ϵ is sufficiently small. Similarly by Jensen's inequality and Cauchy's inequality $(a+b)^2 \le 2a^2 + 2b^2$ we write

$$
\begin{aligned}
J_K &\le \mathbb{E}_n \left\{ \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \left[\sum_{|\mathbf{s}|=0}^{\lceil \alpha-1 \rceil} \left(\xi_{\mathbf{ks}} - \frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}_{\mathbf{k}}^{\star})}{s_1! \dots s_p!} \right) (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star})^{\mathbf{s}} + \sum_{|\mathbf{s}|>\lceil \alpha-1 \rceil}^{m} \xi_{\mathbf{ks}} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star})^{\mathbf{s}} \right]^2 \right\} \\
&\le 2\mathbb{E}_n \left\{ \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \left[\sum_{|\mathbf{s}|=0}^{\lceil \alpha-1 \rceil} \left(\xi_{\mathbf{ks}} - \frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}_{k}^{\star})}{s_1! \dots s_p!} \right) (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star})^{\mathbf{s}} \right]^2 \right\} \\
&\quad + 2\mathbb{E}_n \left\{ \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \left[\sum_{|\mathbf{s}|>\lceil \alpha-1 \rceil}^{m} \xi_{\mathbf{ks}} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star})^{\mathbf{s}} \right]^2 \right\} \\
&\le 2\epsilon^2 \mathbb{E}_n \left\{ \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \left[\sum_{|\mathbf{s}|=0}^{\lceil \alpha-1 \rceil} \|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty}^{|\mathbf{s}|} \right]^2 \right\} \\
&\quad + 2B^2 \mathbb{E}_n \left\{ \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \left[\sum_{|\mathbf{s}|>\lceil \alpha-1 \rceil}^{m} \|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty}^{|\mathbf{s}|} \right]^2 \right\}.
\end{aligned}
$$

The first term on the right-hand side is upper bounded by ϵ^2 up to a constant. Now we analyze the second term. Write

$$
\begin{aligned}
B^2 \mathbb{E}_n & \left\{ \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \left[\sum_{|\mathbf{s}|>\lceil \alpha-1 \rceil}^{m} \|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty}^{|\mathbf{s}|} \right]^2 \right\} \\
& \lesssim \mathbb{E}_n \left\{ \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \left[\sum_{|\mathbf{s}|=\lceil \alpha-1 \rceil+1}^{m} h^{|\mathbf{s}|} \mathbf{1}(\|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty} \le h) \right]^2 \right\}.
\end{aligned}
$$

Now that $h + 1/2K \lesssim 1/K$, $h + 1/2K \le 1$ for sufficiently large K, and $|s| \ge \alpha$, it follows that

$$
\begin{aligned}
B^2 \mathbb{E}_n & \left\{ \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \left[\sum_{|\mathbf{s}|>\lceil \alpha-1 \rceil}^{m} \|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star}\|_{\infty}^{|\mathbf{s}|} \right]^2 \right\} \\
& \lesssim \mathbb{E}_n \left\{ \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \left[\sum_{|\mathbf{s}|>\lceil \alpha-1 \rceil}^{m} \left(\frac{1}{K} \right)^{|\mathbf{s}|} \right]^2 \right\} \lesssim \left(\frac{1}{K} \right)^{2|\mathbf{s}|} \le \epsilon^2.
\end{aligned}
$$

We conclude that $J_K \lesssim \epsilon^2$, and hence $2I_K + 2J_K \lesssim \epsilon^2$. To sum up, there exists a constant C_1, such that for sufficiently small $\epsilon > 0$, $\|f - f_0\|_{L_2(\mathbb{P}_n)}^2 \le 2I_K + 2J_K \le C_1 \epsilon^2$. The proof is thus completed. $\square$

We also provide the prior concentration with respect to the empirical L_2-distance.

Proposition E.1 (Prior concentration, empirical L_2-distance)**.** *Assume that f_0 is in the α-Hölder function class $\mathfrak{C}^{\alpha,B}(\mathcal{X})$ with envelope B and the design space $\mathcal{X}$ is the unit hypercube $[0,1]^p$. Under the prior $\Pi^n_{[\underline{\sigma},\overline{\sigma}]} = \Pi^n_f(\cdot \mid \sigma) \times \Pi_\sigma(\cdot)$, there exists some constant $C_2 > 0$ such that*

$$\Pi^n_{[\underline{\sigma},\overline{\sigma}]}\left(\|f - f_0\|_{L_2(\mathbb{P}_n)} + |\sigma - \sigma_0| < \epsilon\right) \geq \exp\left[-C_2\epsilon^{-p/\alpha}\left(\log n + \log\frac{1}{\epsilon}\right)\right]$$

for all sufficiently small $\epsilon > 0$ when $K_n \geq \epsilon^{-1/\alpha}$.

Proof. Without loss of generality we may assume that $C_1'\epsilon < 1$. Then

$$\begin{aligned}\inf\left\{\pi_\xi(\xi) : \left|\xi - \frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}^\star_{\mathbf{k}})}{s_1!\cdots s_p!}\right|\right\} &\geq \frac{1}{\sqrt{2\pi n^2\sigma^2}}\exp\left[-\frac{1}{2n^2\sigma^2}\left(\left|\frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}^\star_{\mathbf{k}})}{s_1!\cdots s_p!}\right| + 1\right)^2\right] \\ &\geq \frac{1}{\sqrt{2\pi n^2\overline{\sigma}^2}}\exp\left[-\frac{1}{2\underline{\sigma}^2}(B+1)^2\right].\end{aligned}$$

Hence by Lemma E.1, we can directly compute

$$\begin{aligned}&\Pi^n_{[\underline{\sigma},\overline{\sigma}]}\left(\|f - f_0\|_{L_2(\mathbb{P}_n)} + |\sigma - \sigma_0| < \epsilon\right) \\ &\geq \int_{\{|\sigma-\sigma_0|\leq C_1''\epsilon^2\}} \prod_{\mathbf{k}\in[K_n]^p}\left[\prod_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil\alpha-1\rceil}\Pi\left(\left|\xi_{\mathbf{ks}} - \frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}^\star_{\mathbf{k}})}{s_1!\ldots s_p!}\right| \leq C_1''\epsilon \mid \sigma\right)\right] \\ &\quad\times\left[\prod_{|\mathbf{s}|>\lceil\alpha-1\rceil}^{m}\Pi(|\xi_{\mathbf{ks}}| \leq B \mid \sigma)\right]\pi_{[\underline{\sigma},\overline{\sigma}]}(\sigma)\mathrm{d}\sigma \\ &= \prod_{\mathbf{k}\in[K_n]^p}\left[\prod_{\mathbf{s}:|\mathbf{s}|=0}^{\lceil\alpha-1\rceil}\int_{\frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}^\star_{\mathbf{k}})}{s_1!\ldots s_p!}-C_1''\epsilon}^{\frac{D^{\mathbf{s}} f_0(\boldsymbol{\mu}^\star_{\mathbf{k}})}{s_1!\ldots s_p!}+C_1''\epsilon}\frac{1}{\sqrt{2\pi n^2\overline{\sigma}^2}}\mathrm{e}^{-\xi^2/(2\pi\underline{\sigma}^2)}\mathrm{d}\xi\right] \\ &\quad\times\left[\prod_{|\mathbf{s}|>\lceil\alpha-1\rceil}^{m}\int_{-B}^{B}\frac{1}{\sqrt{2\pi n^2\overline{\sigma}^2}}\mathrm{e}^{-\xi^2/(2\pi\underline{\sigma}^2)}\mathrm{d}\xi\right]\int_{\max(\underline{\sigma},\sigma_0-C_1''\epsilon)}^{\min(\overline{\sigma},\sigma_0+C_1''\epsilon)}\pi_{[\underline{\sigma},\overline{\sigma}]}(\sigma)\mathrm{d}\sigma \\ &\geq \exp\left[-\tilde{C}_3(K_n m)^p(\log n) - \tilde{C}_2(K_n\lceil\alpha-1\rceil + K_n)^p\left(\log\frac{1}{\epsilon}\right)\right] \\ &\geq \exp\left[-C_2\epsilon^{-p/\alpha}\left(\log n + \log\frac{1}{\epsilon}\right)\right]\end{aligned}$$

for some constants $C_1'', \tilde{C}_2, C_2, \tilde{C}_3 > 0$. The proof is thus completed. □

Proposition E.2 (Generalized metric entropy bound)**.** *There exists some constant $c_2 > 0$, such that for sufficiently small $\epsilon > 0$,*

$$\log\mathcal{N}(\epsilon, \mathcal{F}^{B_n}, \|\cdot\|_\infty) \leq c_2 K_n^p\left(\log\frac{B_n}{\epsilon}\right).$$

Proof. Let

$$f_j(\mathbf{x}) = \sum_{\mathbf{k}} \sum_{\mathbf{s}:|\mathbf{s}|=0}^{m} \xi_{\mathbf{ks}}^{(j)} w_{\mathbf{k}}(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star})^{\mathbf{s}}$$

be functions in $\mathcal{F}_K$, $j = 1, 2$.

Denote $\theta_{j\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{s}:|\mathbf{s}|=0}^{m} \xi_{\mathbf{ks}}^{(j)} w_{\mathbf{k}}(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star})^{\mathbf{s}}$, $j = 1, 2$. We proceed to compute

$$\begin{aligned} \|f_1 - f_2\|_\infty &\leq \max_{\mathbf{x}\in\mathcal{X}} \left| \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x})(\theta_{1\mathbf{k}}(\mathbf{x}) - \theta_{2\mathbf{k}}(\mathbf{x})) \right| \\ &\leq \max_{\mathbf{k}\in[K]^p} \max_{0\leq|\mathbf{s}|\leq m} \left| \xi_{\mathbf{ks}}^{(1)} - \xi_{\mathbf{ks}}^{(2)} \right| \max_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{k}} w_{\mathbf{k}}(\mathbf{x}) \max_{\mathbf{l}} \sum_{\mathbf{s}:|\mathbf{s}|=0}^{m} \|(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{l}}^{\star})^{\mathbf{s}}\|_\infty \quad (7) \end{aligned}$$

Since $\max_{|\mathbf{s}|=0,1,\dots,m} |\xi_{\mathbf{ks}}^{(2)}| \leq B_n$ for all $\mathbf{k}$ and $\|(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{k}}^{\star})^{\mathbf{s}}\|_\infty$ is upper bounded by a universal constant for all $\mathbf{k} \in [K]^p, |\mathbf{s}| = 0, 1, \dots, m$, it follows that

$$\|f_1 - f_2\|_\infty \lesssim \max_{\mathbf{k}\in[K_n]^p} \max_{0\leq|\mathbf{s}|\leq m} \left| \xi_{\mathbf{ks}}^{(1)} - \xi_{\mathbf{ks}}^{(2)} \right|. \quad (8)$$

For any $\mathbf{x} \in \mathcal{X}$, there exists a unique $\mathbf{k}_{\mathbf{x}} \in [K_n]^p$ such that $\mathbf{x} \in \mathcal{X}_K(\mathbf{k}_{\mathbf{x}})$. Observe that $\sum_{\mathbf{k}\in[K_n]^p} \mathbf{1}(w_{j\mathbf{k}}(\mathbf{x}) > 0)$ is the same as the cardinality of the index set $\{\mathbf{l} \in [K_n]^p : \|\mathbf{x} - \boldsymbol{\mu}_{j\mathbf{l}}^{\star}\|_\infty < h\}$. Similarly to the proof of Lemma 1, the cardinality of this index set is upper bounded by $\lceil 8\overline{h} + 2 \rceil^p$. Therefore to construct an ϵ-net for $\mathcal{F}^{B_n}$, it suffices to construct ϵ-net for $\xi_{\mathbf{ks}}$ in $[-B_n, B_n]$. Simple algebra shows that

$$\mathcal{N}\left(C_1\epsilon, \mathcal{F}^{B_n}, \|\cdot\|_\infty\right) \lesssim [K_n(m+1)]^p \left(\log \frac{B_n}{\epsilon}\right).$$

when ϵ is sufficiently small, and the proof is thus completed. $\square$

We will need the following modified Ghosal-and-van-der-Vaart theorem for deriving rate of contraction with respect to the empirical L_2-distance $\|\cdot\|_{L_2(\mathbb{P}_n)}$ for fixed-design regression, which may also of independent interest.

Theorem E.1. *Let $y_i = f(\mathbf{x}_i) + e_i$, where $e_i \sim \mathrm{N}(0, \sigma^2)$, and the true distribution of $(\mathbf{x}_i, y_i)_{i=1}^n$ is given by $y_i = f_0(\mathbf{x}_i) + e_i$ with $e_i \sim \mathrm{N}(0, \sigma_0^2)$. Suppose f is imposed a sequence of priors $\Pi_f^n(\cdot \mid \sigma)$ supported on $\mathcal{G}_n$ given σ, and σ is imposed a marginal prior Π_σ with density π_σ supported on $[\underline{\sigma}, \overline{\sigma}]$ such that $\sigma_0 \in [\underline{\sigma}, \overline{\sigma}]$. Denote $\Pi^n(\cdot,\cdot) = \Pi_f^n(\cdot \mid \sigma) \times \Pi_\sigma(\cdot)$. If there exists a sequence of sieves $(\mathcal{F}_n)_{n=1}^\infty$ in $\mathcal{G}_n$ such that*

$$-\log \Pi^n\left(\|f - f_0\|_{L_2(\mathbb{P}_n)} + |\sigma - \sigma_0| < \underline{\epsilon}_n\right) \lesssim n\underline{\epsilon}_n^2, \quad (9)$$

$$\log \mathcal{N}\left(\epsilon_n/2, \mathcal{F}_n, \|\cdot\|_{L_2(\mathbb{P}_n)}\right) \lesssim n\epsilon_n^2, \quad (10)$$

$$-\log \Pi^n(f \in \mathcal{F}_n^c)/(n\underline{\epsilon}_n^2) \to \infty, \quad (11)$$

for two sequences $(\epsilon_n)_{n=1}^\infty$, $(\underline{\epsilon}_n)_{n=1}^\infty$ with $n\epsilon_n^2 \geq n\underline{\epsilon}_n^2 \to \infty$ when n is sufficiently large, then there exists a large constant M such that

$$\Pi^n\left(\|f - f_0\|_{L_2(\mathbb{P}_n)} + |\sigma - \sigma_0| > M\epsilon_n \mid \mathcal{D}_n\right) \to 0$$

in $\mathbb{P}_0$-probability.

Proof. Denote $\mathcal{U}_n = \{\|f-f_0\|_{L_2(\mathbb{P}_n)}+|\sigma-\sigma_0| > M\epsilon_n\}$ and write $\mathcal{U}_n = \bigcup_{j=M}^{\infty} \mathcal{V}_{jn}$, where $\mathcal{V}_{jn} = \{j\epsilon_n < \|f - f_0\|_{L_2(\mathbb{P}_n)} + |\sigma - \sigma_0| \le (j+1)\epsilon_n\}$. Denote $\mathcal{N}_{jn} = \mathcal{N}(j\epsilon_n/2, \mathcal{V}_{jn}, d_n)$, where $d_n((f_1,\sigma_1),(f_2,\sigma_2)) = \|f_1 - f_2\|_{L_2(\mathbb{P}_n)} + |\sigma_1 - \sigma_2|$. It follows that

$$\begin{aligned}
\mathcal{N}_{jn} &\le \mathcal{N}\left(j\epsilon_n/4, \{\|f - f_0\|_{L_2(\mathbb{P}_n)} \le (j+1)\epsilon_n\}, \|\cdot\|_{L_2(\mathbb{P}_n)}\right) \times \mathcal{N}(j\epsilon_n/4, [\underline{\sigma}, \overline{\sigma}], |\cdot|) \\
&\le \frac{4(\overline{\sigma} - \underline{\sigma})}{\epsilon_n} \mathcal{N}\left(\epsilon_n/2, \mathcal{F}_n, \|\cdot\|_{L_2(\mathbb{P}_n)}\right) \\
&\lesssim \exp(\bar{C} n\epsilon_n^2)
\end{aligned}$$

for some constant $\bar{C} > 0$. Hence there exists pairs $(f_{njl}, \sigma_{njl})_{l=1}^{\mathcal{N}_{jn}} \subset \mathcal{V}_{jn}$, such that

$$\begin{aligned}
\mathcal{V}_{jn} &= \bigcup_{l=1}^{\mathcal{N}_{jn}} \{\|f - f_{njl}\|_{L_2(\mathbb{P}_n)} + |\sigma - \sigma_{njl}| < j\epsilon_n/4\} \\
&\subset \bigcup_{l=1}^{\mathcal{N}_{jn}} \{\|f - f_{njl}\|_{L_2(\mathbb{P}_n)} < j\epsilon_n/4, |\sigma - \sigma_{njl}| < j\epsilon_n/4\} \\
&\subset \bigcup_{l=1}^{\mathcal{N}_{jn}} \left\{\|f - f_{njl}\|_{L_2(\mathbb{P}_n)} < \frac{1}{2}\|f - f_0\|_{L_2(\mathbb{P}_n)}, |\sigma - \sigma_{njl}| < \frac{1}{2}|\sigma - \sigma_0|\right\} \\
&:= \bigcup_{l=1}^{\mathcal{N}_{jn}} \mathcal{W}_{njl}.
\end{aligned}$$

For each n, j, l, we can take ψ_{njl} to be the test function given by Lemma F.3, and define $\psi_n = \sup_{j\ge M} \max_{l=1,\cdots,\mathcal{N}_{jn}} \psi_{njl}$. Since for all $(f,\sigma) \in \mathcal{W}_{njl}$,

$$\begin{aligned}
\max\{\mathbb{E}_0\psi_{njl}, \mathbb{E}_{f,\sigma}(1-\psi_{njl})\} &\le \exp\left[-\frac{C}{\overline{\sigma}^2}\left(n\|f_0 - f_{njl}\|^2_{L_2(\mathbb{P}_n)} + n|\sigma_0 - \sigma_{njl}|^2\right)\right] \\
&\le \exp\left(-\frac{C}{2\overline{\sigma}^2} nj^2\epsilon_n^2\right),
\end{aligned}$$

then by the union bound we have

$$\mathbb{E}_0\psi_n = \sum_{j=M}^{\infty}\sum_{\ell=1}^{\mathcal{N}_{jl}} \mathbb{E}_0\psi_{njl} \lesssim \epsilon_n^{-1}\exp(\bar{C}n\epsilon_n^2)\sum_{j=M}^{\infty}\exp\left(-\frac{C}{2\overline{\sigma}^2}nj^2\epsilon_n^2\right) \to 0,$$

for sufficiently large M, and

$$\sup_{(f,\sigma)\in\mathcal{U}_n} \mathbb{E}_{f,\sigma}(1-\psi_n) \le \sup_{j\ge M}\sup_{\ell=1,\cdots,\mathcal{N}_{jl}}\sup_{(f,\sigma)\in\mathcal{W}_{njl}} \mathbb{E}_{f,\sigma}(1-\psi_{njl}) \le \exp\left(-\frac{CM^2}{2\overline{\sigma}^2}n\epsilon_n^2\right) \to 0.$$

Consider the event

$$\mathcal{H}_n = \left\{\int\prod_{i=1}^n \frac{p_{f,\sigma}(y_i,\mathbf{x}_i)}{p_0(y_i,\mathbf{x}_i)}\Pi(\mathrm{d}f\mathrm{d}\sigma) \le \Pi^n(\mathcal{B}_n)\exp\left[-(1+c)Cn\underline{\epsilon}_n^2\right]\right\}$$

for some constant $C > 0$, where $\mathcal{B}_n = \{\|f - f_0\|_{L_2(\mathbb{P}_n)} + |\sigma - \sigma_0| < \underline{\epsilon}_n\}$ and c is a constant to be specified later. Then $\mathbb{P}_0(\mathcal{H}_n) \to 0$ according to Lemma F.2. Now denote

$$\ell_n(f, \sigma) = \sum_{i=1}^{n} \log p_{f,\sigma}(y_i, \mathbf{x}_i)$$

to be the log-likelihood function of (f, σ). Then

$$\begin{aligned}
&\mathbb{E}_0 \left[\Pi(\mathcal{U}_n \mid \mathcal{D}_n)\right] \\
&\quad\le \mathbb{E}_0\psi_n + \mathbb{E}_0[(1-\psi_n)\mathbb{1}(\mathcal{H}_n)] + \mathbb{E}_0\left\{(1-\phi_n)\mathbb{1}(\mathcal{H}_n^c)\Pi(\mathcal{U}_n \mid \mathcal{D}_n)\right\} \\
&\quad= o(1) + \mathbb{E}_0\left[(1-\phi_n)\mathbb{1}(\mathcal{H}_n^c)\frac{\int_{\mathcal{U}_n}\exp(\ell_n(f,\sigma)-\ell_n(f_0,\sigma_0))\Pi^n(\mathrm{d}f\mathrm{d}\sigma)}{\int\exp(\ell_n(f,\sigma)-\ell_n(f_0,\sigma_0))\Pi^n(\mathrm{d}f\mathrm{d}\sigma)}\right] \\
&\quad= o(1) + \mathbb{E}_0\left[(1-\phi_n)\frac{\exp[(1+c)Cn\underline{\epsilon}_n^2]}{\Pi^n(\mathcal{B}_n)}\int_{\mathcal{U}_n}\prod_{i=1}^{n}\frac{p_{f,\sigma}(y_i,\mathbf{x}_i)}{p_0(y_i,\mathbf{x}_i)}\Pi^n(\mathrm{d}f\mathrm{d}\sigma)\right] \\
&\quad\le o(1) + \frac{\exp[(1+c)Cn\underline{\epsilon}_n^2]}{\Pi^n(\mathcal{B}_n)}\int_{(\mathcal{F}_n^c)\times[\underline{\sigma},\overline{\sigma}]}\mathbb{E}_0\left[\prod_{i=1}^{n}\frac{p_{f,\sigma}(y_i,\mathbf{x}_i)}{p_0(y_i,\mathbf{x}_i)}\right]\Pi^n(\mathrm{d}f\mathrm{d}\sigma) \\
&\qquad+ \frac{\exp[(1+c)Cn\underline{\epsilon}_n^2]}{\Pi^n(\mathcal{B}_n)}\int_{\mathcal{U}_n\cap(\mathcal{F}_n\times[\underline{\sigma},\overline{\sigma}])}\mathbb{E}_0\left[(1-\psi_n)\prod_{i=1}^{n}\frac{p_{f,\sigma}(y_i,\mathbf{x}_i)}{p_0(y_i,\mathbf{x}_i)}\right]\Pi^n(\mathrm{d}f\mathrm{d}\sigma) \\
&\quad\le o(1) + \frac{\exp[(1+c)Cn\underline{\epsilon}_n^2]}{\Pi^n(\mathcal{B}_n)}\left[\sup_{(f,\sigma)\in\mathcal{U}_n}\mathbb{E}_{f,\sigma}(1-\psi_n)\right] \\
&\qquad+ \frac{\exp[(1+c)Cn\underline{\epsilon}_n^2]}{\Pi^n(\mathcal{B}_n)}\Pi^n(f\in\mathcal{F}_n^c) \\
&\quad\le o(1) + \exp\left[(1+c)Cn\underline{\epsilon}_n^2 + C_1 n\underline{\epsilon}_n^2 - \frac{CM^2}{2\overline{\sigma}^2}n\epsilon_n^2\right] \\
&\qquad+ \exp[(1+c)Cn\underline{\epsilon}_n^2 + \tilde{C}n\underline{\epsilon}_n^2 - M_n n\underline{\epsilon}_n^2] \to 0,
\end{aligned}$$

for some sequence $M_n \to \infty$ and some constant $\tilde{C} > 0$. □

The strategy to prove Theorem 5 is to first prove the rate of contraction under the conditional prior $\Pi^n_{[\underline{\sigma},\overline{\sigma}]}$, and then argue that $\Pi^n(\sigma \in [\underline{\sigma},\overline{\sigma}] \mid \mathcal{D}_n) \to 1$.

Theorem E.2. *Assume that f_0 is in the α-Hölder function class $\mathfrak{C}^{\alpha,B}(\mathcal{X})$ with envelope B and the design space $\mathcal{X}$ is the unit hypercube $[0,1]^p$. Under the prior $\Pi^n_{[\underline{\sigma},\overline{\sigma}]}(\cdot,\cdot) = \Pi^n_f(\cdot \mid \sigma) \times \Pi_{[\underline{\sigma},\overline{\sigma}]}(\cdot)$, where $[\underline{\sigma},\overline{\sigma}]$ contains σ_0, it holds that*

$$\Pi^n_{[\underline{\sigma},\overline{\sigma}]}(\|f - f_0\|^2_{L_2(\mathbb{P}_n)} > M\epsilon_n^2 \mid \mathcal{D}_n) \to 0$$

in $\mathbb{P}_0$-probability for some constant $M > 0$, where $\epsilon_n = n^{-\alpha/(2\alpha+p)}(\log n)^{\alpha/(2\alpha+p)}$.

Proof. Take $B_n = n^2$, $\mathcal{F}_n = \mathcal{F}^{B_n}$, $\underline{\epsilon}_n = \epsilon_n = n^{-\alpha/(2\alpha+p)}(\log n)^{\alpha/(2\alpha+p)}$, and we proceed by verifying the three conditions in Theorem E.1. For condition (9) by Proposition

E.1 we have

$$-\log \Pi^n_{[\underline{\sigma},\overline{\sigma}]}\left(\|f-f_0\|_{L_2(\mathbb{P}_n)}+|\sigma-\sigma_0|<\underline{\epsilon}_n\right)\lesssim \underline{\epsilon}_n^{-p/\alpha}\left(\log n+\log\frac{1}{\underline{\epsilon}_n}\right)=n\underline{\epsilon}_n^2.$$

For condition (10) we derive using Lemma E.2 to obtain

$$\log\mathcal{N}(\epsilon_n,\mathcal{F}^{B_n},\|\cdot\|_{L_2(\mathbb{P}_n)})\le\log\mathcal{N}(\epsilon_n,\mathcal{F}^{B_n},\|\cdot\|_\infty)\lesssim K_n^p\left(\log\frac{B_n}{\epsilon_n}\right)=n\epsilon_n^2.$$

Finally for condition (11) we use the union bound and tail probability of Gaussian to derive

$$\Pi^n(f\in\mathcal{F}_n^c)\le\int_{\underline{\sigma}}^{\overline{\sigma}}\sum_{\mathbf{k}\in[K_n]^p}\sum_{\mathbf{s}:|\mathbf{s}|=0}^{n}\Pi_f^n(|\xi_{\mathbf{ks}}|>B_n\mid\sigma)\pi_{[\underline{\sigma},\overline{\sigma}]}(\sigma)\mathrm{d}\sigma\lesssim K_n^p\exp\left(-\frac{n^2}{4\underline{\sigma}^2}\right),$$

and hence for sufficiently large n

$$\frac{-\log\Pi^n(f\in\mathcal{F}_n^c)}{n\underline{\epsilon}_n^2}=\frac{1}{n\underline{\epsilon}_n^2}\left(\frac{n^2}{4\underline{\sigma}^2}-\log K_n^p\right)=\infty.$$

The proof is thus completed. □

Now we are in a position to prove Theorem 5.

Proof of Theorem 5. Let

$$\mathcal{G}_n=\left\{f(\mathbf{x})=\sum_{\mathbf{k}\in[K_n]^p}\sum_{\mathbf{s}:|\mathbf{s}|=0}^{m}\xi_{\mathbf{ks}}\psi_{\mathbf{ks}}(\mathbf{x}):\xi_{\mathbf{ks}}\in\mathbb{R},\mathbf{k}\in[K_n]^p,|\mathbf{s}|=0,\cdots,m\right\},$$

and $f_n^\star=\arg\inf_{f\in\mathcal{G}_n}\|f-f_0\|_{L_2(\mathbb{P}_n)}$. Let $L=K_n^p\sum_{s=0}^m\binom{p+s-1}{s}$, $\mathbf{b}(\mathbf{x})=(\psi_{\mathbf{ks}}(\mathbf{x}):\mathbf{k}\in[K_n]^p,|\mathbf{s}|=0,1,\cdots,m)^{\mathrm{T}}\in\mathbb{R}^L$, $\boldsymbol{\xi}=(\xi_{\mathbf{ks}}:\mathbf{k}\in[K_n]^p,|\mathbf{s}|=0,1,\cdots,m)^{\mathrm{T}}\in\mathbb{R}^L$, and $\mathbf{B}=[\mathbf{b}(\mathbf{x}_1),\cdots,\mathbf{b}(\mathbf{x}_n)]^{\mathrm{T}}\in\mathbb{R}^{n\times L}$. Now denote

$$\widehat{\sigma}^2=\frac{1}{n}\mathbf{y}^{\mathrm{T}}\left(n^2\mathbf{B}\mathbf{B}^{\mathrm{T}}+\mathbf{I}_n\right)^{-1}\mathbf{y}.$$

By Lemma E.1 there exists $\boldsymbol{\xi}^\star\in\mathbb{R}^L$ such that $\|f_0-\mathbf{b}(\cdot)^{\mathrm{T}}\boldsymbol{\xi}^\star\|_{L_2(\mathbb{P}_n)}=O(\epsilon_n)$, and $\|\boldsymbol{\xi}^\star\|_\infty\le B$. Therefore we proceed to compute

$$\begin{aligned}|\mathbb{E}_0\widehat{\sigma}^2-\sigma_0^2|&\le\frac{1}{n}\mathbf{F}_0^{\mathrm{T}}(n^2\mathbf{B}\mathbf{B}^{\mathrm{T}}+\mathbf{I}_n)^{-1}\mathbf{F}_0+\sigma_0^2\left|\mathrm{tr}\left(\frac{1}{n}\left(\mathbf{I}_n+n^2\mathbf{B}\mathbf{B}^{\mathrm{T}}\right)^{-1}\right)-1\right|\\&=\frac{1}{n}\mathbf{F}_0^{\mathrm{T}}(n^2\mathbf{B}\mathbf{B}^{\mathrm{T}}+\mathbf{I}_n)^{-1}\mathbf{F}_0+\sigma_0^2\frac{1}{n}\left|\mathrm{tr}\left(\mathbf{I}_n-\left(\mathbf{I}_n+n^2\mathbf{B}\mathbf{B}^{\mathrm{T}}\right)^{-1}\right)\right|\\&=\frac{1}{n}\mathbf{F}_0^{\mathrm{T}}(n^2\mathbf{B}\mathbf{B}^{\mathrm{T}}+\mathbf{I}_n)^{-1}\mathbf{F}_0+\sigma_0^2\frac{1}{n}\left|\mathrm{tr}\left(\mathbf{B}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}+\frac{1}{n^2}\mathbf{I}_L\right)^{-1}\mathbf{B}^{\mathrm{T}}\right)\right|\end{aligned}$$

$$\lesssim \frac{1}{n}\mathbf{F}_0^{\mathrm{T}}(n^2\mathbf{B}\mathbf{B}^{\mathrm{T}} + \mathbf{I}_n)^{-1}\mathbf{F}_0 + \sigma_0^2\frac{L}{n},$$

where $\mathbf{F}_0 = [f_0(\mathbf{x}_1), \cdots, f_0(\mathbf{x}_n)]^{\mathrm{T}}$. In addition,

$$\begin{aligned}
\frac{1}{n}\mathbf{F}_0^{\mathrm{T}}(n^2\mathbf{B}\mathbf{B}^{\mathrm{T}} + \mathbf{I}_n)^{-1}\mathbf{F}_0 &\lesssim \frac{1}{n}(\mathbf{F}_0 - \mathbf{B}\boldsymbol{\xi}^\star)^{\mathrm{T}}(n^2\mathbf{B}\mathbf{B}^{\mathrm{T}} + \mathbf{I}_n)^{-1}(\mathbf{F}_0 - \mathbf{B}\boldsymbol{\xi}^\star)\\
&\quad + \frac{1}{n}(\boldsymbol{\xi}^\star)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}(n^2\mathbf{B}\mathbf{B}^{\mathrm{T}} + \mathbf{I}_n)^{-1}\mathbf{B}\boldsymbol{\xi}^\star\\
&\lesssim \|f_0 - \mathbf{b}(\cdot)^{\mathrm{T}}\boldsymbol{\xi}^\star\|_{L_2(\mathbb{P}_n)}^2 + \frac{1}{n^2}(\boldsymbol{\xi}^\star)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}} + \frac{1}{n^2}\mathbf{I}_n\right)^{-1}\mathbf{B}\boldsymbol{\xi}^\star\\
&\leq \|f_0 - \mathbf{b}(\cdot)^{\mathrm{T}}\boldsymbol{\xi}^\star\|_{L_2(\mathbb{P}_n)}^2 + \frac{L}{n^2}\|\boldsymbol{\xi}^\star\|_\infty^2 \operatorname{tr}\left(\mathbf{B}^{\mathrm{T}}(\mathbf{B}\mathbf{B}^{\mathrm{T}})^{-1}\mathbf{B}\right)\\
&\lesssim O(\epsilon_n^2) + O(L/n).
\end{aligned}$$

Therefore we conclude that $\widehat{\sigma}^2 = \sigma_0^2 + o_{\mathbb{P}_0}(1)$.

Since the marginal posterior of σ^2 is again inverse-Gamma under Π^n

$$\Pi(\sigma^2 \in \cdot \mid \mathcal{D}_n) = \mathrm{IG}\left(\frac{a_\sigma + n}{2}, \frac{b_\sigma}{2} + \frac{n\widehat{\sigma}^2}{2}\right),$$

then the posterior mean and variance are

$$\begin{aligned}
\mathbb{E}_\Pi(\sigma^2 \mid \mathcal{D}_n) &= \frac{b_\sigma}{(a_\sigma + n - 2)} + \frac{n\widehat{\sigma}^2}{(a_\sigma + n - 2)} = \sigma_0^2 + o_{\mathbb{P}_0}(1),\\
\mathrm{var}_\Pi(\sigma^2 \mid \mathcal{D}_n) &= \frac{4}{(a_\sigma + n - 4)}\left[\frac{b_\sigma}{(a_\sigma + n - 2)} + \frac{n\widehat{\sigma}^2}{(a_\sigma + n - 2)}\right]^2 = o_{\mathbb{P}_0}(1).
\end{aligned}$$

Thus we proceed to compute by Markov's inequality

$$\begin{aligned}
\Pi\left(|\sigma^2 - \sigma_0^2| > \epsilon \mid \mathcal{D}_n\right) &\leq \frac{1}{\epsilon^2}\mathbb{E}_\Pi\left[(\sigma^2 - \sigma_0^2)^2 \mid \mathcal{D}_n\right]\\
&\leq \frac{1}{\epsilon^2}\left[\mathbb{E}_\Pi(\sigma^2 \mid \mathcal{D}_n) - \sigma_0^2\right]^2 + \frac{1}{\epsilon^2}\mathrm{var}_\Pi(\sigma \mid \mathcal{D}_n)\\
&= o_{\mathbb{P}_0}(1).
\end{aligned}$$

Now take ϵ be sufficiently small and $\underline{\sigma} = \sqrt{\sigma_0^2 - \epsilon^2}, \overline{\sigma} = \sqrt{\sigma_0^2 + \epsilon^2}$. Then we apply Theorem E.2 to derive

$$\begin{aligned}
&\Pi^n\left(\|f - f_0\|_{L_2(\mathbb{P}_n)} > M\epsilon_n \mid \mathcal{D}_n\right)\\
&\quad = \Pi^n\left(\|f - f_0\|_{L_2(\mathbb{P}_n)} > M\epsilon_n \mid \sigma \in [\underline{\sigma}, \overline{\sigma}], \mathcal{D}_n\right)\Pi(\sigma \in [\underline{\sigma}, \overline{\sigma}] \mid \mathcal{D}_n)\\
&\qquad + \Pi^n\left(\|f - f_0\|_{L_2(\mathbb{P}_n)} > M\epsilon_n \mid \sigma \notin [\underline{\sigma}, \overline{\sigma}], \mathcal{D}_n\right)\Pi(\sigma \notin [\underline{\sigma}, \overline{\sigma}] \mid \mathcal{D}_n)\\
&\quad \leq \Pi^n_{[\underline{\sigma}, \overline{\sigma}]}\left(\|f - f_0\|_{L_2(\mathbb{P}_n)} > M\epsilon_n \mid \mathcal{D}_n\right) + \Pi(|\sigma^2 - \sigma_0^2| > \epsilon^2 \mid \mathcal{D}_n) = o_{\mathbb{P}_0}(1).
\end{aligned}$$

The proof is thus completed. $\square$

F Cited theorems and results

The following theorem concerning the rate of contraction with respect to the Hellinger distance is extensively used throughout.

Theorem F.1 (Kruijer et al. (2010), Theorem 3)**.** *Let $\mathcal{M}$ be a statistical model, i.e. a class of density functions with respect to some underlying σ-finite measure over $\mathcal{Y}$. Equip $\mathcal{M}$ with the Hellinger distance H and the Borel σ-field generated by H. Suppose $\mathcal{M}$ is imposed with a prior distribution Π. Let $(y_i)_{i=1}^n$ be i.i.d. according to some density function p_0. If there exist sequences $(\epsilon_n)_{n=1}^\infty, (\underline{\epsilon}_n)_{n=1}^\infty$ with $\epsilon_n \geq \underline{\epsilon}_n$, $\underline{\epsilon}_n \to 0$ and $n\underline{\epsilon}_n^2 \to \infty$, a sequence of (measurable) sub-models $(\mathcal{M}_n)_{n=1}^\infty$, each of which is contained in $\mathcal{M}$, and for each $\mathcal{M}_n$ a partition $(\mathcal{M}_{nm})_{m=1}^\infty$ (with $\mathcal{M}_n = \bigcup_{m=1}^\infty \mathcal{M}_{nm}$), such that*

$$\Pi\left(B_{\mathrm{KL}}(p_0, \underline{\epsilon}_n)\right) \geq \exp(-n\underline{\epsilon}_n^2), \tag{12}$$

$$\Pi\left(\mathcal{M}_n^c\right) \leq \exp(-4n\underline{\epsilon}_n^2), \tag{13}$$

$$\exp(-n\epsilon_n^2) \sum_{m=1}^\infty \sqrt{\mathcal{N}(\epsilon_n, \mathcal{M}_{nm}, H)}\sqrt{\Pi(\mathcal{M}_{nm})} \to 0, \tag{14}$$

then $\Pi\left(H(p, p_0) > M\epsilon_n \mid y_1, \ldots, y_n\right) = o_{\mathbb{P}_0}(1)$ for some constant $M > 0$.

The following result is immediate by Lemma 1 in Ghosal et al. (2007).

Lemma F.1. *Assume that the sequence of submodels $(\mathcal{M}_n)_{n=1}^\infty$ in $\mathcal{M}$ and the sequence $(\underline{\epsilon}_n)_{n=1}^\infty$ satisfies (12) and (13) in Theorem F.1 with $\underline{\epsilon}_n \to 0$, $n\underline{\epsilon}_n^2 \to \infty$. Then $\mathbb{E}_0\left[\Pi\left(\mathcal{M}_n | y_1, \ldots, y_n\right)\right] \to 1$.*

The following classical theorem originally due to Wong and Shen (1995) is used to study the convergence rate of sieve maximum likelihood estimator.

Theorem F.2. *Let $(\mathbf{y}_i)_{i=1}^n$ be independent and identically distributed observations following a distribution $\mathbb{P}_0$ with density p_0, and $(\mathcal{P}_n)_{n=1}^\infty$ be a sequence of classes of densities (referred to as the sieves). Suppose $(\epsilon_n)_{n=1}^\infty$ is a sequence decreasing to 0 such that*

$$\int_0^{\epsilon_n} \sqrt{\log_{[\cdot]} \mathcal{N}(\epsilon, \mathcal{P}_n, H)} \mathrm{d}\epsilon \lesssim \sqrt{n}\epsilon_n^2.$$

Let $\widehat{p}_n = \arg\max_{p \in \mathcal{P}_n} \sum_{i=1}^n \log p(\mathbf{y}_i)$ be the sieve maximum likelihood estimator on $\mathcal{P}_n$ and be well-defined, and define

$$\delta_n = \inf_{q \in \mathcal{P}_n} D_{\mathrm{KL}}(p_0 \| q), \quad \tau_n = \lim_{k \to \infty} \mathbb{E}_0 \left[\log \frac{p_0(\mathbf{y})}{p_k(\mathbf{y})}\right]^2$$

for some sequence $(q_k)_{k=1}^\infty \subset \mathcal{P}_n$ such that $D_{\mathrm{KL}}(p_0 \| q_k) \to \delta_n$. If $\max(\delta_n, \tau_n) \lesssim \epsilon_n^2$, then there exists some constant $M > 0$ such that

$$\mathbb{P}_0\left(H\left(p_0, \widehat{p}_n\right) > M\epsilon_n\right) \lesssim \exp(-n\epsilon_n^2) + \frac{1}{n}.$$

The following semiparametric Bernstein von-Mises theorem presents a set of sufficient conditions for the asymptotic normality of the marginal posterior of the parametric component to occur in a semiparametric Bayesian model.

Theorem F.3 (Yang et al. (2015)). *Let $\mathcal{P} = \{p_{\boldsymbol{\theta},\eta} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q, \eta \in \mathcal{F}\}$ be a class of density functions with respect to some underlying σ-finite measure over $\mathcal{Y}$ parametrized on $\Theta \times \mathcal{F}$, Θ is open, and $\mathcal{F}$ is equipped with metric $d_H(\eta_1, \eta_2) = H(p_{\boldsymbol{\theta}_0,\eta_1}, p_{\boldsymbol{\theta}_0,\eta_2})$. Let $(\mathbf{y}_i)_{i=1}^n$ be i.i.d. according to $p_0 = p_{\boldsymbol{\theta}_0,\eta_0}$ for some $\boldsymbol{\theta}_0 \in \Theta$ and η_0. Assume that the least-favorable submodel $\left\{p_{\boldsymbol{\theta},\eta^*_{\boldsymbol{\theta}}} : \boldsymbol{\theta} \in \Theta\right\}$ defined through the least-favorable curve $\eta^*_{\boldsymbol{\theta}} = \arg\inf_{\eta\in\mathcal{F}} D_{\mathrm{KL}}(p_0 || p_{\boldsymbol{\theta},\eta})$ exists for all $\boldsymbol{\theta} \in \Theta$, and denote the semiparametric bias $\Delta\eta_{\boldsymbol{\theta}} = \eta^*_{\boldsymbol{\theta}} - \eta_0$. Suppose the following conditions hold:*

Condition F.1. *$\Theta \times \mathcal{F}$ is endowed with a product prior $\Pi_{\boldsymbol{\theta}} \times \Pi_\eta$, and $\Pi_{\boldsymbol{\theta}}$ yields a density with respect to the Lebesgue measure on Θ that is positive at $\boldsymbol{\theta}_0$*

Condition F.2. *There exists a sequence $\epsilon_n \to 0$ satisfying $n\epsilon_n^2 \to \infty$ and a sequence of submodels $(\widehat{\mathcal{F}}_n)_{n=1}^\infty$ in $\mathcal{F}$, such that as $n \to \infty$, $\Pi(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \epsilon_n, \eta \in \widehat{\mathcal{F}}_n \mid \mathbf{y}_1, \ldots, \mathbf{y}_n) = 1 - o_{\mathbb{P}_0}(1)$.*

Condition F.3. *There exists an increasing function $G_n : \mathbb{R} \to [0, \infty)$ such that for every sequence $(\boldsymbol{\theta}_n)_{n=1}^\infty$ with $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + o_{\mathbb{P}_0}(1)$,*

$$\begin{aligned}&\sup_{\eta\in\widehat{\mathcal{F}}_n} \left| n\mathbb{P}_n \log\frac{p_{\boldsymbol{\theta}_n,\eta+\Delta\eta_{\boldsymbol{\theta}_n}}}{p_{\boldsymbol{\theta}_0,\eta}} - n(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbb{P}_n\boldsymbol{\ell}_0 + \frac{1}{2}n(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} I_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right| \\ &\quad = O_{\mathbb{P}_0}\left(G_n\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\right)\right),\end{aligned}$$

where $\widehat{\mathcal{F}}_n$ is defined in I, $\boldsymbol{\ell}_0$ and $\mathbf{I}_0$ are the score function and the Fisher information matrix of the least-favorable submodel $\{p_{\boldsymbol{\theta},\eta^*_{\boldsymbol{\theta}}} : \boldsymbol{\theta} \in \Theta\}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Condition F.4. *There exists an increasing function $\widetilde{G}_n : \mathbb{R} \to [0, \infty)$, such that for every sequence $(\boldsymbol{\theta}_n)_{n=1}^\infty$ with $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + o_{\mathbb{P}_0}(1)$,*

$$\frac{\int_{\widehat{\mathcal{F}}_n} \prod_{i=1}^n p_{\boldsymbol{\theta}_0,\eta-\Delta\eta_{\boldsymbol{\theta}_n}}(\mathbf{y}_i)\Pi_\eta\,(\mathrm{d}\eta)}{\int_{\widehat{\mathcal{F}}_n} \prod_{i=1}^n p_{\boldsymbol{\theta}_0,\eta}(\mathbf{y}_i)\Pi_\eta\,(\mathrm{d}\eta)} = 1 + O_{\mathbb{P}_0}\left(\widetilde{G}_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|)\right).$$

If $(G_n + \widetilde{G}_n)\left(n^{-1/2}\log n\right) = o(1)$, then the sequence of marginal posteriors for $\boldsymbol{\theta}$ is asymptotically normal in total variation

$$\sup_F \left|\Pi(\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in F | y_1, \ldots, y_n) - \Phi\left(F \mid \boldsymbol{\Delta}_n, \mathbf{I}_0^{-1}\right)\right| = o_{\mathbb{P}_0}(1),$$

where the supremum is taken over all measurable sets in $\mathbb{R}^q$, $\Phi(\cdot|\Delta_n, I_0^{-1})$ is the $\mathrm{N}(\Delta_n, I_0^{-1})$ probability measure, and

$$\boldsymbol{\Delta}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{I}_0^{-1}\boldsymbol{\ell}_0(\mathbf{y}_i).$$

The following maximum inequality for empirical process plays a fundamental role in the verification of III in Theorem F.3.

Theorem F.4 (van der Vaart (2000), Lemma 19.36). *Let $(\mathbf{y}_i)_{i=1}^n$ be i.i.d. according to a distribution $\mathbb{P}_{\mathbf{y}}$ over $\mathcal{Y}$, and let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{Y} \to \mathbb{R}$. If $\int_{\mathcal{Y}} f^2(\mathbf{y})\mathbb{P}_{\mathbf{y}}(\mathrm{d}\mathbf{y}) < \delta^2$ and $\|f\|_\infty \leq M$ for all $f \in \mathcal{F}$, where δ and M does not depend on $\mathcal{F}$, then*

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_{\mathbf{y}}}\|\mathbb{G}_n\|_{\mathcal{F}} := \mathbb{E}_{\mathbb{P}_{\mathbf{y}}}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[f(\mathbf{y}_i) - E_{\mathbb{P}_{\mathbf{y}}}f(\mathbf{y})\right]\right|\right\} \\ \lesssim J_{[\cdot]}\left(\delta, \mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_{\mathbf{y}})}\right)\left[1 + \frac{M}{\delta^2\sqrt{n}}J_{[\cdot]}\left(\delta, \mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_{\mathbf{y}})}\right)\right],\end{aligned}$$

where

$$J_{[\cdot]}(\delta, \mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_{\mathbf{y}})}) = \int_0^\delta \sqrt{\log\mathcal{N}_{[\cdot]}(\epsilon, \mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_{\mathbf{y}})})}\mathrm{d}\epsilon$$

is the bracketing integral.

The following theorem guarantees the lower bound of the marginal likelihood in terms of the prior concentration.

Lemma F.2 (Ghosal and van der Vaart, 2017). *Suppose $(\underline{\epsilon}_n)_{n=1}^\infty$ is a sequence with $\underline{\epsilon}_n \to 0$ and $n\underline{\epsilon}_n \to \infty$. Denote $\mathcal{B}_n = \{\|f - f_0\|_{L_2(\mathbb{P}_n)} + |\sigma - \sigma_0| < \epsilon_n\}$. There there exists a constant C such that for all $c > 0$,*

$$\mathbb{P}_0\left(\int\prod_{i=1}^n\frac{p_{f,\sigma}(y_i, \mathbf{x}_i)}{p_0(y_i, \mathbf{x}_i)}\Pi(\mathrm{d}f\mathrm{d}\sigma) \leq \Pi^n(\mathcal{B}_n)\exp\left[-(1+c)Cn\epsilon_n^2\right]\right) \to 0.$$

provided that the prior on σ is compactly supported on $[\underline{\sigma}, \overline{\sigma}] \subset (0, \infty)$.

We also need the following local testing property of the empirical L_2-distance for Gaussian regression to ensure the existence of test in establishing Theorem E.1.

Lemma F.3 (Ghosal and van der Vaart (2017), Lemma 8.27). *For $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\sigma > 0$, let $\mathbb{P}_{\boldsymbol{\theta},\sigma} = \mathrm{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I}_n)$ and let $\|\boldsymbol{\theta}\|$ be the Euclidean norm. Then for any $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \mathbb{R}^n$ and $\sigma_0, \sigma_1 > 0$, there exists a test function ψ such that*

$$\max\left\{\mathbb{E}_{\boldsymbol{\theta}_0,\sigma_0}\psi, \mathbb{E}_{\boldsymbol{\theta},\sigma}(1-\psi)\right\} \leq \exp\left[-\frac{C}{\max(\sigma_0^2, \sigma_1^2)}\left(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_1\|^2 + n|\sigma_0 - \sigma_1|^2\right)\right]$$

for any $\boldsymbol{\theta}$, σ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_1\| \leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|/2$ and $|\sigma - \sigma_1| \leq |\sigma_0 - \sigma_1|/2$.

References

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press. 29

Ghosal, S., van der Vaart, A., et al. (2007). "Convergence rates of posterior distributions for non-i.i.d observations." *The Annals of Statistics*, 35(1): 192–223. 16, 27

Kruijer, W., Rousseau, J., van der Vaart, A., et al. (2010). "Adaptive Bayesian density estimation with location-scale mixtures." *Electronic Journal of Statistics*, 4: 1225–1257. 13, 27

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press. 18, 29

Wong, W. H. and Shen, X. (1995). "Probability inequalities for likelihood ratios and convergence rates of sieve MLEs." *The Annals of Statistics*, 23(2): 339–362. 10, 27

Yang, Y., Cheng, G., and Dunson, D. B. (2015). "Semiparametric Bernstein-von Mises Theorem: Second Order Studies." *arXiv preprint arXiv:1503.04493*. 15, 28