Adaptive Bayesian nonparametric regression using kernel mixture of polynomials with application to partial linear model

Fangzheng Xie∗ Yanxun Xu∗†

October 24, 2017

Abstract

We propose a kernel mixture of polynomials prior for Bayesian nonparametric regression. The regression function is modeled by local average of polynomials with kernel weights. We obtain the minimax-optimal rate of contraction up to a logarithmic factor that adapts to the smoothness level of the true function by estimating metric entropies of certain function classes. We also provide a frequentist sieve maximum likelihood estimator with a near-optimal convergence rate. We further investigate the application of the kernel mixture of polynomials to the partial linear model and obtain both the near-optimal rate of contraction for the nonparametric component and the Bernstein von-Mises limit (i.e., asymptotic normality) of the parametric component. These results are based on the development of convergence theory for the kernel mixture of polynomials. The proposed method is illustrated with numerical examples and shows superior performance in terms of accuracy and uncertainty quantification compared to the local polynomial regression, DiceKriging, and the robust Gaussian stochastic process.

Key Words: Bayesian nonparametric regression; Bernstein von-Mises limit; Metric entropies; Partial linear model; Rate of contraction

1 Introduction

The standard nonparametric regression model is of the form $y_i = f(x_i) + e_i$, where $y_i$’s are observations at given design points, $x_i$’s are in the design space $\mathcal{X} \subset \mathbb{R}^p$, $e_i$’s are independently $N(0, \sigma^2)$ distributed noises, $i = 1, \ldots, n$. The inference task is to use the noisy measurements $(x_i, y_i)_{i=1}^n$ to estimate the unknown function $f : \mathcal{X} \rightarrow \mathbb{R}$. Nonparametric regression methods have been widely used in a variety of applications, such as pattern recognition (Györfi et al., 2006; Devroye et al., 2013), image processing and reconstruction (Takeda et al., 2007), electronic healthcare records (Xu et al., 2016b), and semiparametric econometrics (Robinson, 1988; Klein and Spady, 1993).

∗Department of Applied Mathematics and Statistics, Johns Hopkins University
†Correspondence should be addressed to Yanxun Xu (yanxun.xu@jhu.edu)
Frequentist methods for nonparametric regression typically compute a fixed estimated function through the given data \((x_i, y_i)_{i=1}^n\). In contrast, Bayesian nonparametric techniques first impose a carefully-selected prior on the unknown function \(f\) and then find the posterior distribution of \(f\) given the observed data \((x_i, y_i)_{i=1}^n\), providing natural ways for uncertainty quantification (full posterior) instead of the point estimates given by the frequentist approaches. One of the most popular Bayesian nonparametric regression methods is the Gaussian process (Rasmussen and Williams, 2006) due to its tractability. Nevertheless, the computational burden of the Gaussian process in likelihood function evaluation resulted from the inversion of the covariance matrix prevents its scalability to big data.

In this paper, we propose a kernel mixture of polynomials prior for nonparametric regression, which does not involve the cumbersome \(O(n^3)\) inversion of the large covariance matrix. Leaving the Bayesian framework for a moment, let us consider the frequentist Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964) of the form \(\hat{f}(x) = \frac{\sum_i \varphi_h(x-x_i)y_i}{\sum_i \varphi_h(x-x_i)}\), where \(\varphi_h : \mathbb{R}^p \rightarrow [0, +\infty)\) is the kernel function parametrized by the bandwidth parameter \(h \in (0, +\infty)\) and is assumed to decrease when \(\|x\|\) increases, so that the Nadaraya-Watson estimator is a local averaging estimator (Györfi et al., 2006). As the Nadaraya-Watson estimator does not yield an optimal rate of convergence when the true regression function satisfies the \(\alpha\)-Hölder condition for \(\alpha \geq 2\) (Devroye et al., 2013), Fan and Gijbels (1996) considers the more general local polynomial regression to capture higher-order curvature information of the unknown regression function and gain optimal rate of convergence.

Inspired by the Nadaraya-Watson estimator and the local polynomial regression, we seek to develop a kernel mixture of polynomials prior by assuming the following form of \(f\),

\[
\hat{f}(x) = \sum_k \left\{ \frac{\varphi_h(x - \mu_k)}{\sum_l \varphi_h(x - \mu_l)} \right\} \sum_s \xi_{ks}(x - \mu_k^s)^s,
\]

where \(k\) and \(s\) lie in some index sets, and \(\{\xi_{ks}(x - \mu_k^s)^s\}_{ks}\) are some suitably selected polynomial functions that mimic the behavior of \(f\) locally around \(\mu_k\), where \(\mu_k\) and \(\mu_k^s\) are sufficiently close. Furthermore, in order for \(\mu_k\)'s to detect both the local behavior of \(f\) and the global behavior, they need to be well-spread. To achieve this, we partition the design space \(\mathcal{X}\) into \(K^p\) disjoint hypercubes \(\mathcal{X} = \bigcup_k \mathcal{X}_K(k)\), where \(\mathcal{X}_K(k) \cap \mathcal{X}_K(k') = \emptyset\) if \(k \neq k'\), and let \(\mu_k \in \mathcal{X}_K(k)\). This leads to an identifiable sub-model \((\mu_k, \xi_{ks})_{k,s} \rightarrow f\) for each fixed \(K\). The partition restriction is preferred because an unidentifiable model may result in poor mixing of the Markov chain Monte Carlo sampler for the posterior inference (Xie and Carlin, 2006). Alternatively, repulsive prior (Affandi et al., 2013; Xu et al., 2016a; Xie and Xu, 2017) can be incorporated to gain well-spread kernel centers \(\mu_k\)'s, but the identifiability issue remains unsolved. We formulate the setup rigorously in section 2.2.

The proposed kernel mixture of polynomials not only avoids the inversion of large covariance matrix, but also features nice frequentist asymptotic behaviors. The major contribution of this paper is that the proposed method is, to the best of our knowledge, the first one in the literature that simultaneously achieves the following theoretical goals: (i) It yields the near-optimal rate of contraction with respect to the \(L_2\)-topology with adaptation to the smoothness level of the true function; (ii) It provides a frequentist sieve maximum likelihood estimator with a near-optimal convergence rate; (iii) When used to model the nonparametric component in the partial linear model, it leads to optimal convergence results for both the nonparametric
component and the parametric component.

For goal (i), it is worth mentioning that most literatures concerning posterior convergence for Bayesian nonparametric regression only discuss the rate of contraction with respect to the weaker empirical $L_2$-norm (van der Vaart and van Zanten, 2008; De Jonge et al., 2010; van der Vaart and van Zanten, 2009; Bhattacharya et al., 2014), i.e., the convergence of the function at given design points. There is little discussion about the rate of contraction with respect to the exact $L_2$-norm for general Bayesian nonparametric regression methods. van der Vaart and Zanten (2011), Yang et al. (2017), and Yoo et al. (2016) address this issue only in the context of Gaussian process regression. We show that the rate of contraction is minimax-optimal (Stone, 1982; Györfi et al., 2006) (up to a logarithmic factor) and is adaptive to the smoothness level in the sense that the prior does not depend on the smoothness level of the true function. The main technical tools we use for establishing the rate of contraction are the upper bounds for metric entropies of certain function classes. We also obtain a frequentists sieve maximum likelihood estimator with near-optimal convergence rate, fulfilling goal (ii).

For goal (iii), we further study the application of the kernel mixture of polynomials to the partial linear model as a natural semiparametric extension. The partial linear model is of the form $y_i = z_i^T \beta + \eta(x_i) + e_i,$ where $z_i, x_i$’s are design points, $\beta$ is the linear coefficient, $\eta$ is some unknown function to which the kernel mixture of polynomials prior is imposed, and $e_i$’s are independent $N(0, 1)$ noises, $i = 1, \ldots, n$. The literatures of the partial linear model from both the frequentist perspective (Engle et al., 1986; Chen et al., 1988; Speckman, 1988; Hastie and Tibshirani, 1990; Fan and Li, 1999) and Bayesian approaches (Tang et al., 2015; Lenk, 1999; Bickel et al., 2012; Yang et al., 2015) are rich. However, there is little discussion regarding the theoretical properties of the Bayesian partial linear model. To the best of our knowledge, only Bickel et al. (2012) and Yang et al. (2015) discuss the asymptotic behavior of the marginal posterior distribution of $\beta$ with Gaussian process priors for $\eta$ under the Bayesian partial linear model. We impose the kernel mixture of polynomials prior on $\eta$ and obtain both the near-optimal rate of contraction for $\eta$ and the Bernstein von-Mises limit (i.e., asymptotic normality) of the marginal posterior of $\beta$. The technical tools developed for the kernel mixture of polynomials play a fundamental role in the proofs of the convergence properties of $\beta$ and $\eta$.

The layout of this paper is as follows. Section 2 introduces necessary notations and presents the kernel mixture of polynomials prior for nonparametric regression. Section 3 elaborates the convergence properties of the kernel mixture of polynomials regression. In particular, we provide upper bounds for metric entropies of certain function classes, derive an adaptive rate of contraction with respect to the $L_2$-topology, and obtain a sieve maximum likelihood estimator with a near-optimal convergence rate. Section 4 presents a semiparametric application of the kernel mixture of polynomials to the partial linear model. We explore the asymptotic behavior and obtain both a near-optimal rate of contraction for the nonparametric component and the Bernstein von-Mises limit of the parametric component. Section 5 illustrates the proposed methodology using numerical examples. Further discussions are included in section 6.
2 Preliminaries

2.1 Notations

For $1 \leq r \leq \infty$, we use $\| \cdot \|_r$ to denote both the $\ell_r$-norm on any finite dimensional Euclidean space and the $L_r$-norm of a measurable function. We follow the convention that when $r = 2$, the subscript is omitted, i.e., $\| \cdot \|_2 = \| \cdot \|$. For any integer $n$, denote $[n] = \{1, \ldots, n\}$. For $x^* \in \mathbb{R}^p$ and $\epsilon > 0$, denote $B_r(x^*, \epsilon) = \{ x \in \mathbb{R}^p : \| x - x^* \|_r < \epsilon \}$ for $1 \leq r \leq \infty$. We use $[x]$ to denote the maximal integer no greater than $x$, and $\lfloor x \rfloor$ to denote the minimum integer no less than $x$. The notations $a \lesssim b$ and $a \gtrsim b$ denote the inequalities up to a positive multiplicative constant. Given an integer vector $s = [s_1, \ldots, s_p]^T \in \mathbb{N}^p$ and a real vector $x = [x_1, \ldots, x_p]^T \in \mathbb{R}^p$, we denote the monomial $x^s = x_1^{s_1} \cdots x_p^{s_p}$ and the mixed partial derivative operator $D^s = \partial |s| / \partial x_1^{s_1} \cdots \partial x_p^{s_p}$, where $|s| = \sum_{j=1}^p s_j$. When $|s| = 0$, by convention $D^0 f(x) = f(x)$. Given $\alpha, L > 0$ and a compact $\mathcal{X} \subset \mathbb{R}^p$, we say that a function $f : \mathcal{X} \to \mathbb{R}$ satisfies the $\alpha$-H"older condition with envelope $B$, if $f$ is $[\alpha - 1]$-times continuously differentiable with

$$\max_{|s| \leq [\alpha - 1]} \|D^s f\|_\infty + \max_{|s| = [\alpha - 1]} \sup_{x_1 \neq x_2} \frac{|D^s f(x_1) - D^s f(x_2)|}{\|x_1 - x_2\|^{\alpha - |s|}} \leq B$$

for all $x_1, x_2 \in \mathcal{X}$. The class of functions on $\mathcal{X}$ that satisfies the $\alpha$-H"older condition with envelope $B$, denoted by $\mathcal{C}^{\alpha,B} (\mathcal{X})$, is referred to as the $\alpha$-H"older function class (with envelope $B$). Given a distribution $\mathbb{P}_x$ on $\mathcal{X}$, we denote the $L_r(\mathbb{P}_x)$-norm of a measurable function $f$ $\| f \|_{L_r(\mathbb{P}_x)} = \{ \int_\mathcal{X} |f(x)|^r \mathbb{P}_x(dx) \}^{1/r}$ for any $r \in [1, \infty)$. The notation $1(A)$ denotes the indicator of the event $A$.

We slightly abuse the notation and do not distinguish between a random variable and its realization. We refer to $\mathcal{P}$ as a statistical model if it consists of a class of densities on a sample space $\mathcal{X}$ with respect to some underlying $\sigma$-finite measure. Given a (frequentist) statistical model $\mathcal{P}$ and the independent and identically distributed data $\mathcal{D}_n = (x_i)_{i=1}^n$ from some $P \in \mathcal{P}$, the prior and the posterior distribution on $\mathcal{P}$ are always denoted by $\Pi(\cdot)$ and $\Pi(\cdot | \mathcal{D}_n)$, respectively. Given a function $f : \mathcal{X} \to \mathbb{R}$, we use $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(x_i)$ to denote the empirical measure of $f$, and $\mathbb{G}_n f = n^{-1/2} \sum_{i=1}^n (f(x_i) - Ef(x_i))$ to denote the empirical process of $f$, given the independent and identically distributed data $(x_i)_{i=1}^n$. We use $p_x(x)$ or $p(x)$ to denote the density of $x$, $\mathbb{P}_x$ to denote the distribution of $x$, and $E_x$ for the corresponding expected value. In particular, $\phi$ denotes the probability density function of the (univariate) standard normal distribution, and we use the shorthand notation $\phi_\sigma(y) = \phi(y/\sigma)/\sigma$ to denote the density of $\mathcal{N}(0, \sigma^2)$. The Hellinger distance between two densities $p_1, p_2$ is denoted by $H(p_1, p_2)$, and the Kullback-Leibler divergence is denoted by $D_{KL}(p_1||p_2) = \int p_1(x) \log(p_1(x)/p_2(x)) dx$. For a metric space $(\mathcal{F}, d)$, for any $\epsilon > 0$, the $\epsilon$-covering of $(\mathcal{F}, d)$, denoted by $\mathcal{N}(\epsilon, \mathcal{F}, d)$, is defined to be the minimum number of $\epsilon$-balls of the form $\{ g \in \mathcal{F} : d(f, g) < \epsilon \}$ that are needed to cover $\mathcal{F}$. The $\epsilon$-bracketing of $(\mathcal{F}, d)$, denoted by $\mathcal{N}_[\epsilon](\mathcal{F}, d)$, is defined to be the minimum number of brackets of the form $[l_i, u_i]$ that are needed to cover $\mathcal{F}$ with $l_i, u_i \in \mathcal{F}$ such that $d(l_i, u_i) < \epsilon$. We refer to $\log \mathcal{N}(\epsilon, \mathcal{F}, d)$ as the metric entropy, and $\log \mathcal{N}_[\epsilon](\mathcal{F}, d)$ as the bracketing (metric) entropy. The bracketing integral $\int_0^\epsilon \sqrt{\log \mathcal{N}_[\epsilon](\mathcal{U}, \mathcal{F}, d)} du$ is used extensively and is denoted by $J_\epsilon(\mathcal{F}, d)$.

Without loss of generality we assume that the design space $\mathcal{X}$ is the unit hypercube $[0,1]^p$ in $\mathbb{R}^p$, and $p_x$ is a continuous density yielding a distribution on $\mathcal{X}$. We partition the design space $\mathcal{X}$ into $K^p$ disjoint
hypercubes \( \mathcal{X} = \bigcup_{k} \mathcal{X}_K(k) \), where \( \mathcal{X}_K(k) \cap \mathcal{X}_K(k') = \emptyset \) if \( k \neq k' \). For each \( K \in \mathbb{N}_+ \), define \( \mathcal{X}_K(k) = \prod_{j=1}^{p} ((kj - 1)/K, kj/K) \), where \( k = [k_1, \ldots, k_p]^T \in [K]^p \) is the multivariate index.

### 2.2 Setup

We begin with the notion of boxed kernel functions. We say that a continuous function \( \varphi : \mathbb{R}^p \rightarrow [0,1] \) is a boxed kernel function if it is supported on \( \{ x : \| x \|_\infty \leq 1 \} \), does not increase when \( \| x \|_\infty \) increases, and \( \varphi(x) \leq 1(\| x \|_\infty \leq 1) \). Examples of such kernels include the triangle kernel \( \varphi(x) = (1 - \| x \|_\infty)1(\| x \|_\infty \leq 1) \), the Epanechnikov kernel \( \varphi(x) = (1 - \| x \|_\infty^2)1(\| x \|_\infty < 1) \), and the univariate bump kernel \( \varphi(x) = \exp\{-(1-x^2)^{-1}\}1(|x| < 1) \), etc. For convenience we denote \( \varphi_h(x) = \varphi(x/h) \), where \( h > 0 \) is the bandwidth parameter.

For each \( K \in \mathbb{N}_+ \) and each \( k \in [K]^p \), denote the kernel mixture weight as

\[
\psi_{ks}(x) = w_k(x)(x - \mu_k^s)^s, \quad k \in [K]^p, \quad s \in \{ s \in [m]^p : |s| = 0, 1, \ldots, m \}.
\]

We next describe the setup. The data \( \mathcal{D}_n = (x_i, y_i)_{i=1}^n \) are assumed to be i.i.d. according to the joint density \( p_0(x, y) = \phi_{\sigma_0}(y - f_0(x))p_x(x) \) for some \( \sigma_0 \in [\underline{\sigma}, \overline{\sigma}] \) and some unknown function \( f_0 \in \mathcal{C}^{a,B}(\mathcal{X}) \). In nonparametric regression the estimation of \( f_0 \) is of interest, and therefore the marginal density \( p_x \) of the design points \((x_i)_{i=1}^n \) is assumed known and fixed. We use \( p_{r0} \) and \( E_0 \) to denote the probability and expected value under \( p_{r0} \), respectively.

Formally, the kernel mixture of polynomials nonparametric regression model is described by the following statistical model \( \mathcal{M} = \{ p_{f,\sigma}(x, y) : f \in \bigcup_{K=1}^{\infty} \mathcal{F}_K, \sigma \in [\underline{\sigma}, \overline{\sigma}] \} \), where \( p_{f,\sigma}(x, y) = \phi_{\sigma}(y - f(x))p_x(x) \) is the joint density for \((x_1, y_1)_{i=1}^n \). \( \mathcal{F}_K = \bigcup_{K \hbar \in [\underline{h}, \overline{h}] \cap \mathbb{N}} \mathcal{F}_K(h) \) for some constants \( \underline{h}, \overline{h} \) with \( 1 < \underline{h} < \overline{h} < \infty \), and

\[
\mathcal{F}_K(h) = \left\{ \sum_{k \in [K]^p} \sum_{s : |s| = 0}^m \xi_{ks} \psi_{ks}(x) : \mu_k \in \mathcal{X}_K(k), \max_{|s| = 0, \ldots, m} |\xi_{ks}| \leq B, k \in [K]^p \right\}.
\]

Here \( k = [k_1, \ldots, k_p]^T \) and \( \mu_k^s = [(2k_1 - 1)/2K, \ldots, (2k_p - 1)/2K] \) is the center of \( \mathcal{X}_K(k) \). The parameters \( \sigma, K, (\mu_k) : k \in [K]^p \), and \( (\xi_{ks}) : k \in [K]^p, |s| = 0, 1, \ldots, m \) are to be assigned a hierarchical prior in section 2.3. The problem is nonparametric, and therefore \( f_0 \) is not necessarily in \( \mathcal{F}_K \) for any \( K \in \mathbb{N}_+ \). The only assumption we make for \( f_0 \) is that \( f_0 \in \mathcal{C}^{a,B}(\mathcal{X}) \) for some \( \alpha, B > 0 \). The function \( f \) described in (3) is referred to as the kernel mixture of polynomials. Clearly there exists some constant \( A \) such that \( \| f \|_\infty \leq A \) for all \( f \in \bigcup_{K=1}^{\infty} \mathcal{F}_K \), i.e., \( \bigcup_{K=1}^{\infty} \mathcal{F}_K \) is a uniformly bounded function class.

**Remark 1.** Denote \( D(x) = \sum_{t \in [K]^p} \varphi_h(x - \mu_t) \) to be the denominator of the weights in (1). We remark that the kernel weights in (1) are well-defined, i.e., the denominator \( D(x) > 0 \). To see why this is true, notice that \( \varphi(x) > 0 \) when \( \| x \| < 1 \) by assumption, it suffices to show that \( \mathcal{X} \subset \bigcup_{k \in [K]^p} B_\infty(\mu_k, h) \). In fact, for all
There exists some $k = k(x) \in [K]^p$ with $x \in \mathcal{X}_K(k) = B_{\infty}(\mu^*_k, 1/(2K)) \subset B_{\infty}(\mu^*_k, h/2)$, and it follows by triangle inequality that $\|x - \mu_k\| \leq \|x - \mu^*_k\| + \|\mu^*_k - \mu_k\| < h$, i.e., $x \in B_{\infty}(\mu_k, h) \subset \bigcup_{k \in [K]^p} B_{\infty}(\mu_k, h)$. Hence $D(x) > 0$ and $w_k(x)$’s are well-defined.

Remark 2. We require that $K$ ranges over all positive integers and that $Kh$ stays bounded as $K \to \infty$, so that $\bigcup_{k=1}^{\infty} \mathcal{F}_K$ is rich enough to provide good approximation to arbitrary $f_0 \in \mathcal{C}^{\alpha, B}(\mathcal{X})$ (Györfi et al., 2006). The kernel bandwidth parameter $h$, which is typically endowed with a prior in the literature of Bayesian kernel methods (Ghosal et al., 2007; Shen et al., 2013), also plays a key role in the establishing the convergence properties of the kernel mixture of polynomials. It turns out that requiring $1 < h \leq Kh \leq \bar{h} \leq \infty$ yields the near-optimal rate of contraction.

2.3 Prior specification

Let $\pi_* \mu$ be a continuous non-vanishing density on the hypercube $[-1, 1]^p$, and $\pi_\xi$ be a continuous non-vanishing density on $[-B, B]$. For each $f \in \mathcal{F}_K$, we write

$$f(x) = \sum_{k \in [K]^p} \sum_{s:|s|=0}^{m} \xi_{ks} \psi_{ks}(x),$$

where $\{\psi_{ks}(x) : k \in [K]^p, s \in [m]^p, |s| = 0, \ldots, m\}$ is the kernel mixture of polynomials system given by (2), $\mu_k \in \mathcal{X}_K(k), \xi_{ks} \in [-B, B]$, $k \in [K]^p, |s| = 0, 1, \ldots, m$. We define a prior distribution $\Pi$ for $(f, \sigma)$ over $\mathcal{F} \times [\sigma, \bar{\sigma}]$ through (4) by imposing the hierarchical priors for the parameters $(\mu_k : k \in [K]^p)$, $(\xi_{ks} : k \in [K]^p, |s| = 0, 1, \ldots, m)$, $\sigma$, $h$, and $K$ as follows. The standard deviation $\sigma$ of the noises $(e_i)_{i=1}^n$ follows the distribution with a density $\pi_\sigma$ that is continuous and non-vanishing on $[\sigma, \bar{\sigma}]$, independent of $K$, $(\mu_k : k \in [K]^p)$, and $(\xi_{ks} : k \in [K]^p, |s| = 0, 1, \ldots, m)$. Other parameters follow a hierarchical prior, where $K$ is first sampled, and then $(\mu_k : k \in [K]^p)$, $(\xi_{ks} : k \in [K]^p, |s| = 0, 1, \ldots, m)$, and $h$ are conditionally independent given $K$. The prior for $K$ satisfies the following condition:

$$\exp \{-b_0 x^p (\log x^p)^{r_0} \} \preceq \Pi(K \geq x) \preceq \exp \{-b_1 x^p (\log x^p)^{r_0} \}$$

for some constants $b_0, b_1 > 0$, $b_1 \leq b_0$, and $r_0 \geq 0$. The conditional prior for the rest parameters given $K$ satisfies the following conditions: The kernel centers $\mu_k = \mu^*_k + \bar{\mu}_k/(2K)$, where $\bar{\mu}_k$ independently follows $\pi_\mu$ for each $k \in [K]^p$. The coefficients $\xi_{ks}$ for $\psi_{ks}(x)$ follows $\pi_\xi$ independently for each $s \in [m]^p$ with $|s| = 0, 1, \ldots, m$. The bandwidth $h$ is supported on $[\bar{h}/K, \bar{\bar{h}}/K]$ with a non-vanishing density $\pi(h \mid K)$.

Kruijier et al. (2010) adopts the same tail probability condition (5) for the number of support points in the context of nonparametric density estimation. Special cases of (5) include the geometric distribution when $r_0 = 0$, and the Poisson distribution when $r_0 = 1$. In section 3, we show that both $r_0 = 0$ and $r_0 = 1$ yield the same rate of contraction, but any $r_0 > 1$ (i.e., thinner tail) leads to a slower rate of contraction.
3 Convergence properties of kernel mixture of polynomials regression

In this section, we establish the convergence results of the kernel mixture of polynomials regression and obtain a frequentist sieve maximum likelihood estimator with the corresponding convergence rate. For nonparametric regression sieve problems, when the true regression function \( f_0 \) is in \( \mathcal{C}^{\alpha,L}(\mathcal{X}) \), \( p_{\bar{x}}(x) = 1 \), and \( \epsilon_i \sim N(0,1), i = 1, \ldots, n \), it has been shown that the minimax rate of convergence for any estimator \( \hat{f} \) with respect to the \( L_2 \)-norm is \( n^{-\alpha/(2\alpha+p)} \) (Stone, 1982; Györfi et al., 2006). The optimal rate of contraction cannot be faster than the minimax rate of convergence. Theorem 1 below, which is one of the main results of this section, asserts that the rate of contraction with respect to the \( L_2(p_x) \)-topology is minimax-optimal up to a logarithmic factor. Furthermore, the rate of contraction is adaptive to the smoothness \( \alpha \) of \( f_0 \) in the sense that the prior specification does not require knowledge of \( \alpha \). In this section the proofs of lemma 1, proposition 1, proposition 2, and theorem 2 are deferred to the supplementary material.

**Theorem 1** (Rate of contraction). Suppose \( \Pi \) is the prior constructed in section 2.3. Then \( \Pi(\|f-f_0\|_{L_2(p_x)}^2 > M\epsilon_n^2 | \mathcal{D}_n) \to 0 \) in \( \Pr \)-probability for some constant \( M > 0 \), where \( \epsilon_n = n^{-\alpha/(2\alpha+p)}(\log n)^{t/2} \), and \( t > 2\alpha \max(r_0,1)/(2\alpha+p) + \max(0,1-r_0) \).

The sketch of the proof goes as follows. First of all lemma 1 below guarantees that in order to prove theorem 1, it suffices to show that

\[
\Pi(H^2(p_{f,\sigma},p_0) > M\epsilon_n^2|\mathcal{D}_n) \to 0
\]

in \( \Pr \)-probability. Then we prove (6) by verifying a set of sufficient conditions presented in Kruijer et al. (2010). For convenience we describe these conditions in our context. Let \( \mathcal{M} \) be a statistical model, i.e., a class of density functions, equipped a prior \( \Pi \). Define the Kullback-Leibler ball by \( B_{KL}(p_0,\epsilon) = \{p_{f,\sigma} : D_{KL}(p_0||p_{f,\sigma}) < \epsilon, E_0(\log(p_0(x,y)/p_{f,\sigma}(x,y)))^2 < \epsilon^2 \} \). Kruijer et al. (2010) proved that in order that (6) holds, it suffices to find another sequence \( (\xi_n)_{n=1}^\infty \) with \( \xi_n \leq \epsilon_n \), a sequence of sub-classes of densities \( (\mathcal{M}_n)_{n=1}^\infty, \mathcal{M}_n \subset \mathcal{M} \), 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 \( \Pi(p \in B_{KL}(p_0,\xi_n)) \geq \exp(-n\xi_n^2), \Pi(p \in \mathcal{M}_n) \leq \exp(-4n\epsilon_n^2), \) and \( \exp(-n\epsilon_n^2) \sum_{m=1}^\infty \sqrt{N(\epsilon_n, \mathcal{M}_{nm},H)} \) \( \Pi(p \in \mathcal{M}_{nm}) \to 0 \).

The sequence of sub-classes of densities \( (\mathcal{M}_n)_{n=1}^\infty \) is referred to as a sieve in the literature (Shen and Wong, 1994).

The following lemma serves as the first step of proving theorem 1.

**Lemma 1.** Suppose \( \mathcal{G} \subset \{ f : \mathcal{X} \rightarrow \mathbb{R} \} \) is a function class. Then

\[
\|f-g\|_{L_2(p_x)}^2 + |\sigma_1 - \sigma_2|^2 \lesssim H^2(p_{f,\sigma_1},p_{g,\sigma_2}) \lesssim \|f-g\|_{L_1(p_x)} + |\sigma_1 - \sigma_2|^2,
\]

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

\[
\mathcal{N}(\epsilon, \mathcal{M}_{\leq}^2(\mathcal{G}),H) \leq \mathcal{N}(C_1 \epsilon^2, \mathcal{G}, \| \cdot \|_{L_1(p_x)}) \left( \frac{\| \sigma - \hat{\sigma} \|}{C_1 \epsilon} + 1(\hat{\sigma} = \sigma) \right),
\]

7
where $\mathcal{M}_\sigma^\tau(\mathcal{G}) = \{ p_{f,\sigma} : f \in \mathcal{G}, \sigma \in [\sigma, \overline{\sigma}] \}$.

We now provide a lower bound for the prior concentration $\Pi(p \in B_{KL}(p_0, \xi_n))$.

**Proposition 1** (Prior concentration). Suppose $\Pi$ is the prior constructed in section 2.3. Then there exists some constant $C_2 > 0$ such that for all sufficiently small $\epsilon > 0$,

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

We next turn to the estimation of the metric entropies of $\mathcal{F}_K$, which is one of the major technical contributions of this paper.

**Proposition 2** (Metric entropy bound). There exists some constant $c_2 > 0$, such that for sufficiently small $\epsilon > 0$ and any $r \in [1, \infty)$,

$$
\log \mathcal{N}_r(2\epsilon, \mathcal{F}_K, \| \cdot \|_{L_r(\mathcal{F}_K)}) \leq \log \mathcal{N}(\epsilon, \mathcal{F}_K, \| \cdot \|_{\infty}) \leq 2K^p \{ (m+1)^p + p + 1 \} \left( \log \frac{1}{\epsilon} \right).
$$

of theorem 1. We first prove that $\epsilon_n = n^{-\alpha/(2\alpha+p)} (\log n)^{t/2}$ is the contraction rate for the density estimation, i.e., (6) holds with $\epsilon_n = n^{-\alpha/(2\alpha+p)} (\log n)^{t/2}$. For any function class $\mathcal{G}$, denote the class of densities associated with $\mathcal{G}$ by $\mathcal{M}_\sigma^\tau(\mathcal{G}) = \{ p_{f,\sigma} : f \in \mathcal{G}, \sigma \in [\sigma, \overline{\sigma}] \}$. 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 $K_n = [n^{1/(2\alpha+p)} (\log n)^{d}]$, $\xi_n = n^{-\alpha/(p+2\alpha)} (\log n)^{\gamma}$, and $\mathcal{M}_n = \bigcup_{K=1}^{K_n} \mathcal{M}_\sigma^\tau(\mathcal{F}_K)$. We see that $\epsilon_n > \xi_n$ 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 1, proposition 2, and the fact that $\| f \|_{L_r(\mathcal{F}_K)} \leq \| f \|_{\infty}$ for any $r \geq 1$, we have

$$
\exp(-n\epsilon_n^2) \sum_{K=1}^{K_n} \sqrt{\mathcal{N}(\epsilon_n, \mathcal{M}_\sigma^\tau(\mathcal{F}_K), H)} \sqrt{\Pi(p_{f,\sigma} \in \mathcal{M}_\sigma^\tau(\mathcal{F}_K))}
\lesssim \exp \left\{ -n\epsilon_n^2 + \log \frac{1}{\epsilon_n} \right\} \sum_{K=1}^{K_n} \exp \left\{ K^p \{ (m+1)^p + p + 1 \} \left( \log \frac{1}{\epsilon_n^2} \right) \right\}
\leq \exp \left\{ -n\epsilon_n^2 + \log \frac{1}{\epsilon_n} \right\} K_n \exp \left\{ 3K^p \{ (m+1)^p + p + 1 \} \left( \log \frac{1}{\epsilon_n} \right) \right\}
\lesssim \exp(-n\epsilon_n^2) \exp \left\{ 4K^p \{ (m+1)^p + p + 2 \} \left( \log \frac{1}{\epsilon_n} \right) \right\}
\lesssim \exp \left\{ -n\epsilon_n^2 \frac{p^2}{2\alpha} (\log n)^t + C'_1 n^{p/2\alpha} (\log n)^{p\delta+1} \right\} \to 0
$$

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
\[ \Pi( \rho, \sigma \in M_n^c) \leq B_1 \exp\{-b_1 K_n^p (\log K_n^p)^r_0\} \]
\[ \leq \exp\{-b_1' n^{p/(p+2\alpha)} (\log n)^{p\delta +r_0}\} \leq \exp(-4n\xi_n^2), \]
where the fact \( p\delta + r_0 > 2\gamma \) is applied. Lastly, for the prior concentration, by proposition 1
\[ \Pi(\rho, \sigma \in B_{KL}(p_0, \xi_n)) \geq \exp\{-C_2\xi_n^2\left(\log\frac{1}{\xi_n}\right)^{\max(r_0,1)}\} \]
\[ \geq \exp\{-C_2' n^{p/(p+2\alpha)} (\log n)^{\max(r_0,1)} - p\gamma/\alpha\} \geq \exp(-n\xi_n^2) \]
for some constant \( C_2' > 0 \), where we use the fact \( 2\gamma > \max(r_0,1) - p\gamma/\alpha \) in the last inequality. Hence we conclude that \( \Pi(H(\rho, \sigma, p_0) > M\epsilon_n| D_n) \to 0 \) in \( \text{pr}_0 \)-probability for some constant \( M > 0 \). The proof is completed by applying lemma 1.

Besides the rate of contraction, which is a frequentist large sample evaluation of Bayesian posterior distribution, we also obtain a frequentist sieve maximum likelihood estimator with the convergence rate as a result of the metric entropy bounds. This convergence rate is also minimax optimal up to a logarithmic factor. Interestingly, this convergence rate is faster compared to the rate of contraction of the full posterior, although the construction of the sieve depends on the smoothness level \( \alpha \) and the rate is non-adaptive.

**Theorem 2.** Consider the sieve maximum likelihood estimator \( \hat{f}_K(x) \) defined by
\[ \hat{f}_K(x) = \arg\max_{f \in \mathcal{G}_K} \sum_{i=1}^n \log \phi_{\sigma_0}(y_i - f(x_i)), \]
where
\[ \mathcal{G}_K = \left\{ \sum_{s=0}^m \sum_{|s| = 0}^K \xi_{ks} \psi_{ks}(x) : \text{Kh} \in [h, H], \mu_k \in \mathcal{X}_K(k), \max_{|s| = 0, \ldots, m} |\xi_{ks}| \leq B, k \in [K]^p \right\}. \]

If \( K_n = \left[(n/\log n)^{1/(2\alpha + p)}\right] \), then \( \text{pr}_0(\|f_0 - \hat{f}_K_n\|_{L_2(p_i)} \geq M(\log n/n)^{\alpha/(2\alpha + p)}) = 0 \) for some large constant \( M > 0 \).

### 4 Application to partial linear model

In this section we focus on a natural semiparametric application: we use the kernel mixture of polynomials to model the nonparametric component in the partial linear model. Specifically, we consider the partial linear model of the form \( y_i = z_i^T \beta + \eta(x_i) + e_i \), where \( z_i, x_i \)'s are design points, \( \beta \) is the linear coefficient, \( \eta : \mathcal{X} \to \mathbb{R} \) is an unknown function modeled by the kernel mixture of polynomials, and \( e_i \)'s are independent \( N(0, 1) \) noises. The main focus of this section is the convergence results for both the nonparametric component \( \eta \) (theorem 3) and the parametric component \( \beta \) (theorem 4). Furthermore, as a consequence of the metric
entropy results (proposition 2), we obtain the Bernstein von-Mises limit of the marginal posterior distribution of the parametric linear coefficient $\beta$.

4.1 Setup and prior specification

We begin with the detailed description of the model. Let $\mathcal{X} = [0, 1]^p \subset \mathbb{R}^p$ be the design space of the nonparametric component, $\mathcal{Z} \subset \mathbb{R}^q$ be the design space of the parametric component, and $p(x,z) : \mathcal{X} \times \mathcal{Z} \to (0, \infty)$ be a continuous density function supported on $\mathcal{X} \times \mathcal{Z}$. The partial linear model is of the form

$$y_i = z_i^T \beta + \eta(x_i) + \epsilon_i,$$

where $(x_i, z_i)_{i=1}^n$ are independently sampled from $p(x,z)$, $(\epsilon_i)_{i=1}^n$ are independent $N(0,1)$ noises, and $\beta$ is the linear coefficient. In many applications, the estimation of the nonparametric component $\eta$ is of great interest. For example, in Xu et al. (2016b), the parametric term $z_i^T \beta$ models the baseline disease progression and the nonparametric term $\eta(x_i)$ models the individual-specific treatment effect deviations over time. When the regression coefficient $\beta$ is of more interest, the estimation of $\eta$ can still be critical since it could affect the inference of $\beta$.

We incorporate the partial linear model with the kernel mixture of polynomials prior for the nonparametric component $\eta$ through the following statistical model: $P = \{p_{\beta, \eta}(x, z, y) : \beta \in \mathbb{R}^q, \eta \in \bigcup_{K=1}^{\infty} F_K\}$, where $p_{\beta, \eta}(x, z, y) = \phi(y - z^T \beta - \eta(x))p_{x,z}(x, z)$ and $F_K = \bigcup_{h \in [h, h]} F_K(h)$ with $F_K(h)$ given by (3). We assume that the data $D_n = (x_i, z_i, y_i)_{i=1}^n$ are independently sampled from $p_0(x, z, y) = \phi(y - z^T \beta_0 - \eta_0(x))p_{x,z}(x, z)$ for some $\beta_0 \in \mathbb{R}^q$ and some function $\eta_0 \in C^{\alpha,L}(\mathcal{X})$. We make several additional assumptions: The design space $\mathcal{Z} \subset \mathbb{R}^q$ for the linear component is compact with $\sup_{z \in \mathcal{Z}} \|z\|_1 \leq \overline{B}$; The sampling distribution for $z$ satisfies $Ez = 0$ and $Ezz^T$ being non-singular; The density of the design points $(x_i, z_i)_{i=1}^n$ factorizes as $p_{(x,z)}(x, z) = p_x(x)p_z(z)$, i.e., $x$ and $z$ are independent.

For the prior specification, we assume $\eta$ follows the kernel mixture of polynomials prior $\Pi_\eta$ constructed in section 2.3 with $\sigma = 1$. For the parametric component $\beta$, we impose a standard Gaussian prior $\Pi_\beta = N(0, I_q)$, independent of $\Pi_\eta$. The joint prior is denoted by $\Pi = \Pi_\eta \times \Pi_\beta$.

4.2 Convergence results

Before exploring the asymptotic behavior of the posterior distribution of $\beta$, we need to establish the convergence result for $\eta$. The following theorem not only addresses the rate of contraction of the marginal posterior of $\eta$, but also serves as one of the building blocks for proving the Bernstein von-Mises limit of the marginal posterior of $\beta$. The proofs of theorem 3 and theorem 4 are deferred to the supplementary material.

**Theorem 3** (Nonparametric rate). Under the setup and prior specification in section 4.1, $\Pi(\|\eta - \eta_0\|^2_{L_2(p_x)}) > Mc_n^2 | D_n) \to 0$ in $p_{\eta_0}$-probability for some constant $M > 0$, where $c_n = n^{-\alpha/(2\alpha + p)}(\log n)^{t/2}$, and $t > 2\alpha \max(r_0, 1)/(2\alpha + p) + \max(0, 1 - r_0)$.

Now we turn to the convergence results for the parametric component. The focus is the asymptotic normality of marginal posterior distribution of the linear regression parameter $\beta$, i.e., the Bernstein von-Mises limit (Doob, 1949). To achieve this, we need the notion of the least favorable model for semiparametric models (Bickel et al., 1998). For each fixed $\beta \in \mathbb{R}^q$, the least favorable curve $\eta^*_\beta$ is defined by the minimizer of the KL-divergence: $\eta^*_\beta(x) = \arg\inf_{\eta \in F} D_{KL}(p_0 \| p_{\beta, \eta})$. It is easy to see that $\eta^*_\beta$ coincides with $\eta_0(x)$ in our
In fact, for each $\beta$, we have
\[
\eta_\beta^*(x) = \arg\inf_{\eta \in \mathcal{F}} D_{\text{KL}}(p_0 \| p_{\beta,\eta}) \\
= \arg\inf_{\eta \in \mathcal{F}} E \left\{ (z^T(\beta - \beta_0) + \eta(x) - \eta_0(x))^2 \right\} \\
= \arg\inf_{\eta \in \mathcal{F}} E \left\{ \eta^2(x) - 2\eta(x) \left( \eta_0(x) - E(z \mid x)^T(\beta - \beta_0) \right) \right\} \\
= \arg\inf_{\eta \in \mathcal{F}} E \left\{ \eta_0(x) - \eta(x))^2 \right\}.
\]

The least favorable submodel is defined to be \( \{p_{\beta,\eta^*_\beta} : \beta \in \mathbb{R}^q\} \), which coincides with \( \{p_{\beta,\eta_0} : \beta \in \mathbb{R}^q\} \) in our context.

**Theorem 4.** Under the setup and prior specification in section 4.1, if \( \alpha > p/2 \), then
\[
\sup_{\mathcal{F}} |\Pi(\sqrt{n}(\beta - \beta_0) \in \mathcal{F} \mid \mathcal{D}_n) - \Phi(F \mid \Delta, (Ezz^T)^{-1})| \to 0
\]
in \( \mathcal{P}_0 \)-probability, where \( \Phi(\cdot \mid \Delta, (Ezz^T)^{-1}) \) is the \( \mathcal{N}(\Delta, (Ezz^T)^{-1}) \) probability measure and
\[
\Delta_n = n^{-1/2} \sum_{i=1}^n (Ezz^T)^{-1} z\{y_i - \eta_0(x_i) - z_i^T \beta_0\}.
\]

The proof of theorem 4 is based on verifying a set of sufficient conditions in Yang et al. (2015), which are provided in section D of the supplementary material. However, we remark that the metric entropy results (proposition 2) we obtain in section 3 and the previous rate of contraction for \( \eta \) (theorem 3) are also of fundamental interest in the verification process.

## 5 Numerical studies

In this section we perform numerical studies of the kernel mixture of polynomials for nonparametric regression and the partial linear model. The posterior inference for all examples is carried out by a Markov chain Monte Carlo sampler, where the number \( B \) of iterations for the burn-in period is set to 1000. To determine \( K \), we implement the standard Markov chain Monte Carlo to collect posterior samples for each fixed \( K \in \{K_{\min}, \ldots, K_{\max}\} \) and find the optimal \( K \) by minimizing the deviance information criterion (Gelman et al., 2014) over \( K \). We collect 1000 post-burn-in Markov chain Monte Carlo samples for posterior analysis. Numerical evidence shows that all Markov chains converge within 1000 iterations. The kernel we use for the kernel mixture of polynomials is the bump kernel \( \varphi(x) = \exp\{-(1 - x^2)^{-1}\} \mathbf{1}(|x| < 1) \).

### 5.1 A synthetic example for nonparametric regression

We first consider a synthetic example for nonparametric regression. Following Knapik et al. (2011) and Yoo et al. (2016), we consider the true function to be \( f_0(x) = \sqrt{2} \sum_{s=1}^{\infty} s^{-3/2} \sin(s) \cos((s - 1/2)\pi x) \), which has smoothness level \( \alpha = 1 \). We generate \( n = 1000 \) observations \((x_i, y_i)_{i=1}^n\) given by \( y_i = f_0(x_i) + e_i \), where the design points \((x_i)_{i=1}^n\) are independently and uniformly sampled over \( X' = [0, 1) \), and \((e_i)_{i=1}^n\) are independent \( \mathcal{N}(0, 0.1^2) \) noises. For the prior specification, we assume \( \pi_\mu = \text{Unif}(-1,1) \), \( \pi_\beta = \mathcal{N}(0,10^2) \), and \( \pi_\eta = \mathcal{N}(0,1^2) \).
\( \pi_\xi = \mathcal{N}(0, 10^2) \mathbf{1}(|\xi| \leq 50) \), and \( K h \sim \text{Unif}(h, \bar{h}) \). The hyperparameters are set as \( h = 1.2, \bar{h} = 2, B = 50, K = 15, \) and \( m = 2 \). The range of \( K \) is set to be \( \{6, 7, \ldots, 15\} \).

For comparison we consider three competitors for estimating \( f_0 \): the local polynomial regression (Fan and Gijbels, 1996), implemented in the \texttt{locpol} package (Cabrera, 2012), DiceKriging method (Roustant et al., 2012), and the robust Gaussian stochastic process emulation (Gu et al., 2017), implemented in the \texttt{RobustGaSP} package (Gu et al., 2016). The point-wise posterior means and 95%-credible/confidence intervals for \( f(x) \) using the four nonparametric regression methods are plotted in Figure 1, respectively. We also compute the mean-squared error of the posterior means of the four methods, where the ground true \( f_0 \) is evaluated at 1000 equidistant design points. In terms of accuracy measured by mean-squared error and marked in the bottom-left corner of each panel, the kernel mixture of polynomials performs better than DiceKriging and similarly to the local polynomials and the robust Gaussian stochastic process. For the uncertainty quantification measured by the width of posterior credible/confidence band, the kernel mixture of polynomials outperforms the other three competitors.

![Figure 1](image_url)

Figure 1: Synthetic example for nonparametric regression. Shaded regions are point-wise 95%-credible/confidence intervals. The solid lines are point-wise posterior means/point estimators of \( f \) and the dot-dashed lines are the ground true \( f_0 \). Scatter points are the observations.
5.2 Partial linear model for the wage data

We further analyze the cross-sectional data on wages (Wooldridge, 2015), a benchmark dataset for partial linear model. This dataset is also available in the np package (Hayfield et al., 2008). The data consist of 526 observations with 24 variables and are taken from U.S. Current Population Survey for the year 1976. In particular, we are interested in modeling the hourly wage on the logarithm scale as the response with respect to 5 variables: years of education (educ), years of potential experience (exper), years with current employer, gender, and marital status. Choi and Woo (2015) and Hayfield et al. (2008) suggest the following form of the model:

\[ y_i = \beta_{1}\text{female}_i + \beta_{2}\text{married}_i + \beta_{3}\text{educ}_i + \beta_{4}\text{tenure}_i + \eta(x_{i}^{\text{exper}}) + e_i, \]

where \(e_i\)'s are independent \(N(0, \sigma^2)\) noises. The \(\text{female}_i\) are \(\pm 1\)-valued, where \(\text{female}_i = 1\) indicates that the \(i\)th observation is a female, and \(-1\) otherwise. We set \(\text{married}_i = 1\) to represent that the \(i\)th observation is married, and \(-1\) otherwise. We centralize \(\text{educ}_i\) and \(\text{tenure}_i\) before applying the partial linear model, i.e., \(\sum_{i=1}^{n}\text{educ}_i = \sum_{i=1}^{n}\text{tenure}_i = 0\). The \(x_{i}^{\text{exper}}\) are re-scaled so that they lie in \((0, 1)\). To evaluate the performance of the proposed method, we use 300 observations as training data, with the rest 226 observations left as testing data to compute the prediction mean-squared error. The prior specification and hyperparameters for the Markov chain Monte Carlo sampler are set as follows: \(h = 1.2, \bar{h} = 2, m = 3, \pi(\sigma^2) \propto [\sigma^2]^{-2} \exp(-1/\sigma^2)1(\sigma \leq \sigma \leq \bar{\sigma})\) (i.e., the truncated inverse-Gamma density), \(\pi_\beta = N(0, 10^2), \pi_\xi = N(0, 10^2)1(|\xi| \leq 100)\). The range of \(K\) is set to be \(\{11, 12, \ldots, 20\}\).

We calculate the posterior means and posterior 95%-credible intervals for \(\beta\). For comparison, we also provide the least-squared estimate of \(\beta\) and the estimate computed by the np package (Hayfield et al., 2008). The results are summarized in Table 1 (the “KMP” column represents the posterior means of \(\beta\)), showing that the kernel mixture of polynomials estimate is closer to the np package estimate compared to the least-squared estimate, and all three point estimates of \(\beta\) lie in the posterior 95%-credible intervals. For the nonparametric component, we compute the kernel mixture of polynomials prediction on the testing dataset. The comparison to the true testing responses is demonstrate in Panel (a) of Figure 2. For comparison, three alternatives based on the np estimate of \(\beta\) for estimating the nonparametric component \(\eta\) are: the local polynomial regression, DiceKriging, and the robust Gaussian stochastic process. The performance of these three competitors are visualized in Panels (b), (c), and (d) in Figure 2, respectively. The local polynomial regression estimate does not outperform the kernel mixture of polynomials in terms of the prediction mean-squared error, and the prediction curve is highly non-smooth. DiceKriging does not work in this scenario: the prediction mean-squared error is large, the prediction curve is highly non-smooth, and the point-wise confidence intervals show singularity in estimating the covariance matrix. The robust stochastic Gaussian process, though gives similar

|  | KMP | 95% credible intervals | np package | \(\hat{\beta}_{LS}\) |
|---|---|---|---|---|
| female -0.1214 | (-0.2534, -0.0071) | -0.1287 | -0.0921 |
| married 0.0249 | (-0.0943, 0.1629) | 0.0279 | 0.3209 |
| educ 0.0903 | (0.0373, 0.1404) | 0.0891 | 0.1257 |
| tenure 0.0175 | (-0.0032, 0.0372) | 0.0167 | 0.0152 |
mean-squared error compared to the kernel mixture of polynomials, does not capture the local nonlinearity of the nonparametric component. In addition, the point-wise 95% confidence/credible intervals for the local polynomial regression and the robust Gaussian stochastic process are wider than those given by the kernel mixture of polynomials when \( x \in (0, 0.6) \). When the data are dense in the region \((0, 0.6)\), the point-wise credible intervals estimated by the kernel mixture of polynomials are thinner; when the design points are sparser in the region \( x \geq 0.6 \), there exists larger uncertainty in estimating \( \eta \). Namely, the uncertainty of the kernel mixture of polynomials is adaptive to the distribution of the design points.

![Figure 2: Wage data example. Shaded regions are point-wise 95%-credible/confidence intervals. The solid lines are point-wise posterior means/point estimators of \( \eta \) and the dot-dashed lines are \( \eta_0 \). Circle scatter points are the training observations, and triangle scatter points are testing responses. Panels (b), (c), and (d) are computed using the discrepancy data \((x_i, y_i - z_i^T \hat{\beta}_{np})_{i=1}^n\), where \( \hat{\beta}_{np} \) is the estimate of \( \beta \) computed using the np package.](image)

6 Discussion

There are several potential extensions of the current work. Firstly, we develop the theoretical results under the assumption that the noises \((e_i)_{i=1}^n\) are Gaussian. In cases where the noises are only assumed to be sub-Gaussian, further exploration of the convergence properties can be investigated. Secondly, the design
points are assumed to be random in the present paper. In cases where the design points are fixed, which is also a common phenomenon in many physical experiments (Tuo and Wu, 2015), theoretical results for the kernel mixture of polynomials can be further extended using the techniques developed for non-independent nor-identically distributed observations by Ghosal et al. (2007). Thirdly, the kernel mixture of polynomials prior is constructed over a uniformly bounded function space. Dropping such a requirement may require significant work. In addition, when applying the kernel mixture of polynomials to the partial linear model, we only consider the case where $Ez = 0$ and $x$ is independent of $z$, indicating that the linear component and the nonparametric component are orthogonal. On one hand, the idea of orthogonality has been explored in the literature of calibration of inexact computer models (Plumlee and Joseph, 2016; Plumlee, 2017), and therefore exploring the application of the kernel mixture of polynomials to calibration of orthogonal computer models is a promising extension. On the other hand, it is also interesting to investigate the convergence theory when the two components are not orthogonal to each other. Finally, we have developed a theoretical support for the sieve maximum likelihood estimator with compact restrictions on the parameter spaces. In particular, the loss function is of the least-square form. From the computational perspective, an efficient optimization technique can be designed to obtain the frequentist estimator in light of the rich literature of solving nonlinear least-square problems (Nocedal and Wright, 2006).

**Supplementary material**

The supplementary material contains additional numerical studies, additional technical results, the remaining proofs, and the cited theorems that are used in the proofs.

**References**

Affandi, R. H., Fox, E., and Taskar, B. (2013). Approximate inference in continuous determinantal processes. In *Advances in Neural Information Processing Systems*, pages 1430–1438.

Bhattacharya, A., Pati, D., and Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Annals of Statistics*, 42(1):352.

Bickel, P., Kleijn, B., et al. (2012). The semiparametric Bernstein-von Mises theorem. *The Annals of Statistics*, 40(1):206–237.

Bickel, P. J., Klaassen, C. A., Ritov, Y., Wellner, J. A., et al. (1998). Efficient and adaptive estimation for semiparametric models.

Cabrera, J. (2012). *locpol: Kernel local polynomial regression*. R package version 0.4-0.

Chen, H. et al. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, 16(1):136–146.

Choi, T. and Woo, Y. (2015). A partially linear model using a Gaussian process prior. *Communications in Statistics-Simulation and Computation*, 44(7):1770–1786.
De Jonge, R., Van Zanten, J., et al. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *The Annals of Statistics*, 38(6):3300–3320.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilités et ses applications*, pages 23–27.

Engle, R. F., Granger, C. W., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association*, 81(394):310–320.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Fan, Y. and Li, Q. (1999). Root-n-consistent estimation of partially linear time series models. *Journal of Nonparametric Statistics*, 11(1-3):251–269.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

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.

Gu, M., Palomo, J., and Berger, J. O. (2016). Robust GaSP: an R package for objective Bayesian emulation of complex computer model codes. *Technical Report*.

Gu, M., Wang, X., and Berger, J. O. (2017). Robust Gaussian stochastic process emulation. *arXiv preprint arXiv:1708.04738*.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Wiley Online Library.

Hayfield, T., Racine, J. S., et al. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5):1–32.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421.

Knapik, B. T., van der Vaart, A. W., van Zanten, J. H., et al. (2011). Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, 39(5):2626–2657.

Kruijjer, 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.
Lenk, P. J. (1999). Bayesian inference for semiparametric regression using a Fourier representation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):863–879.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Plumlee, M. (2017). Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, pages 1–12.

Plumlee, M. and Joseph, V. R. (2016). Orthogonal Gaussian process models. *arXiv preprint arXiv:1611.00203*.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization. *Journal of Statistical Software*, 51(1):54p.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 413–436.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053.

Takeda, H., Farsiu, S., and Milanfar, P. (2007). Kernel regression for image processing and reconstruction. *IEEE Transactions on image processing*, 16(2):349–366.

Tang, Y., Sinha, D., Pati, D., Lipsitz, S., and Lipshultz, S. (2015). Bayesian partial linear model for skewed longitudinal data. *Biostatistics*, 16(3):441–453.

Tuo, R. and Wu, C. J. (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352.

van der Vaart, A. and Zanten, H. v. (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119.

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.
van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, pages 1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, pages 2655–2675.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

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.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Nelson Education.

Xie, F. and Xu, Y. (2017). Bayesian repulsive Gaussian mixture model. *arXiv preprint arXiv:1703.09061*.

Xie, Y. and Carlin, B. P. (2006). Measures of Bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10):3458–3477.

Xu, Y., Mueller, P., and Telesca, D. (2016a). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, 72(3):955–964.

Xu, Y., Xu, Y., and Saria, S. (2016b). A Bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine Learning for Healthcare Conference*, pages 282–300.

Yang, Y., Bhattacharya, A., and Pati, D. (2017). Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*.

Yang, Y., Cheng, G., and Dunson, D. B. (2015). Semiparametric Bernstein-von Mises theorem: Second order studies. *arXiv preprint arXiv:1503.04493*.

Yoo, W. W., Ghosal, S., et al. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics*, 44(3):1069–1102.
Abstract

This supplementary material contains additional numerical studies, additional technical results, the remaining proofs, and the cited theorems that are used in the proofs.

A Additional numerical studies: a synthetic example for partial linear model

We consider a synthetic example to evaluate the performance of the kernel mixture of polynomials for the partial linear model. We simulate $n = 500$ observations according to the model $y_i = z_i^T \beta_0 + \eta_0(x_i) + e_i$, where $\beta_0$ is provided in Table 2, $(e_i)_{i=1}^n$ are independent $N(0,1)$ noises that are independent of $(x_i, z_i)_{i=1}^n$, and $\eta_0(x) = 2.5 \exp(-x) \sin(10\pi x)$. The nonparametric function $\eta_0$ is highly nonlinear and hence brings natural challenge to estimation. The design points $(z_i)_{i=1}^n$ for the linear component follow Unif([-1,1]) independently, and the design points $(x_i)_{i=1}^n$ for $\eta$ are independently sampled from Unif(0,1). The hyper-parameters for the the kernel mixture of polynomials prior are set as follows: $h_1 = 1.2, h_2 = 2, B = 100$, and $m = 3$. For the prior specification, we assume $\pi_\mu = \text{Unif}(-1,1), \pi_\beta = N(0,10^2), \pi_\xi = N(0,10^2)1(|\xi| \leq 100)$, and $Kh \sim \text{Unif}(h, \tilde{h})$. The range of $K$ is set to be $\{6, 7, \ldots, 15\}$.

For the parametric component, we compute the posterior means and the posterior 95% credible intervals for $\beta$. For comparison, we calculate the least-square estimate of $\beta$: $\hat{\beta}_{LS} = \arg \min_{\beta \in \mathbb{R}^q} \sum_{i=1}^n (y_i - z_i^T \beta)^2$. The comparison is provided in Table 2, where the column “KMP” stands for the posterior means of $\beta$ under the kernel mixture of polynomials prior. From the posterior summary for $\beta$ we see that the underlying true $\beta_0$ lie in the posterior 95% credible intervals under the kernel mixture of polynomials prior, and the corresponding posterior means outperform the least-squared estimate in terms of accuracy. For the nonparametric component, we plot the point-wise posterior means of $\eta$ along with 95% credible intervals. The mean-squared error for the posterior means at 500 equidistant points on $(0,1)$ is 0.0351. This is depicted in Panel (a) of Figure 3. We also consider the local polynomial regression, DiceKriging method, and the robust Gaussian stochastic process to estimate the nonparametric component based on the LS estimate of $\beta$ as three alternatives. The numerical comparisons are illustrated in Panels (b), (c), and (d) in Figure 3, respectively. The kernel mixture of polynomials outperforms the local polynomial in terms of both the mean-squared error and

| $\beta$ | $\beta_0$ | KMP | 95%-credible intervals | $\hat{\beta}_{LS}$ |
|---------|-----------|-----|------------------------|-----------------|
| $\beta_1$ | 1.0338 | 1.0579 | (0.8949, 1.2150) | 1.1976 |
| $\beta_2$ | 0.1346 | 0.1003 | (-0.0560, 0.2516) | 0.1733 |
| $\beta_3$ | 0.2854 | 0.3481 | (0.1832, 0.5090) | 0.4427 |
| $\beta_4$ | 0.6675 | 0.6449 | (0.5007, 0.7887) | 0.8386 |
| $\beta_5$ | 0.6732 | 0.7212 | (0.5630, 0.8838) | 0.7427 |
| $\beta_6$ | 0.5293 | 0.5433 | (0.3971, 0.6866) | 0.6274 |
| $\beta_7$ | -0.5073 | -0.4759 | (-0.6337, -0.3158) | -0.3464 |
| $\beta_8$ | -3.3942 | -3.3031 | (-3.4492, -3.1450) | -3.5253 |
the uncertainty quantification of the unknown regression function (characterized by the point-wise posterior 95% credible intervals). Both DiceKriging and the robust Gaussian stochastic process fail to detect the nonlinearity of \( \eta_0 \), giving rise to significantly larger mean-squared error of the predictive means. All three alternative competitors yield larger uncertainty (i.e., wider point-wise confidence intervals) in estimating \( \eta \) compared to the kernel mixture of polynomials.

![Graphs of synthetic examples](image)

Figure 3: Synthetic example for the partial linear model. Shaded regions are point-wise 95% credible/confidence intervals. The solid lines are point-wise posterior means/point estimators of \( \eta \) and the dotted lines are \( \eta_0 \). Scatter points are the observations. Panels (b), (c), and (d) are computed using the discrepancy data \((x_i, y_i - z_i^T \hat{\beta}_{LS})^n_{i=1}\), where \( \hat{\beta}_{LS} \) is the least-squared estimate of \( \beta \).

### B Additional technical results

To estimate the prior concentration \( \Pi(p_{f,\sigma} \in B_{KL}(p_0, \epsilon)) \), we need to bound the Kullback-Leibler numbers \( D_{KL}(p_0 \| p_{f,\sigma}) \) and \( E_0\{\log(p_0(x)/p_{f,\sigma}(x))\}^2 \). The following lemma provides upper bounds for these two quantities in terms of \( ||f - f_0||_{L_2(P_x)} \), and hence connects the Kullback-Leibler ball \( B_{KL}(p_0, \epsilon) \) with the \( L_2(P_x) \) neighborhood of \( f_0 \).
Lemma B.1. Suppose \( p_0(x, y) = \phi_{\sigma_0}(y - f_0(x))p_x(x) \) and \( p_f(x, y) = \phi_{\sigma}(y - f(x))p_x(x) \) are two joint densities on \( \mathcal{X} \times \mathbb{R} \), where \( f \) and \( f_0 \) lie in some uniformly bounded function class with \( \|f\|_{\infty}, \|f_0\|_{\infty} < A \). Assume that \( \sigma_0, \sigma \in [\sigma, \overline{\sigma}] \subset (0, \infty) \). Then

\[
\max \left[ D_{KL}(p_0\|p_f, \sigma), E_0 \left\{ \log \frac{p_0(x, y)}{p_f(x, y)} \right\}^2 \right] \lesssim |\sigma_0 - \sigma| + \|f - f_0\|_{L_2(p, \sigma)}^2.
\]

Proof. Directly compute

\[
\log \frac{p_0(x, y)}{p_f(x, y)} = \frac{1}{2\sigma_0^2}(y - f(x))^2 - \frac{1}{2\sigma_0^2}(y - f_0(x))^2 + \frac{1}{2} \log \frac{\sigma^2}{\sigma_0^2}.
\]

Therefore

\[
D_{KL}(p_0\|p_f, \sigma) \leq \frac{1}{2} \left\{ \frac{|\sigma_0^2 - \sigma^2|}{\sigma^2} + \log \frac{\sigma^2}{\sigma_0^2} + \frac{1}{\sigma^2} \|f - f_0\|_{L_2(p, \sigma)}^2 \right\}.
\]

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

\[
E_0 \left\{ \log \frac{p_0(x, y)}{p_f(x, y)} \right\}^2 \leq 2E_0 \left\{ \frac{(y - f(x))^2}{2\sigma^2} - \frac{(y - f_0(x))^2}{2\sigma_0^2} \right\}^2 + 2 \left( \frac{1}{2} \log \frac{\sigma^2}{\sigma_0^2} \right)^2.
\]

Since

\[
E_0 \left\{ \frac{(y - f(x))^2}{2\sigma^2} - \frac{(y - f_0(x))^2}{2\sigma_0^2} \right\}^2 = E_x \left\{ \frac{(f_0 - f)^4 + 6\sigma_0^2(f_0 - f)^2 - 2\sigma^2(f - f_0)^2 + 3(\sigma^2 - \sigma_0^2)^2}{4\sigma^4} \right\} \leq E_x \left\{ \frac{4A^2 + 6\sigma_0 + 2\sigma^2}{4\sigma^4}(f_0 - f)^2 + \frac{6\sigma^4}{\sigma^4}\left|\sigma - \sigma_0\right| \right\},
\]

we conclude that \( E_0 \left\{ \log p_0(x, y)/p_f, \sigma(x, y) \right\}^2 \lesssim |\sigma - \sigma_0| + \|f - f_0\|_{L_2(p, \sigma)}^2 \). \( \square \)

The following lemma demonstrates the approximation power of the kernel mixture of polynomials.

Lemma B.2 (Approximation lemma). Let \( f \) be of the form (4) and \( \mu_k \in \mathcal{X}_K(k) \). Then there exists some constant \( C_1 \) such that for sufficiently small \( \epsilon \) the following holds whenever \( K \geq \epsilon^{-1/\alpha} \),

\[
B^*_K := \left\{ f : \max_{\|s\| = 0, 1, \ldots, \left\lfloor \alpha - 1 \right\rfloor} \left| \xi_{ks} - \frac{D^s f_0(\mu_k^*)}{s_1! \cdots s_p!} \right| \leq \epsilon, k \in [K]^p \right\} \subset \left\{ f : \|f - f_0\|_{L_2(p, \sigma)}^2 < C_1 \epsilon^2 \right\}.
\]

Proof. Suppose \( f \in B^*_K \). Define \( \theta_k(x) = \sum_{s:|s| = 0}^{m} \xi_{ks} \psi_{ks}(x), k \in [K]^p \), where \( \psi_{ks}(x) \)'s are the kernel mixture of polynomial system, and \( \bar{\theta}_k(x) \) to be the Taylor polynomial of \( f_0 \) at \( \mu_k^* \):

\[
\bar{\theta}_k(x) = \sum_{s:|s| = 0}^{m-1} (s_1! \cdots s_p!)^{-1} D^s f_0(\mu_k^*)(x - \mu_k^*)^s. \]

Notice that \( \|(x - \mu_k^*)^s\|_{\infty} \) is bounded uniformly over
\( \mu^*_k, \alpha, k \). By the Taylor’s expansion, for all \( x \in \mathcal{X} \) we have

\[
\left| f_0(x) - \bar{\theta}_k(x) \right| = \left| f_0(x) - \sum_{s:|s|=0}^{[\alpha-1]} \frac{D^s f_0(\mu^*_k)}{s!} (x - \mu^*_k)^s \right| \leq \tilde{C}_1 \|x - \mu^*_k\|^\alpha
\]

for some constant \( \tilde{C}_1 > 0 \). Since we assume that \( f_0 \) satisfies the \( \alpha \)-Hölder condition globally over \( \mathcal{X} \), the constant \( \tilde{C}_1 \) does not depend on \( \mu^*_k \). By the Cauchy-Schwart inequality \( (a + b)^2 \leq 2a^2 + 2b^2 \), we write

\[
\|f - f_0\|^2_{L_2(\mathbb{P}_x)} \leq 2E_x \left\{ \sum_k w_k(x) \left( f_0(x) - \bar{\theta}_k(x) \right) \right\}^2 + 2E_x \left\{ \sum_k w_k(x) \left( \bar{\theta}_k(x) - \theta_k(x) \right) \right\}^2 = 2I_K + 2J_K.
\]

By the Jensen’s inequality, for any \( a > 0 \), we proceed to derive

\[
I_K \leq E_x \left\{ \sum_{k \in [K]^p} w_k(x) \left( f_0(x) - \bar{\theta}_k(x) \right)^2 \right\}
= E_x \left\{ \sum_{k \in [K]^p} w_k(x) \left( f_0(x) - \bar{\theta}_k(x) \right)^2 1(\|x - \mu_k\| > a) \right\}
+ E_x \left\{ \sum_{k \in [K]^p} w_k(x) \left( f_0(x) - \bar{\theta}_k(x) \right)^2 1(\|x - \mu_k\| \leq a) \right\}.
\]

Since \( \|f_0 - \bar{\theta}_k\|_{\infty} \leq A + B \) 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

\[
I_K \lesssim E_x \left\{ \sum_{k \in [K]^p} w_k(x) 1(\|x - \mu_k\|_{\infty} > a) \right\} + E_x \left\{ \sum_{k \in [K]^p} w_k(x) \|x - \mu_k\|^\alpha \|x - \mu_k\|_{\infty} 1(\|x - \mu_k\|_{\infty} \leq a) \right\}
\leq E_x \left\{ \sum_{k \in [K]^p} w_k(x) 1(\|x - \mu_k\|_{\infty} > a) \right\}
+ E_x \left\{ \sum_{k \in [K]^p} w_k(x) (\|x - \mu_k\|_{\infty} + \|\mu_k - \mu_k^*\|_{\infty})^{2\alpha} 1(\|x - \mu_k\|_{\infty} \leq a) \right\}
\leq E_x \left\{ \sum_{k \in [K]^p} w_k(x) 1(\|x - \mu_k\|_{\infty} > a) \right\} + (a + h)^{2\alpha}.
\]

Now pick \( a = h \). Since \( \varphi(x) \leq 1(\|x\| \leq 1) \), then \( w_k(x) 1(\|x - \mu_k\|_{\infty} > a) = 0 \), and hence

\[
E_x \left\{ \sum_{k \in [K]^p} w_k(x) 1(\|x - \mu_k\|_{\infty} > a) \right\} = 0. \quad \text{It follows that } I_K \leq \tilde{C}_1^2 (a + h)^{2\alpha} \lesssim h^{2\alpha} \lesssim \epsilon^2 \text{ when } \epsilon \text{ is sufficiently}
\]

22
small. Similarly by Jensen’s inequality and Cauchy’s inequality \((a + b)^2 \leq 2a^2 + 2b^2\) we write

\[
J_K \leq E_x \left[ \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1} \left( \xi_k - \frac{D^s f_0(\mu^*_k)}{s_1! \ldots s_p!} (x - \mu^*_k)^s \right) + \sum_{s:|s|=\alpha - 1 + 1} \xi_k(x - \mu^*_k)^s \right\}^2 \right]
\]

\[
\leq 2E_x \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1} \left( \xi_k - \frac{D^s f_0(\mu^*_k)}{s_1! \ldots s_p!} (x - \mu^*_k)^s \right) \right\}^2
\]

\[
+ 2E_x \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1 + 1} \xi_k(x - \mu^*_k)^s \right\}^2
\]

\[
\leq 2\epsilon^2 E_x \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1 + 1} \|x - \mu^*_k\|_\infty^2 \right\}
\]

\[
+ 2B^2 E_x \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1 + 1} \|x - \mu^*_k\|_\infty^2 \right\}.
\]

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

\[
B^2 E_x \left[ \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1 + 1} \|x - \mu^*_k\|_\infty^2 \right\} \right]
\]

\[
\leq E_x \left[ \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1 + 1} \left( h + \|\mu_k - \mu^*_k\|_\infty \right)^{|s|} 1(\|x - \mu_k\|_\infty \leq h) \right\}^2 \right]
\]

\[
\leq E_x \left[ \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1 + 1} \left( h + \frac{1}{2K} \right)^{|s|} 1(\|x - \mu_k\|_\infty \leq h) \right\}^2 \right].
\]

Now that \(h + 1/2K \lesssim 1/K, h + 1/2K \leq 1\) for sufficiently large \(K\), and \(|s| \geq \alpha\), it follows that

\[
B^2 E_x \left[ \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1 + 1} \|x - \mu^*_k\|_\infty^2 \right\} \right]
\]

\[
\leq E_x \left[ \sum_{k \in [K]^p} w_k(x) \left\{ \sum_{s:|s|=\alpha - 1 + 1} \left( \frac{1}{K} \right)^{|s|} \right\}^2 \right] \lesssim \left( \frac{1}{K} \right)^{2s} \leq \epsilon^2.
\]

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 \(\epsilon > 0\), \(\|f - f_0\|_{L^2(P_x)}^2 \leq 2I_K + 2J_K \leq C_1 \epsilon^2\). The proof is thus completed. □
The following lemma that quantifies the local averaging behavior of kernel mixing weights \((w_k(x))_{k \in [K]^p}\) plays a fundamental role in estimating the metric entropies of \(F_K\).

**Lemma B.3.** Let \(w_{jk}(x) = \varphi_{h_j}(x - \mu_{jk}) / D_j(x)\), where \(D_j(x) = \sum_{l \in [K]^p} \varphi_{h_l}(x - \mu_{jl})\), \(\mu_{jk} \in \mathcal{X}(k)\), and \(Kh_j \in [h, H]\), \(j = 1, 2\), \(k \in [K]^p\). Then

\[
\|w_{1k}(x) - w_{2k}(x)\|_{\infty} \lesssim \frac{1}{h_1} - \frac{1}{h_2} + \frac{1}{h_2} \max_{i \in [K]^p} \|\mu_{1i} - \mu_{2i}\|_{\infty}.
\]

**Proof.** Let \(\hat{\nu} = \left[\frac{1}{h_1}, \ldots, \frac{1}{h_2}\right]^T \in \mathbb{R}^p\), and \(b = \varphi(\hat{\nu}) > 0\). Suppose \(x \in \mathcal{X}\) is a fixed point. Then there exists a unique \(k_x \in [K]^p\) such that \(x \in \mathcal{X}(k_x)\). Since

\[
\frac{|x - \mu_{jk_x}|}{h_j} \leq \frac{K}{h} \|x - \mu_k^*\|_\infty + \frac{K}{h} \|\mu_{k_x}^* - \mu_{jk_x}\|_\infty \leq \frac{1}{h} = \|\hat{\nu}\|_\infty,
\]

it follows that \(D_j(x) = \sum_{l \in [K]^p} \varphi_{h_j}(x - \mu_{jl}) \geq \varphi_{h_j}(x - \mu_{jk}) \geq \varphi(\hat{\nu}) = b\), since \(\varphi(x)\) decreases as \(\|x\|_\infty\) increases. On the other hand, observe that for non-negative \((u_l, v_l)_{l \in [K]^p}\) with \(\sum_{l \in [K]^p} u_l > 0\) and \(v_k > 0\) for some \(k \in [K]^p\),

\[
\left| \frac{u_k}{\sum_{l \in [K]^p} u_l} - \frac{v_k}{\sum_{l \in [K]^p} v_l} \right| \leq \frac{|u_k - v_k|}{\sum_{l \in [K]^p} u_l} + \frac{1}{\sum_{l \in [K]^p} u_l} \left| \sum_{l \in [K]^p} \frac{u_l}{v_l} - \frac{1}{\sum_{l \in [K]^p} v_l} \right|
\]

\[
\leq \frac{|u_k - v_k|}{\sum_{l \in [K]^p} u_l} + \frac{1}{\sum_{l \in [K]^p} u_l} \left| \sum_{l \in [K]^p} \frac{v_l}{u_l} - \frac{1}{\sum_{l \in [K]^p} v_l} \right|
\]

\[
\leq \frac{|u_k - v_k|}{\sum_{l \in [K]^p} u_l} + \frac{1}{\sum_{l \in [K]^p} u_l} \left| \sum_{l \in [K]^p} \frac{v_l}{u_l} - \frac{1}{\sum_{l \in [K]^p} v_l} \right| \sum_{l \in [K]^p} \left| \frac{v_l}{u_l} - \frac{1}{\sum_{l \in [K]^p} v_l} \right|.
\]

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

\[
|w_{1k}(x) - w_{2k}(x)| \lesssim \sum_{l \in [K]^p} \left| \varphi_{h_1}(x - \mu_{1l}) - \varphi_{h_2}(x - \mu_{2l}) \right|
\]

\[
\leq \frac{1}{b} \sum_{l \in K_1(x) \cup K_2(x)} \left| \varphi\left(\frac{x - \mu_{1l}}{h_1}\right) - \varphi\left(\frac{x - \mu_{2l}}{h_2}\right) \right|
\]

where \(K_j(x) = \{l \in [K]^p : \|x - \mu_{jl}\|_\infty \leq h_j\}, j = 1, 2\). This is because when \(l \notin K_1(x) \cup K_2(x), \varphi_{h_1}(x - \mu_{1l}) = \varphi_{h_2}(x - \mu_{2l}) = 0\). Furthermore when \(w_{1k}(x) = w_{2k}(x) = 0\), the above inequality also holds. Now let \(k_x \in [K]^p\) to be the unique index such that \(x \in \mathcal{X}(k_x)\). We claim that

\[
K_1(x) \cup K_2(x) \subset K(k_x) := \left\{ l \in [K]^p : \|\mu_{k_x}^* - \mu_l^*\|_\infty \leq \frac{2\delta}{K} \right\}.
\]

In fact, if \(l \notin K(k_x)\), then we have \(\|x - \mu_{jl}\|_\infty \geq \|\mu_{k_x}^* - \mu_l^*\|_\infty - \|x - \mu_{k_x}^*\|_\infty - \|\mu_l^* - \mu_{jl}\|_\infty > h_j\) by triangle inequality for \(j = 1, 2\). Namely, \(l \notin K_1(x) \cup K_2(x), j = 1, 2\). Hence \(K_1(x) \cup K_2(x) \subset K(k_x)\). Since \(\varphi\) is continuous and is compactly supported, then by defining \(L_\varphi\) to be the Lipschitz constant of \(\varphi\) we proceed
to compute
\[ |w_{1k}(x) - w_{2k}(x)| \lesssim \sum_{l' \in \mathcal{K}(k)} \left| \varphi \left( \frac{x - \mu_{1l'}}{h_1} \right) - \varphi \left( \frac{x - \mu_{2l'}}{h_2} \right) \right| \]
\[ \leq \sum_{l' \in \mathcal{K}(k)} L_\varphi \max_{l \in [K]^p} \left\| \frac{x - \mu_{1l}}{h_1} - \frac{x - \mu_{2l}}{h_2} \right\|_\infty \]
\[ \leq L_\varphi |\mathcal{K}(k)| \max_{l \in [K]^p} \left\{ (\|x\|_\infty + \|\mu_{1l}\|_\infty) \left| \frac{1}{h_1} - \frac{1}{h_2} \right| + \frac{1}{h_2} \|\mu_{1l} - \mu_{2l}\|_\infty \right\} \]
\[ \lesssim \frac{1}{h_1} - \frac{1}{h_2} + \frac{1}{h_2} \max_{l \in [K]^p} \|\mu_{1l} - \mu_{2l}\|_\infty, \]
where we have used the fact that $|\mathcal{K}(k)| \leq [4k + 2]^p$ for all $x \in \mathcal{X}$, $\|x\|_\infty \leq 1$, and $\|\mu_{1k}\|_\infty \leq 1$. To see why $|\mathcal{K}(k)| \leq [4k + 2]^p$, notice that $\mathcal{K}(k) = \{l \in [K]^p : \|k - l\|_\infty \leq 2k\}$, and hence has cardinality at most $[4k + 2]^p$. Now taking the supremum over $x \in \mathcal{X}$ to the above display completes the proof.

\[ \square \]

C  Remaining proofs

Remaining proofs for section 3

This section contains the proofs of lemma 1, proposition 1, proposition 2, and theorem 2 in the manuscript.

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

\[
\frac{1}{2} \int_{\mathbb{R}} \left\{ \sqrt{\phi_{\sigma_1}(y - \mu_1)} - \sqrt{\phi_{\sigma_2}(y - \mu_2)} \right\}^2 dy
= 1 - \left\{ 1 - \frac{(\sigma_1 - \sigma_2)^2}{\sigma_1^2 + \sigma_2^2} \right\} \frac{1}{4\sigma_1^2 + 4\sigma_2^2} (\mu_1 - \mu_2)^2,
\]
where $\phi_{\sigma}(\cdot)$ is the density of $N(0, \sigma^2)$. Now we derive

\[
H^2(p_f, \sigma_1, p_g, \sigma_2) \leq \int_{\mathcal{X}} \left[ 1 - \left\{ 1 - \frac{(\sigma_1 - \sigma_2)^2}{\sigma_1^2 + \sigma_2^2} \right\} \frac{1}{4\sigma_1^2 + 4\sigma_2^2} (f(x) - g(x))^2 \right] p_x(x) dx
\]
\[ \leq \int_{\mathcal{X}} \left[ 1 - \left\{ 1 - \frac{(\sigma_1 - \sigma_2)^2}{\sigma_1^2 + \sigma_2^2} \right\} \frac{1}{4\sigma_1^2 + 4\sigma_2^2} (f(x) - g(x))^2 \right] p_x(x) dx \]
\[ \leq \int_{\mathcal{X}} \left[ 1 - \left\{ 1 - \frac{(\sigma_1 - \sigma_2)^2}{\sigma_1^2 + \sigma_2^2} \right\} \frac{1}{4\sigma_1^2 + 4\sigma_2^2} \right] + \frac{(f(x) - g(x))^2}{8\sigma^2} p_x(x) dx \]
\[ \leq \frac{(\sigma_1 - \sigma_2)^2}{2\sigma^2} + \frac{1}{8\sigma^2} \int_{\mathcal{X}} (f(x) - g(x))^2 p_x(x) dx \]
\[ \lesssim |\sigma_1 - \sigma_2|^2 + \|f - g\|_{L_1(p_x)}, \]
and hence the right inequality of the first assertion holds. It follows that there exists some constant $C_3 > 0$, 

25
such that for any \( f, g \in \mathcal{G} \), \( \sigma_1, \sigma_2 \in [\sigma, \overline{\sigma}] \), \( 2C_2 H^2 (p_{f, \sigma_1}, p_{g, \sigma_2}) \leq |\sigma_1 - \sigma_2| + \| f - g \|_{L_1(p_{\sigma})} \), and hence, \( \{ p_{g, \sigma'} : \sigma' \in B(\sigma, C_3 \epsilon), g \in B_{\|L_1(p_{\sigma})}\} (f, C_3 \epsilon) \} \subset B_H (p_{f, \sigma}, \epsilon^2) \) for any \( \sigma \in [\sigma, \overline{\sigma}] \) and \( f \in \mathcal{G} \). The second entropy inequality naturally follows. On the other hand,

\[
H^2 (p_{f, \sigma_1}, p_{g, \sigma_2}) \geq \int_X \left[ \frac{(f(x) - g(x))^2}{(8 \sigma^2) + 1} - \frac{(f(x) - g(x))^2}{(8 \sigma^2) + 1} \right] p_x(x) dx \\
\geq \int_X \left[ \frac{(f(x) - g(x))^2}{(8 \sigma^2) + 1} + \frac{(f(x) - g(x))^2}{(8 \sigma^2) + 1} \right] p_x(x) dx \\
\geq \left\{ 2|\sigma_1 - \sigma_2|^2 + \int_X (f(x) - g(x))^2 p_x(x) dx \right\} \\
\geq \| f - g \| _{L_2(p_{\sigma})}^2 + |\sigma_1 - \sigma_2|^2.
\]

Therefore the left inequality of the first assertion holds. \( \square \)

**of proposition 1.** By lemma B.1 there exists some constant \( C'_1 \), such that for sufficiently small \( \epsilon > 0 \), \( B_{\text{KL}}(p_0, \epsilon) \supset \{ p_{f, \sigma} : \| f - f_0 \|_{L_2(p_{\sigma})}^2 \leq C'_1 \epsilon^2, |\sigma - \sigma_0| \leq C'_1 \epsilon \} \). By lemma B.2, we can directly compute

\[
\Pi (B_{\text{KL}}(p_0, \epsilon)) \geq \sum_{K \geq \epsilon^{-1/\alpha}} \pi_K (K) \prod_{k \in [K]^p} \prod_{s: |s| = 0} \Pi \left( \left\{ \xi_{ks} - \frac{D f_0 (\mu^*_k)}{s_1 \ldots s_p} \right\} \leq C''_1 \epsilon |K| \right) \\
\times \Pi (\sigma : |\sigma - \sigma_0| \leq C''_2 \epsilon^2) \\
= \sum_{K \geq \epsilon^{-1/\alpha}} \pi_K (K) \prod_{k \in [K]^p} \prod_{s: |s| = 0} \int_{\min (\sigma, \sigma_0 + C''_1 \epsilon)} \int_{\max (\sigma, \sigma_0 - C''_1 \epsilon)} \pi_\sigma (\sigma) d\sigma \\
\geq \sum_{K \geq \epsilon^{-1/\alpha}} \exp \left\{ -b_0 K^p (\log K)^{r_0} - C_2 (K [\alpha - 1] + K)^p \left( \log \frac{1}{\epsilon} \right) \right\} \\
\geq \exp \left\{ -C_2 \epsilon^{-\frac{2}{\alpha}} \left( \log \frac{1}{\epsilon} \right)^{\max (r_0, 1)} \right\}
\]

for some constants \( C''_1, C_2, C_2 > 0 \). The proof is thus completed. \( \square \)

**of proposition 2.** Let \( f_j (x) = \sum_{k \in [K]^p} \sum_{s: |s| = 0} \xi_k^{(j)} w_{jk} (x) (x - \mu^*_k)^s \) be functions in \( \mathcal{F}_K \), where \( w_{jk} (x) = \varphi_{h_j} (x - \mu_{jk}) / D_j (x) \) and \( D_j (x) = \sum_{i \in [K]^p} \varphi_{h_i} (x - \mu_{ji}) \), \( j = 1, 2 \). Denote \( \theta_{jk} (x) = \sum_{s: |s| = 0} \xi_k^{(j)} w_{jk} (x) (x - \mu^*_k)^s \).
$\mu^*_x$, $j = 1, 2$. We proceed to compute

$$\|f_1 - f_2\|_\infty \leq \max_{x \in \mathcal{X}} \left| \sum_k w_{1k}(x)(\theta_{1k}(x) - \theta_{2k}(x)) \right| + \max_{x \in \mathcal{X}} \left| \sum_k (w_{1k}(x) - w_{2k}(x))\theta_{2k}(x) \right|$$

$$\leq \max_{k \in [K]^p} \max_{0 \leq |\alpha| \leq m} \left| \xi^{(1)}_{ks} - \xi^{(2)}_{ks} \right| \max_{x \in \mathcal{X}} \sum_k w_{1k}(x) \max_{l \in [K]^p} \sum_{s: |s| = 0}^m \| (x - \mu^*_l)^{\alpha} \|_\infty$$

$$+ \max_{x \in \mathcal{X}} \sum_k |w_{1k}(x) - w_{2k}(x)| \sum_{s: |s| = 0}^m \xi^{(2)}_{ks} \| (x - \mu^*_k)^{\alpha} \|_\infty.$$

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

$$\|f_1 - f_2\|_\infty \leq \max_{k \in [K]^p} \max_{0 \leq |\alpha| \leq m} \left| \xi^{(1)}_{ks} - \xi^{(2)}_{ks} \right| + \max_{x \in \mathcal{X}} \sum_k |w_{1k}(x) - w_{2k}(x)|$$

$$\leq \max_{k \in [K]^p} \max_{0 \leq |\alpha| \leq m} \left| \xi^{(1)}_{ks} - \xi^{(2)}_{ks} \right|$$

$$+ \max_{x \in \mathcal{X}} \sum_k |w_{1k}(x) - w_{2k}(x)| \left[ 1(w_{1k}(x) > 0) + 1(w_{2k}(x) > 0) \right]$$

$$\leq \max_{k \in [K]^p} \max_{0 \leq |\alpha| \leq m} \left| \xi^{(1)}_{ks} - \xi^{(2)}_{ks} \right| + \max_{x \in \mathcal{X}} \sum_{j=1}^{2} \left\| \sum_{l \in [K]^p} 1(w_{j1}(x) > 0) \right\|_\infty.$$

For any $x \in \mathcal{X}$, there exists a unique $k_x \in [K]^p$ such that $x \in \mathcal{X}_K(k_x)$. Observe that $\sum_{k \in [K]^p} 1(w_{jk}(x) > 0)$ is the same as the cardinality of the index set $\{ l \in [K]^p : \| x - \mu_l \|_\infty < \epsilon \}$. We now argue that

$$\{ l \in [K]^p : \| x - \mu_l \|_\infty < \epsilon \} \subset \left\{ l \in [K]^p : \| \mu^*_x - \mu^*_l \|_\infty \leq \frac{4\tilde{h}}{K} \right\},$$

where the cardinality of the right-hand side set of the last display is upper bounded by $[8\tilde{h} + 2]^p$. Suppose $l$ is in the complement of the right-hand side of the last display. Then the (reverse) triangle inequality yields $\| x - \mu_l \|_\infty \geq \| \mu^*_x - \mu^*_l \|_\infty - \| x - \mu^*_x \|_\infty - \| \mu^*_l - \mu_l \|_\infty \geq \frac{4\tilde{h}}{K} - 1/(2K) - \frac{\tilde{h}}{K} > \epsilon$, finishing the argument for the claim that $\sum_{k \in [K]^p} 1(w_{jk}(x) > 0)$ can be upper bounded by a constant only depending on \( \tilde{h} \).

Therefore, we obtain

$$\|f_1 - f_2\|_\infty \leq \max_{k \in [K]^p} \max_{0 \leq |\alpha| \leq m} \left| \xi^{(1)}_{ks} - \xi^{(2)}_{ks} \right|$$

$$\leq K \max_{k \in [K]^p} \| \mu_{1k} - \mu_{2k} \|_\infty + K^2|\epsilon_1 - \epsilon_2| + \max_{k \in [K]^p, 0 \leq |\alpha| \leq m} \left| \xi^{(1)}_{ks} - \xi^{(2)}_{ks} \right|,$$

where we have applied lemma B.3 in the last inequality.

We now construct the $\epsilon$-net for $\mathcal{X}_K$. Let $\mathcal{E}_\epsilon$ be an $\epsilon$-net of $[-B, B]$, $\mathcal{E}_\mu_k(\epsilon)$ be an $\epsilon$-net of $\mathcal{X}_K(k)$ with respect to $\| \cdot \|_\infty$, $k \in [K]^p$, and $\mathcal{E}_h(\epsilon)$ be an $\epsilon$-net of $[h/K, \tilde{h}/K]$. Then $|\mathcal{E}_\epsilon(\epsilon)| \leq 3B/\epsilon$, $|\mathcal{E}_\mu_k(\epsilon)| \leq (2/K)^p$, $|\mathcal{E}_h(\epsilon)| \leq (2/\tilde{h})^p$.
and $|\mathcal{E}_h(\epsilon)| \leq (2\bar{h} - 2h)/(K\epsilon)$ for sufficiently small $\epsilon$. We claim that

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

is an $\mathcal{E}_1\epsilon$-net of $\mathcal{F}_K$ with $\| \cdot \|_{L_1(\varphi_\epsilon)}$ for some constant $\tilde{c}_1 > 0$, where $\psi'_{ks}(x) = w'_{ks}(x)(x - \mu'_k)^s$, $w'_k(x) = \varphi'_{h'}(x - \mu'_k)/D'(x)$, and $D'(x) = \sum_{\ell=1}^K \varphi'_{h'}(x - \mu'_\ell)$. In fact, for all $f(x) \in \mathcal{F}_K$ of the form (4), there exist some $h' \in \mathcal{E}_h(\epsilon/K^2)$, $\xi'_{ks} \in \mathcal{E}_{\xi_{ks}}(\epsilon)$, $\mu'_k \in \mathcal{E}_{\mu_k}(\epsilon/K)$ for each $k$ and each $\alpha$, such that $|h - h'| < \epsilon/K^2$, $|\xi_{ks} - \xi'_{ks}| < \epsilon$, $|\mu_k - \mu'_k| < \epsilon/K$. Let $f'(x) = \sum_{k \in [K]^p} \sum_{|s|=0}^m \xi'_k \psi'_k(x)(x - \mu'_k)^s \in \mathcal{F}_K$. It follows by (8) that $\|f - f'\|_\infty \leq C_1\epsilon$. Hence

$$
\mathcal{N}(C_1\epsilon, \mathcal{F}_K, \| \cdot \|_\infty) \leq \| \mathcal{E}_k(\epsilon) \|_{\mathcal{F}_K} \mathcal{N}_{\| \cdot \|_\infty}(\epsilon/K^2)^p \left\{ \mathcal{E}_\mu(\epsilon/K) \right\} \leq \exp \left[ 2K \left\{ (m + 1)^p + p + 1 \right\} \log \frac{1}{\epsilon} \right]
$$

when $\epsilon$ 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 $\epsilon$-net of $\mathcal{F}_K$ with respect to $\| \cdot \|_\infty$ such that $N = \mathcal{N}(\epsilon, \mathcal{F}_K, \| \cdot \|_\infty)$. Then define $l_j(x) = \max(f_j(x) - \epsilon, -A)$ and $u_j(x) = \min(f_j(x) + \epsilon, A)$, yielding the brackets $(|l_j - u_j|)_{j=1}^N$ such that $\mathcal{F}_K \subset \bigcup_{j=1}^N [l_j, u_j]$. Furthermore, $\|l_j - u_j\|_{L_1(\varphi_\epsilon)} \leq \|l_j - u_j\|_\infty \leq 2\epsilon$. The proof is completed by the fact that $\log N = \log \mathcal{N}(\epsilon, \mathcal{F}_K, \| \cdot \|_\infty)$. \hfill \Box

**af theorem 2.** Note that we take the closure of $\mathcal{X}_K(k)$ so that the sieve maximum likelihood estimator exists. Also, when $\sigma_0$ is unknown, the computation of $\hat{f}_K$ is not affected since the maximum likelihood estimator of $\sigma_0$ 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 2 also applies to $\mathcal{G}_K$.

We follow the notation $\mathcal{M}_{\mathcal{G}}^\varphi(\mathcal{G}) = \{ p_{f, \varphi} : f \in \mathcal{G}, \varphi \in [\mathcal{G}, \mathcal{F}] \}$ used in the proof of theorem 1. Denote $\tilde{c}_n = (\log n/n)^{\alpha/(2\alpha + p)}$. We first give an upper bound for the bracketing integral $J_{[\gamma]}(\tilde{c}_n, \mathcal{M}_{\sigma_0}^\varphi(\mathcal{G}_{K_n}), H)$. For convenience denote $\gamma = \alpha/(p + 2\alpha)$. By lemma 1 and proposition 2, we have by simple algebra:

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

Observe the following fact $\lim_{x \to \infty} \int_0^x u^2 e^{-u^2} du/(xe^{-x^2}) = 1/2 < 1$. Then change of variable $\epsilon \mapsto \sqrt{\log(1/\epsilon)}$ yields

$$
\int_0^{\tilde{c}_n} \sqrt{\log \frac{1}{\epsilon}} d\epsilon = 2 \int_0^{\tilde{c}_n} u^2 e^{-u^2} du \leq \tilde{c}_n \sqrt{\log \frac{1}{\tilde{c}_n}} \lesssim \tilde{c}_n \sqrt{\log n}.
$$
Hence $J_{[\epsilon]}(\epsilon_n, M_{\alpha_0}(G_{K_n}), H) \lesssim n^p/(2p+4\alpha) (\log n)^{1/2} p\gamma/(2\alpha) \epsilon_n = \sqrt{n\epsilon_n^2}$. Now define

$$\tilde{f}_n(x) = \arg\min_{f \in \mathcal{G}_{K_n}} D_{KL}(p_0\|p_{f,\sigma_0}), \quad \delta_n = D_{KL}(p_0\|p_{\tilde{f}_n,\sigma_0}), \quad \tau_n = E_0 \left\{ \log \frac{p_0(x,y)}{p_{\tilde{f}_n,\sigma_0}(x,y)} \right\}^2.$$

Direct computation yields that for any function $f$, $D_{KL}(p_0\|p_{f,\sigma_0}) = (2\sigma_0^2)^{-1} \| f - f_0 \|_{L_2(\nu)}^2$. Since $K_n \geq \lceil \epsilon_n^{-1/\alpha} \rceil$, then there exists some $\tilde{g}_n \in \mathcal{G}_{K_n} \cap B_{K_n}^\sigma$ such that

$$\delta_n = D_{KL}(p_0\|p_{f,\sigma_0}) \leq D_{KL}(p_0\|p_{\tilde{g}_n,\sigma_0}) \lesssim \| f - f_0 \|_{L_2(\nu)}^2 \leq \epsilon_n^2,$$

where $B_{K_n}^\sigma$ is defined in lemma B.2. Since $\tau_n \lesssim \| f - f_0 \|_{L_2(\nu)}^2 \lesssim D_{KL}(p_0\|p_{\tilde{f}_n,\sigma_0}) = \delta_n \lesssim \epsilon_n^2$ by lemma B.1, it follows that $\max(\delta_n, \tau_n) \lesssim \epsilon_n^2$. Now we apply theorem 4 (ii) in Wong and Shen (1995) (also see theorem D.2) to conclude that $\Pr_0(\| f_0 - \tilde{f}_{K_n} \|_{L_2(\nu)} > M \epsilon_n) \leq \Pr_0(H(p_0, p_{\tilde{f}_{K_n}}) > M \epsilon_n) \to 0$ for some constant $M > 0$.

\[ \square \]

**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. For any $t > 2\alpha \max(r_0, 1)/(p + 2\alpha) - \max(1 - r_0, 0) > 0$, define

$$\epsilon = \frac{t}{3} - \frac{2\alpha \max(r_0, 1)}{3(p + 2\alpha)} - \frac{\max(1 - r_0, 0)}{3}, \quad \delta = \frac{2\alpha \max(r_0, 1)}{p(p + 2\alpha)} - \frac{r_0 + 2\epsilon}{p}, \quad \gamma = \frac{alpha \max(r_0, 1)}{p + 2\alpha} + \frac{\epsilon}{2}. \quad (10)$$

Then simple algebra shows

$$t > p\delta + 1, \quad p\delta + r_0 > 2\gamma, \quad 2\gamma > \max(\gamma_0, 1) - \frac{p\gamma_0}{2}. \quad (11)$$

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 $\Theta_J = \{ \beta \in \mathbb{R}^q : \| \beta \| \leq J \}$. For any functions $f, g : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, define the $\rho_r$ distance between $f$ and $g$ to be

$$\rho_r(f, g) = \left\{ \int_{\mathcal{X} \times \mathcal{Z}} \| f(x,z) - g(x,z) \|^r \rho(x,z) \, dx \, dz \right\}^{1/r}$$

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 C.1.** Under the setup and prior specification in Section 4.1, there exists some constant $M > 0$ such that $\Pi(H(p_{\beta,\eta}, p_0) > M \epsilon_n | D_n) \to 0$ in $\Pr_0$-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 $B_{KL}(p_0, \epsilon) = \{ p_{\beta,\eta} : D_{KL}(p_0\|p_{\beta,\eta}) < \epsilon^2, E_0 \{ 2\{ \log p_0(x,z,y)/p_{\beta,\eta}(x,z,y) \} \} < \epsilon^2 \}$. Let $K_n = n^{1/(2\alpha + p)} (\log n)\delta, J_n = K_n, M_n = \bigcup_{K=1}^{K_n} M_{nK}$ where $M_{nK} = \mathcal{P}(\mathcal{X}, \Theta_J), \epsilon_n = n^{-\alpha/(2\alpha + p)} (\log n)^{t/2}$.
and \( \xi_n = n^{-\alpha/(p+2\alpha)} \log n \), where \( \delta \) and \( \gamma \) are defined in (10). We complete the proof by verifying (12), (13), and (14), which are originally presented in Kruijer et al. (2010).

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

\[
\Pi(p_{\beta, \eta} \in \mathcal{M}_n^c) \leq \Pi_q(K > K_n) + \Pi_{\beta}(\beta \in \Theta_{j_n}^c)
\]

\[
\leq B_1 \exp \left\{ -b_1 K_n^p \log K_n^p \right\} + \Pi \left( \|\beta\|_2^2 \geq J_n^2 \right)
\]

\[
\leq B_1 \exp \left\{ -b_1 K_n^p \log K_n^p \right\} + \sqrt{2} \exp \left( -\frac{J_n^2}{4} \right)
\]

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

\[
\Pi(p_{\beta, \eta} \in \mathcal{M}_n^c) \leq \exp \left\{ -b_1' n \frac{\log n}{\epsilon_n} (\log n)^{p+\gamma} \right\} \leq \exp(-4n\xi_n^2)
\]

for some constant \( b_1' > 0 \) when \( n \) is sufficiently large, where (11) is applied.

We next verify (14). Simple algebra leads to \( H^2(p_{\beta_1, \eta_1}, p_{\beta_2, \eta_2}) \leq \| \beta_1 - \beta_2 \|^2 + \| \eta_1 - \eta_2 \|_{L_1(\mathcal{F}_K)} \). Hence \( \mathcal{N}(\epsilon, \mathcal{M}_{n,K}, H) \leq \mathcal{N}(C_4 \epsilon^2, \mathcal{F}_K, \| \cdot \|_{L_1(\mathcal{F}_K)}) \times \mathcal{N}(C_4 \epsilon, \Theta_{j_n}, \| \cdot \|) \) for some constant \( C_4 > 0 \). For the finite dimensional \( \Theta_{j_n} \), the \( \epsilon \)-covering number can be easily upper bounded: \( \mathcal{N}(C_4 \epsilon, \Theta_{j_n}, \| \cdot \|) \leq (J_n/\epsilon)^q \). Hence by lemma 1 and proposition 2, for sufficiently large \( K \) and sufficiently small \( \epsilon_n \),

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

\[
\leq \exp(-n\xi_n^2) \sqrt{\mathcal{N}(C_4 \epsilon^2, \Theta_{j_n}, \| \cdot \|)} \sum_{K=1}^{K_n} \sqrt{\mathcal{N}(C_4 \epsilon^2, \mathcal{F}_K, \| \cdot \|_{L_1(\mathcal{F}_K)})}
\]

\[
\leq \exp\left( -n\xi_n^2 + q \log \frac{J_n}{\epsilon_n} \right) \sum_{K=1}^{K_n} \exp \left( K^p (m+1)^p + p + 1 \right) \left( \log \frac{1}{C_4 \epsilon_n^2} \right)
\]

\[
\leq \exp\left( -n\xi_n^2 + C_4' \log n \right) \exp \left\{ 4K_n^p (m+1)^p + p + 1 \right\} \left( \log \frac{1}{\epsilon_n} \right)
\]

\[
\leq \exp \left\{ -n \frac{\log n}{\epsilon_n^2} (\log n)^{q} + C_4'' \frac{\log n}{\epsilon_n^2} (\log n)^{p+1} \right\} \to 0
\]

for some constants \( C_4', C_4'' > 0 \), where (11) is used in the last inequality.

Lastly, we verify (12). Since \( \rho_2^2(\beta^T z + \eta(x), \beta_0^T z + \eta_0(x)) \leq \| \beta - \beta_0 \|^2 \| E z z^T \| + \| \eta - \eta_0 \|^2_{L_2(\mathcal{F}_K)} \), which can be derived by direct computation, then as a consequence of lemma B.2, we see that

\[
B^*_K \times B_2(\beta_0, \epsilon) \subset \{ (\eta, \beta) : \rho_2^2(\beta^T z + \eta(x), \beta_0^T z + \eta_0(x)) < (C_1 + \| E z z^T \|)^2 \}
\]

where \( K = \left\lceil \epsilon^{-1/\alpha} \right\rceil \), and \( B^*_K \) is defined in lemma B.2. Furthermore by simple algebra \( D_{KL}(p_{\theta}||p_{\beta, \eta}) = \rho_2^2(\beta^T z + \eta(x), \beta_0^T z + \eta_0(x))/2 \), and the second moment of the likelihood ratio can be upper bounded using

30
the Cauchy-Schwartz inequality \((a + b)^2 \leq 2a^2 + 2b^2\):

\[
E_0 \left\{ \log \frac{p_0(x, z, y)}{p_{\beta, \eta}(x, z, y)} \right\}^2 = \frac{1}{4} E \left\{ \left( z^T (\beta_0 - \beta) + (\eta_0 - \eta)(x) \right)^2 \left\{ (z^T (\beta_0 - \beta) + (\eta_0 - \eta)(x))^2 + 4 \right\} \right\}
\leq \frac{1}{4} E \left\{ \left( z^T (\beta_0 - \beta) + (\eta_0 - \eta)(x) \right)^2 \left\{ 2 \max_{z \in \mathcal{Z}} ||z||^2 ||\beta - \beta_0||^2 + 8A^2 + 4 \right\} \right\}
\lesssim (||\beta - \beta_0||^2 + 1) \rho_0^2 (\beta^T z + \eta(x), \beta_0^T z + \eta_0(x)).
\]

When \(\beta \in B_2(\beta_0, \epsilon)\) for \(\epsilon \leq 1\), the second moment of the log-likelihood ratio is upper bounded by \(\rho_0^2 (\beta^T z + \eta(x), \beta_0^T z + \eta_0(x))\) up to a multiplicative constant. It follows that \(\{p_{\beta, \eta} : (\eta, \beta) \in B_K^r \times B_2(\beta_0, \epsilon)\} \subset B_{KL}(p_0, C_5\epsilon)\) for some constant \(C_5 > 0\) when \(K = [\epsilon^{-1/\alpha}]\). For the prior mass of \(\beta\) in \(B_2(\beta_0, \epsilon)\), we have

\[
\Pi_{\beta}(B_2(\beta_0, \epsilon)) \geq \frac{\text{Vol}(B_2(0, \epsilon))}{\sqrt{(2\pi)^n}} \exp \left( - \frac{||\beta_0|| + 1}{2} \right) \geq \exp \left\{ -C_6 \left( \log \left( \frac{1}{\epsilon} \right) \right) \right\}
\]

for some constant \(C_6 > 0\) when \(\epsilon < 1\). Hence by proposition 1, for sufficiently small \(\epsilon\),

\[
\Pi (p_{\beta, \eta} \in B_{KL}(p_0, C_6\epsilon)) \geq \Pi_{\eta}(\eta \in B_K^r) \Pi_{\beta}(\beta \in B_2(\beta_0, \epsilon)) \geq \exp \left\{ -2C_2\epsilon^{-\frac{r}{2}} \left( \log \left( \frac{1}{\epsilon} \right) \right)^{\max(r_0, 1)} \right\}.
\]

Substituting \(C_6\epsilon\) by \(\xi_n\), we obtain

\[
\Pi (p_{\beta, \eta} \in B_{KL}(p_0, \xi_n)) \geq \exp \left[ -C_7\xi_n^{-\frac{r}{2}} \left( \log \left( \frac{1}{\xi_n} \right) \right)^{\max(r_0, 1)} \right] \geq \exp(-n\xi_n^2)
\]

for some constant \(C_7 > 0\), where (11) is used in the last inequality.

\(\square\)

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 \(((\beta_1 - \beta_2)^T z + (\eta_1 - \eta_2))^2 \leq 2B^2 ||\beta_1 - \beta_2||^2 + 8A^2\), it follows that

\[
H^2(p_{\beta_1, \eta_1}, p_{\beta_2, \eta_2}) = E_{x,z} \left[ \frac{1}{2} \int_{\mathbb{R}} \left\{ \sqrt{\phi(y - z^T \beta_1 - \eta_1(x))} - \sqrt{\phi(y - z^T \beta_2 - \eta_2(x))} \right\}^2 dy \right]
\geq E_{x,z} \left\{ (\beta_1 - \beta_2)^T z + (\eta_1(x) - \eta_2(x)) \right\}^2 \frac{2B^2 ||\beta_1 - \beta_2||^2 + 8 + 8A^2}{2B^2 ||\beta_1 - \beta_2||^2 + 8 + 8A^2}
\geq \frac{\|\beta_1 - \beta_2\|^2 \lambda_{min}(E_{zz}^T) + \|\eta_1 - \eta_2\|^2_{zz(p_x)}}{2B^2 ||\beta_1 - \beta_2||^2 + 8 + 8A^2}
\geq \frac{\|\beta_1 - \beta_2\|^2 + \|\eta_1 - \eta_2\|^2_{zz(p_x)}}{||\beta_1 - \beta_2||^2 + 1}.
\]
where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a real symmetric matrix. Hence

$$\{(\beta, \eta) : H(p_{\beta, \eta}, p_0) \leq M \epsilon_n\} \subset \left\{ (\beta, \eta) : \frac{\|\beta - \beta_0\|^2 + \|\eta - \eta_0\|^2_{L_2(p_x)}}{\|\beta - \beta_0\|^2 + 1} \leq M_1^2 \epsilon_n^2 \right\}$$

for some large constant $M' > 0$. Since for sufficiently large $n$, $M_1^2 \epsilon_n^2 < 1/2$, we see that

$$\left\{ (\beta, \eta) : \frac{\|\beta - \beta_0\|^2 + \|\eta - \eta_0\|^2_{L_2(p_x)}}{\|\beta - \beta_0\|^2 + 1} \leq M_1^2 \epsilon_n^2 \right\}$$

$$\subset \left\{ (\beta, \eta) : \|\beta - \beta_0\|^2 + \|\eta - \eta_0\|^2_{L_2(p_x)} \leq \frac{1}{2} \|\beta - \beta_0\|^2 + M_1^2 \epsilon_n^2 \right\}$$

$$\subset \left\{ (\beta, \eta) : \frac{1}{2} \|\beta - \beta_0\|^2 + \|\eta - \eta_0\|^2_{L_2(p_x)} \leq M_1^2 \epsilon_n^2 \right\}.$$ 

It follows that \{$(\beta, \eta) : H(p_{\beta, \eta}, p_0) \leq M \epsilon_n$\} $\subset$ \{$(\beta, \eta) : \|\eta - \eta_0\|^2_{L_2(p_x)} \leq M_1^2 \epsilon_n^2$\}. Now applying proposition C.1 completes the proof.

of theorem 4. The proof consists of verifying the conditions D.2, D.3, and D.4 of theorem D.3, which was originally proved in Yang et al. (2015). Let $t > 2 \alpha \max(r_0, 1)/(p+2\alpha) + \max(1-r_0, 0)$, $\epsilon_n = n^{-\alpha/(2\alpha + p)}(\log n)^{t/2}$, $\xi_n = n^{-\alpha(p+2\alpha)}(\log n)^\gamma$, $K_n = \lfloor n^{1/(2\alpha + p)}(\log n)^\delta \rfloor$, and $J_n = K_n^p$, where $\gamma$ and $\delta$ are given by (10).

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

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

where $B_{KL}(p_0, \epsilon) = \{p_{\beta, \eta} : D_{KL}(p_0 || p_{\beta, \eta}) < \epsilon^2, E\{\log p_0(x, z, y) / p_{\beta, \eta}(x, z, y)\}\}$ is the Kullback-Leibler ball. Then lemma 1 in Ghosal et al. (2007) (also see lemma D.1) yields

$$E_0 \left\{ \Pi \left( \|\beta\| \leq J_n, \eta \in \bigcup_{K=1}^{K_n} \mathcal{F}_K \bigg| \mathcal{D}_n \right) \right\} \leq E_0 \left\{ \Pi \left( \eta \in \bigcup_{K=1}^{K_n} \mathcal{F}_K \bigg| \mathcal{D}_n \right) \right\} \to 1.$$

By theorem 3, $\Pi(\eta : \|\eta - \eta_0\|_{L_2(p_x)} \leq M \epsilon_n | \mathcal{D}_n) = 1 - o_{pr}(1)$ for some constant $M > 0$. Let $\hat{\mathcal{F}}_n = \{\eta : \|\eta - \eta_0\|_{L_2(p_x)} \leq M \epsilon_n\} \cap \bigcup_{K=1}^{K_n} \mathcal{F}_K$. It follows that $\Pi(\eta \in \hat{\mathcal{F}}_n | \mathcal{D}_n) = 1 - o_{pr}(1)$. Now we consider bounding $\|\beta - \beta_0\|$. By the proof of theorem 3, there exists some constant $M_1 > 0$ such that for sufficiently large $n$,

$$\Pi \left( \frac{\|\beta - \beta_0\|^2 + \|\eta - \eta_0\|^2_{L_2(p_x)}}{\|\beta - \beta_0\|^2 + 1} \leq M_1^2 \epsilon_n^2 \bigg| \mathcal{D}_n \right) = 1 - o_{pr}(1).$$

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

$$\left\{ (\beta, \eta) : \frac{\|\beta - \beta_0\|^2 + \|\eta - \eta_0\|^2_{L_2(p_x)}}{\|\beta - \beta_0\|^2 + 1} \leq M_1^2 \epsilon_n^2 \right\} \subset \left\{ (\beta, \eta) : \frac{1}{4} \|\beta - \beta_0\|^2 \leq M_1^2 \epsilon_n^2 \right\},$$

32
we conclude that $\Pi (\|\beta - \beta_0\| \leq M \epsilon_n) = 1 - o_{pr_n} (1)$ by replacing $2M_1$ by $M$. Therefore

$$\Pi \left( \|\beta - \beta_0\| \leq M \epsilon_n, \eta \in \tilde{F}_n \mid D_n \right) = 1 - o_{pr_n} (1).$$

We now turn to condition D.4. Since the semiparametric bias $\Delta \eta = 0$ for all $\beta$ as the least favorable submodel coincides with $\eta_0$ for all $\beta \in \mathbb{R}^q$, condition D.4 automatically holds with $\tilde{G}_n = 0$.

Now we are left with the verification of condition D.3. Since the least favorable curve $\eta$ for all $\beta \in \mathbb{R}^q$ is given by $\{ p_{\beta, \eta_0} : \beta \in \mathbb{R}^q \}$. The score function and Fisher information of $p_{\beta, \eta_0}$ at $\beta = \beta_0$ are $\ell_0(x, z, y) = z [y - \eta_0(x) - z^T \beta_0]$ and $I_0 = E_0 \left[ \ell_0(x, z, y) \ell_0(x, z, y)^T \right] = Ezz^T$, respectively. Let $\Delta \beta_n = \beta_n - \beta_0$. Then direct computation yields

$$n \beta_n \log \frac{p_{\beta_n, \eta + \Delta \eta_n}}{p_{\beta_0, \eta}} = n \Delta \beta_n \ell_0 - n \frac{1}{2} n \Delta \beta_n I_0 \Delta \beta_n + \sqrt{n} \Delta \beta_n z (\eta_0 - \eta)$$

$$= n (\beta_n - \beta_0) \ell_0 - n \frac{1}{2} n (\beta_n - \beta_0) I_0 (\beta_n - \beta_0) + \sqrt{n} (\beta_n - \beta_0) z (\eta_0 - \eta) + o_{pr_n} (\sqrt{n} \| \beta_n - \beta_0 \|^2).$$

It follows that

$$\sup_{\eta \in \tilde{F}_n} \left| n \beta_n \log \frac{p_{\beta_n, \eta + \Delta \eta_n}}{p_{\beta_0, \eta}} - n (\beta_n - \beta_0) \ell_0 - n \frac{1}{2} n (\beta_n - \beta_0) I_0 (\beta_n - \beta_0) \right|$$

$$\leq \sup_{\eta \in \tilde{F}_n} \left| \sqrt{n} (\beta_n - \beta_0) z (\eta_0 - \eta) \right| + o_{pr_n} (\sqrt{n} \| \beta_n - \beta_0 \|^2)$$

$$\leq \sqrt{n} \| \beta_n - \beta_0 \| \sum_{j=1}^q \sup_{\eta \in \tilde{F}_n} |G_n z_j (\eta_0 - \eta)| + o_{pr_n} (\sqrt{n} \| \beta_n - \beta_0 \|^2),$$

where $z = [z_1, \ldots, z_q] \in \mathbb{R}^q$. 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 $H_{n,j} = \{ z_j (\eta_0 - \eta)(x) : \eta \in \tilde{F}_n \}$ for $j = 1, \ldots, q$. It follows that $\| z_j (\eta_0 - \eta_1)(x) - z_j (\eta_0 - \eta_2)(x) \|_{L_2 (\mathbb{P}_x, z)} \lesssim \| \eta_1 - \eta_2 \|_{L_2 (\mathbb{P}_x)}$ whenever $\eta_1, \eta_2 \in \tilde{F}_n$.

Namely, we obtain the following metric entropy relation:

$$\log \mathcal{N} \left( \epsilon, H_{n,j}, \| \cdot \|_{L_2 (\mathbb{P}_x, z)} \right) \leq \log \mathcal{N} \left( \tilde{C} \epsilon, \tilde{F}_n, \| \cdot \|_{L_2 (\mathbb{P}_x)} \right) \leq \log \left( \sum_{K=1}^{K_n} \mathcal{N} \left( \tilde{C} \epsilon, \mathcal{F}_K, \| \cdot \|_{L_2 (\mathbb{P}_x)} \right) \right)$$

for some constant $\tilde{C} > 0$. By proposition 2 we obtain for sufficiently large $n$:

$$\sum_{K=1}^{K_n} \mathcal{N} \left( \tilde{C} \epsilon, \mathcal{F}_K, \| \cdot \|_{L_2 (\mathbb{P}_x)} \right) \leq \exp \left[ 5K_n \left( (m + 1)^p + p + 1 \right) \left( \log \frac{1}{\epsilon} \right) \right].$$
Therefore we estimate the bracketing integral
\[ J_{[\cdot]} \left( \epsilon_n, \mathcal{H}_{n_{\cdot}}, \| \cdot \|_{L_2(\mathbb{P}_{x,z})} \right) \leq K^n \int_0^{\epsilon_n} \left( \log \frac{1}{\epsilon} \right)^{\frac{1}{2}} \, d\epsilon \leq K^n \epsilon_n (\log n)^{\frac{1}{2}}, \]
here (9) is applied. Applying (11) yields
\[ J_{[\cdot]} \left( \epsilon_n, \mathcal{H}_{n_{\cdot}}, \| \cdot \|_{L_2(\mathbb{P}_{x,z})} \right) \leq n^{-\frac{\alpha-p/2}{\alpha+p}} (\log n)^{t} = \sqrt{n\epsilon_n^2}. \]
Now we apply lemma 19.36 in van der Vaart (2000) (see theorem D.4). Since \( E \{ z_j(\eta_0 - \eta) \}^2 = \| \eta - \eta_0 \|_{L_2(\mathbb{P}_x)}^2 (E z_j^2) \leq (BM)^2 \epsilon_n^2 \) whenever \( \eta \in \hat{F}_n \), we obtain
\[ E \left( \sup_{\eta \in \hat{F}_n} |G_n z_j(\eta_0 - \eta)| \right) = E\|G_n\|_{\mathcal{H}_{n_{\cdot}}} \leq J_{[\cdot]} \left( \epsilon_n, \mathcal{H}_{n_{\cdot}}, \| \cdot \|_{L_2(\mathbb{P}_{x,z})} \right) \lesssim \sqrt{n\epsilon_n^2}. \]
Markov’s inequality thus implies that \( \sup_{\eta \in \hat{F}_n} |G_n z_j(\eta_0 - \eta)| = O_{pr_0}(\sqrt{n\epsilon_n^2}) \). Together with the previous results, we conclude that condition D.3 is satisfied with \( G_n(s) = n\epsilon_n^2 s + \sqrt{n}\epsilon_n^2 \).

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

### D  Cited theorems and results

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

**Theorem D.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 \( \sigma \)-finite measure over \( \mathcal{Y} \). Equip \( \mathcal{M} \) with the Hellinger distance \( H \) and the Borel \( \sigma \)-field generated by \( H \). Suppose \( \mathcal{M} \) is imposed with a prior distribution \( \Pi \). Let \((y_i)_{i=1}^n\) be i.i.d. according to some density function \( p_0 \). If there exist sequences \((\epsilon_n)_{n=1}^\infty, (\xi_n)_{n=1}^\infty\) with \( \epsilon_n \leq \xi_n \), \( \xi_n \to 0 \) and \( n\epsilon_n^2 \to \infty \), 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

\[
\Pi(B_{KL}(p_0, \xi_n)) \geq \exp(-n\epsilon_n^2), \tag{12}
\]
\[
\Pi(\mathcal{M}_{nm}) \leq \exp(-4n\epsilon_n^2), \tag{13}
\]
\[
\exp(-n\epsilon_n^2) \sum_{m=1}^\infty N(\epsilon_n, \mathcal{M}_{nm}, H) \sqrt{\Pi(\mathcal{M}_{nm})} \to 0, \tag{14}
\]
then \( \Pi (H(p, p_0) > M\epsilon_n \mid y_1, \ldots, y_n) = o_{pr_0}(1) \) for some constant \( M > 0 \).

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

**Lemma D.1.** Assume that the sequence of submodels \((\mathcal{M}_n)_{n=1}^\infty\) in \( \mathcal{M} \) and the sequence \((\xi_n)_{n=1}^\infty\) satisfies (12) and (13) in theorem D.1 with \( \xi_n \to 0 \), \( n\epsilon_n^2 \to \infty \). Then \( E_0 [\Pi(\mathcal{M}_n[y_1, \ldots, y_n])] \to 1. \)
The following classical theorem originally due to Wong and Shen (1995) is used to study the convergence rate of sieve maximum likelihood estimator.

**Theorem D.2.** Let \((y_i)_{i=1}^n\) be independent and identically distributed observations following a distribution \(p_0\) 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 \(\int_0^{\epsilon_n} \sqrt{\log(1/N(\epsilon, \mathcal{P}_n, H))} d\epsilon \lesssim \epsilon_n^{2/3}\). Let \(\hat{\theta}_n = \arg\max_{\theta \in \mathcal{P}_n} \sum_{i=1}^n \log p(y_i)\) be the sieve maximum likelihood estimator on \(\mathcal{P}_n\) and be well-defined, and define \(\delta_n = \inf_{q \in \mathcal{P}_n} D_{KL}(p_0 || q)\), \(\tau_n = \lim_{k \to \infty} E_0 \left(\log p_0(y)/p_k(y)\right)^2\) for some sequence \((q_k)_{k=1}^\infty \subset \mathcal{P}_n\) such that \(D_{KL}(p_0 || q_k) \to \delta_n\). If \(\max(\delta_n, \tau_n) \lesssim \epsilon_n^2\), then there exists some constant \(M > 0\) such that
\[
pr_0 \left( H(p_0, \hat{\theta}_n) > M\epsilon_n \right) \lesssim \exp(-n\epsilon_n^2) + \frac{1}{n}.
\]

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 D.3** (Yang et al. (2015)). Let \(\mathcal{P} = \{p_{0, \eta} : \theta \in \Theta \subset \mathbb{R}^d, \eta \in \mathcal{F}\}\) be a class of density functions with respect to some underlying \(\sigma\)-finite measure over \(\mathcal{Y}\) parametrized on \(\Theta \times \mathcal{F}\), \(\Theta\) is open, and \(\mathcal{F}\) is equipped with metric \(d_H(\eta_1, \eta_2) = H(p_{0, \eta_1}, p_{0, \eta_2})\). Let \((y_i)_{i=1}^n\) be i.i.d. according to \(p_0 = p_{0, \eta_0}\) for some \(\theta_0 \in \Theta\) and \(\eta_0\). Assume that the least-favorable submodel \(\{p_{0, \eta^*_0} : \theta \in \Theta\}\) defined through the least-favorable curve \(\eta^*_0 = \arg\inf_{\eta \in \mathcal{F}} D_{KL}(p_0 || p_{0, \eta})\) exists for all \(\theta \in \Theta\), and denote the semiparametric bias \(\Delta \eta_\theta = \eta^*_0 - \eta_0\). Suppose the following conditions hold:

**Condition D.1.** \(\Theta \times \mathcal{F}\) is endowed with a product prior \(\Pi_\theta \times \Pi_\eta\), and \(\Pi_\theta\) yields a density with respect to the Lebesgue measure on \(\Theta\) that is positive at \(\theta_0\).

**Condition D.2.** There exists a sequence \(\epsilon_n \to 0\) satisfying \(n\epsilon_n^2 \to \infty\) and a sequence of submodels \((\hat{\mathcal{F}}_n)_{n=1}^\infty\) in \(\mathcal{F}\), such that as \(n \to \infty\), \(\Pi(||\theta - \theta_0|| \leq \epsilon_n, \eta \in \hat{\mathcal{F}}_n | y_1, \ldots, y_n) = 1 - o_{pr_0}(1)\).

**Condition D.3.** There exists an increasing function \(G_n : \mathbb{R} \to [0, \infty)\) such that for every sequence \((\theta_n)_{n=1}^\infty\) with \(\theta_n = \theta_0 + o_{pr_0}(1)\),
\[
\sup_{\eta \in \hat{\mathcal{F}}_n} \frac{n \log \frac{p_{0, \eta^*_0} + \Delta \eta_{\theta_n}}{p_{0, \eta}} - n(\theta - \theta_0)^T \mathcal{P}_n \ell_0 + \frac{1}{2} n(\theta - \theta_0)^T I_0 (\theta - \theta_0)} = O_{pr_0}(G_n(||\theta - \theta_0||)),
\]
where \(\hat{\mathcal{F}}_n\) is defined in \(I\), \(\ell_0\) and \(I_0\) are the score and Fisher information matrix of the least-favorable submodel \(\{p_{0, \eta^*_0} : \theta \in \Theta\}\) at \(\theta = \theta_0\).

**Condition D.4.** There exists an increasing function \(\bar{G}_n : \mathbb{R} \to [0, \infty)\), such that for every sequence \((\theta_n)_{n=1}^\infty\) with \(\theta_n = \theta_0 + o_{pr_0}(1)\),
\[
\frac{\int_{\hat{\mathcal{F}}_n} \prod_{i=1}^n p_{0, \eta} (y_i \Pi_\eta (d\eta))}{\int_{\hat{\mathcal{F}}_n} \prod_{i=1}^n p_{0, \eta} (y_i \Pi_\eta (d\eta))} = 1 + O_{pr_0}\left(\bar{G}_n(||\theta - \theta_0||)\right).
\]
If \((G_n + \tilde{G}_n)(n^{-1/2} \log n) = o(1)\), then the sequence of marginal posteriors for \(\theta\) is asymptotically normal in total variation

\[
\sup_F |\Pi(\sqrt{n}(\theta - \theta_0) \in F| y_1, \ldots, y_n) - \Phi(F| \Delta_n, I_0^{-1})| = o_{\text{pr}}(1),
\]

where the supremum is taken over all measurable sets in \(\mathbb{R}^q\), \(\Phi(\cdot| \Delta_n, I_0^{-1})\) is the \(N(\Delta_n, I_0^{-1})\) probability measure, and \(\Delta_n = n^{-1/2} \sum_{i=1}^n I_0^{-1} I_0(y_i)\).

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

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

\[
E_{P_y} \left[ \sup_{f \in \mathcal{F}} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \{f(y_i) - E_{P_y} f(y)\} \right\|_E \right] \lesssim J_{\mathcal{F}}(\delta, I_0, \mathcal{F}, \|\cdot\|_{L_2(P_y)}) \left\{ 1 + \frac{M}{\delta^2 \sqrt{n}} J_{\mathcal{F}}(\delta, \mathcal{F}, \|\cdot\|_{L_2(P_y)}) \right\},
\]

where \(J_{\mathcal{F}}(\delta, I_0, \mathcal{F}, \|\cdot\|_{L_2(P_y)}) = \int_0^\delta \sqrt{\log N(\epsilon, \mathcal{F}, \|\cdot\|_{L_2(P_y)})} d\epsilon\) is the bracketing integral.