Journal of Multivariate Analysis 100 (2009) 126–136

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Variance function estimation in multivariate nonparametric regression with fixed design

T. Tony Cai [a], Michael Levine [b,*], Lie Wang [a]

[a] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, United States

[b] Purdue University, 250 N. University Street, West Lafayette, IN 47907, United States

ARTICLE INFO

Article history:
Received 11 May 2007
Available online 3 April 2008

AMS 2000 subject classifications:
primary 62G08
62G20

Keywords:
Minimax estimation
Nonparametric regression
Variance estimation

ABSTRACT

Variance function estimation in multivariate nonparametric regression is considered and the minimax rate of convergence is established in the iid Gaussian case. Our work uses the approach that generalizes the one used in [A. Munk, Bissantz, T. Wagner, G. Freitag, On difference based variance estimation in nonparametric regression when the covariate is high dimensional, J. R. Stat. Soc. B 67 (Part 1) (2005) 19–41] for the constant variance case. As is the case when the number of dimensions $d = 1$, and very much contrary to standard thinking, it is often not desirable to base the estimator of the variance function on the residuals from an optimal estimator of the mean. Instead it is desirable to use estimators of the mean with minimal bias. Another important conclusion is that the first order difference based estimator that achieves minimax rate of convergence in the one-dimensional case does not do the same in the high dimensional case. Instead, the optimal order of differences depends on the number of dimensions.

1. Introduction

We consider the multivariate nonparametric regression problem

$$y_i = g(\boldsymbol{x}_i) + V^{\frac{1}{2}}(\boldsymbol{x}_i)z_i \tag{1}$$

where $y_i \in \mathbb{R}$, $\boldsymbol{x}_i \in S = [0, 1]^d \subset \mathbb{R}^d$ while z_i are iid random variables with zero mean and unit variance and have bounded absolute fourth moments: $E|z_i| \leq \mu_4 < \infty$. We use the bold font to denote any d-dimensional vectors with $d > 1$ (except d-dimensional indices) and regular font for scalars. The design is assumed to be a fixed equispaced d-dimensional grid; in other words, we consider $\boldsymbol{x}_i = \{x_{i_1}, \ldots, x_{i_d}\}' \in \mathbb{R}^d$ where $i_k = 0, 1, \ldots, m$ for $k = 1, \ldots, d$. Each coordinate is defined as

$$x_{i_k} = \frac{i_k}{m} \tag{2}$$

for $k = 1, \ldots, d$. The overall sample size is $n = m^d$. The index i used in the model (1) is a d-dimensional index $i = (i_1, \ldots, i_d)$. Both $g(\mathbf{x})$ and $V(\mathbf{x})$ are unknown functions supported on $S = [0, 1]^d$; we also assume that $V(\mathbf{x}) > 0$. The minimax rate of convergence for the estimator $\hat{V}$ under different smoothness assumptions on g is the main subject of interest. The estimation accuracy for $\hat{V}$ is measured both globally by the mean integrated squared error (MISE)

$$R(\hat{V}, V) = E\int_{R^d} (\hat{V}(\boldsymbol{x}) - V(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \tag{3}$$

* Corresponding author.
E-mail address: mlevins@stat.purdue.edu (M. Levine).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2008.03.007

and locally by the mean squared error at a point (pointwise risk)

$$R(\hat{V}(\boldsymbol{x}_*), V(\boldsymbol{x}_*)) = E(\hat{V}(\boldsymbol{x}_*) - V(\boldsymbol{x}_*))^2. \quad (4)$$

We are particularly interested in finding how the difficulty of estimating V depends on the smoothness of the mean function g as well as the smoothness of the variance function V itself.

Variance function estimation in heteroskedastic nonparametric regression is important in many contexts. Previous work has mainly focused on the univariate regression model. See, for example, [12,13,6,17,4]. More recent work includes [1,21]. In the multidimensional setup of (1), the problem has been considered in [18,15] in the special case of a constant variance function $V(x) \equiv \sigma^2$. Spokoiny [18] investigated the effect of the dimensionality d on the estimation accuracy of σ^2 while assuming that the regression function g is twice continuously differentiable. A rate optimal procedure is constructed using residuals of a local linear fit. Munk et al. [15] used a difference based approach to variance estimation and studied the effects of both the smoothness of g and dimensionality d on the optimal rate of convergence for estimating σ^2.

Munk et al. [15] noted that "... Difference estimators are only applicable when homogeneous noise is present, i.e. the error variance does not depend on the regressor" ([15], p. 20). In the present paper we extend the difference based approach of Munk et al. [15] to the case of the non-homogeneous (heteroskedastic) situation where the variance V is a function of the regressor $\mathbf{x}$. This paper is also closely connected to Wang et al. [21] where a first order difference based procedure for variance function estimation was studied in the one-dimensional case. The present paper considers variance function estimation in the multidimensional case which has some different characteristics from those in the one-dimensional case. In particular, first order differences are inadequate in the high dimensional case. In fact, as in the constant variance case, it is no longer possible to use any fixed order differences and achieve asymptotically a minimax rate of convergence for an arbitrary number of dimensions $d > 1$. The order of differences needs to grow with the number of dimensions d.

We show that the minimax rate of convergence for estimating the variance function V under both the pointwise squared error and global integrated mean squared error is

$$\max\left\{n^{-\frac{4\alpha}{d}}, n^{-\frac{2\beta}{2\beta+d}}\right\} \quad (5)$$

if g has α derivatives, V has β derivatives and d is the number of dimensions; these results are obtained in the iid Gaussian case. So the minimax rate depends on the smoothness of both V and g. The minimax upper bound is obtained by using kernel smoothing of the squared differences of observations. The order of the difference scheme used depends on the number of dimensions d. The minimum order needs to be $\gamma = \lceil d/4 \rceil$, the smallest integer larger than or equal to $d/4$. With such a choice of the difference sequence our estimator is adaptive with respect to the smoothness of the mean function g. The derivation of the minimax lower bound is based on a moment matching technique and a two-point testing argument. A key step is studying a hypothesis testing problem where the alternative hypothesis is a Gaussian location mixture with a special moment matching property.

It is also interesting to note that, if V is known to belong to a regular parametric model, such as the set of positive polynomials of a given order, the cutoff for the smoothness of g on the estimation of V is $d/4$. That is, if g has at least $d/4$ derivatives then the minimax rate of convergence for estimating V is solely determined by the smoothness of V as if g were known. On the other hand, if g has less than $d/4$ derivatives then the minimax rate depends on the relative smoothness of both g and V and, for sufficiently small α, will be completely determined by it. The larger d is, the smoother the mean function g has to be in order not to influence the minimax rate of convergence for estimating the variance function V.

The paper is organized as follows. Section 2 presents an upper bound for the minimax risk while Section 3 derives a rate-sharp lower bound for the minimax risk under both global and local losses. The lower and upper bounds together yield the minimax rate of convergence. Section 4 contains a detailed discussion of results obtained and their implications for practical variance estimation in the nonparametric regression. The proofs are given in Section 5.

2. Upper bound

In this section we shall construct a kernel variance estimator based on squared differences of observations given in (1). Note that it is possible to consider a more general design where not all $m_k \equiv m$, $k = 1, \dots, d$ and x_{i_k} is defined as a solution of the equation $\frac{i_k}{m_k} = \int_{-\infty}^{x_{i_k}} f_k(s)\mathrm{d}s$ for a set of strictly positive densities $f_k(s)$. We will adhere to a simpler design (2) throughout this paper.

Difference based estimators have a long history for estimating a constant variance in univariate nonparametric regression. See, for example, [19,20,16,7,9,2]. The multidimensional case was first considered when the dimensionality $d = 2$ in [8]. The general case of estimating a constant variance in arbitrary dimension has only recently been investigated in [15]. The estimation of the variance function $V(\boldsymbol{x})$ that depends on the covariate is a more recent topic. For the one-dimensional case, we can mention [12,13,1]. The multidimensional case, to the best of our knowledge, has not been considered before.

The following notation will be used throughout the paper. Define a multi-index $J = \{j_1, \dots, j_d\}$ as a sequence of nonnegative integers $j_1, \dots, j_d$. For a fixed positive integer l, let $J(l) = \{J = (j_1, j_2, \dots, j_d : |J| = j_1 + j_2 + \cdots + j_d = l)\}$. Note that $|J|$ stands for the sum of all elements of vector J. For an arbitrary function f, we define $D^{J(l)}f = \frac{\partial^l f(\cdot)}{\partial x_1^{j_1} \dots \partial x_d^{j_d}}$ for all J such that $|J| = l$. For an arbitrary $\boldsymbol{x} \in \mathbb{R}^d$ we define $\boldsymbol{x}^J = x_1^{j_1} \dots x_d^{j_d}$. Also, for any vector $\boldsymbol{u}$ and real number v, the set $B = \boldsymbol{u} + vA$ is the set of all vectors $\{\boldsymbol{y} \in \mathbb{R}^d : \boldsymbol{y} = \boldsymbol{u} + v\boldsymbol{a}$ for some $\boldsymbol{a} \in A \subset \mathbb{R}^d\}$. For any positive integer α, let $\lfloor \alpha \rfloor$ denote the largest integer that is strictly less than α, $\lceil \alpha \rceil$ the smallest integer that is greater than α, and $\alpha' = \alpha - \lfloor \alpha \rfloor$. Now we can state the functional class definition that we need.

Definition 1. For any $\alpha > 0$ and $M > 0$, we define the Lipschitz class $\Lambda^\alpha(M)$ as the set of all functions $f(\boldsymbol{x}) : [0, 1]^d \to \mathbb{R}$ such that $|D^{J(l)} f(\boldsymbol{x})| \le M$ for $l = 0, 1, \dots, \lfloor \alpha \rfloor$, and

$$|D^{J(\lfloor \alpha \rfloor)} f(\boldsymbol{x}) - D^{J(\lfloor \alpha \rfloor)} f(\boldsymbol{y})| \le M \parallel \boldsymbol{x} - \boldsymbol{y} \parallel^{\alpha'} .$$

We assume that $g \in \Lambda^\alpha(M_g)$ and $V \in \Lambda^\beta(M_V)$. We will say for the sake of simplicity that "g has α continuous derivatives" while "V has β continuous derivatives". In this definition, $|\cdot|$ stands for the absolute value and $\|\cdot\|$ is the usual ℓ_2 norm.

In this section we construct a kernel estimator based on differences of raw observations and derive the rate of convergence for the estimator. Special care must be taken to define differences in the multivariate case. When $d = 1$ and there is a set of difference coefficients d_j, $j = 0, \dots, r$, such that $\sum_{j=0}^r d_j = 0$, $\sum_{j=0}^r d_j^2 = 1$ we define the difference "anchored" around the point y_i as $\sum_{j=0}^r d_j y_{i+j}$. When $d > 1$, there are multiple ways to enumerate observations lying around y_i. An example that explains how to do it in the case $d = 2$ is given in [15]. For a general $d > 1$, we first select a d-dimensional index set $J \in \mathbb{Z}^d$ that contains 0. Next, we define the set R consisting of all d-dimensional vectors $i = (i_1, \dots, i_d)$ such that

$$R + J \equiv \{(i + j) | j \in J, i \in R\} \subseteq \{1, \dots, m\}^d. \tag{6}$$

Again, a subset of $R + J$ corresponding to a specific $i^* \in R$ is denoted as $i^* + J$. Then, the difference "anchored" around the point y_{i^*} is defined by

$$D_{i^*} = \sum_{j \in J} d_j y_{i^*+j}. \tag{7}$$

The cardinality of the set J is called the order of the difference. For a good example that illustrates this notation style when $d = 2$ see [15].

Now we can define the variance estimator $\hat{V}(\boldsymbol{x})$. To do this, we use kernel-based weights $K_i^h(\boldsymbol{x})$ that are generated by either the regular kernel function $K(\cdot)$ or the boundary kernel function $K_*(\cdot)$, depending on the location of the point $\boldsymbol{x}$ in the support set S. The kernel function $K(\cdot) : \mathbb{R}^d \to \mathbb{R}$ has to satisfy the following set of assumptions:

$$\begin{aligned} &K(\boldsymbol{x}) \text{ is supported on } T = [-1, 1]^d, \quad \int_T K(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1 \\ &\int_T K(\boldsymbol{x}) \boldsymbol{x}^J \mathrm{d}\boldsymbol{x} = 0 \quad \text{for } 0 < |J| < \lfloor \beta \rfloor \text{ and} \\ &\int_T K^2(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = k_1 < \infty. \end{aligned} \tag{8}$$

Specially designed boundary kernels are needed to control the boundary effects in kernel regression. In the one-dimensional case boundary kernels with special properties are relatively easy to describe. See, for example, [5]. It is, however, more difficult to define boundary kernels in the multidimensional case because not only the distance from the boundary of S but also the local shape of the boundary region plays a role in defining the boundary kernels when $d > 1$. In this paper we use the d-dimensional boundary kernels given in [14]. We only briefly describe the basic idea here. Recall that we work with a nonnegative kernel function $K : T \to \mathbb{R}$ with support $T = [-1, 1]^d \subset \mathbb{R}^d$. For a given point $\boldsymbol{x} \in S$ consider a "moving" support set $S_n(\boldsymbol{x}) = \boldsymbol{x} + h(S - \boldsymbol{x})$ which changes with $\boldsymbol{x}$ and depends on n through the bandwidth h. For example, if $d = 1$, the set $S_n(\boldsymbol{x})$ becomes an interval $[x - hx, x + h(1 - x)] = [(1 - h)x, h + x(1 - h)]$. Plugging in $x = 0$ as the left boundary and $x = 1$ as the right boundary brings us back to the regular support $S = [0, 1]$. Using this varying support set $S_n(\boldsymbol{x})$, it is possible to define the support $T_{\boldsymbol{x}}$ of the boundary kernel that is independent of n. To do this, first define the set $T_n(\boldsymbol{x}) = \boldsymbol{x} - hT$; the subscript n again stresses that this set depends on n through the bandwidth h. This is the set of all points that form an h-neighborhood of x. Using $T_n(\boldsymbol{x})$ and the moving support $S_n(\mathbf{x})$, we have the transposed and rescaled support of the boundary kernel as

$$T_{\boldsymbol{x}} = h^{-1}[\boldsymbol{x} - \{T_n(\boldsymbol{x}) \cap S_n(\mathbf{x})\}] = h^{-1}(\boldsymbol{x} - \{\boldsymbol{x} + h(S - \boldsymbol{x})\} \cap (\boldsymbol{x} - hT)) = (\boldsymbol{x} - S) \cap T. \tag{9}$$

The subscript n has been omitted since $T_{\boldsymbol{x}}$ is, indeed, independent of n. Thus, the support of the boundary kernel has been stabilized. The boundary kernel $K_*(\cdot)$ with support on $T_{\boldsymbol{x}}$ can then be defined as a solution of a certain variational problem in much the same way as a regular kernel $K(\cdot)$. For more details, see [14].

Using this notation, we can define the general variance estimator as

$$\widehat{V}(\boldsymbol{x}) = \sum_{i \in R} K_i^h(\boldsymbol{x}) D_i^2 = \sum_{i \in R} K_i^h(\boldsymbol{x}) \left(\sum_{j \in J} d_j y_{i+j} \right)^2 . \tag{10}$$

The kernel weights are defined as

$$K_i^h(\boldsymbol{x}) = \begin{cases} n^{-1}h^{-d}K\left(\dfrac{\boldsymbol{x}_i - \boldsymbol{x}}{h}\right) & \text{when } \boldsymbol{x} - hT \subset S, \\ n^{-1}h^{-d}K_*\left(\dfrac{\boldsymbol{x}_i - \boldsymbol{x}}{h}\right) & \text{when } \boldsymbol{x} - hT \not\subseteq S. \end{cases}$$

It can also be described by the following algorithm:

1. Choose a d-dimensional index set J.
2. Construct the set R.
3. Define the estimator $\sum_{i\in R} K_i^h(\boldsymbol{x}) \left(\sum_{j\in J} d_j y_{i+j}\right)^2$ as a local average using kernel-generated weights $K_i^h(\boldsymbol{x})$.

In this paper we will use the index set J selected to be a sequence of γ points on the straight line in the d-dimensional space that includes the origin:

$$J = \{(0, 0, \dots, 0), (1, 1, \dots, 1), \dots, (\gamma, \gamma, \dots, \gamma)\}. \tag{11}$$

In addition, we use normalized binomial coefficients as the difference coefficients. This is the so-called *polynomial sequence* (see, e.g., [15]) and is defined as

$$d_k = \binom{\gamma}{k}(-1)^k \Big/ \binom{2\gamma}{\gamma}^{1/2}$$

where $k = 0, 1, \dots, \gamma$. It is clear that $\sum_{k=0}^{\gamma} d_k = 0$, $\sum_{k=0}^{\gamma} d_k^2 = 1$, and $\sum_{k=0}^{\gamma} k^q d_k = 0$ for any $q = 1, 2, \dots, \gamma$.

Remark 1. It is also possible to use the local linear regression estimator instead of the kernel estimator. In this case, the boundary kernel adjustment is not necessary as it is well known that the local linear regression adjusts automatically in boundary regions, preserving the asymptotic order of the bias intact. However, the proof is slightly more technically involved when using the local linear regression estimator; in particular, the local linear regression estimator has to be represented as the "kernel" estimator where the shape of the function $K(\cdot)$ used to define the local weights now depends on the location of the design points, the number of observations n and the point of estimation x. For details, see, for example, [3].

Remark 2. It is possible to define a more general estimator by considering averaging over several possible d-dimensional index sets J_l, $l = 1, \dots, L$, and defining a set R_l for each one of them according to (6). In other words, we define

$$\widehat{V}(\boldsymbol{x}) = \sum_{l=1}^{L} \mu_l \sum_{i\in R_l} K_i^h(\boldsymbol{x}) D_i^2 = \sum_{l=1}^{L} \mu_l \sum_{i\in R_l} K_i^h(\boldsymbol{x}) \left(\sum_{j\in J_l} d_j y_{i+j}\right)^2 \tag{12}$$

where μ_l is a set of weights such that $\sum_l \mu_l = 1$. The proof of the main result in the general case is completely analogous to the case $L = 1$. If some information about the geometry of the surface of $V(\mathbf{x})$ is known, we may be able to choose the collection of index sets J_l as described above in order to minimize the constant factor in the asymptotic variance of the estimator of $V(\mathbf{x})$. In this paper we limit ourselves to the discussion of the case $L = 1$ and the definition (10) will be used with the set J selected as in (11).

Like in the mean function estimation problem, the optimal bandwidth h_n can be easily found to be $h_n = O(n^{-1/(2\beta+d)})$ for $V \in \Lambda^\beta(M_V)$. For this optimal choice of the bandwidth, we have the following theorem.

Theorem 1. *Under the regression model (1) with z_i being independent random variables with zero mean, unit variance and uniformly bounded fourth moments, we define the estimator $\widehat{V}$ as in (10) with the bandwidth $h = O(n^{-1/(2\beta+d)})$ and the order of the difference sequence $\gamma = \lceil d/4 \rceil$. Then there exists some constant $C_0 > 0$ depending only on α, β, M_g, M_V and d such that for sufficiently large n,*

$$\sup_{g\in\Lambda^\alpha(M_g), V\in\Lambda^\beta(M_V)} \sup_{\boldsymbol{x}_*\in S} E(\widehat{V}(\boldsymbol{x}_*) - V(\boldsymbol{x}_*))^2 \le C_0 \cdot \max\{n^{-\frac{4\alpha}{d}}, n^{-\frac{2\beta}{2\beta+d}}\} \tag{13}$$

and

$$\sup_{g\in\Lambda^\alpha(M_g), V\in\Lambda^\beta(M_V)} E\int_{R^d} (\widehat{V}(\boldsymbol{x}) - V(\boldsymbol{x}))^2 d\boldsymbol{x} \le C_0 \cdot \max\{n^{-\frac{4\alpha}{d}}, n^{-\frac{2\beta}{2\beta+d}}\}. \tag{14}$$

Remark 3. The uniform rate of convergence given in (13) yields immediately the pointwise rate of convergence for any fixed point $\boldsymbol{x}_* \in S$,

$$\sup_{g\in\Lambda^\alpha(M_g), V\in\Lambda^\beta(M_V)} E(\widehat{V}(\boldsymbol{x}_*) - V(\boldsymbol{x}_*))^2 \le C_0 \cdot \max\{n^{-\frac{4\alpha}{d}}, n^{-\frac{2\beta}{2\beta+d}}\}.$$

3. Lower bound

Theorem 1 gives the upper bounds for the minimax risks of estimating the variance function $V(\boldsymbol{x})$ under the multivariate regression model (1). In this section we shall show that the upper bounds are in fact rate-optimal. We derive lower bounds for the minimax risks which are of the same order as the corresponding upper bounds given in Theorem 1. In the lower bound argument we shall assume that the errors are normally distributed, i.e., $z_i \overset{iid}{\sim} N(0, 1)$.

Theorem 2. *Under the regression model (1) with $z_i \overset{iid}{\sim} N(0, 1)$,*

$$\inf_{\widehat{V}} \sup_{g \in \Lambda^\alpha(M_g), V \in \Lambda^\beta(M_V)} E\|\widehat{V} - V\|_2^2 \geq C_1 \cdot \max\{n^{-\frac{4\alpha}{d}}, n^{-\frac{2\beta}{d+2\beta}}\} \tag{15}$$

and for any fixed $\boldsymbol{x}_* \in [0, 1]^d$

$$\inf_{\widehat{V}} \sup_{g \in \Lambda^\alpha(M_g), V \in \Lambda^\beta(M_V)} E(\widehat{V}(\boldsymbol{x}_*) - V(\boldsymbol{x}_*))^2 \geq C_1 \cdot \max\{n^{-\frac{4\alpha}{d}}, n^{-\frac{2\beta}{d+2\beta}}\} \tag{16}$$

where $C_1 > 0$ is a constant.

Combining Theorems 1 and 2 yields immediately the minimax rate of convergence,

$$\max\left\{n^{-\frac{4\alpha}{d}}, n^{-\frac{2\beta}{d+2\beta}}\right\},$$

for estimating V under both the global and pointwise losses.

Theorem 2 is proved in Section 5. The proof is based on a moment matching technique and a two-point testing argument. One of the main steps is studying a hypothesis testing problem where the alternative hypothesis is a Gaussian location mixture with a special moment matching property.

4. Discussion

The first important observation that we can make on the basis of reported results is that the unknown mean function g does not have any first order effect on the minimax rate of convergence of the estimator $\hat{V}$ as long as the function g has at least $d/4$ derivatives. When this is true, the minimax rate of convergence for $\hat{V}$ is $n^{-2\beta/2\beta+d}$, which is the same as if the mean function g had been known. Therefore the variance estimator $\hat{V}$ is adaptive over the collection of the mean functions g that belong to Lipschitz classes $\Lambda^\alpha(M_g)$ for all $\alpha \geq d/4$. On the other hand, if the function g has less than $d/4$ derivatives, the minimax rate of convergence for $\hat{V}$ is determined by the relative smoothness of both g and V. When $4\alpha/d < 2\beta/(2\beta + d)$, the roughness of g becomes the dominant factor in determining the convergence rate for $\hat{V}$. In other words, when $\alpha < d\beta/(2(2\beta+d))$, the rate of convergence becomes $n^{-4\alpha/d}$ and thus is completely determined by α. To make better sense of this statement, let us consider the situation when β increases and can become arbitrarily large. Clearly, in this case the cutoff $d\beta/(2(2\beta+d))$ approaches $d/4$. Thus, when $d = 2$, any mean function g with less than half of a derivative will completely determine the rate of convergence for $\hat{V}$; when $d = 4$, any mean function with less than one derivative will do and so on. As the number of dimensions d grows and the function V becomes smoother, the rate of convergence of $\hat{V}$ becomes more and more dependent on the mean function. In other words, an ever increasing set of possible mean functions will completely "overwhelm" the influence of the variance function in determining the minimax convergence rate.

As opposed to many common variance estimation methods, our approach does not estimate the mean function first. Instead, we estimate the variance as the local average of squared differences of observations. Taking a difference of a set of observations is, in a sense, an attempt to "average out" the influence of the mean. It is possible to say then that we use an implicit "estimator" of the mean function g that is effectively a linear combination of all $y_j, j \in J$, except y_0. Such an estimator is, of course, not optimal since its squared bias and variance are not balanced. The reason that it has to be used is because the bias and variance of the mean estimator $\hat{g}$ have very different influences on $\hat{V}$. As is the case when $d = 1$ (again, see [21]), the influence of the bias of $\hat{g}$ is impossible to reduce at the second stage of variance estimation. Therefore, at the first stage we use an "estimator" of g that provides for the maximal reduction in bias possible under the assumption of $g \in \Lambda^\alpha(M_g)$, down to the order $n^{-2\alpha/d}$. In fact, the variance of the "estimator" $\hat{g}$ is high but this is of little concern; it is incorporated easily into the variance estimation procedure. Thus, in practical terms, subtracting optimal estimators of the mean function g first may not be the most desirable course of action.

Note also that it is not enough to use here a simple first order difference as has been done in the case of $d = 1$ by Wang et al. [21]. The reason is that this does not allow us to reduce the mean-related bias of the variance estimator $\hat{V}$ to the fullest extent possible. It is not enough to consider only $\alpha < 1/4$ as is the case when $d = 1$. Instead, when proving the upper bound result, we have to consider mean functions with $\alpha < d/4$. Thus, higher order differences are needed in order to reduce the mean-related bias to the order of $n^{-2\alpha/d}$ and to ensure the minimax rate of convergence.

5. Proofs

5.1. Upper bound: Proof of Theorem 1

We will use M to denote a generic positive constant throughout this section. We shall only prove (13). Inequality (14) is a direct consequence of (13). Recall that $T=[-1,1]^d$ is the support of the kernel K. The following notation will be useful: for any two vectors $\boldsymbol{x}=(x_1,\ldots,x_d)'$ and $\boldsymbol{y}=(y_1\ldots,y_d)'$ we define the differential operator

$$D_{\boldsymbol{x},\boldsymbol{y}}=\sum_{k=1}^{d}(y_k-x_k)\frac{\partial}{\partial z_k}=\langle \boldsymbol{y}-\boldsymbol{x},\nabla\rangle \tag{17}$$

where z_k is a generic kth argument of a d-dimensional function while ∇ is a gradient operator in $\mathbb{R}^d$.

Using the notation that we introduced earlier, we can write the difference D_i as

$$D_i=\sum_{j\in J}d_jg(\boldsymbol{x}_{i+j})+\sum_{j\in J}d_jV^{1/2}(\boldsymbol{x}_{i+j})z_{i+j}=\delta_i+V_i^{\frac{1}{2}}\epsilon_i \tag{18}$$

where $\delta_i=\sum_{j\in J}d_jg(\boldsymbol{x}_{i+j})$, $V_i^{\frac{1}{2}}=\sqrt{\sum_{j\in J}d_j^2V(\boldsymbol{x}_{i+j})}$ and

$$\epsilon_i=\left(\sum_{j\in J}d_j^2V(\boldsymbol{x}_{i+j})\right)^{-1/2}\left(\sum_{j\in J}d_jV^{\frac{1}{2}}(\boldsymbol{x}_{i+j})z_{i+j}\right)$$

has zero mean and unit variance. Thus,

$$D_i^2=\delta_i^2+V_i+V_i(\epsilon_i^2-1)+2\delta_iV_i^{\frac{1}{2}}\epsilon_i.$$

Without loss of generality, suppose $h=n^{-1/(2\beta+d)}$. Because the kernel $K(\cdot)$ has a bounded support $T=[-1,1]^d$, we have

$$\left(\sum_{i\in R}|K_i^h(\boldsymbol{x}_*)|\right)^2\le 2^dnh^d\sum_{i\in R}(K_i^h(\boldsymbol{x}_*))^2\le 2^d\int_{[-1,1]^d}K_{**}^2(\boldsymbol{u})\mathrm{d}\boldsymbol{u}\le 2^dk \tag{19}$$

where $k=\max(k_1,k_2)$. The first step follows from the fact that the h-neighborhood of $\boldsymbol{x}_*$ has $2^d\lfloor nh^d\rfloor$ points while the second step follows from approximating the Riemann sum by the appropriate integral. Also, $K_{**}(\boldsymbol{u})=K(\boldsymbol{u})$ when $\boldsymbol{u}\in T_n(\boldsymbol{u})\cap S$ and $K_{**}(\boldsymbol{u})=K_*(\boldsymbol{u})$ when $\boldsymbol{u}\not\subseteq T_n(\boldsymbol{u})\cap S$ with constants k_1 and k_2 resulting from one of these two respective choices. Recall that $\hat{V}(\boldsymbol{x}_*)-V(\boldsymbol{x}_*)=\sum_{i\in R}K_i^h(\boldsymbol{x}_*)D_i^2-V(\boldsymbol{x}_*)$. For all $g\in\Lambda^\alpha(M_g)$ and $V\in\Lambda^\beta(M_V)$, the mean squared error of $\hat{V}$ at $\boldsymbol{x}_*$ satisfies

$$\begin{aligned}
E(\widehat{V}(\boldsymbol{x}_*)-V(\boldsymbol{x}_*))^2 &= E\left(\sum_{i\in R}K_i^h(\boldsymbol{x}_*)\left(D_i^2-V(\boldsymbol{x}_*)\right)+o(n^{-1}h^{-d})\right)^2\\
&= E\left\{\sum_{i\in R}K_i^h(\boldsymbol{x}_*)\delta_i^2+\sum_{i\in R}K_i^h(\boldsymbol{x}_*)(V_i-V(\boldsymbol{x}_*))\right.\\
&\quad\left.+\sum_{i\in R}K_i^h(\boldsymbol{x}_*)V_i(\epsilon_i^2-1)+2\sum_{i\in R}K_i^h(\boldsymbol{x}_*)\delta_iV_i^{\frac{1}{2}}\epsilon_i+o(n^{-1}h^{-d})\right\}^2\\
&\le 5\left(\sum_{i\in R}K_i^h(\boldsymbol{x}_*)\delta_i^2\right)^2+5\left(\sum_{i\in R}K_i^h(\boldsymbol{x}_*)(V_i-V(\boldsymbol{x}_*))\right)^2\\
&\quad+5E\left(\sum_{i\in R}K_i^h(\boldsymbol{x}_*)V_i(\epsilon_i^2-1)\right)^2+20E\left(\sum_{i\in R}K_i^h(\boldsymbol{x}_*)\delta_iV_i^{\frac{1}{2}}\epsilon_i\right)^2+o(n^{-2}h^{-2d}).
\end{aligned}$$

Recall that it is enough to consider only $\alpha<d/4$. Define $\gamma=\lceil d/4\rceil$. Thus defined, γ will be the same as the maximum possible value of $\lfloor\alpha\rfloor$ for all $\alpha<d/4$. Defining $0\le u\le 1$ and using Taylor expansion of $g(\boldsymbol{x}_{i+j})$ around $\boldsymbol{x}_i$, we have for a difference sequence of order γ

$$\begin{aligned}
|\delta_i| &= \left|\sum_{j\in J}d_jg(\boldsymbol{x}_{i+j})\right|=\left|\sum_{j\in J}d_j\left(g(\boldsymbol{x}_i)+\sum_{m=1}^{\lfloor\alpha\rfloor}\frac{(D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_i})^mg(\boldsymbol{x}_i)}{m!}\right.\right.\\
&\quad\left.\left.+\int_0^1\frac{(1-u)^{\lfloor\alpha\rfloor-1}}{(\lfloor\alpha\rfloor-1)!}((D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_i})^{\lfloor\alpha\rfloor}g(\boldsymbol{x}_i+u(\boldsymbol{x}_{i+j}-\boldsymbol{x}_i))-(D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_i})^{\lfloor\alpha\rfloor}g(\boldsymbol{x}_i)\mathrm{d}u)\right)\right|.
\end{aligned}$$

The first two terms in the above expression are zero by definition of the difference sequence d_j of order γ. Using the notation x_i^k for the kth coordinate of $\boldsymbol{x}_i$, the explicit representation of the operator $(D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_i})^{\lfloor\alpha\rfloor}$ gives

$$|(D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_i})^{\lfloor\alpha\rfloor} g(\boldsymbol{x}_i + u(\boldsymbol{x}_{i+j} - \boldsymbol{x}_i)) - (D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_i})^{\lfloor\alpha\rfloor} g(\boldsymbol{x}_i)|$$
$$= \left| \sum_{1\le t_1\le\cdots\le t_{\lfloor\alpha\rfloor}\le d} \left[\left(\prod_{r=1}^{\lfloor\alpha\rfloor} (x_{i+j}^{t_r} - x_i^{t_r}) \right) D^{\lfloor\alpha\rfloor} g(\boldsymbol{x}_i + u(\boldsymbol{x}_{i+j} - \boldsymbol{x}_i)) \right] - \sum_{1\le t_1\le\cdots\le t_{\lfloor\alpha\rfloor}\le d} \left[\left(\prod_{r=1}^{\lfloor\alpha\rfloor} (x_{i+j}^{t_r} - x_i^{t_r}) \right) D^{\lfloor\alpha\rfloor} g(\boldsymbol{x}_i) \right] \right|.$$

Now we use the definition of Lipschitz space $\Lambda^\alpha(M_g)$, Jensen's and Hölder's inequalities to find that

$$\begin{aligned}|(D^{\lfloor\alpha\rfloor})g(\boldsymbol{x}_i + u(\boldsymbol{x}_{i+j} - \boldsymbol{x}_i)) - (D)^{\lfloor\alpha\rfloor} g(\boldsymbol{x}_i)| &\le M_g \|u(\boldsymbol{x}_{i+j} - \boldsymbol{x}_i)\|^{\alpha'} \left| \sum_{1\le t_1\le\cdots\le t_{\lfloor\alpha\rfloor}\le d} \left(\prod_{r=1}^{\lfloor\alpha\rfloor} (x_{i+j}^{t_r} - x_i^{t_r}) \right) \right| \\ &\le M_g \|\boldsymbol{x}_{i+j} - \boldsymbol{x}_i\|^{\alpha'} \sum_{1\le t_1\cdots\le t_{\lfloor\alpha\rfloor}\le d} \sum_{r=1}^{\lfloor\alpha\rfloor} \frac{|x_{i+j}^{t_r} - x_i^{t_r}|^{\lfloor\alpha\rfloor}}{\lfloor\alpha\rfloor} \\ &\le M \|\boldsymbol{x}_{i+j} - \boldsymbol{x}_i\|^{\alpha'} \|\boldsymbol{x}_{i+j} - \boldsymbol{x}_i\|^{\lfloor\alpha\rfloor} = M\|\boldsymbol{x}_{i+j} - \boldsymbol{x}_i\|^{\alpha};\end{aligned}$$

as a consequence, we have $|\delta_i| \le Mn^{-\alpha/d}$. Thus,

$$4\left(\sum_{i\in R} K_i^h(\boldsymbol{x}_*)\delta_i^2\right)^2 \le 4\left(\sum_{i\in R} |K_i^h(x_*)| M^2 n^{-2\alpha/d}\right)^2 \le 2^{d+2} k M^4 n^{-4\alpha/d} = O(n^{-4\alpha/d}).$$

In exactly the same way as above, for any $\boldsymbol{x}, \boldsymbol{y} \in [0, 1]^d$, Taylor's theorem yields

$$\begin{aligned}&\left| V(\boldsymbol{x}) - V(\boldsymbol{y}) - \sum_{j=1}^{\lfloor\beta\rfloor} \frac{(D_{\boldsymbol{x},\boldsymbol{y}})^j V(\boldsymbol{y})}{j!} \right| = \left| \int_0^1 \frac{(1-u)^{\lfloor\beta\rfloor-1}}{\lfloor\beta\rfloor - 1} ((D_{\boldsymbol{x},\boldsymbol{y}})^{\lfloor\beta\rfloor} V(\boldsymbol{y} + u(\boldsymbol{x} - \boldsymbol{y})) - (D_{\boldsymbol{x},\boldsymbol{y}})^{\lfloor\beta\rfloor} V(\boldsymbol{y}))\mathrm{d}u \right| \\ &\le M\|\boldsymbol{x} - \boldsymbol{y}\|^\beta \int_0^1 \left| \frac{(1-u)^{\lfloor\beta\rfloor-1}}{\lfloor\beta\rfloor - 1} \right| \mathrm{d}u \le M\|\boldsymbol{x} - \boldsymbol{y}\|^\beta. \end{aligned} \tag{20}$$

So,

$$\begin{aligned}V_i - V(\boldsymbol{x}_*) &= \sum_{j\in J} d_j^2 V(\boldsymbol{x}_{i+j}) - V(\boldsymbol{x}_*) = \sum_{j\in J} d_j^2 \left[V(\boldsymbol{x}_{i+j}) - V(\boldsymbol{x}_*)\right] \\ &= \sum_{j\in J} d_j^2 \sum_{k=1}^{\lfloor\beta\rfloor} \frac{(D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_*})^k V(\boldsymbol{x}_*)}{k!} + \sum_{j\in J} d_j^2 \int_0^1 \frac{(1-u)^{\lfloor\beta\rfloor-1}}{\lfloor\beta\rfloor - 1} ((D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_*})^{\lfloor\beta\rfloor} V(\boldsymbol{x}_{i+j}) - (D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_*})^{\lfloor\beta\rfloor} V(\boldsymbol{x}_*))\mathrm{d}u.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\sum_{i\in R} K_i^h(\boldsymbol{x}_*)(V_i - V(\boldsymbol{x}_*)) &= \sum_{i\in R} K_i^h(\boldsymbol{x}_*) \sum_{j\in J} d_j^2 \sum_{k=1}^{\lfloor\beta\rfloor} \frac{(D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_*})^k V(\boldsymbol{x}_*)}{k!} \\ &\quad + \sum_{i\in R} K_i^h(\boldsymbol{x}_*) \sum_{j\in J} d_j^2 \int_0^1 \frac{(1-u)^{\lfloor\beta\rfloor-1}}{\lfloor\beta\rfloor - 1} ((D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_*})^{\lfloor\beta\rfloor} V(\boldsymbol{x}_{i+j}) - (D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_*})^{\lfloor\beta\rfloor} V(\boldsymbol{x}_*))\mathrm{d}u.\end{aligned}$$

It is fairly straightforward to find out that the first term is bounded by

$$\begin{aligned}&\left| \sum_{i\in R} K_i^h(\boldsymbol{x}_*) \sum_{j\in J} d_j^2 \sum_{k=1}^{\lfloor\beta\rfloor} \frac{(D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_*})^k V(\boldsymbol{x}_*)}{k!} \right| \\ &= \left| n^{-1}h^{-d} \sum_{i\in R} K\left(\frac{\boldsymbol{x}_i - \boldsymbol{x}_*}{h}\right) \sum_{j\in J} d_j^2 \sum_{k=1}^{\lfloor\beta\rfloor} \frac{1}{k!} \sum_{1\le t_1\le\cdots\le t_k\le d} \prod_{r=1}^{k} (x_{i+j}^{t_r} - x_i^{t_r}) D^k V(\boldsymbol{x}_*) \right| \\ &\le \left| Mn^{-1}h^{-d} \sum_{k=1}^{\lfloor\beta\rfloor} h^k \sum_{i\in R} K(\boldsymbol{u}_i)\boldsymbol{u}_i^k \right| = o(n^{-1}h^{-(d-1)}).\end{aligned}$$

To establish the last inequality it is important to remember the fact that $V \in \Lambda^\beta(M_V)$ and therefore $|D^k V(\boldsymbol{x}_*)| \le M_V$. To handle the product $\prod_{r=1}^k (x_{i+j}^{t_r} - x_i^{t_r})$ the inequality $\prod_{i=1}^n x_i \le n^{-1}\sum_{i=1}^n x_i^n$, that is true for any positive numbers $x_1, \ldots, x_n$, must be used. The equality that follows is based on the fact that kernel K has $\lfloor\beta\rfloor$ vanishing moments. After taking the square the above will become $o(n^{-2}h^{-2(d-1)})$; comparing to the optimal rate of $n^{-2\beta/2\beta+d}$, it is easy to check that this term is always of smaller order, $o(n^{-2\beta/(2\beta+d)-(2\beta+2)/(2\beta+d)})$.

Using (20), we find that the absolute value of the second term gives us

$$\begin{aligned}&\left| \sum_{i\in R} K_i^h(\boldsymbol{x}_*) \sum_{j\in J} d_j^2 \int_0^1 \frac{(1-u)^{\lfloor\beta\rfloor-1}}{\lfloor\beta\rfloor - 1} ((D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_*})^{\beta} V(\boldsymbol{x}_{i+j}) - (D_{\boldsymbol{x}_{i+j},\boldsymbol{x}_*})^{\lfloor\beta\rfloor} V(\boldsymbol{x}_*))\mathrm{d}u \right| \\ &\le Mh^{-\beta} \sum_{i\in R} |K_h^i(\boldsymbol{x}_*)| \sum_{j\in J} d_j^2 = O(n^{-\beta/2\beta+d}).\end{aligned}$$

From here it follows by taking squares that $5\left(\sum_{i\in R} K_i^h(\boldsymbol{x}_*)(V_i - V(\boldsymbol{x}_*))\right)^2$ is of the order $O(n^{-2\beta/(2\beta+d)})$.

On the other hand, since $V \le M_V$, we have due to (19)

$$5E\left(\sum_{i\in R} K_i^h(\boldsymbol{x}_*)\delta_i V_i^{\frac{1}{2}}\epsilon_i\right)^2 = 5\mathrm{Var}\left(\sum_{i\in R} K_i^h(\boldsymbol{x}_*)\delta_i V_i^{\frac{1}{2}}\epsilon_i\right) = 5\sum_{i\in R}\left(K_i^h(\boldsymbol{x}_*)\right)^2\delta_i^2 V_i$$
$$\le 5M_V n^{-2\alpha/d-2\beta/(2\beta+d)} \times k$$

and

$$20E\left(\sum_{i\in R} K_i^h(\boldsymbol{x}_*)V_i(\epsilon_i^2-1)\right)^2 = 20\mathrm{Var}\left(\sum_{i\in R} K_i^h(\boldsymbol{x}_*)V_i(\epsilon_i^2-1)\right) \le 20M_V^2\mu_4\sum_{i\in R}\left(K_i^h(\boldsymbol{x}_*)\right)^2$$
$$\le 20M_V^2\mu_4\frac{1}{nh^d}k = 20M_V^2\mu_4 n^{-2\beta/(2\beta+d)} \times k.$$

Putting the four terms together we have, uniformly for all $\boldsymbol{x}_* \in [0,1]^d$, $g \in \Lambda^\alpha(M_g)$ and $V \in \Lambda^\beta(M_V)$,

$$E(\widehat{V}(\boldsymbol{x}_*) - V(\boldsymbol{x}_*))^2 \le C_0 \cdot \max\{n^{-4\alpha/d}, n^{-2\beta/(2\beta+d)}\}$$

for some constant $C_0 > 0$. This proves (13). ■

6. Proof of Theorem 2

The proof of this theorem can be naturally divided into two parts. The first step is to show

$$\inf_{\widehat{V}} \sup_{g\in\Lambda^\alpha(M_g),V\in\Lambda^\beta(M_V)} E(\widehat{V}(x_*) - V(x_*))^2 \ge C_1 n^{-\frac{2\beta}{d+2\beta}}. \tag{21}$$

This part is standard and relatively easy. The proof of the second step,

$$\inf_{\widehat{V}} \sup_{g\in\Lambda^\alpha(M_g),V\in\Lambda^\beta(M_V)} E(\widehat{V}(x_*) - V(x_*))^2 \ge C_1 n^{-\frac{4\alpha}{d}}, \tag{22}$$

is based on a moment matching technique and a two-point testing argument. More specifically, let $X_1, \ldots, X_n \overset{iid}{\sim} P$ and consider the following hypothesis testing problem, between

$$H_0 : P = P_0 = N(0, 1+\theta_n^2)$$

and

$$H_1 : P = P_1 = \int N(\theta_n \nu, 1)G(\mathrm{d}\nu)$$

where $\theta_n > 0$ is a constant and G is a distribution of the mean ν with compact support. The distribution G is chosen in such a way that, for some positive integer q depending on α, the first q moments of G match exactly with the corresponding moments of the standard normal distribution. The existence of such a distribution is given in the following lemma from Karlin and Studden [10].

Lemma 1. *For any fixed positive integer q, there exist a $B < \infty$ and a symmetric distribution G on $[-B, B]$ such that G and the standard normal distribution have the same first q moments, i.e.*

$$\int_{-B}^{B} x^j G(\mathrm{d}x) = \int_{-\infty}^{+\infty} x^j \varphi(x)\mathrm{d}x, \quad j = 1, 2, \ldots, q$$

where φ denotes the density of the standard normal distribution.

We shall only prove the lower bound for the pointwise squared error loss. The same proof with minor modifications immediately yields the lower bound under integrated squared error. Note that, to prove inequality (22), we only need to focus on the case where $\alpha < d/4$; otherwise $n^{-2\beta/(d+2\beta)}$ is always greater than $n^{-4\alpha/d}$ for sufficiently large n and then (22) follows directly from (21).

For a given $0 < \alpha < d/4$, there exists an integer q such that $(q+1)\alpha > d$. For convenience we take q to be an odd integer. From Lemma 1, there is a positive constant $B < \infty$ and a symmetric distribution G on $[-B, B]$ such that G and $N(0, 1)$ have the same first q moments. Let r_i, $i = 1, \ldots, n$, be independent variables with the distribution G. Set $\theta_n = \frac{M_g}{2B} m^{-\alpha}$, $g_0 \equiv 0$, $V_0(x) \equiv 1 + \theta_n^2$ and $V_1(x) \equiv 1$. Let $h(x) = 1 - 2m|x|$ for $|x| \in [-\frac{1}{2m}, \frac{1}{2m}]$ and 0 otherwise (here $|x| \triangleq \sqrt{x_1^2 + \cdots + x_d^2}$). Define the random function g_1 by

$$g_1(x) = \sum_{i=1}^{n} \theta_n r_i h(x - x_i) I \quad (x \in [0, 1]^d).$$

Then it is easy to see that g_1 is in $\Lambda^\alpha(M_g)$ for all realizations of r_i. Moreover, $g_1(x_i) = \theta_n r_i$ are independent and identically distributed.

Now consider testing the following hypotheses:

$$H_0 : y_i = g_0(x_i) + V_0^{\frac{1}{2}}(x_i)\epsilon_i, \quad i = 1, \ldots, n,$$
$$H_1 : y_i = g_1(x_i) + V_1^{\frac{1}{2}}(x_i)\epsilon_i, \quad i = 1, \ldots, n,$$

where ϵ_i are independent $N(0, 1)$ variables which are also independent of the r_i's. Denote by P_0 and P_1 the joint distributions of y_i's under H_0 and H_1, respectively. Note that for any estimator $\widehat{V}$ of V,

$$\begin{aligned}\max\{E(\widehat{V}(x_*) - V_0(x_*))^2, E(\widehat{V}(x_*) - V_1(x_*))^2\} &\geq \frac{1}{16}\rho^4(P_0, P_1)(V_0(x_*) - V_1(x_*))^2 \\ &= \frac{1}{16}\rho^4(P_0, P_1)\frac{M_g^4}{16B^4}m^{-4\alpha} \end{aligned} \tag{23}$$

where $\rho(P_0, P_1)$ is the Hellinger affinity between P_0 and P_1. See, for example, [11]. Let p_0 and p_1 be the probability density functions of P_0 and P_1 with respect to the Lebesgue measure μ; then $\rho(P_0, P_1) = \int \sqrt{p_0 p_1}\mathrm{d}\mu$. The minimax lower bound (22) follows immediately from the two-point bound (23) if we show that for any n, the Hellinger affinity $\rho(P_0, P_1) \geq C$ for some constant $C > 0$. (Note that $m^{-4\alpha} = n^{-4\alpha/d}$.)

Note that under H_0, $y_i \sim N(0, 1 + \theta_n^2)$ and its density d_0 can be written as

$$d_0(t) \triangleq \frac{1}{\sqrt{1+\theta_n^2}}\varphi\left(\frac{t}{\sqrt{1+\theta_n^2}}\right) = \int \varphi(t - v\theta_n)\varphi(v)\mathrm{d}v.$$

Under H_1, the density of y_i is $d_1(t) \triangleq \int \varphi(t - v\theta_n)G(\mathrm{d}v)$.

It is easy to see that $\rho(P_0, P_1) = (\int \sqrt{d_0 d_1}\mathrm{d}\mu)^n$, since the y_i's are independent variables. Note that the Hellinger affinity is bounded below by the total variation affinity,

$$\int \sqrt{d_0(t)d_1(t)}\mathrm{d}t \geq 1 - \frac{1}{2}\int |d_0(t) - d_1(t)|\,\mathrm{d}t.$$

Taylor expansion yields $\varphi(t - v\theta_n) = \varphi(t)\left(\sum_{k=0}^\infty v^k\theta_n^k \frac{H_k(t)}{k!}\right)$ where $H_k(t)$ is the corresponding Hermite polynomial. And from the construction of distribution G, $\int v^i G(\mathrm{d}v) = \int v^i\varphi(v)\mathrm{d}v$ for $i = 0, 1, \ldots, q$. So,

$$\begin{aligned}|d_0(t) - d_1(t)| &= \left|\int \varphi(t - v\theta_n)G(\mathrm{d}v) - \int \varphi(t - v\theta_n)\varphi(v)\mathrm{d}v\right| \\ &= \left|\int \varphi(t)\sum_{i=0}^\infty \frac{H_i(t)}{i!}v^i\theta_n^i G(\mathrm{d}v) - \int \varphi(t)\sum_{i=0}^\infty \frac{H_i(t)}{i!}v^i\theta_n^i\varphi(v)\mathrm{d}v\right| \\ &= \left|\int \varphi(t)\sum_{i=q+1}^\infty \frac{H_i(t)}{i!}v^i\theta_n^i G(\mathrm{d}v) - \int \varphi(t)\sum_{i=q+1}^\infty \frac{H_i(t)}{i!}v^i\theta_n^i\varphi(v)\mathrm{d}v\right| \\ &\leq \left|\int \varphi(t)\sum_{i=q+1}^\infty \frac{H_i(t)}{i!}v^i\theta_n^i G(\mathrm{d}v)\right| + \left|\int \varphi(t)\sum_{i=q+1}^\infty \frac{H_i(t)}{i!}v^i\theta_n^i\varphi(v)\mathrm{d}v\right|. \end{aligned} \tag{24}$$

Suppose $q + 1 = 2p$ for some integer p; it can be seen that

$$\begin{aligned}\left|\int \varphi(t)\sum_{i=q+1}^\infty \frac{H_i(t)}{i!}v^i\theta_n^i G(\mathrm{d}v)\right| &= \left|\int \varphi(t)\sum_{i=p}^\infty \frac{H_{2i}(t)}{(2i)!}\theta_n^{2i}v^{2i}G(\mathrm{d}v)\right| \\ &\leq \varphi(t)\sum_{i=p}^\infty\left|\frac{H_{2i}(t)}{(2i)!}\theta_n^{2i}\right|\left|\int v^{2i}G(\mathrm{d}v)\right| \leq \varphi(t)\sum_{i=p}^\infty\left|\frac{H_{2i}(t)}{(2i)!}\right|\theta_n^{2i}B^{2i}\end{aligned}$$

and

$$\begin{aligned}&\left|\int \varphi(t)\sum_{i=q+1}^\infty \frac{H_i(t)}{i!}v^i\theta_n^i\varphi(v)\mathrm{d}v\right| = \left|\int \varphi(t)\sum_{i=p}^\infty \frac{H_{2i}(t)}{(2i)!}\theta_n^{2i}v^{2i}\varphi(v)\mathrm{d}v\right| \\ &\leq \varphi(t)\sum_{i=p}^\infty\left|\frac{H_{2i}(t)}{(2i)!}\theta_n^{2i}\right|\left|\int v^{2i}\varphi(v)\mathrm{d}v\right| = \left|\varphi(t)\sum_{i=p}^\infty H_{2i}(t)\theta_n^{2i}\frac{1}{2^i\cdot i!}\right| \leq \varphi(t)\sum_{i=p}^\infty\left|\frac{H_{2i}(t)}{2^i\cdot i!}\right|\theta_n^{2i}\end{aligned}$$

where $(2i-1)!! \triangleq (2i-1)\times(2i-3)\times\cdots 3\times 1$. So from (24),

$$|d_0(t) - d_1(t)| \leq \varphi(t)\sum_{i=p}^\infty\left|\frac{H_{2i}(t)}{(2i)!}\right|\theta_n^{2i}B^{2i} + \varphi(t)\sum_{i=p}^\infty\left|\frac{H_{2i}(t)}{2^i\cdot i!}\right|\theta_n^{2i}$$

and then

$$\int \sqrt{d_0(t)d_1(t)}\mathrm{d}t \geq 1 - \frac{1}{2}\int \left(\varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{(2i)!}\right|\theta_n^{2i}B^{2i} + \varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{2^i\cdot i!}\right|\theta_n^{2i}\right)\mathrm{d}t$$

$$= 1 - \frac{1}{2}\int \varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{(2i)!}\right|\theta_n^{2i}B^{2i}\mathrm{d}t - \frac{1}{2}\int \varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{2^i\cdot i!}\right|\theta_n^{2i}\mathrm{d}t. \tag{25}$$

For the Hermite polynomial H_{2i}, we have

$$\begin{aligned}\int \varphi(t)\,|H_{2i}(t)|\,\mathrm{d}t &= \int \varphi(t)\left|(2i-1)!!\times\left[1+\sum_{k=1}^{i}\frac{(-2)^k i(i-1)\cdots(i-k+1)}{(2k)!}t^{2k}\right]\right|\mathrm{d}t\\ &\leq \int \varphi(t)\left[(2i-1)!!\times\left(1+\sum_{k=1}^{i}\frac{2^k i(i-1)\cdots(i-k+1)}{(2k)!}t^{2k}\right)\right]\mathrm{d}t\\ &= (2i-1)!!\times\left(1+\sum_{k=1}^{i}\frac{2^k i(i-1)\cdots(i-k+1)}{(2k)!}\int t^{2k}\varphi(t)\mathrm{d}t\right)\\ &= (2i-1)!!\times\left(1+\sum_{k=1}^{i}\frac{2^k i(i-1)\cdots(i-k+1)}{(2k)!}(2k-1)!!\right)\\ &= (2i-1)!!\times\left(1+\sum_{k=1}^{i}\frac{i(i-1)\cdots(i-k+1)}{k!}\right)\\ &= 2^i\times(2i-1)!!.\end{aligned}$$

For sufficiently large n, $\theta_n < 1/2$ and it then follows from the above inequality that

$$\begin{aligned}\int \varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{(2i)!}\right|\theta_n^{2i}B^{2i}\mathrm{d}t &\leq \sum_{i=p}^{\infty}\frac{\theta_n^{2i}B^{2i}}{(2i)!}\int \varphi(t)\,|H_{2i}(t)|\,\mathrm{d}t \leq \sum_{i=p}^{\infty}\frac{\theta_n^{2i}B^{2i}}{(2i)!}2^i\times(2i-1)!!\\ &= \theta_n^{2p}\sum_{i=p}^{\infty}\frac{B^{2i}\theta_n^{2i-2p}}{i!} \leq \theta_n^{2p}\times \mathrm{e}^{B^2}\end{aligned}$$

and

$$\begin{aligned}\int \varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{2^i\cdot i!}\right|\theta_n^{2i}\mathrm{d}t &\leq \sum_{i=p}^{\infty}\frac{\theta_n^{2i}}{2^i\cdot i!}\int \varphi(t)\,|H_{2i}(t)|\,\mathrm{d}t\\ &\leq \sum_{i=p}^{\infty}\frac{\theta_n^{2i}}{2^i\cdot i!}2^i\times(2i-1)!! = \theta_n^{2p}\sum_{i=p}^{\infty}\frac{(2i-1)!!}{i!}\theta_n^{2i-2p}\\ &\leq \theta_n^{2p}\sum_{i=p}^{\infty}2^i\times\theta_n^{2i-2p} \leq \theta_n^{2p}\sum_{i=p}^{\infty}2^i\times\left(\frac{1}{2}\right)^{2i-2p}\\ &= \theta_n^{2p}\times 2^{2p+1}.\end{aligned}$$

Then from (25)

$$\int \sqrt{d_0(t)d_1(t)}\mathrm{d}t \geq 1 - \frac{1}{2}\theta_n^{2p}\times\mathrm{e}^{B^2} - \frac{1}{2}\theta_n^{2p}\times 2^{2p+1} = 1 - \theta_n^{2p}\left(\frac{1}{2}\mathrm{e}^{B^2} + 2^{2p}\right) \triangleq 1 - c\theta_n^{q+1}$$

where c is a constant that only depends on q. So

$$\rho(P_0, P_1) = \left(\int \sqrt{d_0(t)d_1(t)}\mathrm{d}t\right)^n \geq (1 - c\theta_n^{q+1})^n = \left(1 - cn^{-\frac{\alpha(q+1)}{d}}\right)^n.$$

Since $\frac{\alpha(q+1)}{d} \geq 1$, $\lim_{n\to\infty}(1 - cn^{-\frac{\alpha(q+1)}{d}})^n \geq \mathrm{e}^{-c} > 0$ and the theorem then follows. ■

Acknowledgments

The research of Tony Cai was supported in part by NSF Grants DMS-0306576 and DMS-0604954.

References

[1] L.D. Brown, M. Levine, Variance estimation in nonparametric regression via the difference sequence method, Ann. Statist. 35 (5) 2219–2232.
[2] H. Dette, A. Munk, T. Wagner, Estimating the variance in nonparametric regression—what is a reasonable choice? J. R. Statist. Soc. B 60 (1998) 751–764.
[3] J. Fan, I. Gijbels, Local Polynomial Modelling and its Applications, 1995.
[4] J. Fan, Q. Yao, Efficient estimation of conditional variance functions in stochastic regression, Biometrika 85 (1998) 645–660.
[5] T. Gasser, H.G. Müller, Kernel estimation of regression functions, in: Smoothing Techniques for Curve Estimation, in: Lecture Notes in Mathematics, vol. 757, Springer, Berlin, 1979, pp. 23–68.
[6] P. Hall, R.J. Carroll, Variance function estimation in regression: The effect of estimating the mean, J. R. Statist. Soc. B 51 (1989) 3–14.
[7] P. Hall, J.W. Kay, D.M. Titterington, Asymptotically optimal difference based estimation of variance in nonparametric regression, Biometrika 77 (1990) 521–528.
[8] P. Hall, J.W. Kay, D.M. Titterington, On estimation of noise variance in two-dimensional signal processing, Adv. Appl. Probab. 23 (1991) 476–495.
[9] P. Hall, J.S. Marron, On variance estimation in nonparametric regression, Biometrika 77 (1990) 415–419.
[10] S. Karlin, W.J. Studden, Tchebycheff Systems: With Applications In Analysis And Statistics, Interscience, New York, 1966.
[11] L. Le Cam, Asymptotic Methods in Statistical Decision Theory, Springer-Verlag, New York, 1986.
[12] H.-G. Müller, U. Stadtmüller, Estimation of heteroskedasticity in regression analysis, Ann. Statist. 15 (1987) 610–625.
[13] H.-G. Müller, U. Stadtmüller, On variance function estimation with quadratic forms, J. Stat. Plan. Inference 35 (1993) 213–231.
[14] H.G. Müller, U. Stadtmüller, Multivariate boundary kernels and a continuous least squares principle, J. R. Stat. Soc. B 61 (Part 2) (1999) 439–458.
[15] A. Munk, Bissantz, T. Wagner, G. Freitag, On difference based variance estimation in nonparametric regression when the covariate is high dimensional, J. R. Stat. Soc. B 67 (Part 1) (2005) 19–41.
[16] J. Rice, Bandwidth choice for nonparametric kernel regression, Ann. Statist. 12 (1984) 1215–1230.
[17] D. Ruppert, M.P. Wand, U. Holst, O. Hössjer, Local polynomial variance-function estimation, Technometrics 39 (1997) 262–273.
[18] V. Spokoiny, Variance estimation for high-dimensional regression problems, J. Multivariate Anal. 82 (2002) 111–133.
[19] J. von Neumann, Distribution of the ratio of the mean squared successive difference to the variance, Ann. Math. Statist. 12 (1941) 367–395.
[20] J. von Neumann, A further remark concerning the distribution of the ratio of the mean squared successive difference to the variance, Ann. Math. Statist. 13 (1942) 86–88.
[21] L. Wang, L.D. Brown, T. Cai, M. Levine, Effect of mean on variance function estimation on nonparametric regression, Ann. Statist. 36 (2008) 646–664.