**Noname manuscript No.**
(will be inserted by the editor)

# Posterior Contraction Rates of Density Derivative Estimation

**Weining Shen · Subhashis Ghosal**

Received: date / Accepted: date

**Abstract** In this paper we study the problem of Bayesian estimation of derivatives of a density function on the unit interval. We use a finite random series prior based on B-splines and study the asymptotic properties of the posterior distribution under the setting of fixed smoothness of the true function. We obtain the posterior contraction rate under both the $L_2$- and $L_\infty$-distances. The rate under $L_2$-distance agrees with the minimax optimal rate. This result is then extended to the estimation of a multivariate density function on the unit cube and its mixed partial derivatives using tensor product B-splines.

**AMS subject classification** Primary: 62G20, Secondary: 62G05

**Keywords** B-spline · density derivative estimation · nonparametric Bayes · posterior contraction rate · tensor product

## 1 Introduction

Estimating density or regression curves by nonparametric smoothing has a rich literature in both frequentist and Bayesian setting. Some key features of these curves are described by their derivatives. For example, the first order derivative is directly related with the mode of the function. Under an appropriate setting, the accuracy of estimating the mode of a function can be shown to be of the same order of estimating the first derivative. The number of

Weining Shen
Department of Statistics, University of California, Irvine, CA, 92697, USA
Tel.: +1-949-824-9795
E-mail: weinings@uci.edu

Subhashis Ghosal
Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
E-mail: sghosal@ncsu.edu

modes in a density is an important surrogate for the number of mixture components in mixture models with nonparametric components (Donoho, 1988). The mode is also used by mean shift algorithm for clustering and image processing (Fukunaga and Hostetler, 1975; Chacón and Duong, 2013). The second order derivative can be used for testing the density mode (Genovese et al, 2016) and deciding the optimal bandwidth for kernel density estimation (Silverman, 1986). Density derivatives are also closely connected with other fundamental problems in statistics such as regression, Fisher information estimation and hypothesis testing (Singh, 1977a). In higher dimensions, a filament curve or a ridge-line, describes useful geometrical structure on the surface of the function. Estimation of the filament curve has been especially important in recent astronomical applications such as identifying the cosmic web and its relation with dark matter (Qiao and Polonik, 2016).

The kernel method is arguably the most popular approach for density derivative estimation in the literature. Some early work includes Rosenblatt (1956) for univariate density estimation, and Parzen (1962) and Bartlett (1963) for studying the asymptotic properties of these estimators. Results on convergence rates were first obtained by Bhattacharya (1967) and Schuster (1969), and then improved by Farrell (1972) and Singh (1977b). Recently, several authors considered data-driven kernel methods and bandwidth selection for density derivative estimation, and related to the applications in engineering, economy and machine learning (Hall and Yatchew, 2007; Chacón and Duong, 2013; Sasaki et al, 2016). Besides kernels, wavelet is another popular tool for estimating derivatives; see Prakasa Rao (1996) and Hosseinioun et al (2011, 2012) for examples. A spline based method was used in Zhou and Wolfe (2000).

Although convergence theory for Bayesian density estimation has been well developed, e.g., by Ghosal et al (1999); Ghosal and van der Vaart (2001, 2007b); Shen et al (2013) for Dirichlet mixture of normal prior, by Tokdar and Ghosh (2007); van der Vaart and van Zanten (2008, 2009) for Gaussian process based priors and by Rivoirard and Rousseau (2012); de Jonge and van Zanten (2012); Shen and Ghosal (2015) for finite random series prior, convergence theory for Bayesian density derivative estimation has not been addressed despite its importance. A possible reason for this is that posterior contraction rates are typically obtained either by analyzing explicit expression of the posterior characteristics or by applying the general theory of posterior contraction (Ghosal et al, 2000; Ghosal and van der Vaart, 2007a). For density derivatives explicit expressions of posterior characteristics are not available, while the general theory is hard to apply for posterior contraction problem with respect to metrics stronger than the Hellinger distance such as those on the derivatives. The main difficulty lies in constructing uniformly powerful tests against complements of shrinking neighborhoods of the true distribution. Such tests are generally not possible with respect to stronger metrics without excluding parts of the parameter space, and hence the prior plays an even more significant role in the analysis. Some tests with respect to $L_p$- and supremum metrics obtained recently by Giné and Nickl (2011) using empirical process bounds will turn out to be useful for studying posterior contraction rates for density derivatives.

Derivative estimation can be regarded as an inverse problem where the object of interest is obtained by applying an unbounded operator on the parameter regulating the distribution of the observations. Inference in this setting involves regularization which, in the Bayesian context, is induced from the prior. Posterior contraction and coverage properties of credible regions in the inverse problem setting were studied for the white noise model exploiting conjugacy (Knapik et al, 2011; Szabo et al, 2015). Posterior contraction results in the same problem using non-conjugate priors were obtained by Ray (2013) using the Hilbert space structure of the underlying parameter space. Posterior contraction and coverage of credible regions for nonparametric regression functions and its derivatives were treated by Yoo and Ghosal (2016) using conjugacy. Beyond these two models, posterior contraction rate for derivative estimation has not been studied, ostensibly due to the lack of explicit expressions.

In this paper, we study the problem of contraction rate of posterior distributions of density function and its derivatives under both $L_2$- and supremum norm. We consider a prior obtained by putting a prior on the coefficients of a B-spline basis expansion. The same prior has been used for density estimation in de Jonge and van Zanten (2012); Shen and Ghosal (2015), and an analogous prior in Yoo and Ghosal (2016) for the nonparametric regression problem. Such finite random series priors have been identified as very versatile by Shen and Ghosal (2015), as a variety of bases with suitable approximation properties and many different choices of prior distributions on the coefficients can be used. The prior can maintain additional structural properties such as positivity, monotonicity or normalizing to unity very easily if the B-spline basis is used and the coefficients are restricted appropriately. In our context, the primary reason for adopting the finite random B-spline series prior is the availability of explicit expressions for derivatives in terms of lower order B-spline basis, allowing us to bound distances on derivatives in terms of distances on the density function. Then posterior contraction rates with respect to the later can be handled by the general theory of Ghosal et al (2000) with the tests constructed by Giné and Nickl (2011). The rate under the $L_2$-metric agrees with the minimax optimal rate up to logarithmic factors. We consider both univariate and multivariate density functions. For the later case we also consider anisotropic case in the sense that the smoothness levels of the density function may vary at different directions.

In this paper we consider only the non-adaptive setting where the smoothness of the underlying function is known to us. This allows us to use a deterministic number of terms in the finite random series depending on the sample size and the smoothness level. As a result alternative densities in the test construction of Giné and Nickl (2011) are representable as linear combinations of the same basis functions. Unfortunately, the technique of relating distance on derivatives with that on the function does not apply across splines with different bases. In particular, a technical difficulty is that for a series with few basis functions, the error of approximation is too large for the argument to apply. Such a problem does not arise in the direct problem (i.e. estimation of the density function itself) with respect to the Hellinger distance where the test construction does not depend on the approximation and hence adaptation is relatively easily established. It may be noted that even for the direct problem with respect to stronger metrics $L_p$, $2 < p \leq \infty$, adaptive posterior construction rates have not been obtained in the literature so far.

The paper is organized as follows. We introduce the notation in Section 2. In Sections 3 and 4, we present the main results on derivative estimation for univariate and multivariate density functions. We conduct a simulation study in Section 5 and conclude with a discussion in Section 6. The proof of the theorems are given in Section 7.

## 2 Notation

We introduce some commonly used notation in this section. Let $\mathbb{N} = \{0, 1, \ldots\}$ be the collection of natural numbers. For a real number $x$, we use $\lfloor x \rfloor$ and $\lceil x \rceil$ to denote it floor and ceiling numbers, i.e., the largest (smallest) integer less than (greater than) $x$. For two $d$-dimensional vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, we say $\boldsymbol{a} \leq \boldsymbol{b}$ if $a_k \leq b_k$ for every $k = 1, \ldots, d$. In general, we define $\phi(\boldsymbol{a}) = (\phi(a_1), \ldots, \phi(a_d))^T$ for any univariate function $\phi$.

For any matrix $\boldsymbol{A} = (\!(a_{ij})\!) \in \mathbb{R}^{p \times q}$, we define several matrix norms, including $\|\boldsymbol{A}\|_1 = \max_{1 \leq j \leq q} \sum_{i=1}^{p} |a_{ij}|$, $\|\boldsymbol{A}\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^{q} |a_{ij}|$, and $\|\boldsymbol{A}\|_2 = \sigma_{\max}(\boldsymbol{A})$, which is the largest eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$. We use $\lesssim$ to denote an inequality up to a universal constant multiple, and write $a \asymp b$ if $a \lesssim b \lesssim a$.

For a real-valued univariate function $f$, we use $f^{(r)}$ and $D^r f$ to denote its $r$-th derivative. In particular, we define $f^{(0)} = f$. For multivariate function $f : \mathbb{R}^d \to \mathbb{R}$, define its mixed partial derivative by $f^{(\boldsymbol{r})} = D^{(\boldsymbol{r})} f = \partial^{\sum_{k=1}^{d} r_k} / \partial x_1^{r_1} \cdots \partial x_d^{r_d}$ for some $\boldsymbol{r} \in \mathbb{N}^d$. Similarly, we write $f^{(\boldsymbol{0})} = f$. We define an anisotropic Hölder class of functions on $\mathcal{A}$ with smoothness levels $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ as

$$\Big\{ f : \mathcal{A} \to \mathbb{R},\ \|f^{(\boldsymbol{r})}\|_\infty < \infty,\ \text{and } \sum_{k=1}^{d} \|D^{(\alpha_k - r_k)\boldsymbol{e}_k} f^{(\boldsymbol{r})}\|_\infty < \infty, \qquad \forall \boldsymbol{r} \in \mathbb{N}^d, \sum_{k=1}^{d} r_k/\alpha_k < \alpha \Big\},$$

where $\boldsymbol{e}_k \in \mathbb{N}^d$ with $k$-th element of 1 and the rest of 0. We denote the anisotropic $\boldsymbol{\alpha}$-Hölder class on $[0,1]^d$ by $\mathcal{H}^{\boldsymbol{\alpha}}([0,1]^d)$. As a special case, if $\alpha_1 = \cdots = \alpha_d = \alpha$, then we call it the isotropic class and denote it by $\mathcal{C}^\alpha([0,1]^d)$. For univariate functions, $d = 1$ and the smoothness class is denoted by $\mathcal{C}^\alpha([0,1])$.

## 3 Univariate density function

We first focus on the estimation of a univariate density function and its derivatives using the spline method. We assume that the density is supported on a compact interval, which can be taken as the unit interval $[0, 1]$ without loss of generality. For any $x \in [0, 1]$, we denote the collection of B-spline functions with knots $0 = t_0 < t_1 < \cdots < t_N < t_{N+1} = 1$ of order $q$ by $\boldsymbol{b}_{J,q}(x) = (B_{1,q}(x), \ldots, B_{J,q}(x))^T$, where $J = q + N$ is the total number of basis functions that we will be using in the model. For convenience, we use equal-spaced knots, i.e., $t_i = i/(N + 1)$ for $i = 1, \ldots, N$. It is possible to extend the results for quasi-uniform spaced knots (Schumaker, 2007) using the approaches in Yoo and Ghosal (2016) or consider random knots (Belitser and Serra, 2014).

In the nonparametric estimation literature, it is common to approximate a target function of interest $p$ by a linear combination of splines, say $\boldsymbol{b}_{J,q}^T\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a $J$-dimensional coefficient vector. A commonly used density estimation method is called log-spline model (Stone, 1990). The idea is to introduce an exponential link function that ensures the resulting estimator being a valid probability density function. In other words, one writes

$$p = C_{\boldsymbol{\theta}}^{-1} \exp(\boldsymbol{\theta}^T \boldsymbol{b}_{J,q})$$

for a normalizing constant $C_{\boldsymbol{\theta}} > 0$. In some situations, it is also of interest to use other link functions. For example, Shen and Ghosal (2016) also considered the identity link function where the B-splines are replaced by their normalizations and the vector of coefficients $\boldsymbol{\theta}$ lies in the unit simplex. This prior has the advantage that the pointwise posterior mean and variance can be computed without adopting Markov chain Monte Carlo sampling. In this paper, we consider a general link function $\Psi$ and model the density function $p$ by

$$p(x; \boldsymbol{\theta}, J) = \frac{\Psi\left(\boldsymbol{\theta}^T \boldsymbol{b}_{J,q}\right)}{\int_0^1 \Psi\left\{\boldsymbol{\theta}^T \boldsymbol{b}_{J,q}(u)\right\} du}, \tag{1}$$

where $\boldsymbol{\theta}$ is a $J$-dimensional coefficient vector. The link function $\Psi$ is pre-chosen and should be non-negative and continuously differentiable. Given $p$, we can take its $r$-th derivative as

$$D^r p(x; \boldsymbol{\theta}, J) = \frac{D^r \Psi\left(\boldsymbol{\theta}^T \boldsymbol{b}_{J,q}\right)}{\int_0^1 \Psi\left\{\boldsymbol{\theta}^T \boldsymbol{b}_{J,q}(u)\right\} du}.$$

We assume that the smoothness level of $p$ is given to be $\alpha$, and the number of spline basis functions $J$ is chosen deterministically (depending on $n$). Bayesian inference then can be conducted by putting prior distributions on $\boldsymbol{\theta}$ and studying the posterior behavior. A nice property of B-spline basis is that the $r$-th order derivative of $\boldsymbol{\theta}^T \boldsymbol{b}_{J,q}$ can be expressed in terms of its lower-order basis functions as follows,

$$D^r \left(\boldsymbol{b}_{J,q}^T \boldsymbol{\theta}\right) = \boldsymbol{b}_{J,q-r}^T \boldsymbol{W}_r \boldsymbol{\theta}, \tag{2}$$

where $\boldsymbol{W}_r$ is a $(J - r) \times J$ matrix and we call it a derivative matrix. Its mathematical expression is given by Lemma A.2 in Yoo and Ghosal (2016).

Most of the entries in $\boldsymbol{W}_r$ are zero, and the values for non-zero entries are of order $O(J^r)$. This result is crucial for us to establish posterior contraction rate for derivative estimation. More properties of $\boldsymbol{W}_r$ are summarized in Lemma 3.

One key ingredient in the study of posterior convergence property is the approximation ability of the B-spline basis as summarized in the following lemma. Interestingly, the approximation holds simultaneously for the function and its derivatives.

**Lemma 1** *For any function $p \in \mathcal{C}^\alpha[0,1]$ with $\alpha \in (0,q]$, there exist positive constants $L$ and $C$ such that for any positive integer $J$, we can find a $\boldsymbol{\theta}_0 \in [-L,L]^J$ satisfying*

$$\|\boldsymbol{b}_{J,q}^T\boldsymbol{\theta}_0 - p\|_\infty \leq CJ^{-\alpha}, \tag{3}$$

$$\|\boldsymbol{b}_{J,q-r}^T\boldsymbol{W}_r\boldsymbol{\theta}_0 - p^{(r)}\|_\infty \leq CJ^{-\alpha+r}, \tag{4}$$

*for every integer $r \in (0,\alpha)$.*

Lemma 1 asserts that an $\alpha$-smooth function can be uniformly approximated by a B-spline series at the rate $J^{-\alpha}$ with coefficients chosen to lie within a bounded set. The result also assures that we can use the derivatives of the approximating B-spline series to approximate the corresponding derivatives. The proof is omitted since it essentially follows from the results in Chapter 12 of Schumaker (2007).

*Remark 1* In Lemma 1, if $p$ is a strictly positive probability density function and each B-spline $B_{j,q}$, $j = 1,\ldots,J$, is replaced by the corresponding normalized B-spline $B_{j,q}^* = B_{j,q}/\int B_{j,q}$, then $\boldsymbol{\theta}_0$ can be restricted to a compact subset of the open unit $J$-simplex of the form $\{(\theta_1,\ldots,\theta_J) : \theta_j = x_j/\sum_{k=1}^J x_k, L^{-1} \leq x_j \leq L\}$ for some $L > 0$ without compromising the approximation rate $O(J^{-\alpha})$. This follows from Lemma 1(d) of Shen and Ghosal (2015). This allows us to choose the link function $\Psi$ in (1) to be identity if the coefficients are restricted to the unit simplex and an appropriate prior such as a Dirichlet distribution is put on it. Then Theorem 1 below will hold with minor modifications in its proof.

We assume the following conditions on the true density function $p_0$, the link function $\Psi$ and the prior distribution on the coefficient vector $\boldsymbol{\theta}$.

(A1) $\Psi$ is non-negative, strictly monotonic, and $\lceil\alpha\rceil$-times continuous differentiable. Also, $\Psi^{-1}$ is $\lceil\alpha\rceil$-times continuous differentiable on $[\underline{M},\bar{M}]$ for sufficiently small $\underline{M} > 0$ and large $\bar{M}$.
(A2) The true density function $p_0$ satisfies $\Psi^{-1}(p_0) \in \mathcal{C}^\alpha([0,1])$, where $\alpha$ is assumed known and satisfies $0 < \alpha \leq q$.
(A3) $p_0 \geq 2\underline{M}$ on $[0,1]$, where $\underline{M}$ is defined in (A1).
(A4) $J$ is chosen to be of the order $(n/\log n)^{1/(2\alpha+1)}$.
(A5) The prior distribution for $\boldsymbol{\theta}$ on $\mathbb{R}^J$ satisfies the following conditions:

(a) Uniform prior concentration: given any $L > 0$, there exits positive constants $c_1$ depending only on $L$ such that for any $J$, $\boldsymbol{\theta_0} \in [-L, L]^J$ and $\epsilon > 0$, we have that

$$\mathit{\Pi}_\theta(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|_2 \leq \epsilon) \geq \exp\{-c_1 J \log(1/\epsilon)\}, \tag{5}$$

(b) Tail decay: there exist positive constants $c_1', \kappa_1$ such that for any $J$ and all sufficiently large $M > 0$

$$\mathit{\Pi}_\theta(\boldsymbol{\theta} \notin [-M, M]^J) \leq J \exp\{-c_1' M^{\kappa_1}\}. \tag{6}$$

Conditions (A2)–(A5) are commonly used in the literature; see Shen and Ghosal (2015) for example. In (A1), we need the link function $\mathit{\Psi}$ to be differentiable in order to define its derivatives. In addition, we need $\mathit{\Psi}^{-1}$ to be continuously differentiable in the large compact set $[\underline{M}, \bar{M}]$. This is because our approximation result builds on $\left\{\mathit{\Psi}(\boldsymbol{b}^T\boldsymbol{\theta}) - p\right\}$ (and its derivatives), and we would like to bound $\left\{\boldsymbol{b}^T\boldsymbol{\theta} - \mathit{\Psi}^{-1}(p)\right\}$ as well in the proof. This condition is easily satisfied, for example, for the exponential link function. The condition (A5) holds for priors obtained by putting independent positive and continuous densities on each component, such as normal densities with means and variances lying in a fixed compact set. In the case where the normalized B-spline basis is used, $\boldsymbol{\theta}$ is restricted to unit simplexes of appropriate dimensions and the identity link is used, Condition (A5) should be modified to make the constants uniform over compact subsets of the open unit simplexes in the sense described in Remark 1. The analogous condition is then satisfied by Dirichlet priors with all parameters lying in a fixed compact subset of the positive half-line.

Let $\mathit{\Pi}_n(\cdot)$ be a generic notation for the posterior distribution given $n$ i.i.d observations sampling from $p_0$. Then we have the following posterior contraction rate result. The proof is given in Section 7.

**Theorem 1** *Suppose that Conditions (A1)–(A5) hold. Then for every integer $r \in [0, \alpha)$ and any $M_n \to \infty$,*

$$\lim_{n\to\infty} \mathit{\Pi}_n \left[\{p : \|p^{(r)} - p_0^{(r)}\|_2 \leq M_n \epsilon_{n,r},\ \|p^{(r)} - p_0^{(r)}\|_\infty \leq M_n \zeta_{n,r}\}\right] = 1 \quad a.s., \tag{7}$$

*where $\epsilon_{n,r} = (n/\log n)^{(r-\alpha)/(2\alpha+1)}$ and $\zeta_{n,r} = \epsilon_{n,r}(n\epsilon_{n,0}^2)^{1/2}$ are the contraction rate under $L_2$- and $L_\infty$-metrics, respectively.*

Theorem 1 states that the posterior distribution of $p$ and its derivatives $p^{(r)}$ contracts around the underlying true functions $p_0$ and $p_0^{(r)}$ simultaneously by using the same prior distribution. We derive these results using posterior convergence results of Giné and Nickl (2011) under stronger norms. The posterior contraction rate for derivatives under the $L_2$-metric clearly implies the same rate under the $L_1$-metric. This rate agrees (up to a logarithmic factor) with the one obtained in Corollary 3.3.4 of Singh (1977b), in which the kernel method was used, and with the posterior contraction rate obtained in Yoo and Ghosal (2016) for regression derivative estimation. For the convergence rate $\zeta_{n,r}$ under $L_\infty$-metric, there is an additional factor of $(n\epsilon_{n,0}^2)^{1/2}$ compared with the optimal rate obtained by Yoo and Ghosal (2016) for nonparametric regression with Gaussian residuals and conjugate Gaussian prior on the coefficients of the spline basis. However it must be remembered that the model of Yoo and Ghosal (2016) is conjugate for which explicit expressions are available while in our density estimation model these are not available. The same problem also appeared in Giné and Nickl (2011), where they obtained optimal $L_\infty$-rate for conjugate white noise model while possibly suboptimal rate for the non-conjugate density estimation problem. It is not known whether the suboptimality is due to an artifact of the proof or an artifact of the prior itself. Using an wavelet based prior and Bernstein–von Mises theorems in increasing dimension, Castillo (2014) obtained optimal $L_\infty$-posterior convergence rates also for a density estimation problem. It is possible that this technique can improve the posterior $L_\infty$-convergence rate but we do not pursue the approach here.

## 4 Multivariate density functions

Next we consider estimating a multi-dimensional density function $p$ defined on $[0,1]^d$ using the tensor-product of univariate B-splines, which can be viewed as a generalization of univariate B-splines. More specifically, we define a tensor-product B-spline basis function by $B_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{x}) = \prod_{k=1}^d B_{j_k,q_k}(x_k)$, where the indexes $j_k$ take values between 1 and $J_k$ for $k = 1,\ldots,d$, and $\boldsymbol{x} = (x_1,\ldots,x_d) \in [0,1]^d$. Let $\boldsymbol{J} = (J_1,\ldots,J_d)^T$ be the number of basis functions from directions 1 through $d$, and similarly, $\boldsymbol{q} = (q_1,\ldots,q_d)^T$ be the corresponding order of B-spline basis. Here we allow $q_i$ taking different values in different directions to accommodate possibly different smoothness levels in the true density function.

Similarly with the univariate densities, we define a vector of coefficients $\boldsymbol{\theta} \in \mathbb{R}^{\prod_{i=1}^d J_i}$, and consider

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\Psi\left(\boldsymbol{\theta}^T \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}\right)}{\int_{[0,1]^d} \Psi\left\{\boldsymbol{\theta}^T \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{u})\right\} d\boldsymbol{u}}, \tag{8}$$

and its mixed partial derivative function

$$D^{\boldsymbol{r}}\left\{p_{\boldsymbol{\theta}}(\boldsymbol{x})\right\} = \frac{D^{\boldsymbol{r}}\left\{\Psi\left(\boldsymbol{\theta}^T \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}\right)\right\}}{\int_{[0,1]^d} \Psi\left\{\boldsymbol{\theta}^T \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}(\boldsymbol{u})\right\} d\boldsymbol{u}},$$

where $\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}$ is the collection of tensor-product B-splines, $\boldsymbol{r} = (r_1,\ldots,r_d)^T$ and each $r_i$ is an integer between 0 and $\lfloor \alpha_i \rfloor$ for $i = 1,\ldots,d$ such that $\sum_{i=1}^d r_i/\alpha_i < 1$.

The main reason for using tensor-product B-spline basis is that we can still obtain a closed-form expression for its mixed partial derivatives as follows,

$$D^{\boldsymbol{r}}\left(\boldsymbol{\theta}^T \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}\right) = \boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}-\boldsymbol{r}}^T \boldsymbol{W}_{\boldsymbol{r}} \boldsymbol{\theta}, \tag{9}$$

where $\boldsymbol{W}_{\boldsymbol{r}}$ is a $\prod_{i=1}^{d}(J_i - r_i) \times \prod_{i=1}^{d} J_i$ matrix with each entry of order $O(\prod_{i=1}^{d} J_i^{r_i})$. Similarly with the univariate B-spline basis, tensor-product B-splines also enjoy nice approximation properties to smooth functions and their derivatives. In the following lemma, we restate the approximation results in Lemma 7.1 of Yoo and Ghosal (2016) for anisotropic functions and also include isotropic function as a special case.

**Lemma 2** *Let $p$ belong to an isotropic smoothness class $\mathcal{C}^{\alpha}([0,1]^d)$ with $\alpha \leq \min(q_1, \ldots, q_d)$. Then there exist positive constants $L$ and $C$ such that for any $\boldsymbol{J}$, we can find a $\boldsymbol{\theta}_0 \in [-L, L]^{\prod_{k=1}^{d} J_k}$ satisfying*

$$\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}^T \boldsymbol{\theta}_0 - p\|_\infty \leq C \sum_{k=1}^{d} J_k^{-\alpha} \quad \text{and} \quad \|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}-\boldsymbol{r}}^T \boldsymbol{W}_r \boldsymbol{\theta}_0 - p^{(\boldsymbol{r})}\|_\infty \leq C \sum_{k=1}^{d} J_k^{r_k - \alpha}, \tag{10}$$

*for every integer vector $\boldsymbol{r} = (r_1, \ldots, r_d)^T \in \mathbb{N}^d$ satisfying $\sum_{k=1}^{d} r_k < \alpha$. Moreover, if $p$ belongs to an anisotropic class $\mathcal{H}^{\boldsymbol{\alpha}}([0,1]^d)$ with $\alpha_k$ being an integer in $(0, q_k]$ for $k = 1, \ldots, d$, then the above result still holds while (10) shall be replaced by*

$$\|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}}^T \boldsymbol{\theta}_0 - p\|_\infty \leq C \sum_{k=1}^{d} J_k^{-\alpha_k} \quad \text{and} \quad \|\boldsymbol{b}_{\boldsymbol{J},\boldsymbol{q}-\boldsymbol{r}}^T \boldsymbol{W}_r \boldsymbol{\theta}_0 - p^{(\boldsymbol{r})}\|_\infty \leq C \sum_{k=1}^{d} J_k^{r_k - \alpha_k} \tag{11}$$

*for any $\boldsymbol{r} \in \mathbb{N}^d$ satisfying $\sum_{k=1}^{d} r_k/\alpha_k < 1$.*

We make the following assumptions. These assumptions are essentially extensions of (A1)–(A5) under the multivariate setting.

(B1) $\Psi$ is non-negative, strictly monotonic. and $\lceil \alpha \rceil$-times continuous differentiable. Also, $D^{\lceil \alpha \rceil}\left(\Psi^{-1}\right)$ is continuous on $[\underline{M}, \bar{M}]$ for sufficiently small $\underline{M} > 0$ and large $\bar{M}$.
(B2) The true density function $p_0$ satisfies $\Psi^{-1}(p_0) \in \mathcal{C}^{\alpha}\left([0,1]^d\right)$, where $\alpha$ is assumed known and satisfies $0 < \alpha \leq \min(q_1, \ldots, q_d)$.
(B3) $p_0 \geq 2\underline{M}$ on $[0,1]^d$.
(B4) $J_1 = \cdots = J_d$ are of the order $(n/\log n)^{1/(2\alpha+d)}$.
(B5) The $J$-dimensional prior distribution for $\boldsymbol{\theta}$ satisfies (5) and (6) for $J = \prod_{k=1}^{d} J_k$.

Then we have the following theorem on the posterior contraction rate.

**Theorem 2** *Suppose that Conditions (B1)–(B5) hold. Then for every $\boldsymbol{r} \in \mathbb{N}^d$ satisfying $0 \leq \sum_{k=1}^{d} r_k < \alpha$, and any $M_n \to \infty$,*

$$\lim_{n\to\infty} \Pi_n \left[\{p : \|p^{(\boldsymbol{r})} - p_0^{(\boldsymbol{r})}\|_2 \leq M_n \epsilon_{n,\boldsymbol{r}},\ \|p^{(\boldsymbol{r})} - p_0^{(\boldsymbol{r})}\|_\infty \leq M_n \zeta_{n,\boldsymbol{r}}\}\right] = 1 \quad a.s.,$$

*where $\epsilon_{n,\boldsymbol{r}} = (n/\log n)^{(-\alpha + \sum_{k=1}^{d} r_k)/(2\alpha+d)}$ and $\zeta_{n,r} = \epsilon_{n,\boldsymbol{r}}(n\epsilon_{n,\boldsymbol{0}}^2)^{1/2}$ are the contraction rate under $L_2$- and $L_\infty$-metrics, respectively.*

As in Remark 1, if tensor products of normalized B-splines are used, the coefficient vector may be restricted to the unit $J$-simplex and the link function $\Psi$ may be taken to be identity. For a given strictly positive smooth multivariate probability density function, $\boldsymbol{\theta}_0$ appearing in Lemma 2 will then lie in compact subsets of the respective open unit $J$-simplexes in the sense described in Remark 1. Then the constants in Condition (B5) should be uniform over compact subsets of the open unit simplexes, which for instance, will hold for Dirichlet priors with all parameters from a fixed compact subset of the positive half-line.

Theorem 2 extends the previous result for univariate density derivative estimation to the multivariate setting. By taking $d = 1$, we obtain the rate in Theorem 1. Here we choose $J_1, \ldots, J_d$ to have the same order since the true density is assumed to be isotropic. However, in practice, the smoothness levels across different directions may actually differ. This motivates us to consider extensions for anisotropic density functions. The following set of conditions are based on slight modifications of (B1), (B2) and (B4).

(C1) $\Psi$ is non-negative, strictly monotonic, $\lceil\alpha\rceil$-times continuous differentiable and $D^{\lceil\boldsymbol{\alpha}\rceil}\left(\Psi^{-1}\right)$ is continuous on $[\underline{M}, \bar{M}]$ for sufficiently small $\underline{M} > 0$ and large $\bar{M}$.
(C2) $\Psi^{-1}(p_0) \in \mathcal{H}^{\boldsymbol{\alpha}}\left([0,1]^d\right)$, where $\boldsymbol{\alpha} \in \mathbb{N}^d$ is assumed known and satisfies $\mathbf{0} < \boldsymbol{\alpha} \leq \boldsymbol{q}$.
(C3) $J_k$ is of the order $\left((n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}\right)$ for $k = 1, \ldots, d$, where $\alpha^*$ is the harmonic mean of $(\alpha_1, \ldots, \alpha_d)$ defined by $1/\alpha^* = \left\{\sum_{i=1}^d \alpha_i^{-1}\right\}/d$.

Then we obtain the posterior contraction rate for anisotropic density derivative estimation.

**Theorem 3** *Suppose that Conditions (B3), (B5), and (C1)–(C3) hold. Then for every* $\boldsymbol{r} \in \mathbb{N}^d$ *satisfying* $0 \leq \sum_{k=1}^d r_k/\alpha_k < 1$, *and any* $M_n \to \infty$,

$$\lim_{n\to\infty} \Pi_n \left[\{p : \|p^{(\boldsymbol{r})} - p_0^{(\boldsymbol{r})}\|_2 \leq M_n \epsilon_{n,\boldsymbol{r}},\ \|p^{(\boldsymbol{r})} - p_0^{(\boldsymbol{r})}\|_\infty \leq M_n \zeta_{n,\boldsymbol{r}}\}\right] = 1 \quad a.s.,$$

*where* $\epsilon_{n,\boldsymbol{r}} = (n/\log n)^{-\alpha^*(1-\sum_{k=1}^d r_k/\alpha_k)/(2\alpha^*+d)}$ *and* $\zeta_{n,r} = \epsilon_{n,\boldsymbol{r}}(n\epsilon_{n,\mathbf{0}}^2)^{1/2}$ *are the contraction rates under* $L_2$*- and* $L_\infty$*-metrics, respectively.*

In Theorem 3, if $\alpha_1 = \cdots = \alpha_d$, then we obtain Theorem 2 as a special case. The rate under $L_2$-metric agrees with the one obtained by Yoo and Ghosal (2016) for regression derivatives. Note that for anisotropic case, we will need the smoothness levels $\boldsymbol{\alpha}$ to be integers. This is due to the limitation of the approximation result in Lemma 2.

## 5 Simulation studies

We conduct a simulation study to compare the numerical performance of the proposed method with kernel method implemented using R function “locpoly”.

**Table 1** Mean $L_1$-distance for density derivative estimation using random series prior (RS) and kernel smoothing method (Kernel) for sample sizes $n = 100$, 200, 800.

| | | density | | | 1st derivative | | | 2nd derivative | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Density | Method | 100 | 400 | 800 | 100 | 400 | 800 | 100 | 400 | 800 |
| N(0,$16^{-1}$) | RS($J = 8$) | .073 | .067 | .052 | .150 | .151 | .135 | 2.01 | 1.69 | 1.46 |
| | RS($J = 14$) | .030 | .023 | .022 | .129 | .125 | .101 | 1.65 | 1.64 | 1.13 |
| | Kernel | .017 | .016 | .014 | .138 | .127 | .091 | 1.72 | 1.32 | .99 |
| Beta$(2, 3)$ | RS($J = 8$) | .091 | .076 | .069 | 1.41 | 1.00 | .93 | 12.0 | 7.31 | 4.23 |
| | RS($J = 14$) | .075 | .063 | .048 | 1.07 | .81 | .76 | 7.12 | 4.85 | 2.28 |
| | Kernel | .086 | .051 | .037 | 2.13 | .94 | .64 | 21.6 | 11.4 | 3.19 |

We consider the random series prior for $J = 8$ and $J = 14$ with $q = 4$. To simplify the posterior computation, we use the identity link function and re-scale the B-spline basis such that the normalizing constant in (1) can be removed. Similar method has been used in Shen and Ghosal (2015). We generate data from a normal distribution $N(0, 1/16)$ (truncated between $-2$ and 2) and a beta distribution Beta$(2, 3)$. We vary the sample size $n = 100, 400, 800$ and repeat the procedure for 200 Monte-Carlo replications. We summarize the mean $L_1$-distance of both methods for the density functions and their first two order derivatives in Table 1. It can be seen that increasing the number of terms from $J = 8$ to 14 leads to a significant improvement in estimation accuracy. In general, kernel method and random series spline method have a comparable performance. The estimation error becomes larger as the smoothness of the target function (derivative) decreases.

## 6 Discussion

In this paper, we studied posterior contraction rates of derivatives of a probability density function under different distance metrics. A fundamental difficulty in treating derivatives is that neither explicit posterior expressions are available nor it is easy to construct tests as required by the general theory of posterior contraction. We construct priors using a random series based on B-splines. In the prior distribution, we fix the number of basis terms $J$ to be at the oracle order $n^{1/(2\alpha+d)}$. A key feature of B-splines allow us to relate distance on derivatives with those on the functions, and tests can be constructed using some sharp empirical process bounds. While a spline-based finite random series prior possesses very helpful structural properties useful for a theoretical study of posterior concentration of derivatives, it seems reasonable to believe that other popular choices such a Dirichlet process mixture prior will also lead to good concentration properties of the derivatives. Also, the $L_\infty$-contraction rate established in the paper is possibly sub-optimal. A recent technique for establishing optimal supremum-norm posterior concentration rate based on high-dimensional Bernstein–von Mises theorem may also be applicable in studying posterior contraction of derivatives.

A very important issue in studying posterior contraction rate is Bayesian adaptation. The smoothness level $\alpha$ which guides the contraction rates for the number of terms in the finite random series is typically not known in practice. A fairly rich theory of Bayesian adaptation lets us derive (nearly) the oracle rate of contraction with respect to the Hellinger distance on the density function without knowing the smoothness by simply putting a prior on the number of terms in the B-spline random series. It will be of interest to develop adaptive procedures that enjoy nice convergence properties for a continuous range of $\alpha$. In particular it seems reasonable to expect that the posterior contraction rate for derivatives based on the B-spline random series prior will automatically adapt to the unknown smoothness especially if the $L_2$-distance is used. We shall return to these topics elsewhere.

## 7 Proofs

We first re-state some useful results in Yoo and Ghosal (2016) about the derivative matrix.

**Lemma 3** *For the univariate density derivate estimation, the derivative matrix $\boldsymbol{W}_r$ satisfies*

*(a) Each row and column of $\boldsymbol{W}_r$ only has $(r+1)$ non-zero entries, whose values are of the order $J^r$.*
*(b) $\|\boldsymbol{W}_r^T\boldsymbol{W}_r\|_2 \leq J^{2r}$.*

*For multivariate density derivative estimation, the derivative matrix $\boldsymbol{W}_{\boldsymbol{r}}$ satisfies*

*(c) Each row and column of $\boldsymbol{W}_{\boldsymbol{r}}$ only has $\prod_{k=1}^d(r_k+1)$ non-zero entries, whose values are of order $O(\prod_{k=1}^d J_k^{r_k})$.*
*(d) $\|\boldsymbol{W}_{\boldsymbol{r}}^T\boldsymbol{W}_{\boldsymbol{r}}\|_2 \leq \prod_{k=1}^d J_k^{2r_k}$.*

*Proof (Theorem 1)* Throughout the proof, we add a subscript $n$ to $J$ to indicate its dependence on $n$. We first show that the result holds for $r = 0$, and then extend the results for positive values of $r$. When $r = 0$, the $L_1$-convergence is immediate by using Theorem 2.1 of Ghosal and van der Vaart (2001) with the choice of sieve as $\mathcal{F}_n = \{\boldsymbol{\theta} \in (-n^{1/\kappa_1}, n^{1/\kappa_1})^{J_n}\}$. The rest of the arguments proceed similarly with Corollary 1 in Shen and Ghosal (2015). The only difference is that we consider a known smoothness $\alpha$ and choose $J_n$ at the oracle order already, while Shen and Ghosal (2015) considered an unknown smoothness level $\alpha$ and a prior on $J_n$. The rates under $L_2$- and $L_\infty$-metrics are obtained by Theorem 3 of Giné and Nickl (2011). Note that we have assumed boundedness of $p_0$, so condition (3) in that theorem will hold trivially.

Next we consider $r > 0$. For simplicity, we drop the subscripts in $\boldsymbol{b}_{J,q}$ and denote it by $\boldsymbol{b}$ for the rest of the proof. By Lemma 1, we can find a

$J_n$-dimensional vector $\boldsymbol{\theta}_0$ and a constant $C > 0$ such that

$$\begin{aligned} &\|\boldsymbol{b}^T\boldsymbol{\theta}_0 - \Psi^{-1}(p_0)\|_\infty \le CJ_n^{-\alpha}, \\ &\|\boldsymbol{b}^T\boldsymbol{W}_i\boldsymbol{\theta}_0 - D^i\{\Psi^{-1}(p_0)\}\|_\infty \le CJ_n^{i-\alpha}, \quad i = 1, \ldots, r. \end{aligned} \tag{12}$$

By Faà di Bruno's formula, we have

$$D^r\Psi(g(x)) = \sum \frac{r!}{m_1!m_2!\cdots m_r!}\Psi^{m_1+\cdots+m_r}(g(x)) \cdot \prod_{i=1}^{r}\left(\frac{g^{(i)}(x)}{i!}\right)^{m_i},$$

where the sum is over all $r$-tuples of non-negative integers $(m_1, \ldots, m_r)$ with $\sum_{i=1}^r m_i = r$. Therefore we can expand $D^r\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\right\}$ and $D^r\left\{\Psi(\Psi^{-1}(p_0))\right\}$ accordingly, and bound their difference by a constant multiple of $J_n^{r-\alpha}$, i.e.,

$$\|D^r\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\right\} - p_0^{(r)}\|_\infty = \|D^r\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\right\} - D^r\left\{\Psi(\Psi^{-1}(p_0))\right\}\|_\infty \lesssim J_n^{r-\alpha}, \tag{13}$$

since the derivatives are all bounded from above, and the worst approximation comes from $\|\boldsymbol{b}^T\boldsymbol{W}_i\boldsymbol{\theta}_0 - D^i\{\Psi^{-1}(p_0)\}\|_\infty$, which can be bounded by using (12).

Now we define the following sets for arbitrary large constant $M$:

$$\begin{aligned} A_n(M) = \Big\{\boldsymbol{\theta} :& \|p_0 - c(\boldsymbol{\theta})^{-1}\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\|_2 \le MJ_n^{-\alpha}, \\ & \|p_0 - c(\boldsymbol{\theta})^{-1}\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\|_\infty \le MJ_n^{-\alpha}(nJ_n^{-2\alpha})^{1/2}\Big\} \\ B_n(M) = \Big\{\boldsymbol{\theta} :& \|p_0 - \Psi(\boldsymbol{b}^T\boldsymbol{\theta})\|_2 \le MJ_n^{-\alpha}, \\ & \|p_0 - \Psi(\boldsymbol{b}^T\boldsymbol{\theta})\|_\infty \le MJ_n^{-\alpha}(nJ_n^{-2\alpha})^{1/2}\Big\} \\ C_n(M) = \Big\{\boldsymbol{\theta} :& \|\Psi(\boldsymbol{b}^T\boldsymbol{\theta}) - \Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\|_2 \le MJ_n^{-\alpha}, \\ & \|\Psi(\boldsymbol{b}^T\boldsymbol{\theta}) - \Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\|_\infty \le MJ_n^{-\alpha}(nJ_n^{-2\alpha})^{1/2}\Big\} \\ D_n(M) = \Big\{\boldsymbol{\theta} :& \|D^r\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\right\} - D^r\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\right\}\|_2 \le MJ_n^{r-\alpha}, \\ & \|D^r\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\right\} - D^r\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\right\}\|_\infty \le MJ_n^{r-\alpha}(nJ_n^{-2\alpha})^{1/2}\Big\}, \\ E_n(M) = \Big\{\boldsymbol{\theta} :& \|p_0^{(r)} - c(\boldsymbol{\theta})^{-1}D^r\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\right\}\|_2 \le MJ_n^{r-\alpha}, \\ & \|p_0 - D^r\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\right\}\|_\infty \le MJ_n^{r-\alpha}(nJ_n^{-2\alpha})^{1/2}\Big\}. \end{aligned}$$

Given a large constant $M_1$, the posterior probability of $A_n(M_1)$ goes to one. By integration, $c(\boldsymbol{\theta}) = 1 + O_p(J_n^{-\alpha})$. Hence we can find $M_2 > M_1$ such that $A_n(M_1) \subset B_n(M_2)$. By the $L_\infty$-approximation ability of B-splines to $p_0$, we can replace $p_0$ in $B_n$ by $\boldsymbol{b}^T\boldsymbol{\theta}$ and obtain $C_n$. In other words, $B_n(M_2) \subset C_n(M_3)$ for some large $M_3$. From $C_n$ to $D_n$, note that $\Psi(\boldsymbol{b}^T\boldsymbol{\theta})$ is finite and sufficiently close to $p$, so it is lower bounded by $\underline{M}$, hence $\Psi^{-1}$ is continuously differentiable by Condition (A1). Let

$$C_n'(M) = \left\{\boldsymbol{\theta} : \|\boldsymbol{b}^T\boldsymbol{\theta} - \boldsymbol{b}^T\boldsymbol{\theta}_0\|_2 \le MJ_n^{-\alpha}, \|\boldsymbol{b}^T\boldsymbol{\theta} - \boldsymbol{b}^T\boldsymbol{\theta}_0\|_\infty \le MJ_n^{-\alpha}(nJ_n^{-2\alpha})^{1/2}\right\}.$$

Then we have $C_n(M_3) \subset C'_n(M'_3)$ for some constant $M'_3$. By Lemma (A.8) of Yoo and Ghosal (2016), $\|\boldsymbol{b}^T\boldsymbol{\theta}\|_2 \asymp J_n^{-1/2}\|\boldsymbol{\theta}\|_2$ for $J_n$-dimensional vector $\boldsymbol{\theta}$. Therefore for any $\boldsymbol{\theta} \in C'_n$, we have $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \le M J_n^{1/2-\alpha}$. By using part (b) of Lemma 3, we have $\|\boldsymbol{W}_i(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_2 \lesssim J_n^{i+1/2-\alpha}$ and $\|\boldsymbol{b}^T\boldsymbol{W}_i(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_2 \lesssim J_n^{i-\alpha}$ for $i = 1, \ldots, r$. Similarly, we also have

$$\|\boldsymbol{b}^T\boldsymbol{W}_i(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_\infty \le \|\boldsymbol{b}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_\infty \|\boldsymbol{W}_i\|_1 \lesssim J_n^{i-\alpha}(nJ_n^{-2\alpha})^{1/2}$$

by part (a) of Lemma 3. Using Faà di Bruno's formula and the assumption that $\Psi$ is continuously differentiable, we have $C'_n(M'_3) \subset D_n(M_4)$. From $D_n$ to $E_n$, we use the triangle inequality combining (13) and the fact that $c(\boldsymbol{\theta})^{-1} = 1 + O_p(J_n^{-\alpha})$. We have $D_n(M_4) \subset E_n(M_5)$ for any sufficiently large $M_5$, hence $\Pi_n(E_n(M_n)) \to 1$ for any $M_n \to \infty$. This concludes the proof.

*Proof (Theorem 2 and Theorem 3)* We only prove Theorem 3 since it includes Theorem 2 as a special case. The posterior contraction rate for densities, i.e., $\boldsymbol{r} = \boldsymbol{0}$, can be obtained similarly with Section 3.2 in Shen and Ghosal (2015) by fixing $J_n$ at the oracle order. Hence it is good enough to show the results for non-zero $\boldsymbol{r}$. By Lemma 2, we can find a $\prod_{k=1}^d J_k$-dimensional vector $\boldsymbol{\theta}_0$ and a constant $C > 0$ such that

$$\begin{aligned} \|\boldsymbol{b}^T\boldsymbol{\theta}_0 - \Psi^{-1}(p_0)\|_\infty \le C\epsilon_{n,\boldsymbol{0}}, \quad & \|\boldsymbol{b}^T\boldsymbol{W}_{\boldsymbol{i}}\boldsymbol{\theta}_0 - D^{\boldsymbol{i}}\{\Psi^{-1}(p_0)\}\|_\infty \\ & \le C\sum_{k=1}^d J_k^{i_k-\alpha_k}, \qquad \boldsymbol{i} \le \boldsymbol{r}. \end{aligned} \tag{14}$$

Then by Faà di Bruno's formula, we have

$$\begin{aligned} \|D^{\boldsymbol{r}}\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\right\} - p_0^{(\boldsymbol{r})}\|_\infty = & \|D^{\boldsymbol{r}}\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\right\} - D^{\boldsymbol{r}}\left\{\Psi(\Psi^{-1}(p_0))\right\}\|_\infty \\ \lesssim & \sum_{k=1}^d J_k^{r_k-\alpha_k}. \end{aligned} \tag{15}$$

Similarly with the proof of Theorem 1, we can define the following sets for any large constant $M$,

$$\begin{aligned}
A_n(M) = \Big\{\boldsymbol{\theta} :& \|p_0 - c(\boldsymbol{\theta})^{-1}\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\|_2 \leq M\epsilon_{n,\mathbf{0}}, \\
& \|p_0 - c(\boldsymbol{\theta})^{-1}\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\|_\infty \leq M\epsilon_{n,\mathbf{0}}(n\epsilon_{n,\mathbf{0}}^2)^{1/2}\Big\} \\
B_n(M) = \Big\{\boldsymbol{\theta} :& \|p_0 - \Psi(\boldsymbol{b}^T\boldsymbol{\theta})\|_2 \leq M\epsilon_{n,\mathbf{0}}, \\
& \|p_0 - \Psi(\boldsymbol{b}^T\boldsymbol{\theta})\|_\infty \leq M\epsilon_{n,\mathbf{0}}(n\epsilon_{n,\mathbf{0}}^2)^{1/2}\Big\} \\
C_n(M) = \Big\{\boldsymbol{\theta} :& \|\Psi(\boldsymbol{b}^T\boldsymbol{\theta}) - \Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\|_2 \leq M\epsilon_{n,\mathbf{0}}, \\
& \|\Psi(\boldsymbol{b}^T\boldsymbol{\theta}) - \Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\|_\infty \leq M\epsilon_{n,\mathbf{0}}(n\epsilon_{n,\mathbf{0}}^2)^{1/2}\Big\} \\
D_n(M) = \Big\{\boldsymbol{\theta} :& \|D^{\boldsymbol{r}}\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\right\} - D^{\boldsymbol{r}}\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\right\}\|_2 \leq M(\prod_{k=1}^d J_k^{r_k})\epsilon_{n,\mathbf{0}}, \\
& \|D^{\boldsymbol{r}}\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\right\} - D^{\boldsymbol{r}}\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta}_0)\right\}\|_\infty \leq M(\prod_{k=1}^d J_k^{r_k})\epsilon_{n,\mathbf{0}}(n\epsilon_{n,\mathbf{0}}^2)^{1/2}\Big\}, \\
E_n(M) = \Big\{\boldsymbol{\theta} :& \|p_0^{(\boldsymbol{r})} - c(\boldsymbol{\theta})^{-1}D^{\boldsymbol{r}}\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\right\}\|_2 \leq M(\prod_{k=1}^d J_k^{r_k})\epsilon_{n,\mathbf{0}}, \\
& \|p_0 - D^{\boldsymbol{r}}\left\{\Psi(\boldsymbol{b}^T\boldsymbol{\theta})\right\}\|_\infty \leq M(\prod_{k=1}^d J_k^{r_k})\epsilon_{n,\mathbf{0}}(n\epsilon_{n,\mathbf{0}}^2)^{1/2}\Big\}.
\end{aligned}$$

Now for some positive constants $M_1, \ldots, M_5$, we have $A_n(M_1) \subset B_n(M_2) \subset C_n(M_3)$ since $c(\boldsymbol{\theta}) = 1 + O_p(\epsilon_{n,\mathbf{0}})$ and the approximation result in (14). To get $C_n(M_3) \subset D_n(M_4)$, note that $\Psi(\boldsymbol{b}^T\boldsymbol{\theta})$ is finite and sufficiently close to $p$, so it is lower bounded by $\underline{M}$, hence $\Psi^{-1}$ is continuously differentiable by Condition (A1). Let

$$C_n'(M) = \left\{\boldsymbol{\theta} : \|\boldsymbol{b}^T\boldsymbol{\theta} - \boldsymbol{b}^T\boldsymbol{\theta}_0\|_2 \leq M\epsilon_{n,\mathbf{0}}, \|\boldsymbol{b}^T\boldsymbol{\theta} - \boldsymbol{b}^T\boldsymbol{\theta}_0\|_\infty \leq M\epsilon_{n,\mathbf{0}}(n\epsilon_{n,\mathbf{0}}^2)^{1/2}\right\}.$$

Then we have $C_n(M_3) \subset C_n'(M_3')$ for some constant $M_3'$. By Lemma (A.8) of Yoo and Ghosal (2016), $\|\boldsymbol{b}^T\boldsymbol{\theta}\|_2 \asymp \left(\prod_{k=1}^d J_k^{-1/2}\right)\|\boldsymbol{\theta}\|_2$ for $J_n$-dimensional vector $\boldsymbol{\theta}$. Therefore for any $\boldsymbol{\theta} \in C_n'$, we have $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq M\left(\prod_{k=1}^d J_k^{1/2}\right)\epsilon_{n,\mathbf{0}}$. By using part (d) of Lemma 3, we have $\|\boldsymbol{W}_{\boldsymbol{i}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_2 \lesssim \epsilon_{n,\mathbf{0}}\prod_{k=1}^d J_k^{i_k+1/2}$ and $\|\boldsymbol{b}^T\boldsymbol{W}_{\boldsymbol{i}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_2 \lesssim \epsilon_{n,\mathbf{0}}\prod_{k=1}^d J_k^{i_k}$ for any $\boldsymbol{i} \leq \boldsymbol{r}$. Similarly, we also have $\|\boldsymbol{b}^T\boldsymbol{W}_{\boldsymbol{i}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_\infty \leq \|\boldsymbol{b}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_\infty\|\boldsymbol{W}_{\boldsymbol{i}}\|_1 \lesssim \epsilon_{n,\mathbf{0}}(n\epsilon_{n,\mathbf{0}}^2)^{1/2}\prod_{k=1}^d J_k^{i_k}$ by part (c) of Lemma 3 for any $\boldsymbol{i} \leq \boldsymbol{r}$. Using Faà di Bruno's formula and the assumption that $\Psi$ is continuously differentiable, we have $C_n'(M_3') \subset D_n(M_4)$.

To get $D_n(M_4) \subset E_n(M_5)$, we use (15), $c(\boldsymbol{\theta}) = 1 + O_p(\epsilon_{n,\mathbf{0}})$, and the fact that the rate for approximating $p^{(\boldsymbol{r})}$ by spline derivative series $\sum_{k=1}^d J_k^{r_k - \alpha_k} =$

$\sum_{k=1}^{d} J_k^{r_k} \epsilon_{n,\mathbf{0}}$ is bounded above by a multiple of $(\prod_{k=1}^{d} J_k^{r_k})\epsilon_{n,\mathbf{0}} = \epsilon_{n,\boldsymbol{r}}$. Therefore $\Pi_n\{E_n(M_5)\} \to 1$.

**Acknowledgements** The authors would like to thank a referee for insightful comments. The research of the second author is partially supported by National Science Foundation grant number DMS-1510238.

## References

Bartlett MS (1963) Statistical estimation of density function. Sankhyā, Ser A 25:245–254

Belitser E, Serra P (2014) Adaptive priors based on splines with random knots. Bayesian Anal 9:859–882

Bhattacharya PK (1967) Estimation of a probability density function and its derivatives. Sankhyā Ser A 29:373–382

Castillo I (2014) On Bayesian supremum norm contraction rates. Ann Statist 42:2058–2091

Chacón JE, Duong T (2013) Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. Electron J Stat 7:499–532

Donoho DL (1988) One-sided inference about functionals of a density. Ann Statist 16:1390–1420

Farrell RH (1972) On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. Ann Math Statist 43:170–180

Fukunaga K, Hostetler L (1975) The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inform Theory 21:32–40

Genovese CR, Perone-Pacifico M, Verdinelli I, Wasserman L (2016) Nonparametric inference for density modes. J Roy Statist Soc, Ser B 78:99–126

Ghosal S, van der Vaart A (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann Statist 29:1233–1263

Ghosal S, van der Vaart A (2007a) Convergence rates of posterior distributions for noniid observations. Ann Statist 35:192–223

Ghosal S, van der Vaart A (2007b) Posterior convergence rates of Dirichlet mixtures at smooth densities. Ann Statist 35:697–723

Ghosal S, Ghosh JK, Ramamoorthi RV (1999) Posterior consistency of Dirichlet mixtures in density estimation. Ann Statist 27:143–158

Ghosal S, Ghosh JK, van der Vaart A (2000) Convergence rates of posterior distributions. Ann Statist 28:500–531

Giné E, Nickl R (2011) Rates of contraction for posterior distributions in $L^r$-metrics, $1 \leq r \leq \infty$. Ann Statist 39:2883–2911

Hall P, Yatchew A (2007) Nonparametric estimation when data on derivatives are available. Ann Statist 35:300–323

Hosseinioun N, Doosti H, Niroumand HA (2011) Nonparametric estimation of a multivariate probablity density for mixing sequences by the method of wavelets. Ital J Pure Appl Math 28:31–40

Hosseinioun N, Doosti H, Nirumand HA (2012) Nonparametric estimation of the derivatives of a density by the method of wavelet for mixing sequences. Statist Papers 53:195–203

de Jonge R, van Zanten H (2012) Adaptive estimation of multivariate functions using conditionally gaussian tensor-product spline priors. Electron J Stat 6:1984–2001

Knapik BT, van der Vaart AW, van Zanten JH (2011) Bayesian inverse problems with Gaussian priors. Ann Statist 39

Parzen E (1962) On the estimation of a probability density and mode. Ann Math Statist 33:1065–1076

Prakasa Rao B (1996) Nonparametric estimation of the derivatives of a density by the method of wavelets. Bull Inform Cybernet 28:91–100

Qiao W, Polonik W (2016) Theoretical analysis of nonparametric filament estimation. Ann Statist 44:1269–1297

Ray K (2013) Bayesian inverse problems with non-conjugate priors. Electron J Stat 7:2516–2549

Rivoirard V, Rousseau J (2012) Posterior concentration rates for infinite dimensional exponential families. Bayesian Anal 7:311–334

Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. Ann Math Statist 27:832–837

Sasaki H, Noh YK, Niu G, Sugiyama M (2016) Direct density derivative estimation. Neural Computation 28:1101–1140

Schumaker L (2007) Spline Functions: Basic Theory. Cambridge University Press

Schuster EF (1969) Estimation of a probability density function and its derivatives. Ann Math Statist 40:1187–1195

Shen W, Ghosal S (2015) Adaptive Bayesian procedures using random series priors. Scand J Statist 42:1194–1213

Shen W, Ghosal S (2016) Adaptive Bayesian density regression for high dimesnional data. Bernoulli 22:396–420

Shen W, Tokdar ST, Ghosal S (2013) Adaptive Bayesian multivariate density estimation with dirichlet mixtures. Biometrika 100:623–640

Silverman BW (1986) Density Estimation for Statistics and Data Analysis. CRC press

Singh RS (1977a) Applications of estimators of a density and its derivatives to certain statistical problems. J Roy Statist Soc, Ser B 39:357–363

Singh RS (1977b) Improvement on some known nonparametric uniformly consistent estimators of derivatives of a density. Ann Statist 5:394–399

Stone CJ (1990) Large-sample inference for log-spline models. Ann Statist 18:717–741

Szabo B, van der Vaart AW, van Zanten JH (2015) Frequentist coverage of adaptive nonparametric Bayesian credible sets. Ann Statist 43

Tokdar ST, Ghosh JK (2007) Posterior consistency of logistic gaussian process priors in density estimation. J Statist Plann Inference 137:34–42

van der Vaart A, van Zanten H (2008) Rates of contraction of posterior distributions based on Gaussian process priors. Ann Statist 36:1435–1463

van der Vaart A, van Zanten JH (2009) Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. Ann Statist 37:2655–2675

Yoo WW, Ghosal S (2016) Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. Ann Statist 44:1069–1102

Zhou S, Wolfe DA (2000) On derivative estimation in spline regression. Statist Sinica 10:93–108