Variance Function Estimation in Multivariate Nonparametric Regression

T. Tony Cai[1], Michael Levine* Lie Wang[1]

Abstract

Variance function estimation in multivariate nonparametric regression is considered and the minimax rate of convergence is established. Our work uses the approach that generalizes the one used in Munk et al (2005) for the constant variance case. As is the case when the number of dimensions $d = 1$, and very much contrary to the common practice, 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 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.

Keywords: Minimax estimation, nonparametric regression, variance estimation.
AMS 2000 Subject Classification: Primary: 62G08, 62G20.

[1] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. The research of Tony Cai was supported in part by NSF Grant DMS-0306576.

*Corresponding author. Address: 250 N. University Street, Purdue University, West Lafayette, IN 47907. E-mail: mlevins@stat.purdue.edu. Phone: 765-496-7571. Fax: 765-494-0558

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: $E\,|z_i| \le \mu_4 < \infty$. 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 $\boldsymbol{x}_i = \{x_{i_1}, \dots, x_{i_d}\}' \in \mathbb{R}^d$ where $i_k = 1, \dots, m$ for $k = 1, \dots, d$. Each coordinate is defined as

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

for $k = 1, \dots, d$. The overall sample size is $n = m^d$. The index i used in the model (1) is a d-dimensional index $i = (i_1, \dots, i_d)$. Both $g(\mathbf{x})$ and $V(\mathbf{x})$ are unknown functions supported on $S = [0,1]^d$. 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)

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

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

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

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. This paper is closely related to Munk et al (2005) where the problem of estimating a constant variance $V(\mathbf{x}) \equiv \boldsymbol{\sigma^2}$ in the multidimensional setup (1) is considered. They use a difference-based approach to variance estimation but note that "... Difference estimators are only applicable when homogeneous noise is present, i.e. the error variance does not depend on the regressor" (Munk et al (2005), p.20). We extend their difference-based approach to the case of 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 (2006) 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 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 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

$$\max\{n^{-\frac{4\alpha}{d}},\, n^{-\frac{2\beta}{2\beta+d}}\} \tag{5}$$

if g has α derivatives, V has β derivatives and d is the number of dimensions. 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 to study 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 (which corresponds to $\beta = \infty$), 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 obtained results 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 $m_k \equiv m$, $k = 1, \ldots, d$ and x_{i_k} is defined as a solution of the equation $\frac{i_k}{m_k} = \int_{-\infty}^{x_{i_k}} f_k(s)\, ds$ for a set of strictly positive densities $f_k(s)$. This does not change the conclusion of the paper and only adds a layer of technical complexity to the discussion. 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, von Neumann (1941, 1942), Rice (1984), Hall et al (1990), Hall and Marron (1990), Dette et al (1998). The multidimensional case was first considered when the dimensionality $d = 2$ in Hall et al (1991). The general case of estimating a constant variance in arbitrary dimension has only recently been investigated in Munk et al (2005). The estimation of the variance function $V(\boldsymbol{x})$ that depends on the covariate is a more recent topic. In the one-dimensional case, we can mention Müller and Stadtmüller (1987, 1993) and Brown and Levine (2006). 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 $J = \{j_1, \ldots, j_d\}$ as a sequence of nonnegative integers $j_1, \ldots, j_d$. For a fixed positive integer l, let $J(l) = \{J = (j_1, j_2, \ldots, j_d : |J| = l\}$. For an arbitrary function f, we define $D^l f = \frac{\partial^l f(\cdot)}{\partial x_1^{j_1} \ldots \partial x_d^{j_d}}$, if $|J| = l$. For any two vectors $\boldsymbol{x} = (x_1, \ldots, x_d)^{'}$ and $\boldsymbol{y} = (y_1 \ldots, y_d)^{'}$ we define the differential operator

$$D_{\boldsymbol{x},\boldsymbol{y}} = \sum_{k=1}^{d} (y_k - x_k) \frac{\partial}{\partial z_k} = \langle \boldsymbol{y} - \boldsymbol{x}, \nabla \rangle \tag{6}$$

where z_k is a generic kth argument of a d-dimensional function while ∇ is a gradient operator in $\mathbb{R}^d$. (6) is useful for writing the multivariate Taylor expansion in a concise form. For an arbitrary $\boldsymbol{x} \in \mathbb{R}^d$ we define $\boldsymbol{x}^J = x_1^{j_1} \ldots x_d^{j_d}$. 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 $|D^l f(\boldsymbol{x})| \leq M$ for $l = 0, 1, \ldots, \lfloor \alpha \rfloor$, and,*

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

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 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 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 Munk et al (2005). 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

$$R + J \equiv \{(i + j) | j \in J, i \in R\} \subseteq \otimes_{k=1}^{d} \{1, \dots, m\}. \tag{7}$$

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

$$D_{i^*} = \sum_{j \in J} d_j y_{i^*+j}. \tag{8}$$

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 Munk et al (2005).

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 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:

$$\begin{aligned} & K(\boldsymbol{x}) \text{ is supported on } T = [-1, 1]^d \,, \int_T K(\boldsymbol{x}) d\boldsymbol{x} = 1 \qquad (9) \\ & \int_T K(\boldsymbol{x}) \boldsymbol{x}^J \, d\boldsymbol{x} = 0 \quad \text{for } 0 < |J| < \lfloor \beta \rfloor \quad \text{and} \\ & \int_T K^2(\boldsymbol{x}) d\boldsymbol{x} = k_1 < \infty. \end{aligned}$$

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, Gasser and Müller (1979). It is, however, more difficult to define boundary kernels in 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 Müller and Stadtmüller (1999). 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 $S_n = \boldsymbol{x} + h(S - \boldsymbol{x})$ which changes with $\boldsymbol{x}$ and depends on n through the bandwidth h. Using this varying support set S_n, 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, we have the transposed and rescaled support of the boundary kernel as

$$T_{\boldsymbol{x}} = h^{-1}[\boldsymbol{x} - \{T_n(\boldsymbol{x}) \cap S_n\}] = h^{-1}(\boldsymbol{x} - \{\boldsymbol{x} + h(S - \boldsymbol{x})\} \cap (\boldsymbol{x} - hT)) = (\boldsymbol{x} - S) \cap T. \quad (10)$$

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 Müller and Stadtmüller (1999).

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

$$\widehat{V}(\boldsymbol{x}) = \sum_{i \in R} K_i^h(\boldsymbol{x}) D_i^2 = \sum_{i \in R} K_i^h(\boldsymbol{x}) \left(\sum_{j \in J} d_j y_{i+j} \right)^2 \quad (11)$$

The kernel weights are defined as

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

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:

$$J = \{(0,0,\ldots,0), (1,1,\ldots,1), \ldots, (\gamma,\gamma,\ldots,\gamma)\}. \quad (12)$$

In addition, we use normalized binomial coefficients as the difference coefficients. This is the so-called *polynomial sequence* (see, e.g. Munk et al (2005)) and is defined as

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

where $k = 0, 1, \ldots, \gamma$. 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 $q = 1, 2, \ldots, \gamma$.

Remark 1: It is possible to define a more general estimator by considering averaging over several possible d dimensional index sets J_l, $l = 1, \ldots, L$ and defining a set R_l for each one of them according to (7). In other words, we define

$$\widehat{V}(\boldsymbol{x}) = \sum_{l=1}^{L} \mu_l \sum_{i \in R_l} K_i^h(\boldsymbol{x}) D_i^2 = \sum_{l=1}^{L} \mu_l \sum_{i \in R_l} K_i^h(\boldsymbol{x}) \left(\sum_{j \in J_l} d_j y_{i+j} \right)^2 \tag{13}$$

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 $L = 1$ with an added layer of technical complication. Therefore, in this paper we will limit ourselves to the discussion of the case $L = 1$ and the definition (11) will be used with the set J selected as in (12).

Similarly to 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 (11) 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,*

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

and

$$\sup_{g \in \Lambda^\alpha(M_g), V \in \Lambda^\beta(M_V)} E \int_{R^d} (\widehat{V}(\boldsymbol{x}) - V(\boldsymbol{x}))^2 \, d\boldsymbol{x} \le C_0 \cdot \max\{n^{-\frac{4\alpha}{d}},\, n^{-\frac{2\beta}{2\beta+d}}\}. \tag{15}$$

Remark 2: The uniform rate of convergence given in (14) 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* $z_i \overset{iid}{\sim} N(0,1)$,

$$\inf_{\widehat{V}} \sup_{g\in\Lambda^\alpha(M_g), V\in\Lambda^\beta(M_V)} E\|\widehat{V}-V\|_2^2 \geq C_1 \cdot \max\{n^{-\frac{4\alpha}{d}},\ n^{-\frac{2\beta}{d+2\beta}}\} \tag{16}$$

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

$$\inf_{\widehat{V}} \sup_{g\in\Lambda^\alpha(M_g), V\in\Lambda^\beta(M_V)} E(\widehat{V}(\boldsymbol{x}_*)-V(\boldsymbol{x}_*))^2 \geq C_1 \cdot \max\{n^{-\frac{4\alpha}{d}},\ n^{-\frac{2\beta}{d+2\beta}}\} \tag{17}$$

where $C_1 > 0$ *is a constant.*

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

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

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 to study 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 then $d/4$ 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 case of $\beta = \infty$ which corresponds to the variance function V belonging to a known parametric family (see Hall and Carroll (1989)). Clearly, when $\beta \to \infty$ the cutoff $d\beta/(2(2\beta + d)) \to d/4$. Thus, when $d = 2$, any mean function g with less than 1/2 of a derivative will completely determine the rate of convergence for $\hat{V}$; when $d = 4$, any mean function with less than 1 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 dependant on the mean function. In other words, 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, we do 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 it has to be used is because the bias and variance of the mean estimator $\hat{g}$ have a very different influence on $\hat{V}$. As is the case when $d = 1$ (again, see Wang et al (2006)), 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 $n^{-2\alpha/d}$. On the contrary, 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 the way it has been done in the case of $d = 1$ by Wang et al (2006). 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 (14). Inequality (15) is a direct consequence of (14). Recall that $T = [-1, 1]^d$ is the support of the kernel K. Using the notation we introduced earlier, we can write the difference D_i as

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

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

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

has zero mean and unit variance. Thus,

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

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

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

where $k = max(k_1, k_2)$. In the above, $K_{**}(\boldsymbol{u}) = K(u)$ when $\boldsymbol{u} \in T_n(\boldsymbol{u}) \cap S$ and $K_{**}(\boldsymbol{u}) = K_*(u)$ when $\boldsymbol{u} \not\subseteq T_n(\boldsymbol{u}) \cap S$.

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

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

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

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

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

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

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

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

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

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

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

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

So,

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

Therefore, we have

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

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

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

To establish the last inequality it is important to remember that the fact that $V\in\Lambda^\beta(M_V)$ and therefore $|D^kV(\boldsymbol{x}_*)|\le M_V$. 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 square the above will become $o(n^{-2}h^{-2(d-1)})$; compared 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

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

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)

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

and

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

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

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

for some constant $C_0 > 0$. This proves (14). ∎

6 Proof of Theorem 2

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

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

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

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

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

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

and

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

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 (1966).

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.*

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

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, $i = 1, ..., n$, 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

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

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,

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

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,

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

where $\rho(P_0, P_1)$ is the Hellinger affinity between P_0 and P_1. See, for example, Le Cam (1986). Let p_0 and p_1 be the probability density function of P_0 and P_1 with respect to the Lebesgue measure μ, then $\rho(P_0, P_1) = \int \sqrt{p_0 p_1} d\mu$. 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

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

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

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

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

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, $\int v^i G(dv) = \int v^i \varphi(v)dv$ for $i = 0, 1, \cdots, q$. So,

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

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

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

and

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

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

$$\begin{aligned}\int \sqrt{d_0(t)d_1(t)}dt &\geq 1-\frac{1}{2}\int\left(\varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{(2i)!}\right|\theta_n^{2i}B^{2i}+\varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{2^i\cdot i!}\right|\theta_n^{2i}\right)dt\\ &= 1-\frac{1}{2}\int\varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{(2i)!}\right|\theta_n^{2i}B^{2i}dt-\frac{1}{2}\int\varphi(t)\sum_{i=p}^{\infty}\left|\frac{H_{2i}(t)}{2^i\cdot i!}\right|\theta_n^{2i}dt. \quad (25)\end{aligned}$$

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

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

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

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

and

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

Then from (25)

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

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

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

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

References

Brown, L.D., and Levine, M. (2006) "Variance Estimation in Nonparametric Regression via the Difference Sequence Method". *Ann. Statist.*, to appear.

Dette, H., Munk, A., and Wagner, T. (1998) "Estimating the variance in nonparametric regression-what is a reasonable choice?", *J. R. Statist. Soc. B*, 60, pp. 751-764.

Fan, J. and Yao, Q. (1998) "Efficient estimation of conditional variance functions in stochastic regression", *Biometrika*, 85, pp. 645-660.

Gasser, T. and Müller, H. G. (1979) "Kernel estimation of regression functions", in *Smoothing Techniques for Curve Estimation*, pp. 23-68. Berlin: Springer. (Lecture Notes in Mathematics No. 757)

Hall, P., and Carroll, R.J. (1989) "Variance Function Estimation in Regression: the Effect of Estimating the Mean", *J. R. Statist. Soc. B*, 51, pp. 3-14.

Hall, P., Kay, J.W., and Titterington, D.M. (1990) "Asymptotically optimal difference-based estimation of variance in nonparametric regression", *Biometrika*, 77, pp. 521-528.

Hall, P., Kay, J.W., and Titterington, D.M. (1991) "On Estimation of Noise Variance in Two-Dimensional Signal Processing", *Advances in Applied Probability*, 23, pp. 476-495.

Hall, P., and Marron, J.S. (1990) "On variance estimation in nonparametric regression", *Biometrika*, 77, pp. 415-419.

Härdle, W. and Tsybakov, A. (1997) "Local polynomial estimators of the volatility function in nonparametric autoregression", *Journal of Econometrics*, 81, pp. 223-242.

Karlin, S., and Studden, W. J. (1966) *Tchebycheff Systems: With Applications In Analysis And Statistics*, Interscience, New York.

Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*, Springer- Verlag, New York.

Müller, H.-G., and Stadtmüller, U. (1987) "Estimation of heteroskedasticity in regression analysis", *Ann. Statist.*, 15, pp. 610-625.

Müller, H.-G., and Stadtmüller, U. (1993) "On variance function estimation with quadratic forms", *Journal of Statistical Planning and Inference*, 35, pp. 213-231.

Müller, H.G., and Stadtmüller, U. (1999) "Multivariate Boundary Kernels and a Continuous Least Squares Principle", *J. R. Stat. Soc. B*, 61, Part 2, pp. 439-458.

Munk, A., Bissantz, Wagner, T. and Freitag, G. (2005) "On Difference-based Variance Estimation in Nonparametric Regression when the Covariate is High Dimensional", *J. R. Stat. Soc. B*, 67, Part 1, pp. 19-41.

Rice, J. (1984) "Bandwidth choice for nonparametric kernel regression", *Ann. Statist.*, 12, pp. 1215-1230.

Ruppert, D., Wand, M.P., Holst, U., and Hössjer, O. (1997) "Local Polynomial Variance-Function Estimation", *Technometrics*, 39, pp. 262-273.

von Neumann, J. (1941) "Distribution of the ratio of the mean squared successive difference to the variance", *Ann. Math. Statist.*, 12, pp. 367–395.

von Neumann, J. (1942) "A further remark concerning the distribution of the ratio of the mean squared successive difference to the variance", *Ann. Math. Statist.*, 13, pp. 86-88.

Wang, L., Brown, L.D., Cai, T. and Levine, M. (2006) "Effect of Mean on Variance Function Estimation on Nonparametric Regression" *Ann. Statist.*, to appear.