Convergence Rates for Kernel Regression in Infinite Dimensional Spaces

Joydeep Chowdhury and Probal Chaudhuri

October 31, 2016

Abstract

It is well-known that the rate of convergence of nonparametric regression estimates depends on the dimension of the covariate when it is finite dimensional. When the covariate is infinite dimensional, as it happens in the case of functional data, the derivation of convergence rates for nonparametric regression estimates is a challenging problem that has received attention in the recent literature. We derive the optimum convergence rates for a wide class of kernel regression estimates with infinite dimensional covariates. In our setup, the covariate is a random element in a complete separable metric space, and the function to be estimated takes values in a separable Banach space. The small ball probability function in the covariate space plays a critical role in determining the asymptotic variance of kernel estimates. Unlike what happens in the case of finite dimensional covariates, the optimal asymptotic orders of the bias and the variance of a nonparametric estimate are not same for infinite dimensional covariates.

Keywords and phrases: bias-variance decomposition, Gaussian processes, maximum likelihood regression, optimal bandwidth, separable Banach space, small ball probability

1 Introduction

Consider the regression problem where the covariate $\mathbf{X}$ is a random element in a metric space, and the response $\mathbf{Y}$ is a real valued random variable. Suppose that we want to estimate $\Theta(\mathbf{x}) = \mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$. Let $(\mathbf{X}_1, \mathbf{Y}_1), \cdots, (\mathbf{X}_n, \mathbf{Y}_n)$ be the sample of *i.i.d.* observations from the joint distribution of $(\mathbf{X}, \mathbf{Y})$. When the metric space is the q-dimensional Euclidean space $\mathbb{R}^q$, Stone (1980) proved that the optimal convergence rate of a nonparametric estimate $\widehat{\Theta}_n(\mathbf{x})$ of $\Theta(\mathbf{x})$ is $n^{-(\beta/(2\beta+q))}$. Here, β is a positive constant such that $|\Theta(\mathbf{z}) - \Theta(\mathbf{x})| = O(\|\mathbf{z} - \mathbf{x}\|^{\beta})$ as $\mathbf{z} \to \mathbf{x}$, with $\| \cdot \|$ being the Euclidean norm in $\mathbb{R}^q$. The optimum achievable convergence rate for nonparametric regression with finite dimensional covariate was further investigated in Stone (1982), Ibragimov and Haśminskii (1980), Yatracos (1988), Donoho and Liu (1991a,b) etc. However, when the dimension of the covariate space is infinite, the expressions of the optimum rate of convergence derived by these authors are no longer valid.

Recently nonparametric regression with functional covariates has been studied in Masry (2005), Ferraty et al. (2007), Rachdi and Vieu (2007), etc. These authors investigated nonparametric estimation of the conditional mean when the covariate is functional, and the response is real valued. They studied the consistency and the asymptotic normality of kernel estimates as well as data-driven selection of the bandwidth. But they did not consider the problem of best achievable convergence rate of a nonparametric regression estimate. In Mas (2012) and Chagny and Roche (2014), the problem of optimum convergence rate was explored. In Mas (2012), the usual mean regression problem with a real valued response was considered, and a lower bound for the rate of convergence of the minimax risk was established (see Theorem 3 in Mas (2012)). Mas (2012) tried to choose the bandwidth of the kernel estimate by balancing the asymptotic orders of the bias and the variance (see Lemma 1 and the discussion preceding it in Mas (2012)). However, as it will be shown in this paper, the optimum choice of the bandwidth in a kernel estimate that minimizes the mean square error leads to different asymptotic orders of the bias and the variance when the covariate is infinite dimensional (see subsection 4.2). In Chagny and Roche (2014), the optimum convergence rate was derived for the estimate of the conditional distribution function of a real valued response given a functional covariate. This is a special case of the general mean regression problem with the regression function being real valued and bounded.

In most of the literature on regression with functional data, the authors considered real or multivariate responses and functional covariates. However, one may face regression problems where the response itself may be functional in nature, and one may be interested in parameters of the conditional distribution of the response other than the conditional mean. In this paper, we consider nonparametric regression problems involving a covariate $\mathbf{X}$, which is a random element in a complete separable metric space, a response $\mathbf{Y}$, which lies in some arbitrary measure space, and a parameter $\Theta(\mathbf{x})$ associated with the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Here, $\Theta(\cdot)$ is a function from the covariate space into some separable Banach space. We shall investigate the convergence rate of the kernel regression estimate in this setup.

In Section 2, our regression setup and the kernel regression estimates are described in detail. In Section 3, we discuss an asymptotic bias-variance decomposition of our kernel estimate, and study the asymptotic behavior of the bias and the variance terms. We show that the asymptotic behavior of the variance term critically depends on the small ball probability in the covariate space. The main convergence results for the estimate $\widehat{\Theta}_n(\mathbf{x})$ are presented in Section 4. Section 5 contains concluding remarks and discussion. The proofs and related mathematical details are provided in Section 6.

2 Kernel Estimates

We assume the covariate $\mathbf{X}$ to be a random element in some complete separable metric space $(\mathcal{C}, d)$ with d being the metric, and the response $\mathbf{Y}$ to be a random element in some measure space $\mathcal{R}$. Denote the conditional probability measure of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ as $\mu(\cdot\,|\,\mathbf{x})$. We need to estimate a parameter $\Theta(\mathbf{x})$ of the conditional measure $\mu(\cdot\,|\,\mathbf{x})$ for a fixed $\mathbf{x} \in \mathcal{C}$. In the nonparametric kernel regression method developed in Nadaraya (1964) and Watson (1964), one uses a suitable kernel function $K(\cdot)$ and a bandwidth $h > 0$ to construct an estimate of $\Theta(\mathbf{x})$. One first estimates the weighted empirical probability measure $\mu_n(\cdot\,|\,\mathbf{x})$

that assigns probability mass

$$W_{i,n} = \frac{K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^{n} K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))}$$

to the data point $\mathbf{Y}_i$, for $i = 1, \cdots, n$. The kernel estimate $\widehat{\Theta}_n(\mathbf{x})$ of $\Theta(\mathbf{x})$ is the corresponding parameter associated with $\mu_n(\cdot \,|\, \mathbf{x})$. Some examples of kernel estimates are given below.

Example 2.0.1 (**Mean regression**)**.** In this regression problem, we are interested in estimating $\Theta(\mathbf{x}) = \mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}]$, where $\Psi(\cdot) : \mathcal{R} \to \mathcal{B}$, and $\mathcal{B}$ is a separable Banach space. Here, the estimate $\widehat{\Theta}_n(\mathbf{x})$ is

$$\widehat{\Theta}_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} \Psi(\mathbf{Y}_i) K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^{n} K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))}.$$

Some examples of $\Psi(\cdot)$ are the following. Let the response $\mathbf{Y} \in \mathbb{R}$. Consider $\Psi(\mathbf{Y}) = I(\mathbf{Y} \leq y)$, where $y \in \mathbb{R}$. In this case, we are estimating the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ at y (see Chagny and Roche (2014)). Alternatively, if $\Psi(\mathbf{Y}) = \mathbf{Y}^r$, we get the conditional rth raw moment of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Next, let $\mathbf{Y}$ be a random vector in $\mathbb{R}^q$. For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^q$ with $\mathbf{u} = [u_1, \cdots, u_q]$ and $\mathbf{v} = [v_1, \cdots, v_q]$, let $\mathbf{u} \leq \mathbf{v}$ imply that $u_i \leq v_i$ for $i = 1, \cdots, q$. Here $\Psi(\mathbf{Y}) = I(\mathbf{Y} \leq \mathbf{y})$, where $\mathbf{y} \in \mathbb{R}^q$, leads to the estimation of the conditional multivariate distribution of $\mathbf{Y}$ at $\mathbf{y}$ given $\mathbf{X} = \mathbf{x}$. When $\mathbf{Y} \in \mathcal{B} =$ a separable Hilbert space, the choices $\Psi(\mathbf{Y}) = \mathbf{Y}$ or $\mathbf{Y} \otimes \mathbf{Y}$ (the outer product of $\mathbf{Y}$ with itself) corresponds to the estimation of the first or the second conditional moment of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$, respectively. Note that when $\mathcal{B} = \mathbb{R}^q$, $\mathbf{Y} \otimes \mathbf{Y}$ becomes the $q \times q$ matrix $\mathbf{Y}\mathbf{Y}^t$.

Example 2.0.2 (**Functions of conditional mean**)**.** Let $\mathcal{B}_1, \mathcal{B}_2$ be two separable Banach spaces, $\Psi(\cdot) : \mathcal{R} \to \mathcal{B}_1$ and $\Gamma(\cdot) : \mathcal{B}_1 \to \mathcal{B}_2$ be two functions. Consider $\Theta(\mathbf{x}) = \Gamma(\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}])$. Here, the kernel regression estimate $\widehat{\Theta}_n(\mathbf{x})$ is

$$\widehat{\Theta}_n(\mathbf{x}) = \Gamma\left(\frac{\sum_{i=1}^{n} \Psi(\mathbf{Y}_i) K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^{n} K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))} \right).$$

As an example, let the response space $\mathcal{R}$ be a separable Hilbert space, and $\mathcal{B}_2$ denote the space of Hilbert-Schmidt operators on $\mathcal{R}$. Set $\mathcal{B}_1 = \mathcal{B}_2 \times \mathcal{R}$. Define $\Psi(\mathbf{Y}) = (\mathbf{Y} \otimes \mathbf{Y}, \mathbf{Y})$, and $\Gamma(\mathbf{u}, \mathbf{v}) = \mathbf{u} - \mathbf{v} \otimes \mathbf{v}$. Then, $\Theta(\mathbf{x}) = \mathbb{COV}[\mathbf{Y} \,|\, \mathbf{X} = \mathbf{x}]$, which is the conditional covariance of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$.

Example 2.0.3 (**Maximum likelihood regression**)**.** Nonparametric estimation in a maximum likelihood regression problem with finite dimensional covariate was investigated in Staniswalis (1989), Chaudhuri and Dewanji (1995) and Aerts and Claeskens (1997). Let the covariate $\mathbf{X}$ and the response $\mathbf{Y}$ be random elements in the complete separable metric spaces $\mathcal{C}$ and $\mathcal{R}$, respectively. Suppose $\mathbf{Y}$ given $\mathbf{X}$ has a conditional density with respect to some sigma-finite measure in $\mathcal{R}$, and it is given by $f(\cdot \,|\, \Theta(\mathbf{x}))$ for $\mathbf{X} = \mathbf{x}$, where $\Theta(\cdot) : \mathcal{C} \to \mathbb{R}^q$. We assume that the form of the function $f(\cdot \,|\, \cdot)$ is known, but $\Theta(\cdot)$ is unknown. We are interested in estimating $\Theta(\mathbf{x})$ using maximum weighted likelihood procedure, where $\mathbf{x} \in \mathcal{C}$ is fixed. The kernel estimate $\widehat{\Theta}_n(\mathbf{x})$ of $\Theta(\mathbf{x})$ is given by

$$\widehat{\Theta}_n(\mathbf{x}) = \arg\max_{\mathbf{t}\in\mathbb{R}^q} \prod_{i=1}^{n} [f(\mathbf{Y}_i \,|\, \mathbf{t})]^{W_{i,n}(\mathbf{x})}, \text{ where } W_{i,n}(\mathbf{x}) = \frac{K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^{n} K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))}.$$

So, when $f(\mathbf{y} \,|\, \mathbf{t})$ is a differentiable function of $\mathbf{t} \in \mathbb{R}^q$, $\widehat{\Theta}_n(\mathbf{x})$ is the solution (in $\mathbf{t}$) of the likelihood equation

$$\sum_{i=1}^{n} [\nabla(\log f(\mathbf{Y}_i \,|\, \mathbf{t}))] \, W_{i,n}(\mathbf{x}) = \mathbf{0}.$$

Here ∇ denotes the gradient vector of first partial derivatives with respect to $\mathbf{t}$.

2.1 The Kernel Function and The Bandwidth

It is well known that when the covariate $\mathbf{X}$ is finite dimensional, say $\mathbf{X} \in \mathbb{R}^q$, and $\mathbf{X}$ has a continuous positive density at $\mathbf{x}$, one needs to have a sequence of bandwidths $\{h_n\}$ such that $h_n \to 0$ and $nh_n^q \to \infty$ as $n \to \infty$ to ensure the consistency of the kernel regression estimate $\widehat{\Theta}_n(\mathbf{x})$ (see, e.g., Hardle (1990)). To deal with covariates, which are not necessarily finite dimensional, define $\phi(\mathbf{z}, h) = \mathbb{P}[d(\mathbf{z}, \mathbf{X}) \leq h]$. The function $\phi(\mathbf{z}, h)$ is known as the small ball probability function, and plays an important role in the asymptotic properties of the nonparametric regression procedure. We make the following assumptions on the kernel and the sequence of bandwidths.

- A(i) The kernel $K(\cdot)$ is supported on $[0, 1]$ with $K(u)$ being bounded and bounded away from 0 for $0 \leq u \leq 1$, i.e., there are constants $0 < l \leq L$ such that $l \leq K(u) \leq L$ for all $0 \leq u \leq 1$.
- A(ii) The bandwidth $h_n \to 0$ and $n\phi(\mathbf{x}, h_n) \to \infty$ as $n \to \infty$.

Note that for $\mathbf{X} \in \mathbb{R}^q$ having a continuous positive density at $\mathbf{x}$, the condition $n\phi(\mathbf{x}, h_n) \to \infty$ as $n \to \infty$ is equivalent to $nh_n^q \to \infty$ as $n \to \infty$. *Conditions A(i) and A(ii) will be assumed to be true for the rest of the paper.*

3 Bias-Variance Decomposition

A separable Banach space is called to be of type 2 if there is a positive constant c such that for any finite collection of independent zero-mean random elements $\mathbf{Z}_1, \cdots, \mathbf{Z}_n$ in that Banach space with $\mathbb{E}\|\mathbf{Z}_i\|^2 < \infty$ for $i = 1, \cdots, n$, we have $\mathbb{E}\|\mathbf{Z}_1 + \cdots + \mathbf{Z}_n\|^2 \leq c(\mathbb{E}\|\mathbf{Z}_1\|^2 + \cdots + \mathbb{E}\|\mathbf{Z}_n\|^2)$ (see (Araujo and Giné, 1980, p. 158)). Separable Hilbert spaces and spaces like $L_p[a, b]$ with $p \geq 2$ and $-\infty \leq a < b \leq \infty$ are well-known examples of type 2 Banach spaces. Let $\mathcal{B}$ be a separable type 2 Banach space, and $\Theta(\cdot) : \mathcal{C} \to \mathcal{B}$. For $\mathbf{x} \in \mathcal{C}$, we consider the class of kernel regression estimates, which satisfy

$$\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x}) = B_n(\mathbf{x}) + V_n(\mathbf{x}) + R_n(\mathbf{x}), \tag{3.1}$$

where

$$B_n(\mathbf{x}) = \mathbb{L}_{\mathbf{x}} \left(\frac{\sum_{i=1}^{n} F(\mathbf{X}_i) K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^{n} K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))} - F(\mathbf{x}) \right), \tag{3.2}$$

$$V_n(\mathbf{x}) = \mathbb{L}_{\mathbf{x}}\left(\frac{\sum_{i=1}^{n}[G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i)\,|\,\mathbf{X}_i]]K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{\sum_{i=1}^{n} K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}\right). \tag{3.3}$$

Here, $F(\cdot) : \mathcal{C} \to \mathcal{G}$, $G(\cdot) : \mathcal{R} \to \mathcal{G}$, $\mathbb{L}_{\mathbf{x}}(\cdot) : \mathcal{G} \to \mathcal{B}$ and $\mathcal{G}$ is a separable Banach space. The functions $F(\cdot)$, $G(\cdot)$, $\mathbb{L}_{\mathbf{x}}(\cdot)$ and the remainder term $R_n(\mathbf{x})$ are assumed to satisfy the following conditions.

B(i) Let $\beta > 0$ be a constant. Then, $F(\cdot) \in \mathcal{F}(\mathbf{x},\beta,\mathcal{G})$. Here, $\mathcal{F}(\mathbf{x},\beta,\mathcal{G})$ is a class of functions $H(\cdot) : \mathcal{C} \to \mathcal{G}$ such that for some constant $b_H > 0$, $\|H(\mathbf{z}) - H(\mathbf{x})\| \leq b_H d(\mathbf{x},\mathbf{z})^{\beta}$ for all $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$.

B(ii) $\mathbb{L}_{\mathbf{x}}(\cdot)$ is a continuous linear map.

B(iii) $G(\cdot)$ is such that $\mathbb{E}[\|G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y})\,|\,\mathbf{X} = \mathbf{z}]\|^2\,|\,\mathbf{X} = \mathbf{z}]$ is uniformly bounded for $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$.

B(iv) $R_n(\mathbf{x}) = o_{\mathbb{P}}(\delta_n)$ as $n \to \infty$, where $\delta_n = \max\left\{h_n^{\beta}, \left[n\phi(\mathbf{x},h_n)\right]^{-1/2}\right\}$.

Note that $\mathbb{E}[V_n(\mathbf{x})\,|\,\mathbf{X}_1,\cdots,\mathbf{X}_n] = \mathbf{0}$. We can view $B_n(\mathbf{x})$ as the bias term and $V_n(\mathbf{x})$ as the variance term in kernel regression. This bias-variance decomposition holds in many common regression problems as discussed below.

Example 3.0.1 (**Mean regression**)**.** Recall the mean regression problem described in Example 2.0.1. We now assume $\mathcal{B}$ to be a type 2 Banach space. Here, we have $G(\mathbf{Y}) = \Psi(\mathbf{Y})$, $F(\mathbf{X}) = \mathbb{E}[\Psi(\mathbf{Y})\,|\,\mathbf{X}]$ and $\mathbb{L}_{\mathbf{x}}(\cdot)$ to be the identity map on $\mathcal{B}$. So,

$$B_n(\mathbf{x}) = \frac{\sum_{i=1}^{n}\mathbb{E}[\Psi(\mathbf{Y}_i)\,|\,\mathbf{X}_i]K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{\sum_{i=1}^{n} K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))} - \mathbb{E}[\Psi(\mathbf{Y})\,|\,\mathbf{X} = \mathbf{x}],$$

$$V_n(\mathbf{x}) = \frac{\sum_{i=1}^{n}[\Psi(\mathbf{Y}_i) - \mathbb{E}[\Psi(\mathbf{Y}_i)\,|\,\mathbf{X}_i]]K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{\sum_{i=1}^{n} K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))},$$

and $R_n(\mathbf{x}) = \mathbf{0}$.

Hence, (3.1) holds for any kernel satisfying A(i) and any sequence of bandwidths $\{h_n\}$ satisfying A(ii). Here, conditions B(ii) and B(iv) are trivially satisfied. Note that in this case, $F(\mathbf{z}) = \Theta(\mathbf{z})$, and so condition B(i) is satisfied when $\Theta(\mathbf{z})$ is Holder continuous at $\mathbf{x}$ with exponent β, and the class $\mathcal{F}(\mathbf{x},\beta,\mathcal{B})$ can be taken as the class of all Holder continuous functions. Condition B(iii) is satisfied when the conditional covariance of $\Psi(\mathbf{Y})$ given $\mathbf{X} = \mathbf{z}$ converges to the conditional covariance of $\Psi(\mathbf{Y})$ given $\mathbf{X} = \mathbf{x}$ as $\mathbf{z} \to \mathbf{x}$, and the conditional covariance of $\Psi(\mathbf{Y})$ given $\mathbf{X} = \mathbf{x}$ is bounded and positive definite. In particular, B(iii) holds for the location-scale type model $\Psi(\mathbf{Y}) = l(\mathbf{X}) + s(\mathbf{X})\mathbf{U}$, where $l(\cdot) : \mathcal{C} \to \mathcal{B}$ and $s(\cdot) : \mathcal{C} \to (0,\infty)$ are continuous functions, and $\mathbf{U}$ is a zero-mean random element in $\mathcal{B}$, which is independent of $\mathbf{X}$ having a positive definite covariance operator.

Example 3.0.2 (**Functions of conditional mean**)**.** Now, consider the class of regression problems in Example 2.0.2 with $\mathcal{B}_2$ being a type 2 Banach space. Let the kernel function $K(\cdot)$ satisfy A(i) and the bandwidths $\{h_n\}$ satisfy A(ii). Then, we have the following theorem.

Theorem 3.1. *Recall $\Psi(\cdot)$ and $\Gamma(\cdot)$ in the model described in Example 2.0.2. Assume that $\Gamma(\cdot)$ is Fréchet differentiable with derivative $\Gamma'(\cdot)$. Let $\mathbb{L}(\mathbf{x})(\cdot) = \Gamma'(\mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}])(\cdot)$, $G(\mathbf{Y}) = \Psi(\mathbf{Y})$, $F(\mathbf{z}) = \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}]$, and conditions B(i), B(ii) and B(iii) hold. Then, B(iv) is also satisfied, and consequently the bias-variance decomposition in* (3.1) *holds.*

It is easy to verify that when $\Theta(\mathbf{x})$ is the conditional covariance of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$, which is mentioned in Example 2.0.2, the assumptions in Theorem 3.1 are satisfied if $\mathbb{E}[\mathbf{Y} \otimes \mathbf{Y} \mid \mathbf{X} = \mathbf{z}]$ and $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{z}]$ are both Holder continuous at $\mathbf{x}$ with exponent β.

Example 3.0.3 (**Maximum likelihood regression**)**.** Recall the maximum likelihood regression described in Example 2.0.3. Define $g(\mathbf{y} \mid \mathbf{t}) = \log f(\mathbf{y} \mid \mathbf{t})$, where $\mathbf{t} \in \mathbb{R}^q$. Let $\mathcal{T}$ be an open rectangle in $\mathbb{R}^q$ containing the range of $\Theta(\cdot)$. We now assume some Cramer-type regularity conditions on the log-likelihood $g(\mathbf{y} \mid \mathbf{t})$ that are required for asymptotic analysis of maximum likelihood estimates. The support of $f(\mathbf{y} \mid \mathbf{t})$ is assumed to be same for all $\mathbf{t} \in \mathcal{T}$, and $f(\mathbf{y} \mid \mathbf{t})$ is assumed to be thrice continuously differentiable with respect to $\mathbf{t}$ for $\mathbf{t} \in \mathcal{T}$. Denote the Hessian matrix of all second order partial derivatives of $g(\mathbf{y} \mid \mathbf{t})$ with respect to $\mathbf{t}$ as $\Delta_2(g(\mathbf{y} \mid \mathbf{t}))$, and the array of all third order partial derivatives of $g(\mathbf{y} \mid \mathbf{t})$ with respect to $\mathbf{t}$ as $\Delta_3(g(\mathbf{y} \mid \mathbf{t}))$. Define $\mathbf{I}(\Theta(\mathbf{z})) = \mathbb{E}[\Delta_2(g(\mathbf{Y} \mid \Theta(\mathbf{z}))) \mid \mathbf{X} = \mathbf{z}]$, and assume that $\mathbf{I}(\Theta(\mathbf{z}))$ is positive definite and continuous for $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$. Also, assume that for $\mathbf{t} \in \mathcal{T}$, there exist two non-negative random variables $\mathbf{D}_1(\mathbf{Y} \mid \mathbf{t})$ and $\mathbf{D}_2(\mathbf{Y} \mid \mathbf{t})$ such that $\mathbb{E}[\mathbf{D}_1(\mathbf{Y} \mid \mathbf{t})]^2 < \infty$, $\mathbb{E}[\mathbf{D}_2(\mathbf{Y} \mid \mathbf{t})] < \infty$, and $\|\Delta_2(g(\mathbf{Y} \mid \mathbf{s}))\| \le \mathbf{D}_1(\mathbf{Y} \mid \mathbf{t})$, $\|\Delta_3(g(\mathbf{Y} \mid \mathbf{s}))\| \le \mathbf{D}_2(\mathbf{Y} \mid \mathbf{t})$ for any $\mathbf{s}$ in some neighborhood of $\mathbf{t}$ contained in $\mathcal{T}$.

As in Example 2.0.3, let ∇ denote the gradient vector of first partial derivatives. Then, $\sum_{i=1}^{n} \nabla g(\mathbf{Y}_i \mid \widehat{\Theta}_n(\mathbf{x})) W_{i,n}(\mathbf{x}) = \mathbf{0}$. Note that here $\mathcal{G} = \mathbb{R}^q$.

Theorem 3.2. *Suppose that in the model described in Example 2.0.3, the Cramer type regularity conditions stated above hold. Assume that $\Theta(\mathbf{z}) \in \mathcal{F}(\mathbf{x}, \beta, \mathbb{R}^q)$ for some $\beta > 0$, where $\mathcal{F}(\mathbf{x}, \beta, \mathbb{R}^q)$ is as defined in B(i). Then,* (3.1) *along with conditions B(i) through B(iv) will hold if we choose $\mathbb{L}(\mathbf{x})(\cdot) = [\mathbf{I}(\Theta(\mathbf{x}))]^{-1}(\cdot)$, $G(\mathbf{Y}) = -\nabla g(\mathbf{Y} \mid \Theta(\mathbf{X}))$ and $F(\mathbf{X}) = \mathbf{I}(\Theta(\mathbf{x}))(\Theta(\mathbf{X}))$, where $g(\mathbf{y} \mid \mathbf{t}) = \log f(\mathbf{y} \mid \mathbf{t})$.*

3.1 Asymptotic Behavior of the Bias and the Variance

In this subsection, the order of convergence of the bias term $B_n(\mathbf{x})$ and the variance term $V_n(\mathbf{x})$ in (3.1) are investigated. It follows from assumptions A(ii) and B(i) that $\|B_n(\mathbf{x})\| \le \|\mathbb{L}_{\mathbf{x}}\| b_F h_n^{\beta}$ for all sufficiently large n. So,

$$\mathbb{E}\|B_n(\mathbf{x})\|^2 \le (\|\mathbb{L}_{\mathbf{x}}\| b_F)^2 h_n^{2\beta} \tag{3.4}$$

for all sufficiently large n. The inequality (3.4) leads to an upper bound of the rate of convergence of the bias term, and will be used later to study the asymptotic properties of the estimate $\widehat{\Theta}_n(\mathbf{x})$.

We next discuss the asymptotic behavior of the variance term $V_n(\mathbf{x})$. Under A(i), A(ii) and B(iii), we derive an upper bound of the convergence rate of $\mathbb{E}\|V_n(\mathbf{x})\|^2$ in the theorem below.

Theorem 3.3. *Under A(i), A(ii) and B(iii), $n\phi(\mathbf{x}, h_n)\mathbb{E}\|V_n(\mathbf{x})\|^2$ is uniformly bounded over n.*

Note that $\mathbb{E}\|B_n(\mathbf{x}) + V_n(\mathbf{x})\|^2 \leq 2(\mathbb{E}\|B_n(\mathbf{x})\|^2 + \mathbb{E}\|V_n(\mathbf{x})\|^2)$. When $\mathcal{B}$ is a Hilbert space, we actually have $\mathbb{E}\|B_n(\mathbf{x}) + V_n(\mathbf{x})\|^2 = \mathbb{E}\|B_n(\mathbf{x})\|^2 + \mathbb{E}\|V_n(\mathbf{x})\|^2$. Now, consider the following conditions.

B(v) $\mathcal{B}$ is a separable Hilbert space, and $G(\cdot)$ in (3.3) is such that the covariance operator $\mathbb{D}(\cdot, \cdot \mid \mathbf{z}) : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ defined by $\mathbb{D}(\mathbf{u}, \mathbf{v} \mid \mathbf{z}) = \mathbb{E}[\langle \mathbf{u}, \mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}])\rangle \langle \mathbf{v}, \mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}])\rangle \mid \mathbf{X} = \mathbf{z}]$, where $(\mathbf{u}, \mathbf{v}) \in \mathcal{B} \times \mathcal{B}$, converges to $\mathbb{D}(\cdot, \cdot \mid \mathbf{x})$ in the trace norm as $\mathbf{z} \to \mathbf{x}$, and $\mathbb{D}(\cdot, \cdot \mid \mathbf{x})$ is a bounded positive definite operator.

B(vi) For some $\nu > 2$, $\mathbb{E}[\|\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}])\|^\nu \mid \mathbf{X} = \mathbf{z}]$ is uniformly bounded for $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$.

The conditions B(v) and B(vi) hold in many common models. For example, consider the location-scale type model $\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y})) = l(\mathbf{X}) + s(\mathbf{X})\mathbf{U}$, where $l(\cdot) : \mathcal{C} \to \mathcal{B}$ and $s(\cdot) : \mathcal{C} \to (0, \infty)$ are continuous functions, and $\mathbf{U}$ is a zero-mean random element in $\mathcal{B}$, which is independent of $\mathbf{X}$ having a positive definite covariance operator and satisfies $\mathbb{E}\|\mathbf{U}\|^\nu < \infty$. Then, it is easy to verify that conditions B(v) and B(vi) are satisfied.

From assumption A(i), it follows that $0 < l^j \phi(\mathbf{x}, h) \leq \mathbb{E}K^j(h^{-1}d(\mathbf{x}, \mathbf{X})) \leq L^j \phi(\mathbf{x}, h)$ for any positive integer j and any bandwidth $h > 0$. Define $E_n^{(j)}(\mathbf{x}) = \mathbb{E}K^j(h_n^{-1}d(\mathbf{x}, \mathbf{X}))/\phi(\mathbf{x}, h_n)$ for all j. Note that $0 < L^{-1}l \leq [E_n^{(2)}(\mathbf{x})]^{-1/2} E_n^{(1)}(\mathbf{x}) \leq l^{-1}L < \infty$ for all n. In the next theorem, we establish that $[n\phi(\mathbf{x}, h_n)]^{1/2} [E_n^{(2)}(\mathbf{x})]^{-1/2} E_n^{(1)}(\mathbf{x}) V_n(\mathbf{x})$ has an asymptotic Gaussian distribution.

Theorem 3.4. *Let the kernel function $K(\cdot)$ satisfy A(i), and the sequence of bandwidths $\{h_n\}$ satisfy A(ii). Then, under conditions B(v) and B(vi), we have $[n\phi(\mathbf{x}, h_n)]^{1/2}[E_n^{(2)}(\mathbf{x})]^{-1/2} E_n^{(1)}(\mathbf{x}) V_n(\mathbf{x}) \to \mathbf{W}$ in distribution as $n \to \infty$. Here $\mathbf{W}$ is a zero mean Gaussian random element in $\mathcal{B}$ with covariance $\mathbb{D}(\cdot, \cdot \mid \mathbf{x})$.*

The function $\phi(\mathbf{x}, h)$ plays a central role in determining the convergence rate and the asymptotic distribution of $V_n(\mathbf{x})$, and we discuss it in detail in the next subsection.

3.2 The Small Ball Probability Function

When the covariate $\mathbf{X}$ is finite dimensional, say $\mathbf{X} \in \mathbb{R}^q$, and it has a continuous positive density at $\mathbf{x}$, it is easy to show that $\phi(\mathbf{x}, h) \sim h^q$ as $h \to 0^+$. But when $\mathbf{X}$ is a random element in an infinite dimensional space, getting the asymptotic order of $\phi(\mathbf{x}, h)$ as $h \to 0^+$ is much more difficult. Most of the available results in this area are for the situation where $\mathbf{X}$ is a Gaussian process (see, e.g., Lifshits (2013)). In the literature, the popular approach has been to first derive the limiting behavior of $\log \phi(\mathbf{0}, h)$ as $h \to 0^+$, when $\mathbf{X}$ is a Gaussian random element centered at $\mathbf{0}$. Then, one makes a connection between $\phi(\mathbf{x}, h)$ and $\phi(\mathbf{0}, h)$ for suitable $\mathbf{x}$ and sufficiently small h.

The asymptotic behavior of $\log \phi(\mathbf{0}, h)$ was investigated in Li (2001) for real valued centered Gaussian Markov processes on $[0, 1]$ under the L_p-norm, where $1 \leq p \leq \infty$. It was shown there that in such a case, $h^2 \log \phi(\mathbf{0}, h) \to -c_1$ as $h \to 0^+$, where $c_1 > 0$ is a constant depending on p. For $\mathbf{X}$ being a fractional Brownian motion on $[0, 1]$ with Hurst index $\gamma \in (0, 1)$, it was shown in Theorem 4.6 in Li and Shao (2001) that under the L_∞-norm, $-c_2 h^{-1/\gamma} \leq \log \phi(\mathbf{0}, h) \leq$

$-c_3h^{-1/\gamma}$ for all $0 < h \leq 1$. Here, c_2 and c_3 are positive constants depending on γ. For $\mathbf{X}$ being an integrated fractional Brownian motion with Hurst index $\gamma \in (0,1)$, it was established in Theorem 4.10 of Li and Shao (2001) that under the L_∞-norm, $-c_4h^{-1/(1+\gamma)} \leq \log\phi(\mathbf{0},h) \leq -c_5h^{-1/(1+\gamma)}$ for all $0 < h \leq 1$, where c_4 and c_5 are positive constants depending on γ.

For the Lévy fractional Brownian motion on $[0,1]^q$ with Hurst index $\gamma \in (0,1)$, it was proved in Theorem 5.1 in Li and Shao (2001) that under the L_∞-norm, $-c_6h^{-q/\gamma} \leq \log\phi(\mathbf{0},h) \leq -c_7h^{-q/\gamma}$ for all $0 < h \leq 1$. Here, c_6 and c_7 are positive constants depending on γ and q. For a Brownian sheet on $[0,1]^q$, it follows from Theorem 5.3 in Li and Shao (2001) that under the L_2-norm, $-c_8h^{-2}(\log(1/h))^{(2q-2)} \leq \log\phi(\mathbf{0},h) \leq -c_9h^{-2}(\log(1/h))^{(2q-2)}$ as $h \to 0^+$, where $c_8, c_9 > 0$ are constants depending on q. It was shown in Theorem 5.4 in Li and Shao (2001) that if $\mathbf{X}$ is a Brownian sheet on $[0,1]^2$, we have $-c_{10}h^{-2}(\log(1/h))^3 \leq \log\phi(\mathbf{0},h) \leq -c_{11}h^{-2}(\log(1/h))^3$ under the L_∞-norm, where $c_8, c_9 > 0$ are constants.

3.3 Shifted Small Ball Probability

As we have already mentioned, the asymptotic behavior of $\log\phi(\mathbf{x},h)$ is derived by establishing some relationship between $\phi(\mathbf{x},h)$ and $\phi(\mathbf{0},h)$. As described in subsection 1.2 in Mas (2012), one can establish a relation between $\phi(\mathbf{x},h)$ and $\phi(\mathbf{0},h)$ if the probability measure of $\mathbf{X}-\mathbf{x}$ is absolutely continuous with respect to the probability measure of $\mathbf{X}$, and the density of the measure of $\mathbf{X}-\mathbf{x}$ with respect to the measure of $\mathbf{X}$ is suitably smooth. This approach is inspired by the Cameron-Martin Theorem describing the Radon-Nikodym derivative of a Weiner measure translated by $\mathbf{x}$ with respect to the centered Weiner measure, where $\mathbf{x}$ is an element of the reproducing kernel Hilbert space associated with the centered Weiner measure (see Cameron and Martin (1944)). When $\mathbf{X}$ is a centered Gaussian random element in a separable Banach space and $\mathbf{x}$ is an element of the associated reproducing kernel Hilbert space, from Theorem 3.1 in Li and Shao (2001) we get that $\exp[-(1/2)\|\mathbf{x}\|_\mu^2]\phi(\mathbf{0},h) \leq \phi(\mathbf{x},h) \leq \phi(\mathbf{0},h)$ for all $h > 0$, where $\|\cdot\|_\mu$ is the norm in the reproducing kernel Hilbert space. *But this result is not very useful for our purpose since the probability of the event that an infinite dimensional Gaussian random element lies in its reproducing kernel Hilbert space is zero (see Corollary 7.1 in Lukić and Beder (2001)).* Fortunately, it follows from Remark 2.2 in Dereich and Lifshits (2005) that when $\mathbf{X}$ is a centered Gaussian random element in a separable Banach space, then for almost all $\mathbf{x}$, $(\phi(\mathbf{0},h/2))^2 \leq \phi(\mathbf{x},h) \leq \phi(\mathbf{0},h)$ for all sufficiently small h depending on $\mathbf{x}$. On the other hand, it follows from Theorem 2.1 in Hoffmann-Jorgensen et al. (1979) that for $\mathbf{X}$ being a centered Gaussian random element in a separable infinite dimensional Hilbert space, we have $\exp[-(1/2)\|\mathbf{x}\|^2]\phi(\mathbf{0},h) \leq \phi(\mathbf{x},h) \leq \phi(\mathbf{0},h)$ for all $h > 0$.

Let $\mathbf{X}$ be a centered Gaussian random element in a separable Hilbert space. The Karhunen-Loeve expansion of $\mathbf{X}$ is $\mathbf{X} = \sum_{j=1}^{\infty}\sqrt{\lambda_j}Z_j\boldsymbol{\psi}_j$, where $\{Z_j\}$ is a collection of independent normal random variables with mean 0 and variance 1, $\{\lambda_j\}$ is the sequence of decreasing eigenvalues of the covariance of $\mathbf{X}$, and $\{\boldsymbol{\psi}_j\}$ is an orthonormal basis of the Hilbert space. Here, the small ball probability $\phi(\mathbf{x},h)$ can be related to the rate of decrease of the sequence $\{\lambda_j\}$. As discussed in subsection 4.1 in Chagny and Roche (2014), for certain rates of decrease for $\{\lambda_j\}$, e.g., if for some $\alpha > 1$, $j^\alpha\lambda_j$ is bounded and bounded away from 0 for all j, we may have $c_{12}h^{p_1}\exp(-c_{13}h^{-q_1}) \leq \phi(\mathbf{x}, h) \leq c_{14}h^{p_2}\exp(-c_{15}h^{-q_1})$ for positive constants c_{12}, c_{13}, c_{14}, c_{15}, p_1, p_2 and q_1. Alternatively, for some other rates, e.g., if $j\exp[2j]\lambda_j$ is bounded and bounded away from 0 for all j, we may have $c_{16}h^{p_3}\exp[-c_{17}(\log(1/h))^{q_2}] \leq \phi(\mathbf{x}, h) \leq c_{18}h^{p_4}\exp[-c_{19}(\log(1/h))^{q_2}]$ for positive constants c_{16}, c_{17}, c_{18}, c_{19}, p_3, p_4 and $q_2 > 1$ (see subsection 4.1 in Chagny and Roche (2014)). See also Theorem 4.4, Examples 4.5, 4.6 and 4.7 in Hoffmann-Jorgensen et al. (1979) for a discussion on the relation of the small ball probability $\phi(\mathbf{x}, h)$ and the rate of decrease of $\{\lambda_j\}$.

Define $m(h) = (1/h)^{t_2}(\log(1/h))^{t_3}$ for $0 < h < 1$. From the discussion on the small ball probability functions above, it is now clear that in a diverse collection of cases, we have

$$C_1 h^{t_1}\exp\left[-C_2 m(h)\right] \leq \phi(\mathbf{x}, h) \leq C_3 h^{t_4}\exp\left[-C_4 m(h)\right] \tag{3.5}$$

as $h \to 0^+$. Here, $C_1, C_2, C_3, C_4 > 0$ and $t_1, t_2, t_3, t_4 \geq 0$ are appropriate constants. Further, C_1 and C_3 depend on $\mathbf{x}$, and at least one of t_1, t_2, t_3 as well as at least one of t_2, t_3 and t_4 is positive. The above form was considered in (Ferraty and Vieu, 2006, p. 209) with $C_2 = C_4 = 1$, $t_1 = t_4 = 0$, and they called it the small ball probability function of an exponential type process. For $t_2 = 0$ and $t_3 = 1$ and appropriate values of the parameters C_1, C_2, C_3 and C_4, (3.5) yields the case of a finite dimensional covariate $\mathbf{X}$ with a continuous positive density at $\mathbf{x}$, or a fractal-type process as defined in (Ferraty and Vieu, 2006, p. 207).

4 Convergence Rate

We now derive the optimum achievable convergence rate for kernel estimates satisfying the bias variance decomposition (3.1). As we shall see, the function $m(h)$ defined in the previous section plays a central role in determining the convergence rate of the estimate $\widehat{\Theta}_n(\mathbf{x})$. We shall consider the covariate space to be infinite dimensional. The case of finite dimensional covariates is extensively discussed in the past literature (see, e.g., Stone (1980, 1982), Ibragimov and Haśminskii (1980), Yatracos (1988), Donoho and Liu (1991a,b)). *In order to consider only infinite dimensional covariates, we assume that in* (3.5)*, either* $t_2 > 0$ *or* $t_3 > 1$. In the next theorem, we establish an asymptotic lower bound of the sequence of bandwidths $\{h_n\}$ that leads to consistent kernel regression estimates.

Theorem 4.1. *Suppose that in the upper and the lower bounds in the shifted small ball probability in* (3.5)*, we have either* $t_2 > 0$*, or* $t_3 > 1$*. Then, for any sequence of bandwidths* $\{h_n\}$*, which satisfies assumption A(ii), we have* $h_n/m^{-1}(\log n)$ *bounded away from* 0 *as* $n \to \infty$*.*

Note that if we have either $t_2 > 0$ or $t_3 > 1$ in (3.5), $m(h)$ is a strictly decreasing positive function, and $m^{-1}(\cdot)$ is well-defined. In the next theorem, we shall see that $\left(m^{-1}(\log n)\right)^{\beta}$ is an attainable rate of convergence of $\widehat{\Theta}_n(\mathbf{x})$.

Theorem 4.2. *Suppose that in* (3.5)*, we have either* $t_2 > 0$*, or* $t_3 > 1$*. Then, for any kernel* $K(\cdot)$ *satisfying A(i) and* $\Theta(\mathbf{x})$ *satisfying* (3.1) *along with conditions B(i) through B(iv), there is a sequence of bandwidths* $\{h_n\}$ *satisfying A(ii) such that* $\left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\| = O_{\mathbb{P}}\left(\left(m^{-1}(\log n)\right)^{\beta}\right)$ *as* $n \to \infty$*.*

4.1 Lower Bound on the Convergence Rate

We now proceed to investigate the lower bound of the convergence rate of $\widehat{\Theta}_n(\mathbf{x})$. Let $E_n = \mathbb{E}K(h_n^{-1}d(\mathbf{x},\mathbf{X}))$. Define $\tilde{B}_n(\mathbf{x}) = \mathbb{L}_{\mathbf{x}}\big(\mathbb{E}[(F(\mathbf{X}) - F(\mathbf{x}))E_n^{-1} K(h_n^{-1}d(\mathbf{x},\mathbf{X}))]\big)$. Consider the following assumption.

C(i) There is $\Theta(\cdot) : \mathcal{C} \to \mathcal{B}$ with the corresponding $\mathbb{L}_{\mathbf{x}}(\cdot)$ and $F(\cdot)$ such that for any sequence of bandwidths $\{h_n\}$ satisfying A(ii),

$$h_n^{-\beta}\|\tilde{B}_n(\mathbf{x})\| > b_1 > 0 \tag{4.1}$$

for all sufficiently large n.

The following two conditions are sufficient to ensure that C(i) holds.

(a) There is a constant $s > 1$ such that $\phi(\mathbf{x}, h/s)/\phi(\mathbf{x}, h/s)$ is bounded away from 1 for all sufficiently small $h > 0$.

(b) Let $\mathbb{L}_{\mathbf{x}}(\mathcal{F}(\mathbf{x}, \beta, \mathcal{G}))$ be the class of all functions defined by the composition $\mathbb{L}_{\mathbf{x}} \circ H$, where $H \in \mathcal{F}(\mathbf{x}, \beta, \mathcal{G})$ and $\mathcal{F}(\mathbf{x}, \beta, \mathcal{G})$ is as defined in B(i). Then $\mathbb{L}_{\mathbf{x}}(\mathcal{F}(\mathbf{x}, \beta, \mathcal{G}))$ contains the function $\mathbf{z} \mapsto d(\mathbf{x}, \mathbf{z})^{\beta}\mathbf{v}$, where $\mathbf{v}$ is some vector in $\mathcal{B}$, and $\mathbf{z}$ lies in a neighborhood of $\mathbf{x}$.

Condition (a) is satisfied when in (3.5) $t_2 > 0$, or $t_1 < t_4$, or $C_2 = C_4$. We observe that at least one of these is true in the examples that we have described in subsection 3.2.

Now, we derive the lower bound of the order of convergence of the bias term $B_n(\mathbf{x})$ in (3.1) under B(i), B(ii) and C(i). Note that $B_n(\mathbf{x}) = \tilde{B}_n(\mathbf{x}) + \tilde{R}_n(\mathbf{x})$, where

$$\begin{aligned}
&\tilde{R}_n(\mathbf{x}) \\
&= \mathbb{L}_{\mathbf{x}}\left(\frac{\sum_{i=1}^n (F(\mathbf{X}_i) - F(\mathbf{x}))K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))} - \tilde{B}_n(\mathbf{x})\right) \\
&= \mathbb{L}_{\mathbf{x}}\left(\frac{\sum_{i=1}^n (F(\mathbf{X}_i) - F(\mathbf{x}))K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))} - \sum_{i=1}^n (F(\mathbf{X}_i) - F(\mathbf{x}))\frac{K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{nE_n}\right) \\
&\quad + \mathbb{L}_{\mathbf{x}}\left(\sum_{i=1}^n (F(\mathbf{X}_i) - F(\mathbf{x}))\frac{K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{nE_n} - \mathbb{E}\left[(F(\mathbf{X}) - F(\mathbf{x}))\frac{K(h_n^{-1}d(\mathbf{x},\mathbf{X}))}{E_n}\right]\right).
\end{aligned}$$

It follows form condition A(i) and Markov inequality that $(n^{-1}\sum_{i=1}^n E_n^{-1}K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i)) - 1) = O_{\mathbb{P}}\big(\big[n\phi(\mathbf{x}, h_n)\big]^{-1/2}\big)$ as $n \to \infty$. Hence, from conditions A(i), A(ii) and B(i), we have

$$\begin{aligned}
&\mathbb{L}_{\mathbf{x}}\left(\frac{\sum_{i=1}^n (F(\mathbf{X}_i) - F(\mathbf{x}))K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))} - \sum_{i=1}^n (F(\mathbf{X}_i) - F(\mathbf{x}))\frac{K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{nE_n}\right) \\
&\quad = o_{\mathbb{P}}(h_n^{\beta})
\end{aligned}$$

as $n \to \infty$. Also, from assumptions A(i), A(ii), B(i) and Markov inequality, it follows that

$$\mathbb{L}_{\mathbf{x}}\left(\sum_{i=1}^n (F(\mathbf{X}_i) - F(\mathbf{x}))\frac{K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{nE_n} - \mathbb{E}\left[(F(\mathbf{X}) - F(\mathbf{x}))\frac{K(h_n^{-1}d(\mathbf{x},\mathbf{X}))}{E_n}\right]\right)$$

$$= o_{\mathbb{P}}(h_n^{\beta})$$

as $n \to \infty$. Hence,

$$\tilde{R}_n(\mathbf{x}) = o_{\mathbb{P}}(h_n^{\beta}) \text{ as } n \to \infty. \tag{4.2}$$

Note that inequality (4.1) provides a lower bound of the convergence rate for the bias term $B_n(\mathbf{x})$ in view of (4.2), and this will be used to determine a lower bound of the rate of convergence of $\widehat{\Theta}_n(\mathbf{x})$. From now on, we shall assume the following.

C(ii) The separable Banach space $\mathcal{B}$ has a Schauder basis $\{\mathbf{e}_1, \mathbf{e}_2, \cdots\}$ such that for any $\mathbf{v} \in \mathcal{B}$, $\mathbf{v} = \sum_{n=1}^{\infty} v_n \mathbf{e}_n$ for a sequence of real numbers $\{v_n\}$.

All $L_p[a, b]$ spaces with $1 \le p < \infty$ and $-\infty \le a < b \le \infty$ have Schauder bases. For $-\infty < a < b < \infty$, the space $C[a, b]$ of continuous functions on an interval $[a, b]$ equipped with the supremum norm has a Schauder basis. Let $\tilde{\phi}_i \in \mathcal{B}^*$ be the projection functional corresponding to $\mathbf{e}_i$, i.e., $\mathbf{v} = \sum_{i=1}^{\infty} \tilde{\phi}_i(\mathbf{v})\mathbf{e}_i$ for all $\mathbf{v} \in \mathcal{B}$. Note that $\tilde{\phi}_i$ is a bounded linear functional for all i. We now state another condition.

C(iii) Let $G(\cdot)$ be as in Section 3. For some positive integer i_0, the conditional variance function $\mathbb{V}(\mathbf{z}) : \mathcal{C} \to \mathbb{R}$ defined by $\mathbb{V}(\mathbf{z}) = \mathbb{E}\big[\big(\tilde{\phi}_{i_0}\big(\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{z}])\big)\big)^2 \,\big|\, \mathbf{X} = \mathbf{z}\big]$ converges to $\mathbb{V}(\mathbf{x})$ as $\mathbf{z} \to \mathbf{x}$, and $\mathbb{V}(\mathbf{x}) > 0$.

Note that $\tilde{\phi}_{i_0}\big(\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}))\big)$ is a real valued random variable. So, the convergence condition in C(iii) of the conditional variance may be viewed as a special case of condition B(v). We now state the theorem on the lower bound of the convergence rate of $\big\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\big\|$.

Theorem 4.3. *Suppose the kernel $K(\cdot)$ satisfies A(i), the sequence of bandwidths $\{h_n\}$ satisfies A(ii), and the representation* (3.1) *along with conditions B(i), B(ii), B(iv), B(vi) and C(i) through C(iii) hold. Then, we have* $\liminf_{n\to\infty} \mathbb{P}\big[\big(m^{-1}(\log n)\big)^{-\beta} \big\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\big\| > c\big] > 0$, *for some constant $c > 0$ depending on $\Theta(\mathbf{x})$.*

Combining Theorem 4.2 and Theorem 4.3, we get that $\big(m^{-1}(\log n)\big)^{\beta}$ is the optimum rate of convergence of $\widehat{\Theta}_n(\mathbf{x})$ when all the conditions of the two theorems are satisfied. We now deduce simplified expressions of the optimum rates for various infinite dimensional covariate distributions considered in subsection 3.2.

For $\mathbf{X}$ being a real valued continuous Gaussian Markov process on $[0, 1]$, under the L_p-norm, we have $\big(m^{-1}(\log n)\big)^{2\beta} = (\log n)^{-\beta}$. For fractional Brownian motion with Hurst index $\gamma \in (0, 1)$, under the L_∞-norm, we have $t_2 = 1/\gamma$, and consequently $\big(m^{-1}(\log n)\big)^{2\beta} = (\log n)^{-2\gamma\beta}$. On the other hand, for an integrated fractional Brownian motion with Hurst index γ and under the L_∞-norm, we have $t_2 = 1/(1 + \gamma)$ and $\big(m^{-1}(\log n)\big)^{2\beta} = (\log n)^{-2(1+\gamma)\beta}$. When $\mathbf{X}$ is a Lévy fractional Brownian motion on $[0, 1]^q$ with Hurst index γ, $t_2 = q/\gamma$ and $\big(m^{-1}(\log n)\big)^{2\beta} = (\log n)^{-2\gamma\beta/q}$.

In the class of processes $H_{X,L}$ considered in subsection 4.1 of Chagny and Roche (2014), $t_2 > 0$ and $t_3 = 0$, and we have $\big(m^{-1}(\log n)\big)^{2\beta} = (\log n)^{-2\beta/t_2}$.

On the other hand, for the class of processes $H_{X,M}$ considered by these authors, we have $t_2 = 0$ and $t_3 > 1$, and consequently the optimum convergence rate is $\big(m^{-1}(\log n)\big)^{2\beta} = \exp[-2\beta(\log n)^{1/t_3}]$.

4.2 Asymptotic Dominance of Bias over Variance

Recall that in the case of finite dimensional covariates, the bias and the variance terms in nonparametric regression have the same rate of convergence (see our discussion in Section 1). But, this is no longer true when the covariate $\mathbf{X}$ is an infinite dimensional random element for which either $t_2 > 0$ or $t_3 > 1$ in (3.5).

Theorem 4.4. *Suppose assumption A(i) holds and either* $t_2 > 0$ *or* $t_3 > 1$ *in the bounds in* (3.5). *Also, let the representation* (3.1) *along with conditions B(i) through B(iv) holds. Then for* $\Theta(\mathbf{x})$ *satisfying C(i), the ratio* $\|V_n(\mathbf{x})\|/\|B_n(\mathbf{x})\| \to 0$ *in probability as* $n \to \infty$ *for the optimum choice of bandwidth* $\{h_n\}$ *obtained in Theorem 4.2.*

Theorem 4.4 illustrates that the bandwidth that leads to the best achievable rate of convergence does not balance the convergence rates of the bias and the variance in kernel regression if the covariate is infinite dimensional. Instead, the ratio of the bias and the variance for the optimal choice of bandwidth explodes to infinity as the sample size increases. This phenomenon is due to the exponential decay of the small ball probability function in infinite dimensional spaces. When the covariate is infinite dimensional, we may have very small number of observations in a neighborhood in the covariate space due to exponentially small values of the small ball probability function. To cope with this problem, one has to use relatively larger bandwidths than what is required for finite dimensional covariates. In a sense, this results in an 'over-smoothed' estimate with its bias asymptotically larger than its variance.

5 Concluding Remarks

In this paper, we have derived the optimum convergence rate for a wide class of kernel regression estimates when the covariate as well as the response may be infinite dimensional. It is shown that the convergence rates of such estimates do not depend on the dimension of the response, but they depend critically on the dimension of the covariate. We have seen that, for a wide class of covariates having infinite dimensional Gaussian distributions, the convergence rate is much slower than the optimum achievable rate for finite dimensional covariates. For instance, if the covariate is a real valued continuous Gaussian Markov process in $L_p[0,1]$, the convergence rate is $O((\log n)^{-\delta})$ for some $\delta > 0$. Theorem 3.4 and the proof of Theorem 4.3 imply that if $h_n^{2\beta} n\phi(\mathbf{x}, h_n) \to 0$ as $n \to \infty$, $[n\phi(\mathbf{x}, h_n)]^{1/2} c_n[\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})]$ converges in distribution to a Gaussian random element with zero mean as $n \to \infty$, where $c_n = [E_n^{(2)}(\mathbf{x})]^{-1/2} E_n^{(1)}(\mathbf{x})$ is a sequence of positive numbers bounded and bounded away from 0. Note that this corresponds to an under-smoothed kernel estimate of $\Theta(\mathbf{x})$. On the other hand, if $h_n^{2\beta} n\phi(\mathbf{x}, h_n) \to \infty$ as $n \to \infty$, which includes the case of the optimum bandwidth obtained in Theorem 4.2, we have $h_n^{-\beta}[\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})] - h_n^{-\beta}\tilde{B}_n(\mathbf{x}) \to 0$ *in probability* as $n \to \infty$. Here, $\tilde{B}_n(\mathbf{x})$ is a non-random deterministic object described at the beginning of subsection 4.1.

In Ferraty and Vieu (2006), Ferraty et al. (2006), Ferraty et al. (2010) and Chaouch and Laïb (2013, 2015), asymptotic properties of nonparametric regression estimates of different parameters other than the mean of the conditional distribution of the response were investigated. However, they only considered finite dimensional responses, and they did not investigate the problem of optimum convergence rates of nonparametric regression estimates.

In practice, one has to choose the bandwidth h by some data-driven adaptive procedure. Such adaptive choice of bandwidth, when the covariate is functional, has been investigated in Chagny and Roche (2014, 2016) for the kernel estimates of the conditional distribution and the conditional mean of a real valued response, respectively. Their data based bandwidth selection procedure can be suitably adjusted for more general regression problems considered in this paper, and some relevant technical details are described below. Let $\mathbb{H}_n$ be a finite collection of bandwidths with cardinality less than or equal to n such that, for any $h \in \mathbb{H}_n$, $\phi(\mathbf{x}, h) \geq C(\log n)/n$ for a positive constant C. We assume that the decomposition (3.1) holds with $\mathcal{B}$ being a separable Hilbert space, where $F(\mathbf{z})$ is a Holder continuous function of $\mathbf{z}$ with exponent $\beta \in (0, 1]$ for $d(\mathbf{x}, \mathbf{z}) \leq \max\{h_n \,|\, h_n \in \mathbb{H}_n\}$. This condition is a modified version of B(i). We assume B(ii) holds. We also assume that for every integer $k \geq 2$, $\mathbb{E}[\|G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{z}]\|^k \,|\, \mathbf{X} = \mathbf{z}] \leq (1/2)L^k k!$ for some constant $L > 0$, where $d(\mathbf{x}, \mathbf{z}) \leq \max\{h_n \,|\, h_n \in \mathbb{H}_n\}$. Note that this condition is stronger than condition B(iii). We assume that $\mathbb{E}\|R_n(\mathbf{x})\|^2 = o(\mathbb{E}\|B_n(\mathbf{x}) + V_n(\mathbf{x})\|^2)$, or $\mathbb{E}\|R_n(\mathbf{x})\|^2 = o\big(\max\{h_n^{2\beta}, [n\phi(\mathbf{x}, h_n)]^{-1}\}\big)$ as $n \to \infty$, which is stronger than condition B(iv). Define the empirical shifted small ball probability $\widehat{\phi}(\mathbf{x}, h) = (1/n)\sum_{i=1}^n I(d(\mathbf{x}, \mathbf{X}_i) \leq h)$. An estimate of $\mathbb{E}\|V_n(\mathbf{x})\|^2$ is then defined by

$$\widehat{V}_n(\mathbf{x}, h) = \begin{cases} \zeta \log n / \big(n\hat{\phi}(\mathbf{x}, h)\big) \text{ if } \hat{\phi}(\mathbf{x}, h) > 0 \\ \infty \text{ otherwise,} \end{cases}$$

where ζ is a positive constant independent of h, n or the sample. Let $\widehat{\Theta}_n(\mathbf{x}, h)$ be the estimate $\widehat{\Theta}_n(\mathbf{x})$ constructed with bandwidth h. Define

$$\widehat{B}_n(\mathbf{x}, h) = \max_{h' \in \mathbb{H}_n} (\|\widehat{\Theta}_n(\mathbf{x}, h') - \widehat{\Theta}_n(\mathbf{x}, \max\{h, h'\})\|^2 - \widehat{V}_n(\mathbf{x}, h'))_+.$$

$\widehat{V}_n(\mathbf{x}, h)$ estimates the upper bound of the variance term and $\widehat{B}_n(\mathbf{x}, h)$ approximates the bias term. The data-driven choice of bandwidth is $h_n^* = \arg\min_{h \in \mathbb{H}_n} [\widehat{B}_n(\mathbf{x}, h) + \widehat{V}_n(\mathbf{x}, h)]$. Then, under A(i), one can show that

$$\begin{aligned} &\mathbb{E}\|\widehat{\Theta}_n(\mathbf{x}, h_n^*) - \Theta(\mathbf{x})\|^2 \\ &\leq r_1 \min\{h^{2\beta} + L^2 \log n / (n\phi(\mathbf{x}, h)) \,|\, h \in \mathbb{H}_n\} + r_2/n \end{aligned} \tag{5.1}$$

for some constants $r_1, r_2 > 0$ and for all sufficiently large n. The proof of (5.1) can be carried out using arguments which are closely related to the those in the proof of Theorem 2 in Chagny and Roche (2016). The required conditions will be satisfied when $\Theta(\mathbf{x})$ is the conditional mean $\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}]$, or a smooth function of the conditional mean, like $\Gamma(\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}])$ (see Examples 2.0.1, 2.0.2 in Section 2 and Examples 3.0.1, 3.0.2 in Section 3). However, we need a stronger moment condition than B(iii) that we have pointed out above.

While deriving the optimum rate of convergence, we needed the asymptotic order of the shifted small ball probability function in the covariate space. The relationship between the shifted small ball probability function and the centered small ball probability function for Gaussian random elements in Banach and Hilbert spaces has been studied in detail in the literature (see, e.g., Hoffmann-Jorgensen et al. (1979), Li and Shao (2001) and Dereich and Lifshits (2005)). We have used this relationship to derive the optimum convergence rate when the infinite dimensional covariate has appropriate Gaussian distribution. However, no such result concerning shifted small ball probabilities for non-Gaussian random elements in infinite dimensional spaces is known to us, and consequently we could not derive the optimum convergence rate for non-Gaussian covariates in infinite dimensional spaces.

6 Proofs and Mathematical Details

Proof of Theorem 3.1. It follows from the Fréchet differentiability of $\Gamma(\cdot)$ that

$$
\begin{aligned}
&\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x}) \\
&= \Gamma\left(\frac{\sum_{i=1}^n \Psi(\mathbf{Y}_i)K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}\right) - \Gamma(\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X}=\mathbf{x}]) \\
&= \Gamma'\left(\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X}=\mathbf{x}]\right)\left(\frac{\sum_{i=1}^n \Psi(\mathbf{Y}_i)K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))} - \mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X}=\mathbf{x}]\right) \\
&\quad + o\left(\left\|\frac{\sum_{i=1}^n \Psi(\mathbf{Y}_i)K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))} - \mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X}=\mathbf{x}]\right\|\right).
\end{aligned}
$$

Taking $\mathbb{L}(\mathbf{x})(\cdot) = \Gamma'\left(\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X}=\mathbf{x}]\right)(\cdot)$, $G(\mathbf{Y}) = \Psi(\mathbf{Y})$ and $F(\mathbf{X}) = \mathbb{E}[G(\mathbf{Y}) \,|\, \mathbf{X}]$, we have $\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x}) = V_n(\mathbf{x}) + B_n(\mathbf{x}) + R_n(\mathbf{x})$, where $R_n(\mathbf{x}) = o(\|V_n(\mathbf{x}) + B_n(\mathbf{x})\|)$. So, $\mathbb{E}\|R_n(\mathbf{x})\|^2 = o(\mathbb{E}\|V_n(\mathbf{x}) + B_n(\mathbf{x})\|^2)$ as $n \to \infty$. From the discussion about $\mathbb{E}\|B_n(\mathbf{x})\|^2$ and $\mathbb{E}\|V_n(\mathbf{x})\|^2$ in subsection 3.1, we have $\mathbb{E}\|B_n(\mathbf{x})\|^2 = O(h_n^{2\beta})$ and $\mathbb{E}\|V_n(\mathbf{x})\|^2 = O([n\phi(\mathbf{x},h_n)]^{-1})$ as $n \to \infty$. So, $\mathbb{E}\|R_n(\mathbf{x})\|^2 = o(h_n^{2\beta} + [n\phi(\mathbf{x},h_n)]^{-1}) = o\big(\max\{h_n^{2\beta}, [n\phi(\mathbf{x},h_n)]^{-1}\}\big)$ as $n \to \infty$. So, B(iv) is satisfied from a simple application of Markov inequality. □

Proof of Theorem 3.2. Under the assumptions stated in Example 3.0.3, it is easy to verify that conditions B(i) follows from the Holder continuity assumption on $\Theta(\mathbf{z})$, B(ii) follows from the invertibility of $\mathbf{I}(\Theta(\mathbf{x}))$ and B(iii) follows from the continuity of $\mathbf{I}(\Theta(\mathbf{x}))$ as a function of $\mathbf{x}$. We now proceed to verify condition B(iv). Using a Taylor expansion, it is easy to show that

$$
\begin{aligned}
&\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x}) = \\
&\left[\sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \,|\, \eta_i(\mathbf{x})))W_{i,n}(\mathbf{x})\right]^{-1}\left(\sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \,|\, \eta_i(\mathbf{x})))(\Theta(\mathbf{X}_i) - \Theta(\mathbf{x}))W_{i,n}(\mathbf{x})\right) \\
&\quad - \left[\sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \,|\, \eta_i(\mathbf{x})))W_{i,n}(\mathbf{x})\right]^{-1}\left(\sum_{i=1}^n \nabla g(\mathbf{Y}_i \,|\, \Theta(\mathbf{X}_i))W_{i,n}(\mathbf{x})\right),
\end{aligned}
$$

where $\eta_i(\mathbf{x})$ lies between $\Theta(\mathbf{X}_i)$ and $\widehat{\Theta}_n(\mathbf{x})$. Also, under the assumptions in Example 3.0.3, it is straightforward to verify that $\|\sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \,|\, \eta_i(\mathbf{x})))W_{i,n}(\mathbf{x}) -$

$\mathbf{I}(\Theta(\mathbf{x}))\| \to 0$ *in probability* as $n \to \infty$. Consequently, it follows that

$$\begin{aligned}
&\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x}) \\
&= [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^{n} \mathbf{I}(\Theta(\mathbf{x}))(\Theta(\mathbf{X}_i) - \Theta(\mathbf{x})) W_{i,n}(\mathbf{x}) \right) \\
&\quad - [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^{n} \nabla g(\mathbf{Y}_i \,|\, \Theta(\mathbf{X}_i)) W_{i,n}(\mathbf{x}) \right) \\
&\quad + o_P \left(\left\| [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^{n} \mathbf{I}(\Theta(\mathbf{x}))(\Theta(\mathbf{X}_i) - \Theta(\mathbf{x})) W_{i,n}(\mathbf{x}) \right) \right\| \right) \\
&\quad + o_P \left(\left\| [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^{n} \nabla g(\mathbf{Y}_i \,|\, \Theta(\mathbf{X}_i)) W_{i,n}(\mathbf{x}) \right) \right\| \right).
\end{aligned}$$

Taking

$$\begin{aligned}
V_n(\mathbf{x}) &= -[\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^{n} \nabla g(\mathbf{Y}_i \,|\, \Theta(\mathbf{X}_i)) W_{i,n}(\mathbf{x}) \right) \\
\text{and } B_n(\mathbf{x}) &= [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^{n} \mathbf{I}(\Theta(\mathbf{x}))(\Theta(\mathbf{X}_i) - \Theta(\mathbf{x})) W_{i,n}(\mathbf{x}) \right),
\end{aligned}$$

we have $R_n(\mathbf{x}) = o_{\mathbb{P}}\left(\|B_n(\mathbf{x})\| + \|V_n(\mathbf{x})\|\right)$, and the proof is complete using the upper bounds of $\mathbb{E}\|B_n(\mathbf{x})\|^2$ and $\mathbb{E}\|V_n(\mathbf{x})\|^2$ derived and discussed in subsection 3.1. □

Proof of Theorem 3.3. The arguments used in this proof are closely related to the arguments in the proof of Proposition 1 in Chagny and Roche (2016). Define $W_n(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} E_n^{-1} K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))$, where $E_n = \mathbb{E}[K(h_n^{-1} d(\mathbf{x}, \mathbf{X}))]$. It follows from Bernstein's inequality and condition A(i) that

$$\mathbb{P}[|W_n(\mathbf{x}) - 1| > (1/2)] \leq 2 \exp(-c' n \phi(\mathbf{x}, h_n)), \tag{6.1}$$

where c' is a positive constant (see, e.g., Lemma 4 in Chagny and Roche (2016)). Note that

$$\begin{aligned}
&\mathbb{E}\|V_n(\mathbf{x})\|^2 \\
&= \mathbb{E}[\|V_n(\mathbf{x})\|^2 I(W_n(\mathbf{x}) < (1/2))] + \mathbb{E}[\|V_n(\mathbf{x})\|^2 I(W_n(\mathbf{x}) \geq (1/2))].
\end{aligned} \tag{6.2}$$

For the first term on the RHS in (6.2), using the fact that $\mathcal{B}$ is a type 2 Banach space and condition B(iii), we have

$$\begin{aligned}
&\mathbb{E}[\|V_n(\mathbf{x})\|^2 I(W_n(\mathbf{x}) < (1/2))] \\
&= \mathbb{E}[\mathbb{E}[\|V_n(\mathbf{x})\|^2 I(W_n(\mathbf{x}) < (1/2)) \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_n]] \\
&\leq c' \mathbb{E}\left[\frac{\sum_{i=1}^{n} \mathbb{E}[\|G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \,|\, \mathbf{X}_i]\|^2 \,|\, \mathbf{X}_i] K^2(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{(\sum_{i=1}^{n} K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)))^2} I(W_n(\mathbf{x}) < (1/2)) \right] \\
&\leq c'' \mathbb{P}[|W_n(\mathbf{x}) - 1| > (1/2)],
\end{aligned} \tag{6.3}$$

where c' and c'' are positive constants. From (6.1), (6.3) and the fact that $ue^{-u} \leq e^{-1}$ for $u > 0$, we get that $n\phi(\mathbf{x}, h_n) \mathbb{E}[\|V_n(\mathbf{x})\|^2 I(W_n(\mathbf{x}) < (1/2))]$ is uniformly bounded over n.

Now, for the second term on the RHS in (6.2), we get

$$
\begin{aligned}
&\mathbb{E}[\|V_n(\mathbf{x})\|^2 I(W_n(\mathbf{x}) \geq (1/2))] \\
&\leq c'\mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{E}[\|G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \,|\, \mathbf{X}_i]\|^2 \,|\, \mathbf{X}_i] K^2(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{(nE_n W_n(\mathbf{x}))^2} I(W_n(\mathbf{x}) \geq (1/2))\right] \\
&\leq c''' E_n^{-2} \mathbb{E}[K^2(h_n^{-1} d(\mathbf{x}, \mathbf{X}))], \qquad (6.4)
\end{aligned}
$$

where c''' is some positive constant. From condition A(i), the definition of E_n and (6.4), it follows that $n\phi(\mathbf{x}, h_n)\mathbb{E}[\|V_n(\mathbf{x})\|^2 I(W_n(\mathbf{x}) \geq (1/2))]$ is uniformly bounded over n. Therefore, $n\phi(\mathbf{x}, h_n)\mathbb{E}\|V_n(\mathbf{x})\|^2$ is uniformly bounded over n. □

Proof of Theorem 3.4. Recall that $E_n = \mathbb{E}K(h_n^{-1} d(\mathbf{x}, \mathbf{X}))$. Note that

$$
\begin{aligned}
&[E_n^{(2)}(\mathbf{x})]^{-1/2} E_n^{(1)}(\mathbf{x}) V_n(\mathbf{x}) \\
&= \frac{n^{-1} \sum_{i=1}^n \frac{E_n^{(1)}(\mathbf{x})}{[E_n^{(2)}(\mathbf{x})]^{1/2}} \frac{K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{E_n} \mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \,|\, \mathbf{X}_i])}{n^{-1} \sum_{i=1}^n E_n^{-1} K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}. \qquad (6.5)
\end{aligned}
$$

Denote

$$
\tilde{V}_n(\mathbf{x}) = n^{-1} \sum_{i=1}^n \frac{E_n^{(1)}(\mathbf{x})}{[E_n^{(2)}(\mathbf{x})]^{1/2}} \frac{K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{E_n} \mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \,|\, \mathbf{X}_i]).
$$

If conditions B(v) and B(vi) hold, then using A(i) and A(ii), it is easy to verify that the covariance of $[n\phi(\mathbf{x}, h_n)]^{1/2} \tilde{V}_n(\mathbf{x})$ converges to $\mathbb{D}(\cdot, \cdot \,|\, \mathbf{x})$ in the trace norm as $n \to \infty$. This ensures that the conditions in Theorem 1.1 in Kundu et al. (2000) are satisfied, and consequently $[n\phi(\mathbf{x}, h_n)]^{1/2} \tilde{V}_n(\mathbf{x}) \to \mathbf{W}$ *in distribution* as $n \to \infty$.

Under conditions A(i) and A(ii), the denominator in the RHS of (6.5) $n^{-1} \sum_{i=1}^n E_n^{-1} K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)) \to 1$ *in probability* as $n \to \infty$. Hence the proof is complete by Slutsky's Theorem. □

Proof of Theorem 4.1. For a sequence of bandwidths $\{h_n\}$ satisfying A(ii), it follows from the upper bound of $\phi(\mathbf{x}, h)$ in (3.5) that

$$
\begin{aligned}
&nC_3 h_n^{t_4} \exp\left[-C_4 m(h_n)\right] \geq n\phi(\mathbf{x}, h_n) \longrightarrow \infty \text{ as } n \longrightarrow \infty \\
&\Longrightarrow \log n - t_4 \log(1/h_n) - C_4 m(h_n) \longrightarrow \infty \\
&\Longleftrightarrow m(h_n)\left[\frac{\log n}{m(h_n)} - t_4 \frac{\log(1/h_n)}{m(h_n)} - C_4\right] \longrightarrow \infty \qquad (6.6)
\end{aligned}
$$

as $n \to \infty$. Now, since either $t_2 > 0$ or $t_3 > 1$, we have

$$
\frac{\log(1/h_n)}{m(h_n)} \longrightarrow 0 \text{ as } n \longrightarrow \infty. \qquad (6.7)
$$

Also, since $h_n \to 0$ from A(ii),

$$
m(h_n) \longrightarrow \infty \text{ as } n \longrightarrow \infty. \qquad (6.8)
$$

Hence, for (6.6) to be satisfied, in view of (6.7) and (6.8), we must have

$$
\frac{\log n}{m(h_n)} - C_4 > 0 \text{ for all sufficiently large } n
$$

$$\begin{aligned}
&\iff \frac{\log n}{C_4} > m(h_n) \text{ for all sufficiently large } n \\
&\iff m^{-1}\left(\frac{\log n}{C_4}\right) < h_n \text{ for all sufficiently large } n \\
&\implies \frac{h_n}{m^{-1}(\log n)} > C_4' > 0 \text{ for all sufficiently large } n,
\end{aligned}$$

where C_4' is a constant depending on C_4. □

Proof of Theorem 4.2. From the upper bounds of $\mathbb{E}\|B_n(\mathbf{x})\|^2$ and $\mathbb{E}\|V_n(\mathbf{x})\|^2$ derived in subsection 3.1 and the lower bound of $\phi(\mathbf{x}, h_n)$ in (3.5), it follows that $\mathbb{E}\|B_n(\mathbf{x}) + V_n(\mathbf{x})\|^2$ is asymptotically bounded above by $f_1(h_n)$, where

$$f_1(h_n) = ah_n^{2\beta} + \frac{b}{nC_1}(1/h_n)^{t_1} \exp\left[C_2 m(h_n)\right],$$

and $a, b > 0$ are some constants. Note that $m(h)$ is a differentiable function of h, and

$$m'(h) = -m(h)(1/h)\left(t_2 + \frac{t_3}{\log(1/h)}\right). \tag{6.9}$$

It follows that for each n, $f_1(h)$ is a twice differentiable function of h, and we have

$$\begin{aligned}
&f_1'(h) \\
&= 2\beta a h^{2\beta-1} - \frac{bt_1}{nC_1}(1/h)^{t_1+1}\exp[C_2 m(h)] \\
&\quad - \frac{bC_2}{nC_1}(1/h)^{t_1+1}\exp[C_2 m(h)]m(h)\left(t_2 + \frac{t_3}{\log(1/h)}\right),
\end{aligned} \tag{6.10}$$

and

$$\begin{aligned}
&f_1''(h) \\
&= \left[2\beta(2\beta-1)ah^{2(\beta-1)} + \frac{bt_1(t_1+1)}{nC_1}(1/h)^{t_1+2}\exp[C_2 m(h)]\right] \\
&\quad + \frac{bC_2(2t_1+1)}{nC_1}(1/h)^{t_1+2}\exp[C_2 m(h)]m(h)\left(t_2 + \frac{t_3}{\log(1/h)}\right) \\
&\quad + \frac{bC_2^2}{nC_1}(1/h)^{t_1+2}\exp[C_2 m(h)](m(h))^2\left(t_2 + \frac{t_3}{\log(1/h)}\right)^2 \\
&\quad + \frac{bC_2}{nC_1}(1/h)^{t_1+2}\exp[C_2 m(h)]m(h) \\
&\quad \times \left[\left(t_2 + \frac{t_3}{\log(1/h)}\right)^2 - t_3\left(\frac{1}{\log(1/h)}\right)^2\right].
\end{aligned} \tag{6.11}$$

Now, from (6.10), $f_1'(h_n) = 0$ implies that

$$\begin{aligned}
&2\beta a h_n^{2\beta-1} \\
&= n^{-1}(1/h_n)^{t_1+1}\exp[C_2 m(h_n)]
\end{aligned}$$

$$
\begin{aligned}
&\times \left[\frac{bt_1}{C_1} + \frac{bC_2}{C_1} m(h_n) \left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right] \\
\iff\ & h_n^{2\beta+t_1} \\
&= n^{-1} \exp[C_2 m(h_n)] \\
&\quad \times \left[\frac{bt_1}{2\beta a C_1} + \frac{bC_2}{2\beta a C_1} m(h_n) \left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right] \qquad (6.12) \\
\iff\ & -(2\beta + t_1)\log(1/h_n) \\
&= -\log n + C_2 m(h_n) \\
&\quad + \log\left(\left[\frac{bt_1}{2\beta a C_1} + \frac{bC_2}{2\beta a C_1} m(h_n) \left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right]\right) \\
\iff\ & -(2\beta + t_1)\frac{\log(1/h_n)}{m(h_n)} \\
&= -\frac{\log n}{m(h_n)} + C_2 + \frac{1}{m(h_n)} \qquad (6.13) \\
&\quad \times \log\left(\left[\frac{bt_1}{2\beta a C_1} + \frac{bC_2}{2\beta a C_1} m(h_n) \left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right]\right).
\end{aligned}
$$

Now, if either $t_2 > 0$ or $t_3 > 1$, then for all $\beta > 0$,

$$
\frac{\log(1/h_n)}{m(h_n)} \longrightarrow 0 \qquad (6.14)
$$

and

$$
\frac{1}{m(h_n)} \log\left(\left[\frac{bt_1}{2\beta a C_1} + \frac{bC_2}{2\beta a C_1} m(h_n) \left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right]\right) \longrightarrow 0 \qquad (6.15)
$$

as $n \to \infty$. Combining (6.13), (6.14) and (6.15), we have

$$
\frac{\log n}{m(h_n)} \longrightarrow C_2 \text{ as } n \longrightarrow \infty. \qquad (6.16)
$$

It follows using (6.16) and the lower bound of $\phi(\mathbf{x}, h_n)$ in (3.5) that $\{h_n\}$ satisfies A(ii) if either $t_2 > 0$ or $t_3 > 1$, where $f_1'(h_n) = 0$ for each n. From (6.16) we also have that for all sufficiently large n,

$$
\log n < 2C_2 m(h_n) \implies a h_n^{2\beta} < C_2' (m^{-1}(\log n))^{2\beta}
$$

for some positive constant C_2' depending on C_2 and a.

Since either $t_2 > 0$ or $t_3 > 1$, it follows from (6.11) that $f_1''(h_n) > 0$ whenever $f_1'(h_n) = 0$ for all sufficiently large n. So, for all sufficiently large n, h_n minimizes $f_1(h_n)$ whenever $f_1'(h_n) = 0$, and the minimizer must be unique. Also, from (6.12) it follows that for h_n satisfying $f_1'(h) = 0$, $a h_n^{2\beta} < f_1(h_n) < 2a h_n^{2\beta}$ for all sufficiently large n, and consequently $f_1(h_n) < C_2'(m^{-1}(\log n))^{2\beta}$ for all sufficiently large n and for this choice of the sequence of bandwidths $\{h_n\}$. Hence $\mathbb{E}\|B_n(\mathbf{x}) + V_n(\mathbf{x})\|^2 \le C_2' \left(m^{-1}(\log n)\right)^{2\beta}$ for all sufficiently large n and for the bandwidth sequence $\{h_n\}$ minimizing $f_1(h)$ for every fixed n. So, for this choice of $\{h_n\}$, $\|B_n(\mathbf{x}) + V_n(\mathbf{x})\| = O_{\mathbb{P}}(\left(m^{-1}(\log n)\right)^{\beta})$ as $n \to \infty$. Also, from (6.12) and the lower bound of $\phi(\mathbf{x}, h)$ in (3.5), we get $h_n^{\beta}/[[n\phi(\mathbf{x}, h_n)]^{-1}] \to \infty$

as $n \to \infty$. Hence, $\left(m^{-1}(\log n)\right)^{-\beta} \|R_n(\mathbf{x})\| \to 0$ *in probability* as $n \to \infty$. Therefore, $\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\| = O_{\mathbb{P}}\big(\left(m^{-1}(\log n)\right)^{\beta} \big)$ as $n \to \infty$. $\square$

Proof of Theorem 4.3. Recall that $B_n(\mathbf{x}) = \tilde{B}_n(\mathbf{x}) + \tilde{R}_n(\mathbf{x})$, where $\tilde{R}_n(\mathbf{x}) = o_{\mathbb{P}}(h_n^{\beta})$, and $\tilde{B}_n(\mathbf{x})$ is a non-random quantity. We choose $\Theta(\mathbf{x})$ satisfying C(i), so that for any kernel $K(\cdot)$ satisfying A(i) and any sequence of bandwidths $\{h_n\}$ satisfying A(ii), we have $h_n^{-\beta}\|\tilde{B}_n(\mathbf{x})\| \geq b_1 > 0$ for all sufficiently large n and some constant b_1. So, we have $\mathbb{P}[h_n^{-\beta}\|B_n(\mathbf{x})\| \geq b_1/2] \to 1$ as $n \to \infty$ for any sequence of bandwidths $\{h_n\}$ satisfying A(ii) and any kernel $K(\cdot)$ satisfying A(i).

Denote $a_0 = \|\tilde{\phi}_{i_0}\| > 0$. Then, for all $\mathbf{v} \in \mathcal{B}$, $|\tilde{\phi}_{i_0}(\mathbf{v})| \leq a_0\|v\|$. So,

$$\mathbb{P}[\|V_n(\mathbf{x})\| > c] \geq \mathbb{P}[|\tilde{\phi}_{i_0}(V_n(\mathbf{x}))| > a_0 c] \tag{6.17}$$

for any positive number c.

Using A(i), A(ii), B(vi), C(iii) and arguments similar to those in Theorem 3.4, it is easy to show that $[n\phi(\mathbf{x}, h_n)]^{1/2}[E_n^{(2)}(\mathbf{x})]^{-1/2}E_n^{(1)}(\mathbf{x})\tilde{\phi}_{i_0}(V_n(\mathbf{x}))$ converges *in distribution* to a standard normal random variable as $n \to \infty$. Also, $0 < L^{-1}l \leq [E_n^{(2)}(\mathbf{x})]^{-1/2}E_n^{(1)}(\mathbf{x}) \leq l^{-1}L < \infty$ for all n. Hence, given $\delta < 1$, we can find $c_0 > 0$ such that $\mathbb{P}\big[[n\phi(\mathbf{x}, h_n)]^{1/2}|\tilde{\phi}_{i_0}(V_n(\mathbf{x}))| > 2a_0c_0\big] > \delta$ for all sufficiently large n. Therefore, using (6.17) we have $\mathbb{P}\big[[n\phi(\mathbf{x}, h_n)]^{1/2}\|V_n(\mathbf{x})\| > 2c_0\big] > \delta$ for all sufficiently large n. Now, we consider two cases separately.

Case 1: We assume that $\{h_n\}$ satisfying A(ii) is such that

$$h_n^{2\beta} n\phi(\mathbf{x}, h_n) \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{6.18}$$

Since $\{h_n\}$ satisfies A(ii), from Theorem 4.1, we get $h_n^{\beta}/\left(m^{-1}(\log n)\right)^{\beta} > C_4'' > 0$ for all sufficiently large n and a constant C_4. So, for all sufficiently large n,

$$\begin{aligned} &[n\phi(\mathbf{x}, h_n)]^{-1/2} > h_n^{\beta} > C_4''\left(m^{-1}(\log n)\right)^{\beta} \\ \Longrightarrow\ &\left(m^{-1}(\log n)\right)^{-\beta}[n\phi(\mathbf{x}, h_n)]^{-1/2} > C_4''. \end{aligned}$$

We also have

$$\begin{aligned} &\mathbb{P}\big[\left(m^{-1}(\log n)\right)^{-\beta}\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\| > c\big] \\ &= \mathbb{P}\left[\left(m^{-1}(\log n)\right)^{-\beta}[n\phi(\mathbf{x}, h_n)]^{-1/2}[n\phi(\mathbf{x}, h_n)]^{1/2}\|\tilde{B}_n(\mathbf{x}) + V_n(\mathbf{x}) + Q_n(\mathbf{x})\| > c\right], \end{aligned}$$

where $Q_n(\mathbf{x}) = R_n(\mathbf{x}) + \tilde{R}_n(\mathbf{x}) = o_{\mathbb{P}}\big(\max\{h_n^{\beta}, [n\phi(\mathbf{x}, h_n)]^{-1/2}\}\big)$ as $n \to \infty$. So, in this case, $Q_n(\mathbf{x}) = o_{\mathbb{P}}\big([n\phi(\mathbf{x}, h_n)]^{-1/2}\big)$ as $n \to \infty$. Note that here $[n\phi(\mathbf{x}, h_n)]^{1/2}\tilde{B}_n(\mathbf{x}) \to 0$ as $n \to \infty$, since $h_n^{-\beta}\tilde{B}_n(\mathbf{x})$ is bounded uniformly over n. So, $\mathbb{P}\big[[n\phi(\mathbf{x}, h_n)]^{1/2}\|\tilde{B}_n(\mathbf{x}) + V_n(\mathbf{x}) + Q_n(\mathbf{x})\| > c_0\big] > \delta$ for all sufficiently large n. Consequently, for all sufficiently large n,

$$\begin{aligned} &\mathbb{P}\big[\left(m^{-1}(\log n)\right)^{-\beta}\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\| > C_4''c_0\big] \\ &\geq \mathbb{P}\big[C_4''[n\phi(\mathbf{x}, h_n)]^{1/2}\|\tilde{B}_n(\mathbf{x}) + V_n(\mathbf{x}) + Q_n(\mathbf{x})\| > C_4''c_0\big] \\ &= \mathbb{P}\big[[n\phi(\mathbf{x}, h_n)]^{1/2}\|\tilde{B}_n(\mathbf{x}) + V_n(\mathbf{x}) + Q_n(\mathbf{x})\| > c_0\big] \\ &> \delta. \end{aligned}$$

Hence, taking $c = C_4'' c_0$, the proof is complete when $\{h_n\}$ satisfies (6.18).

Case 2: We now assume that $\{h_n\}$ satisfying A(ii) is such that

$$h_n^{2\beta} n\phi(\mathbf{x}, h_n) > \epsilon_0 > 0 \tag{6.19}$$

for all sufficiently large n and some ϵ_0. Then, from (3.5) we have that for all sufficiently large n,

$$\begin{aligned}
& C_3 n h_n^{t_4+2\beta} \exp\left[-C_4 m(h_n)\right] \geq h_n^{2\beta} n\phi(\mathbf{x}, h_n) > \epsilon_0 > 0 \\
\implies & \log n - (t_1 + 2\beta)\log(1/h_n) - C_4 m(h_n) > \log \frac{\epsilon_0}{C_3} \\
\iff & m(h_n)\left[\frac{\log n}{m(h_n)} - \frac{(t_1+2\beta)\log(1/h_n)}{m(h_n)} - C_4\right] > \log\frac{\epsilon_0}{C_3}. \qquad (6.20)
\end{aligned}$$

Now, since $\{h_n\}$ satisfies A(ii), and either $t_2 > 0$ or $t_2 = 0$ with $t_3 > 0$, we have

$$m(h_n) \longrightarrow \infty \text{ and } \frac{\log(1/h_n)}{m(h_n)} \longrightarrow 0 \tag{6.21}$$

as $n \to \infty$. Therefore, given any $\epsilon > 0$, in view of (6.20) and (6.21), we must have, for all sufficiently large n,

$$\frac{\log n}{m(h_n)} - C_4 > -\epsilon \implies m(h_n) < \frac{\log n}{C_4 - \epsilon}. \tag{6.22}$$

Taking $\epsilon = C_4/2$ and from (6.22), we have, for all sufficiently large n,

$$m(h_n) < \frac{2}{C_4}\log n \implies h_n > m^{-1}\left(\frac{2}{C_4}\log n\right) > C_4''' m^{-1}(\log n),$$

where $C_4''' > 0$ is a constant depending on C_4. So, for all sufficiently large n,

$$(m^{-1}(\log n))^{-\beta} h_n^{\beta} > (C_4''')^{\beta} > 0. \tag{6.23}$$

By (6.19), it follows that $h_n^{-\beta}[n\phi(\mathbf{x}, h_n)]^{-1/2}$ is bounded uniformly on n. Therefore, using similar arguments as in *Case 1* and (6.23), the proof follows in this case also. $\square$

Proof of Theorem 4.4. From (6.12) in the proof of Theorem 4.2 and the lower bound of $\phi(\mathbf{x}, h)$ in (3.5), it follows that

$$h_n^{2\beta} n\phi(\mathbf{x}, h_n) \longrightarrow \infty \tag{6.24}$$

as $n \to \infty$. Now, choose $\Theta(\cdot)$ as described in the proof of Theorem 4.3 such that $h_n^{-\beta}\|\tilde{B}_n(\mathbf{x})\| \geq b_1 > 0$ for a constant b_1 and all sufficiently large n. So, $\mathbb{P}[h_n^{-\beta}\|B_n(\mathbf{x})\| > b_1/2] \to 1$ as $n \to \infty$. Hence, for this choice of $\Theta(\cdot)$ and using Theorem 3.4 and (6.24), we have

$$\begin{aligned}
\frac{\|V_n(\mathbf{x})\|}{\|B_n(\mathbf{x})\|} &= \frac{[n\phi(\mathbf{x}, h_n)]^{1/2}\|V_n(\mathbf{x})\|}{[n\phi(\mathbf{x}, h_n)]^{1/2} h_n^{\beta} h_n^{-\beta}\|B_n(\mathbf{x})\|} \\
&= \frac{1}{[n\phi(\mathbf{x}, h_n)]^{1/2} h_n^{\beta}} \frac{[n\phi(\mathbf{x}, h_n)]^{1/2}\|V_n(\mathbf{x})\|}{h_n^{-\beta}\|B_n(\mathbf{x})\|} \\
&\longrightarrow 0
\end{aligned}$$

in probability as $n \to \infty$. $\square$

References

Aerts, M. and Claeskens, G. (1997). Local polynomial estimation in multiparameter likelihood models. *Journal of the American Statistical Association*, 92(440):1536–1545.

Araujo, A. and Giné, E. (1980). *The central limit theorem for real and Banach valued random variables*. Wiley, New York.

Cameron, R. H. and Martin, W. T. (1944). Transformations of Weiner integrals under translations. *Annals of Mathematics*, 45(2):386–396.

Chagny, G. and Roche, A. (2014). Adaptive and minimax estimation of the cumulative distribution function given a functional covariate. *Electronic Journal of Statistics*, 8(2):2352–2404.

Chagny, G. and Roche, A. (2016). Adaptive estimation in the functional nonparametric regression model. *Journal of Multivariate Analysis*, 146:105–118.

Chaouch, M. and Laïb, N. (2013). Nonparametric multivariate L_1-median regression estimation with functional covariates. *Electronic Journal of Statistics*, 7:1553–1586.

Chaouch, M. and Laïb, N. (2015). Vector-on-function quantile regression for stationary ergodic processes. *Journal of the Korean Statistical Society*, 44(2):161–178.

Chaudhuri, P. and Dewanji, A. (1995). On a likelihood-based approach in nonparametric smoothing and cross-validation. *Statistics & Probability Letters*, 22(1):7–15.

Dereich, S. and Lifshits, M. (2005). Probabilities of randomly centered small balls and quantization in Banach spaces. *Annals of Probability*, 33(4):1397–1421.

Donoho, D. L. and Liu, R. C. (1991a). Geometrizing rates of convergence, II. *The Annals of Statistics*, 19(2):633–667.

Donoho, D. L. and Liu, R. C. (1991b). Geometrizing rates of convergence, III. *The Annals of Statistics*, 19(2):668–701.

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical Planning and Inference*, 140(2):335–352.

Ferraty, F., Laksaci, A., and Vieu, P. (2006). Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statistical Inference for Stochastic Processes*, 9(1):47–76.

Ferraty, F., Mas, A., and Vieu, P. (2007). Nonparametric regression on functional data: inference and practical aspects. *Australian & New Zealand Journal of Statistics*, 49(3):267–286.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, New York.

Hardle, W. (1990). Applied nonparametric regression. *Cambridge University Press, UK.*

Hoffmann-Jorgensen, J., Shepp, L. A., and Dudley, R. M. (1979). On the lower tail of Gaussian seminorms. *The Annals of Probability*, 7(2):319–342.

Ibragimov, I. A. and Haśminskii, R. Z. (1980). On nonparametric estimation of regression. *Soviet Math. Dokl.*, 21:810–814.

Kundu, S., Majumdar, S., and Mukherjee, K. (2000). Central limit theorems revisited. *Statistics & Probability Letters*, 47(3):265–275.

Li, W. V. (2001). Small ball probabilities for Gaussian Markov processes under the L_p-norm. *Stochastic processes and their applications*, 92(1):87–102.

Li, W. V. and Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications. *Stochastic processes: theory and methods*, 19:533–597.

Lifshits, M. A. (2013). *Gaussian random functions*, volume 322. Springer Science & Business Media, Dordrecht.

Lukić, M. and Beder, J. (2001). Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969.

Mas, A. (2012). Lower bound in regression for functional data by representation of small ball probabilities. *Electronic Journal of Statistics*, 6:1745–1778.

Masry, E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Processes and their Applications*, 115(1):155–177.

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

Rachdi, M. and Vieu, P. (2007). Nonparametric regression for functional data: automatic smoothing parameter selection. *Journal of Statistical Planning and Inference*, 137(9):2784–2801.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84(405):276–283.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372.

Yatracos, Y. G. (1988). A lower bound on the error in nonparametric regression type problems. *The Annals of Statistics*, 16(3):1180–1187.