The Annals of Statistics
2020, Vol. 48, No. 6, 3589–3618
https://doi.org/10.1214/20-AOS1944

OPTIMAL ESTIMATION OF VARIANCE IN NONPARAMETRIC REGRESSION WITH RANDOM DESIGN

BY YANDI SHEN[1,*], CHAO GAO[2], DANIELA WITTEN[1,†] AND FANG HAN[1,‡]

[1]*Department of Statistics, University of Washington,* [*]*ydshen@uw.edu;* [†]*dwitten@uw.edu;* [‡]*fanghan@uw.edu*
[2]*Department of Statistics, University of Chicago, chaogao@galton.uchicago.edu*

Consider the heteroscedastic nonparametric regression model with random design

$$Y_i = f(X_i) + V^{1/2}(X_i)\varepsilon_i, \quad i = 1, 2, \ldots, n,$$

with $f(\cdot)$ and $V(\cdot)$ α- and β-Hölder smooth, respectively. We show that the minimax rate of estimating $V(\cdot)$ under both local and global squared risks is of the order

$$n^{-\frac{8\alpha\beta}{4\alpha\beta+2\alpha+\beta}} \vee n^{-\frac{2\beta}{2\beta+1}},$$

where $a \vee b := \max\{a, b\}$ for any two real numbers a, b. This result extends the fixed design rate $n^{-4\alpha} \vee n^{-2\beta/(2\beta+1)}$ derived in (*Ann. Statist.* **36** (2008) 646–664) in a nontrivial manner, as indicated by the appearances of both α and β in the first term. In the special case of constant variance, we show that the minimax rate is $n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$ for variance estimation, which further implies the same rate for quadratic functional estimation and thus unifies the minimax rate under the nonparametric regression model with those under the density model and the white noise model. To achieve the minimax rate, we develop a U-statistic-based local polynomial estimator and a lower bound that is constructed over a specified distribution family of randomness designed for both ε_i and X_i.

1. Introduction.

Consider the model

$$Y_i = f(X_i) + V^{1/2}(X_i)\varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\{X_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) univariate random design points, and $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. with zero mean, unit variance and are independent of $\{X_i\}_{i=1}^n$. In this paper, we study the optimal estimation of $V(\cdot)$ under both local and global squared risks. Variance estimation is a fundamental statistical problem (von Neumann (1941, 1942), Rice (1984), Hall, Kay and Titterington (1990)) with wide applications. It is useful in, for example, construction of confidence bands for the mean function, estimation of the signal-to-noise ratio (Verzelen and Gassiat (2018)), and selection of the optimal kernel bandwidth (Fan (1992)).

When $\{X_i\}_{i=1}^n$ are fixed, estimation of $V(\cdot)$ in (1) has been studied extensively in the literature via residual-based methods (Hall and Carroll (1989), Ruppert et al. (1997), Härdle and Tsybakov (1997), Fan and Yao (1998)) and difference-based methods (Müller and Stadtmüller (1987), Müller, Schick and Wefelmeyer (2003), Brown and Levine (2007), Wang et al. (2008)). One important heuristic from previous studies is that, compared to residual-based methods, difference-based methods are able to achiever a smaller bias and subsequently a smaller mean squared error by avoiding direct estimation of the mean function. More precisely, when $X_i = i/n$, $i = 1, \ldots, n$ and $f(\cdot)$ and $V(\cdot)$ in (1) are α- and β-Hölder smooth,

Received March 2019; revised December 2019.
MSC2020 subject classifications. 62G08, 62G20.
Key words and phrases. Variance estimation, nonparametric regression, random design, minimax rate, U-statistics.

respectively, Wang et al. (2008) proposed a difference estimator which achieved the optimal rate of the order $n^{-4\alpha} \vee n^{-\frac{2\beta}{2\beta+1}}$ under both local and global squared risks.

In contrast, our study focuses on the case where $\{X_i\}_{i=1}^n$ are i.i.d. random design points on the real line. For this, we show that when $f(\cdot)$ and $V(\cdot)$ in (1) are α- and β-Hölder smooth, respectively, the minimax rate of estimating $V(\cdot)$ is of the order $n^{-\frac{8\alpha\beta}{4\alpha\beta+2\alpha+\beta}} \vee n^{-\frac{2\beta}{2\beta+1}}$ under both local and global squared risks. This result has several noteworthy implications:

- The minimax rates in random and fixed design settings share a common component, $n^{-\frac{2\beta}{2\beta+1}}$, as well as the same transition boundary $\alpha = \beta/(4\beta+2)$.
- For $\alpha < \beta/(4\beta+2)$, a faster rate is achievable with a random design.
- Unlike the fixed design setting, for $\alpha < \beta/(4\beta+2)$, α and β are now both present in the first term of the minimax rate in the random design case.

We now discuss in more detail this minimax rate. The upper bound of the minimax rate is achieved by smoothing pairwise differences via local polynomial regression, the former of which is formulated via U-statistics. Our analysis of this estimator hence relies on the four-term Bernstein inequality in Giné, Latała and Zinn (2000), and unlike classic kernel methods, requires no smoothness assumption on the design density.

For the lower bound, due to the appearances of both α and β in the nontrivial $n^{-\frac{8\alpha\beta}{4\alpha\beta+2\alpha+\beta}}$ part of the minimax rate and the additional randomness of $\{X_i\}_{i=1}^n$, the derivation is much more involved than its counterpart in the fixed design setting. We tackle the first difficulty of entangled α and β via a proper localization technique in the construction of the mean function $f(\cdot)$, depicted in Figure 2 in Section 3.2. The second difficulty caused by the randomness of $\{X_i\}_{i=1}^n$ is resolved with a new trapezoid-shaped construction of the mean $f(\cdot)$, aided by a result due to Kolchin, Sevast'yanov and Chistyakov (1978) on the sparse multinomial distribution. This result helps characterize the asymptotic behavior of the locations of $\{X_i\}_{i=1}^n$ and plays a key role in our proof, but to our knowledge has not been well used in the nonparametric statistics literature.

In the special case of constant variance, (1) is reduced to

$$Y_i = f(X_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

and the goal becomes estimation of σ^2. In this case, the problem is linked to estimation of a quadratic functional, which has been studied in depth in the other two benchmark nonparametric models, the density model (Bickel and Ritov (1988), Laurent (1996), Giné and Nickl (2008)) and the white noise model (Donoho and Nussbaum (1990), Fan (1991), Laurent and Massart (2000)). In the density model, one observes an i.i.d. univariate sequence $\{X_i\}_{i=1}^n$ from some unknown density $f(\cdot)$, and the goal is to estimate $\int f^2(x)\,dx$. In the white noise model, one observes a continuous-time process from $dY_t = f(t)\,dt + n^{-1/2}\,dW_t$ for $t \in [0,1]$ with W_t a standard Wiener process. The goal is to estimate $\int_0^1 f^2(t)\,dt$. Under an α-smoothness condition on $f(\cdot)$, the minimax rate in both of the aforementioned two cases is $n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$ (cf. Theorem 1(ii) and 2(ii) in Bickel and Ritov (1988), Theorem 4 in Fan (1991)).

Following Doksum and Samarov (1995), a quadratic functional of interest under (2) with random design is

$$Q := \int f^2(x) p_X(x) w(x)\,dx, \tag{3}$$

where $p_X(\cdot)$ is the unknown design density and $w(\cdot) \geq 0$ is some known weight function. Assuming in (2) that f is α-Hölder smooth, we show that the minimax rate of estimating σ^2

TABLE 1

Summary of minimax rates in (1), (2), (4) and (5). The two types of fixed design considered, (GD) and (DD), are defined in (20) and (21), respectively. For a d-dimensional smoothness index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\top$, $\underline{\alpha} := d/(\sum_{k=1}^d 1/\alpha_k)$, $\alpha_{\min} := \min_{1\le k\le d} \alpha_k$, *and* $\alpha_{\max} := \max_{1\le k\le d} \alpha_k$. *The respective sections contain the definition of the distribution class of* $\{(X_i, \varepsilon_i)\}_{i=1}^n$ *in the random design setting and distribution class of* $\{\varepsilon_i\}_{i=1}^n$ *in the fixed design setting. Our results include all of the random design rates and fixed design rates in (4) and (5). Note results for (4) and (5) have additional requirements; see Sections 4.1 and 4.2 for details*

	Stated in	Minimax rate	Boundary
(1), fixed	Wang et al. (2008)	$n^{-4\alpha} \vee n^{-2\beta/(2\beta+1)}$	$\alpha = \beta/(4\beta+2)$
(1), random	Theorems 3, 4, 5	$n^{-\frac{8\alpha\beta}{4\alpha\beta+\beta+2\alpha}} \vee n^{-\frac{2\beta}{2\beta+1}}$	
(2), fixed	Wang et al. (2008)	$n^{-4\alpha} \vee n^{-1}$	$\alpha = 1/4$
(2), random	Theorems 1, 2	$n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$	
(4), fixed (GD)	Proposition 3	$n^{-4\alpha_{\max}/d} \vee n^{-1}$	$\alpha_{\max} = d/4$
(4), fixed (DD)	Proposition 4	$n^{-4\alpha_{\min}} \vee n^{-1}$	$\alpha_{\min} = 1/4$
(4), random	Propositions 1, 2	$n^{-8\underline{\alpha}/(4\underline{\alpha}+d)} \vee n^{-1}$	$\underline{\alpha} = d/4$
(5), fixed (GD)	Proposition 5	n^{-1}	–
(5), fixed (DD)	Proposition 6	$n^{-4\alpha_{\min}} \vee n^{-1}$	$\alpha_{\min} = 1/4$
(5), random	Propositions 7, 8	$n^{-8\alpha_{\min}/(4\alpha_{\min}+1)} \vee n^{-1}$	

and Q (when σ^2 is unknown) is $n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$, thereby unifying the minimax rate of quadratic functional estimation in all three benchmark nonparametric models.

In this paper, we also provide extensions of (2) to multivariate cases, with a focus on the multivariate nonparametric regression model

$$Y_i = f(\boldsymbol{X}_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{4}$$

and the nonparametric additive model

$$Y_i = \sum_{k=1}^d f_k(X_{i,k}) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{5}$$

in both fixed and random designs. Here, $\boldsymbol{X}_i := (X_{i,1}, \ldots, X_{i,d})^\top$, $i = 1, \ldots, n$, for some fixed positive integer d. Regarding the fixed design, we consider two types, namely, the grid design (GD) and the diagonal design (DD). With a total of n design points, the former places them on a regular grid in the d-dimensional cube $[0, 1]^d$ while the latter only places design points on the diagonal. Details are given in Sections 4.1 and 4.2.

We summarize the minimax rates in all of the aforementioned models in Table 1.

The rest of the paper is organized as follows. Section 2 presents the simple model (2) with constant variance. Section 3 discusses its heteroscedastic extension (1). Section 4 discusses the multivariate nonparametric regression model (4), the additive model (5) and several other extensions of our main results. The essential lower bound proof of the minimax rate $n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$ under model (2) is presented in Section 5, with the rest of the proofs given in the Supplementary Material (Shen et al. (2020)).

The notation used throughout the paper is as follows. For any positive integer n, $[n]$ denotes the set $\{1, 2, \ldots, n\}$. For any real number a, we use $\lceil a \rceil$ to denote the smallest integer greater than or equal to a, and $\lfloor a \rfloor$ the largest integer strictly smaller than a. For any positive integer d, $\mathbf{0}_d$ denotes the zero vector of dimension d and $\mathbf{I}_d$ denotes the identity matrix of dimension d. For a real vector x, $\|x\|$ and $\|x\|_\infty$ denote its Euclidean and infinity norms, respectively. For a real matrix $\mathbf{A}$, we use $\|\mathbf{A}\|$, $\|\mathbf{A}\|_F$, and $|\mathbf{A}|$ to denote its spectral norm, Frobenius norm,

and determinant, respectively. For an m-times differentiable function $f : \mathbb{R} \to \mathbb{R}$ with some positive integer m, we use $f^{(k)}$ to denote its kth derivative for $k = 1, 2, \ldots, m$. For identically distributed random variables X_i and X_j, we use $\mathbb{P}_{X_i}(\cdot)$ and $p_{X_i}(\cdot)$ to denote the distribution and density of X_i, $\widetilde{X}_{ij}$ to denote $X_i - X_j$, and $p_{\widetilde{X}_{ij}}(\cdot)$ to denote the density of $X_i - X_j$. Similar notation $\mathbb{P}_{\boldsymbol{X}_i}(\cdot)$, $p_{\boldsymbol{X}_i}(\cdot)$, $\widetilde{\boldsymbol{X}}_{ij}$, $p_{\widetilde{\boldsymbol{X}}_{ij}}(\cdot)$ applies to identically distributed random vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. For a positive integer d and $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We will drop the subscript d for simplicity when $d = 1$. $\Phi(\cdot)$ and $\varphi(\cdot)$ represent the standard normal distribution and density. More generally, we will write $\varphi_{\mu,\sigma^2}(\cdot)$ as the density for the normal distribution with mean μ and variance σ^2. For two probability measures $\mathbb{P}$, $\mathbb{Q}$ defined on a common space $(\Omega, \mathcal{A})$, $\mathsf{TV}(\mathbb{P}, \mathbb{Q})$ denotes their total variation distance, that is, $\mathsf{TV}(\mathbb{P}, \mathbb{Q}) := \sup_{A\in\mathcal{A}} |\mathbb{P}(A) - \mathbb{Q}(A)|$. For two real sequences $\{a_n\}$ and $\{b_n\}$, $a_n \lesssim b_n$ if $|a_n| \le C|b_n|$ for some positive absolute constant C. We say $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2. Homoscedastic case. To illustrate some of the main ideas developed in this paper, we begin with a discussion of the elementary univariate homoscedastic nonparametric regression model (2):

$$Y_i = f(X_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Here, $\{X_i\}_{i=1}^n$ are i.i.d. copies of a univariate random variable X, $f(\cdot)$ belongs to an α-Hölder class that will be specified soon, and $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. copies of a variable ε with zero mean and unit variance and are independent of $\{X_i\}_{i=1}^n$. Both the mean function $f(\cdot)$ and the distribution of $\{X_i\}_{i=1}^n$ are assumed unknown.

Model (2) has been extensively studied using residual-based and difference-based methods; see, among many others, von Neumann (1941, 1942), Rice (1984), Gasser, Sroka and Jennen-Steinmetz (1986), Hall, Kay and Titterington (1990), Hall and Marron (1990), Thompson, Kay and Titterington (1991), Müller, Schick and Wefelmeyer (2003), Wang et al. (2008). A related functional estimation problem has also been studied in semiparametric models (Robins et al. (2008, 2009)). Most of the previous studies focus on the case of fixed design, especially the equidistant design with $X_i = i/n$, $i \in [n]$, for which the minimax rate of estimating σ^2 under an α-Hölder smoothness constraint on $f(\cdot)$ is known to be $n^{-4\alpha} \vee n^{-1}$ (cf. Theorems 1 and 2 in Wang et al. (2008)).

In detail, let I be a fixed (possibly infinite) interval on the real line. Define the Hölder class $\Lambda_{\alpha,I}(C_{\mathcal{F}})$ on I as follows:

$$\begin{aligned}\Lambda_{\alpha,I}(C_{\mathcal{F}}) := \big\{ f :\ & \text{for all } x, y \in I \text{ and } k = 0, \ldots, \lfloor\alpha\rfloor, \\ & \big|f^{(k)}(x)\big| \le C_{\mathcal{F}} \text{ and } \big|f^{(\lfloor\alpha\rfloor)}(x) - f^{(\lfloor\alpha\rfloor)}(y)\big| \le C_{\mathcal{F}}|x - y|^{\alpha'}\big\},\end{aligned} \tag{6}$$

where $\alpha' := \alpha - \lfloor\alpha\rfloor$. Denote the support of X as $\operatorname{supp}(X)$.

Define the joint distribution class $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$ (where "cv" stands for "constant variance") with the following conditions:

(a) X satisfies $\operatorname{supp}(X) \subset I$.

(b) X has density $p_X(\cdot)$ and there exists a fixed positive constant C_0 such that

$$\sup_{x\in\mathbb{R}} p_X(x) \le C_0.$$

(c) There exist two fixed constants $\delta_0 > 0$ and $c_0 > 0$ such that for any $0 < \delta < \delta_0$, there exists a set $\mathcal{U}_\delta \subset [-1, 1]$ such that

$$\lambda(\mathcal{U}_\delta) \ge c_0 \quad \text{and} \quad \inf_{u\in\mathcal{U}_\delta} p_{\widetilde{X}_{ij}}(u\delta) \ge c_0,$$

where $\lambda(\cdot)$ represents the Lebesgue measure on the real line, and $\widetilde{X}_{ij} = X_i - X_j$.

(d) $\mathbb{E}\varepsilon^4 \le C_\varepsilon$ for some fixed positive constant C_ε.

Note that no smoothness condition is placed on the density of X. Condition (c) essentially requires the density $p_{\widetilde{X}_{ij}}$ to be "dense" around 0, and is strictly weaker than a uniform lower bound of $p_{\widetilde{X}_{ij}}$ over a fixed neighborhood of 0. It also follows from the following sufficient condition on the marginal density $p_X(\cdot)$ (see Lemma A4 in the Supplementary Material (Shen et al. (2020)) for the justification):

(c′) X is compactly supported (taken to be $[0, 1]$ without loss of generality). There exists some positive constant c_0 and subset $S \subset [-1, 1]$ with Lebesgue measure $\lambda(S) \geq 3/4$ such that $p_X(t) \geq c_0$ uniformly over $t \in S$.

In particular, (c′) covers the uniform distribution on $[0, 1]$ and the distribution of X in the lower bound construction in the proof of Theorem 2.

The rest of the section is devoted to proving, for any fixed positive constants $C_{\mathcal{F}}$ and C_σ, the following minimax rate:

$$\inf_{\widetilde{\sigma}^2} \sup_{f \in \Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{\sigma^2 \leq C_\sigma} \sup_{\mathbb{P}_{(X,\varepsilon)} \in \mathcal{P}_{\mathrm{cv},(X,\varepsilon)}} \mathbb{E}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2 \asymp n^{-8\alpha/(4\alpha+1)} \vee n^{-1}, \tag{7}$$

where $\mathbb{P}_{(X,\varepsilon)}$ denotes the joint distribution of (X, ε), and $\widetilde{\sigma}^2$ ranges over all estimators of σ^2.

2.1. *Upper bound.* The upper bound is achieved by a difference estimator based on U-statistics (with convention $0/0 = 0$):

$$\widehat{\sigma}^2 := \frac{\binom{n}{2}^{-1} \sum_{i<j} K_h(X_i - X_j)(Y_i - Y_j)^2/2}{\binom{n}{2}^{-1} \sum_{i<j} K_h(X_i - X_j)}. \tag{8}$$

Here, $K_h(\cdot) := K(\cdot/h)/h$, where $h = h_n$ is a bandwidth parameter satisfying $h_n \downarrow 0$ as $n \to \infty$, and $K(\cdot)$ is a symmetric density kernel supported on $[-1, 1]$ that satisfies

$$\underline{M}_K \leq \inf_{|u| \leq 1} K(u) \leq \sup_{|u| \leq 1} K(u) \leq \overline{M}_K \tag{9}$$

for two fixed constants $\overline{M}_K$ and $\underline{M}_K$; one example is the box kernel $K(u) = \mathbb{1}\{|u| \leq 1\}/2$ which satisfies (9) with $\overline{M}_K = \underline{M}_K = 1/2$.

The following error bound is derived via the exponential inequality for degenerate U-statistics due to Giné, Latała and Zinn (2000).

THEOREM 1. *Suppose the kernel $K(\cdot)$ in $\widehat{\sigma}^2$ is chosen such that (9) is satisfied with constants $\overline{M}_K$ and $\underline{M}_K$, and the bandwidth h_n is chosen as*

$$h_n \asymp \begin{cases} n^{-2/(4\alpha+1)}, & 0 < \alpha < 1/4, \\ n^{-1}, & \alpha \geq 1/4. \end{cases} \tag{10}$$

Then, under (2) with random design, it holds that

$$\sup_{f \in \Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{\sigma^2 \leq C_\sigma} \sup_{\mathbb{P}_{(X,\varepsilon)} \in \mathcal{P}_{\mathrm{cv},(X,\varepsilon)}} \mathbb{E}\big(\widehat{\sigma}^2 - \sigma^2\big)^2 \leq C\big(n^{-8\alpha/(4\alpha+1)} \vee n^{-1}\big),$$

where C is some fixed positive constant that only depends on $\overline{M}_K$, $\underline{M}_K$, α, $C_{\mathcal{F}}$, C_σ and C_0, c_0, C_ε in $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$.

REMARK 1. The error rate in Theorem 1 is achieved by choosing the optimal bandwidth h_n to balance the "bias-variance" decomposition:

$$\big\{\mathbb{E}\big(\widehat{\sigma}^2 - \sigma^2\big)^2\big\}^{1/2} \lesssim h_n^{2(\alpha \wedge 1)} + \frac{1}{nh_n^{1/2}}, \tag{11}$$

where $a \wedge b := \min\{a, b\}$ for any two real numbers a, b. The bias term $h_n^{2(\alpha \wedge 1)}$ reflects the second-order effect of the unknown mean on variance estimation, which has been noted by Hall and Carroll (1989) and Wang et al. (2008). The variance part follows from the fact that there is an average number of $n^2 h_n$ pairs of (i, j) such that $|X_i - X_j| \le h_n$. We note that the same "bias-variance" decomposition has appeared in quadratic functional estimation in the density model and Gaussian sequence model (Bickel and Ritov (1988), Fan (1991), Giné and Nickl (2008)). See Section 4.3 for a more detailed discussion.

REMARK 2. While most of the previous works are in the context of fixed design, Müller, Schick and Wefelmeyer (2003) considered constant variance estimation with random design, and their estimator (formula (1.4) therein) is almost identical to our $\widehat{\sigma}^2$. Under certain assumptions (Assumptions 1 and 2 and (2.4)–(2.7) therein), they show that their estimator is root-n consistent and asymptotically normal. However, as commented in the first paragraph on p. 184 of their paper, their condition (2.7) is only satisfied when the mean function smoothness α is strictly larger than $1/4$, and no analysis is provided below this threshold. Our minimax rate $n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$ therefore confirms that $\alpha \ge 1/4$ is indeed the minimal requirement for any variance estimator to be root-n consistent and we also demonstrate the optimality of $\widehat{\sigma}^2$ for $0 < \alpha < 1/4$.

Finally, in (2), we have assumed that the smoothness index α is known. If it is unknown, then the variance can be estimated adaptively via Lepski-type methods (Lepski (1991, 1992)). This is discussed in more detail in Section 4.5.

2.2. *Lower bound.* The derivation of the lower bound in (7) is much more involved. In particular, the construction in the fixed design setting (cf. Theorem 2 in Wang et al. (2008)) cannot be extended to the random design case, since the spike-type construction of $f(\cdot)$ located at each deterministic design point leads to a suboptimal rate in the random design setting. To achieve a sharp rate, we have to exploit the randomness of $\{X_i\}_{i=1}^n$; this requires us to handle a highly convoluted alternative hypothesis that no longer leads to a product measure of $\{Y_i\}_{i=1}^n$ given each realization of $\{X_i\}_{i=1}^n$ in LeCam's two-point method. This calls for a careful analysis of the locations of $\{X_i\}_{i=1}^n$.

We now sketch a proof of the $n^{-8\alpha/(4\alpha+1)}$ component in (7) for $0 < \alpha < 1/4$, with a particular emphasis on where the difference arises with the fixed design setting. The proof can be roughly divided into two steps. In the first step, we construct a two-point testing problem with the null being a Gaussian (H_0) and the alternative a Gaussian location mixture ($\widetilde{H}_1$). In the second step, we approximate the Gaussian location mixture ($\widetilde{H}_1$) by a location mixture with compact support (H_1), which, unlike the alternative in the first step, belongs to the considered model class.

We start by introducing the construction of $f(\cdot)$, σ^2, ε, and X under the null H_0 and the alternative $\widetilde{H}_1$ in the first step. For each n, let

$$h_n \asymp n^{-2/(4\alpha+1)}, \qquad \theta_n^2 \asymp h_n^{2\alpha} \quad \text{and} \quad N := 1/(6h_n),$$

and divide the unit interval $[0, 1]$ into N intervals of length $6h_n$, with n large enough and h_n chosen such that N is a positive integer.

Choice of $f(\cdot)$: Under H_0, let $f \equiv 0$. Under $\widetilde{H}_1$, let $f(\cdot)$ be a piecewise trapezoidal function on the N intervals. That is, for each $i \in [N]$, f takes on a value of $h_n^{\alpha}\widetilde{r}_i$ on the intervals $[(6i-5)h_n, (6i-1)h_n]$ and then linearly decreases to zero on the two endpoints $6(i-1)h_n$ and $6ih_n$, with $\{\widetilde{r}_i\}_{i=1}^N$ i.i.d. standard normal variables.

Choice of σ^2: Under H_0, let $\sigma^2 = 1 + \theta_n^2$. Under $\widetilde{H}_1$, let $\sigma^2 = 1$.

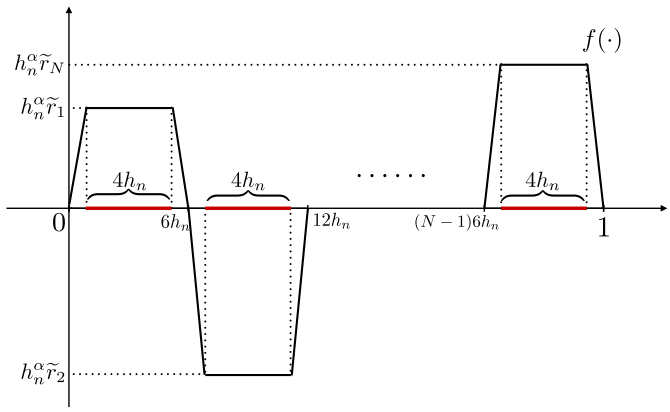

FIG. 1. *The black solid line represents the construction of $f(\cdot)$ under the alternative hypothesis $\widetilde{H}_1$. The thick red segments indicate the support of X under both H_0 and $\widetilde{H}_1$, on which X is uniformly distributed. Here, $h_n \asymp n^{-2/(4\alpha+1)}$ and is chosen such that $N := 1/(6h_n)$ is a positive integer. $\{\widetilde{r}_i\}_{i=1}^N$ are N i.i.d. standard normal variables.*

Choice of ε: Under both H_0 and $\widetilde{H}_1$, let $\varepsilon \sim \mathcal{N}(0,1)$.

Choice of X: Under both H_0 and $\widetilde{H}_1$, let $\{X_i\}_{i=1}^n$ be uniformly distributed over the union of the upper bases of the trapezoids, that is, over $\bigcup_{i=1}^N [(6i-5)h_n, (6i-1)h_n]$.

See Figure 1 for an illustration of the construction.

In contrast to the spike-type construction of $f(\cdot)$ in the fixed design setting, our construction is trapezoid-shaped, which guarantees a maximal variation in the mean to compensate for the difference in the variance under the null and alternative. This is unnecessary in the fixed design setting since the point of maximal variation in the mean (center of each spike) can be directly placed at each fixed $X_i = i/n$, resulting in n evenly spaced spikes in $f(\cdot)$.

Denote the joint distribution of $\{(X_i, Y_i)\}_{i=1}^n$ under H_0 and $\widetilde{H}_1$ by $\mathbb{P}_0$ and $\widetilde{\mathbb{P}}_1$ with respective density p_0 and $\widetilde{p}_1$. Under the above construction, conditional on $\{X_i\}_{i=1}^n$, $\{Y_i\}_{i=1}^n$ are distributed as

$$H_0 : p_0\big(\{Y_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n\big) = \prod_{i=1}^n \varphi_{0,1+\theta_n^2}(Y_i)$$

and

$$\widetilde{H}_1 : \widetilde{p}_1\big(\{Y_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n\big) = \prod_{j=1}^N \int \Big(\prod_{\{i:b_i=j\}} \varphi_{h_n^\alpha v,1}(Y_i) \Big) \varphi(v)\, dv,$$

where $\{b_i\}_{i=1}^n$ is the location index sequence of $\{X_i\}_{i=1}^n$ defined as

$$b_i := j \quad \text{if } X_i \in \big[(6j-5)h_n, (6j-1)h_n\big],$$

which characterizes which trapezoid each X_i falls into. Using Lemma 2 that will be stated in Section 5, one can then upper bound

$$\mathsf{TV}(\mathbb{P}_0, \widetilde{\mathbb{P}}_1) = \mathbb{E}\mathsf{TV}\big(\mathbb{P}_0\big(\{Y_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n\big), \qquad \widetilde{\mathbb{P}}_1\big(\{Y_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n\big)\big) \lesssim \theta_n^2 n h_n^{1/2},$$

which can be made smaller than a sufficiently small constant c by choosing h_n sufficiently small.

The second step of the proof aims to find a sequence of bounded random variables $\{r_i\}_{i=1}^N$ to replace the standard normal sequence $\{\widetilde{r}_i\}_{i=1}^N$ in $\widetilde{\mathbb{P}}_1$, so that for each realization of $\{r_i\}_{i=1}^N$,

the corresponding $f(\cdot)$ in the alternative is α-Hölder smooth with a fixed constant. Then, denoting the distribution of $\{r_i\}_{i=1}^N$ as $\mathbb{G}$, one wishes to approximate the conditional distribution $\widetilde{\mathbb{P}}_1(\{Y_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n)$ in $\widetilde{H}_1$ by $\mathbb{P}_1(\{Y_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n)$ with density

$$p_1(\{Y_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n) = \prod_{j=1}^N \int \Big(\prod_{\{i:b_i=j\}} \varphi_{h_n^\alpha v,1}(Y_i) \Big) \mathbb{G}(dv)$$

in H_1. Even with the aid of moment matching techniques already established in the literature, upper bounding $\mathsf{TV}(\mathbb{P}_1, \widetilde{\mathbb{P}}_1)$ is still nontrivial. Specifically, unlike in the fixed design setting, now with high probability the conditional distribution of $\{Y_i\}_{i=1}^n$ given $\{X_i\}_{i=1}^n$ is no longer a product measure. This is because multiple X_i's could fall into the same trapezoid in the construction of $f(\cdot)$. This can be handled relatively easily in the first step since there we only have to analyze the pairwise correlation of $Y_i \mid X_i$ and $Y_j \mid X_j$ depending on whether X_i and X_j fall into the same trapezoid, but it is much less tractable in the second step. More specifically, in order to match moments, we now have to divide the X_i's into groups based on their memberships among the trapezoids, which naturally requires us to monitor the locations of $\{X_i\}_{i=1}^n$, and in particular the number of X_i's that fall into the same trapezoid. This is possible by observing that the memberships of $\{X_i\}_{i=1}^n$ now follow a sparse multinomial distribution ($n^{2/(4\alpha+1)}$ bins, n balls) so that a result in Kolchin, Sevast'yanov and Chistyakov (1978) can be applied. This allows us to show that with high probability the maximum number of X_i's in each trapezoid is bounded by a fixed constant, which, along with Lemma 1 in Section 5, allows us to calculate

$$\mathsf{TV}(\mathbb{P}_1, \widetilde{\mathbb{P}}_1) \lesssim n\theta_n^{2p}$$

for $p := 1 + \lceil 1/4\alpha \rceil$. This indicates that $\mathsf{TV}(\mathbb{P}_1, \widetilde{\mathbb{P}}_1)$ is smaller than some sufficiently small constant c. Then, by the triangle inequality,

$$\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) \le \mathsf{TV}(\mathbb{P}_0, \widetilde{\mathbb{P}}_1) + \mathsf{TV}(\mathbb{P}_1, \widetilde{\mathbb{P}}_1) \le 2c.$$

Details of the above derivation will be given in Section 5. The resulting lower bound is as follows.

THEOREM 2. *Under* (2) *with random design, it holds that*

$$\inf_{\widetilde{\sigma}^2} \sup_{f \in \Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{\sigma^2 \le C_\sigma} \sup_{\mathbb{P}_{(X,\varepsilon)} \in \mathcal{P}_{\mathrm{cv},(X,\varepsilon)}} \mathbb{E}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2 \ge c\big(n^{-8\alpha/(4\alpha+1)} \vee n^{-1}\big),$$

where c is some fixed positive constant that only depends on α, $C_{\mathcal{F}}$, C_σ and C_0, c_0, C_ε in $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$, and $\widetilde{\sigma}^2$ ranges over all estimators of σ^2.

REMARK 3. It remains an open problem to prove a lower bound rate that is strictly slower than n^{-1} over the sub-class of $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$ with more regular designs, which includes in particular the uniform design on $[0, 1]$. We conjecture that in this case, $n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$ is still the minimax rate in view of analogous results in quadratic functional estimation (Bickel and Ritov (1988), Fan (1991)).

3. Heteroscedastic case. We now study the heteroscedastic model (1),

$$Y_i = f(X_i) + V^{1/2}(X_i)\varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\{X_i\}_{i=1}^n$ are i.i.d. copies of X on the real line, $f(\cdot)$ and $V(\cdot)$ are α- and β-Hölder smooth on the fixed (possibly infinite) interval I, respectively, and $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. copies of ε with zero mean and unit variance and are independent of $\{X_i\}_{i=1}^n$. As in Section 2,

smoothness indices α and β are assumed known, while $f(\cdot)$, $V(\cdot)$, and the distribution of X are unknown. For any estimator $\widetilde{V}(\cdot)$, the estimation accuracy is measured both locally via

$$R_1(\widetilde{V}, V; x^*) := \big(\widetilde{V}(x^*) - V(x^*)\big)^2 \tag{12}$$

at a point x^* in the support of X, $\operatorname{supp}(X)$, and globally via

$$R_2(\widetilde{V}, V) := \int \big(\widetilde{V}(x) - V(x)\big)^2 \mathbb{P}_X(dx) \tag{13}$$

with $\mathbb{P}_X$ the distribution of X.

Model (1) has been studied in, for example, Müller and Stadtmüller (1987), Hall and Carroll (1989), Ruppert et al. (1997), Härdle and Tsybakov (1997), Fan and Yao (1998), Munk and Ruymgaart (2002), Brown and Levine (2007), Wang et al. (2008), with a focus mainly on the fixed design case. An exception is Munk and Ruymgaart (2002), with which we draw a detailed comparison in Remark 8 below. Theorems 1 and 2 in Wang et al. (2008) established a minimax rate of the order $n^{-4\alpha} \vee n^{-2\beta/(2\beta+1)}$ under equidistance design $X_i = i/n$, $i \in [n]$ when $f(\cdot)$ and $V(\cdot)$ are α- and β-Hölder smooth on $[0, 1]$.

Define $\mathcal{P}_{\mathrm{vf},(X,\varepsilon)}$ (where "vf" stands for "variance function") as follows:

(a) X satisfies $\operatorname{supp}(X) \subset I$.

(b) X has density $p_X(\cdot)$, and there exists a fixed positive constant C_0 such that

$$\sup_{x\in\mathbb{R}} p_X(x) \le C_0.$$

(c) There exist fixed positive constants c_0 and δ_0 such that

$$\inf_{x^*\in\operatorname{supp}(X)} p_X(x^*) \ge c_0 \quad \text{and}$$

$$\inf_{0<\delta<\delta_0} \inf_{x^*\in\operatorname{supp}(X)} \lambda\big(\{u \in [-1, 1] : x^* + \delta u \in \operatorname{supp}(X)\}\big) \ge c_0,$$

where $\lambda(\cdot)$ is the Lebesgue measure on the real line.

(d) $\mathbb{E}\varepsilon^4 \le C_\varepsilon$ for some fixed positive constant C_ε.

One can readily verify that $\mathcal{P}_{\mathrm{vf},(X,\varepsilon)} \subset \mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$, with the latter defined in the beginning of Section 2. Compared to $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$, Condition (c) in $\mathcal{P}_{\mathrm{vf},(X,\varepsilon)}$ is posed on the marginal density and support of X, since in the variance function case we require a sufficient number of close pairs (X_i, X_j) around each target x^*. We also note that, as in $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$, no smoothness assumption is posed on the design density in $\mathcal{P}_{\mathrm{vf},(X,\varepsilon)}$.

The rest of the section is devoted to proving, for any fixed positive constants $C_{\mathcal{F}}$ and $C_{\mathcal{V}}$, the following minimax rates:

$$\begin{aligned}
&\inf_{\widetilde{V}} \sup_{f\in\Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{V\in\Lambda_{\beta,I}(C_{\mathcal{V}})} \sup_{\mathbb{P}_{(X,\varepsilon)}\in\mathcal{P}_{\mathrm{vf},(X,\varepsilon)}} \sup_{x^*\in\operatorname{supp}(X)} \mathbb{E}R_1(\widetilde{V}, V; x^*) \\
&\quad \asymp n^{-\frac{8\alpha\beta}{4\alpha\beta+2\alpha+\beta}} \vee n^{-\frac{2\beta}{2\beta+1}}, \qquad (14) \\
&\inf_{\widetilde{V}} \sup_{f\in\Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{V\in\Lambda_{\beta,I}(C_{\mathcal{V}})} \sup_{\mathbb{P}_{(X,\varepsilon)}\in\mathcal{P}_{\mathrm{vf},(X,\varepsilon)}} \mathbb{E}R_2(\widetilde{V}, V) \asymp n^{-\frac{8\alpha\beta}{4\alpha\beta+2\alpha+\beta}} \vee n^{-\frac{2\beta}{2\beta+1}},
\end{aligned}$$

where $\mathbb{P}_{(X,\varepsilon)}$ denotes the joint distribution of (X, ε), and $\widetilde{V}(\cdot)$ ranges over all estimators of $V(\cdot)$.

3.1. *Upper bound.* We now propose an estimator of $V(x^*)$ for some fixed $x^* \in \operatorname{supp}(X)$ by combining pairwise differences with local polynomial regression. We first introduce some notation. Let ℓ be the largest integer strictly smaller than β and

$$\boldsymbol{q}(u) := \big(1, u, u^2/2!, \ldots, u^\ell/\ell!\big)^\top.$$

For any $1 \le i < j \le n$, define

$$D_{ij} := (Y_i - Y_j)^2/2, \qquad X_{ij} := (X_i + X_j)/2 \quad \text{and} \quad K_{ij} := K_{h_1}(X_i - X_j)K_{h_2}(X_{ij} - x^*),$$

where h_1, h_2 are two bandwidths. Define an $(\ell+1) \times (\ell+1)$ matrix

$$\mathbf{B}_n := \binom{n}{2}^{-1} \sum_{i<j} \boldsymbol{q}\Big(\frac{X_{ij} - x^*}{h_2}\Big)\boldsymbol{q}^\top\Big(\frac{X_{ij} - x^*}{h_2}\Big)K_{ij}$$

and $\mathbf{B}_n^*$ as its adjugate such that $\mathbf{B}_n\mathbf{B}_n^* = \mathbf{B}_n^*\mathbf{B}_n = |\mathbf{B}_n|\mathbf{I}_{\ell+1}$. For example, when $\ell = 1$, we have

$$\mathbf{B}_n = \begin{bmatrix} s_0 & s_1 \\ s_1 & s_2 \end{bmatrix}, \qquad \mathbf{B}_n^* = \begin{bmatrix} s_2 & -s_1 \\ -s_1 & s_0 \end{bmatrix} \quad \text{and} \quad |\mathbf{B}_n| = s_0 s_2 - s_1^2,$$

where

$$s_k := \binom{n}{2}^{-1} \sum_{i<j} \Big(\frac{X_{ij} - x^*}{h_2}\Big)^k K_{ij}, \quad k = 0, 1, 2.$$

Following Fan (1993), we propose a robust local polynomial estimator:

$$\widehat{V}_{\mathrm{LP}}(x^*) := \binom{n}{2}^{-1} \sum_{i<j} D_{ij}\big(|\mathbf{B}_n| + \tau_n\big)^{-1}\boldsymbol{q}^\top(0)\mathbf{B}_n^*\boldsymbol{q}\Big(\frac{X_{ij} - x^*}{h_2}\Big)K_{ij}, \tag{15}$$

where τ_n is some sufficiently small positive constant that decays to 0 polynomially with n. Let

$$w_{ij} := \binom{n}{2}^{-1} \boldsymbol{q}^\top(0)\mathbf{B}_n^*\boldsymbol{q}\Big(\frac{X_{ij} - x^*}{h_2}\Big)K_{ij} \quad \text{and} \quad \widetilde{w}_{ij} := w_{ij}/\big(|\mathbf{B}_n| + \tau_n\big).$$

Then it holds that $\widehat{V}_{\mathrm{LP}}(x^*) = \sum_{i<j} \widetilde{w}_{ij} D_{ij}$, $\sum_{i<j} w_{ij} = |\mathbf{B}_n|$, and

$$\sum_{i<j} w_{ij}\big(X_{ij} - x^*\big)^k = \sum_{i<j} \widetilde{w}_{ij}\big(X_{ij} - x^*\big)^k = 0, \quad k = 1, 2, \ldots, \ell. \tag{16}$$

The last property (16) is referred to as the *reproducing property* of local polynomial estimators (cf. Proposition 1.12 in Tsybakov (2009)).

THEOREM 3. *Suppose the kernel $K(\cdot)$ in $\widehat{V}_{\mathrm{LP}}$ is chosen such that (9) holds with constants $\overline{M}_K$ and $\underline{M}_K$, $\tau_n \asymp n^{-\kappa}$ for some fixed constant $\kappa \ge 1$, and the bandwidths h_1, h_2 are chosen as*

$$(h_1, h_2) \asymp \begin{cases} \big(n^{-\frac{2\beta}{4\alpha\beta+\beta+2\alpha}}, n^{-\frac{4\alpha}{4\alpha\beta+\beta+2\alpha}}\big), & 0 < \alpha < \dfrac{\beta}{4\beta+2}, \\ \big(n^{-1}, n^{-\frac{1}{2\beta+1}}\big), & \alpha \ge \dfrac{\beta}{4\beta+2}. \end{cases} \tag{17}$$

Then, under (1) with random design, it holds that

$$\sup_{f \in \Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{V \in \Lambda_{\beta,I}(C_{\mathcal{V}})} \sup_{\mathbb{P}_{(X,\varepsilon)} \in \mathcal{P}_{\mathrm{vf},(X,\varepsilon)}} \sup_{x^* \in \operatorname{supp}(X)} \mathbb{E}R_1\big(\widehat{V}_{\mathrm{LP}}, V; x^*\big) \le C\big(n^{-\frac{8\alpha\beta}{4\alpha\beta+\beta+2\alpha}} \vee n^{-\frac{2\beta}{2\beta+1}}\big)$$

and

$$\sup_{f\in\Lambda_{\alpha,I}(C_{\mathcal{F}})}\ \sup_{V\in\Lambda_{\beta,I}(C_{\mathcal{V}})}\ \sup_{\mathbb{P}_{(X,\varepsilon)}\in\mathcal{P}_{\mathrm{vf},(X,\varepsilon)}} \mathbb{E}R_2(\widehat{V}_{\mathrm{LP}}, V) \le C\big(n^{-\frac{8\alpha\beta}{4\alpha\beta+\beta+2\alpha}} \vee n^{-\frac{2\beta}{2\beta+1}}\big),$$

where C is some fixed positive constant that only depends on $\overline{M}_K$, $\underline{M}_K$, α, β, $C_{\mathcal{F}}$, $C_{\mathcal{V}}$ and C_0, c_0, C_ε in $\mathcal{P}_{\mathrm{vf},(X,\varepsilon)}$.

REMARK 4. Variance function estimation in (1) with fixed design $X_i = i/n$, $i \in [n]$, has been studied in Wang et al. (2008). There the minimax rate is

$$\inf_{\widetilde{V}}\ \sup_{f\in\Lambda_{\alpha,[0,1]}(C_{\mathcal{F}})}\ \sup_{V\in\Lambda_{\beta,[0,1]}(C_{\mathcal{V}})}\ \sup_{\mathbb{E}\varepsilon^4\le C_\varepsilon}\ \sup_{x^*\in[0,1]} \mathbb{E}R_1\big(\widetilde{V}, V; x^*\big) \asymp n^{-4\alpha} \vee n^{-2\beta/(2\beta+1)},$$

$$\inf_{\widetilde{V}}\ \sup_{f\in\Lambda_{\alpha,[0,1]}(C_{\mathcal{F}})}\ \sup_{V\in\Lambda_{\beta,[0,1]}(C_{\mathcal{V}})}\ \sup_{\mathbb{E}\varepsilon^4\le C_\varepsilon} \mathbb{E}R_2(\widetilde{V}, V) \asymp n^{-4\alpha} \vee n^{-2\beta/(2\beta+1)},$$

with the integral in R_2 under the Lebesgue measure on $[0, 1]$. Comparing the above result with the error rate in Theorem 3, we see that the transition boundary in both the fixed and random design settings is $\alpha = \beta/(4\beta + 2)$. When $\alpha \ge \beta/(4\beta + 2)$, $V(\cdot)$ under both R_1 and R_2 can be estimated at the classic nonparametric rate $n^{-2\beta/(2\beta+1)}$ as if the mean function $f(\cdot)$ were known. When $\alpha < \beta/(4\beta + 2)$, a faster rate can be achieved in the random design case. This can be intuitively understood by the fact that, by contrast to the fixed design case, a significant portion of pairs have distance smaller than $1/n$ in the random design setting.

REMARK 5. As has been noted in Wang et al. (2008), in the fixed design setting, estimating the variance (function) by smoothing the squared residuals obtained from pre-estimation of the mean function $f(\cdot)$ is suboptimal. The same conclusion also applies to the random design setting. Since the design being fixed or random has no first-order effect on the estimation of the mean, the above method only achieves the rates $n^{-4\alpha/(2\alpha+1)} \vee n^{-1}$ in variance estimation and $n^{-4\alpha/(2\alpha+1)} \vee n^{-2\beta/(2\beta+1)}$ in variance function estimation, neither of which is minimax optimal.

REMARK 6. Unlike in the fixed design case, once below the threshold $\alpha = \beta/(4\beta + 2)$, α and β are now both present in the minimax rate in the random design case, suggesting that the smoothness of $V(\cdot)$ always has an effect on its estimation. This is because variance function estimation in the random design setting is essentially a "two-dimensional" problem, where we have to jointly choose two optimal neighborhood sizes to characterize the closeness between (i) each X_i and X_j; and (ii) every pair (X_i, X_j) and each target point x^*. By contrast, in the fixed design setting, the distance between X_i and X_j is constrained to be no smaller than $1/n$, and thus cannot be jointly optimized with the distance between (X_i, X_j) and x^*.

REMARK 7. One might wonder whether the following Nadaraya–Watson-type estimator can be used to establish the upper bound in Theorem 3:

$$\widehat{V}_{\mathrm{NW}}(x^*) := \frac{\sum_{i<j} K_{h_1}(X_i - X_j)K_{h_2}(X_{ij} - x^*)D_{ij}}{\sum_{i<j} K_{h_1}(X_i - X_j)K_{h_2}(X_{ij} - x^*)}, \tag{18}$$

where $K(\cdot)$ is now chosen to be a higher-order kernel to further reduce bias when $\beta > 1$. It turns out that the analysis of $\widehat{V}_{\mathrm{NW}}$ requires an extra assumption on the smoothness of the density $p_X(\cdot)$ which can be completely avoided with $\widehat{V}_{\mathrm{LP}}$. Moreover, it is well known that local polynomial estimators have good finite sample properties and boundary performances when X is compactly supported (Fan and Gijbels (1996)).

REMARK 8. Munk and Ruymgaart (2002) considered minimax estimation of the variance function (and more generally, its derivatives) in the context of nonparametric regression with random design. We focus on the comparison of their results on variance function estimation with ours. Their lower bound (Theorem 1 therein) is proved independent of the smoothness level of the mean function and upper bound (Theorem 4 therein) is proved under sufficient smoothness on the mean function. Therefore, their minimax rate is only comparable to the $n^{-2\beta/(2\beta+1)}$ component in ours. In this case, their lower bound of the order $n^{-(2\beta-1)/(2\beta)}$ is proved over the following class of variance function:

$$\mathcal{S}_\beta := \left\{ 1 + \sum_{k=1}^{\infty} \delta_k e_k : |\delta_k| \lesssim k^{-\beta} \right\}$$

for any $\beta > 1$, where $\{e_k\}_{k=1}^{\infty}$ is an arbitrary basis on $L^2([-\pi, \pi])$. Moreover, continuous differentiability of the error density is required in their paper. In contrast, we pose no smoothness conditions on the error density, and neither $\mathcal{S}_\beta$ nor $\mathcal{S}_{\beta+1/2}$ can be embedded in the β-Hölder class Λ_β considered in our setting (e.g., $f(x) = |x|$ with domain $[-\pi, \pi]$ belongs to $\mathcal{S}_2$ but is not 1.5- or 2-Hölder smooth since it is not differentiable at the origin). In summary, the results in Munk and Ruymgaart (2002) neither imply nor contradict the $n^{-2\beta/(2\beta+1)}$ part in our minimax rate, and our results are more refined since they characterize the exact elbow $\alpha = \beta/(4\beta + 2)$ and also the minimax rate below this threshold.

3.2. *Lower bound.* The following are matching lower bounds to Theorem 3.

THEOREM 4. *Under* (1) *with random design, for any* $x^* \in \operatorname{supp}(X)$,

$$\inf_{\widetilde{V}} \sup_{f \in \Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{V \in \Lambda_{\beta,I}(C_{\mathcal{V}})} \sup_{\mathbb{P}_{(X,\varepsilon)} \in \mathcal{P}_{\mathrm{vf},(X,\varepsilon)}} \mathbb{E} R_1(\widetilde{V}, V; x^*) \geq c\big(n^{-\frac{8\alpha\beta}{4\alpha\beta+\beta+2\alpha}} \vee n^{-\frac{2\beta}{2\beta+1}}\big),$$

where c *is some fixed positive constant that only depends on* α, β, $C_{\mathcal{F}}$, $C_{\mathcal{V}}$ *and* C_0, c_0, C_ε *in* $\mathcal{P}_{\mathrm{vf},(X,\varepsilon)}$, *and* $\widetilde{V}$ *ranges over all estimators of* V.

THEOREM 5. *Under* (1) *with random design,*

$$\inf_{\widetilde{V}} \sup_{f \in \Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{V \in \Lambda_{\beta,I}(C_{\mathcal{V}})} \sup_{\mathbb{P}_{(X,\varepsilon)} \in \mathcal{P}_{\mathrm{vf},(X,\varepsilon)}} \mathbb{E} R_2(\widetilde{V}, V) \geq c\big(n^{-\frac{8\alpha\beta}{4\alpha\beta+\beta+2\alpha}} \vee n^{-\frac{2\beta}{2\beta+1}}\big),$$

where c *is some fixed positive constant that only depends on* α, β, $C_{\mathcal{F}}$, $C_{\mathcal{V}}$ *and* C_0, c_0, C_ε *in* $\mathcal{P}_{\mathrm{vf},(X,\varepsilon)}$, *and* $\widetilde{V}$ *ranges over all estimators of* V.

Due to the appearances of both α and β in the nontrivial $n^{-\frac{8\alpha\beta}{4\alpha\beta+\beta+2\alpha}}$ part of the minimax rate, proving the above two results is more involved than proving Theorem 2. In particular, it takes an extra step of localization in the construction of the mean function $f(\cdot)$ as well as $V(\cdot)$. More precisely, for the lower bound at a target point x^* in Theorem 4, our construction of both $f(\cdot)$ and $V(\cdot)$ only has variation within a small neighborhood of x^*. Such localized construction is not necessary in the fixed design setting, since when proving the $n^{-4\alpha}$ component therein (see Remark 4), the variance function can simply be taken as a constant.

In what follows, we give a proof sketch of the nontrivial $n^{-8\alpha\beta/(4\alpha\beta+\beta+2\alpha)}$ component of the lower bound in Theorem 4 for $\alpha < \beta/(4\beta + 2)$; the proof of Theorem 5 can be seen as an extension of Theorem 4 via a standard construction of multiple hypotheses. We assume the support of X is contained in $I = [0, 1]$, and for clarity of illustration, here we present the construction for an interior point $x^* \in (0, 1) \cap \operatorname{supp}(X)$. The proof works for boundary points as well.

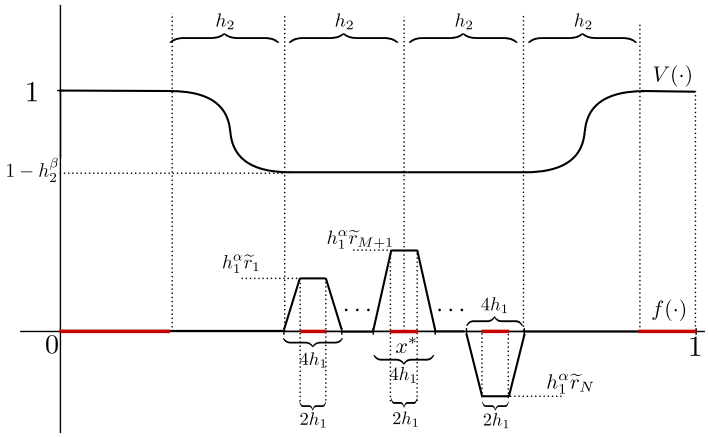

FIG. 2. *The black solid line on the top represents the variance function $V(\cdot)$ in the alternative $\widetilde{H}_1$, and the black solid line on the bottom represents the mean function $f(\cdot)$. The thick red segments mark the support of X under both H_0 and $\widetilde{H}_1$. Here, $h_1 \asymp n^{-\frac{2\beta}{4\alpha\beta+\beta+2\alpha}}$, $h_2 \asymp n^{-\frac{4\alpha}{4\alpha\beta+\beta+2\alpha}}$, and are chosen such that both $M := h_2/(4h_1) - 1/2$ and $N := 2M + 1$ are positive integers. $\{\widetilde{r}_i\}_{i=1}^N$ are N i.i.d. standard normal variables.*

We continue to adopt the two-step approach introduced in the proof sketch of Theorem 2 in Section 2.2. The second step is very similar with the help of Lemmas 1 and 3, so we will focus on the construction under the null H_0 and alternative $\widetilde{H}_1$ in the first step. Choose the parameters

$$h_1 \asymp n^{-\frac{2\beta}{4\alpha\beta+\beta+2\alpha}}, \qquad h_2 \asymp n^{-\frac{4\alpha}{4\alpha\beta+\beta+2\alpha}} \quad \text{and} \quad \theta_n^2 = h_1^{2\alpha} = h_2^\beta$$

so that $h_2/h_1 \to \infty$ as $n \to \infty$.

Choice of $V(\cdot)$: Under H_0 let $V \equiv 1$. Under $\widetilde{H}_1$, let $V(\cdot)$ be one minus a smooth bump function around x^* with width h_2 and height h_2^β so that $V(x^*) = 1 - \theta_n^2$.
Choice of $f(\cdot)$: Under H_0 let $f \equiv 0$. Under $\widetilde{H}_1$, let $f(\cdot)$ be a "local" version of the design in Theorem 2. That is, f takes on a value of 0 outside of $[x^* - h_2, x^* + h_2]$, and inside that h_2-neighborhood of x^*, f is piecewise trapezoidal with upper base length $2h_1$, lower base length $4h_1$ and height $\{h_1^\alpha \widetilde{r}_i\}_{i=1}^N$ for a standard normal sequence $\{\widetilde{r}_i\}_{i=1}^N$ with $N := h_2/(2h_1)$ a positive integer.
Choice of ε: Under both H_0 and $\widetilde{H}_1$, let $\varepsilon \sim \mathcal{N}(0, 1)$.
Choice of X: Under both H_0 and $\widetilde{H}_1$, let X be uniformly distributed on the union of $[0, 1]\backslash[x^* - h_2, x^* + h_2]$ and the upper bases of all the trapezoids inside $[x^* - h_2, x^* + h_2]$.

See Figure 2 for an illustration of $\widetilde{H}_1$.

Under the above construction, the squared distance between the null and alternative hypotheses $(1 - (1 - \theta_n^2))^2 = \theta_n^4 \asymp n^{-\frac{8\alpha\beta}{4\alpha\beta+\beta+2\alpha}}$ is the desired minimax rate. Using Lemma 2, we can show that

$$\mathsf{TV}(\mathbb{P}_0, \widetilde{\mathbb{P}}_1) \lesssim \theta_n^2 n h_1^{1/2} h_2^{1/2} \le c$$

for some sufficiently small c, where $\mathbb{P}_0$ and $\widetilde{\mathbb{P}}_1$ represent the joint distribution of $\{(X_i, Y_i)\}_{i=1}^n$ under H_0 and $\widetilde{H}_1$, respectively. The detailed proof is presented in the Supplementary Material (Shen et al. (2020)).

4. Discussion. The two univariate models (1) and (2) discussed in the previous two sections raise natural questions about possible extensions to the multivariate setting. In what

follows, we first present some partial results in this direction in the sense of (4) and (5). We then establish some connections between our study and quadratic functional estimation and variance estimation in the linear model. Lastly, we discuss two more extensions of (2) in the direction of adaptive estimation and mean function with inhomogeneous smoothness. Throughout, consider $C_{\mathcal{F}}$, C_σ, C_0, c_0, C_ε to be fixed positive constants.

4.1. *Multivariate nonparametric regression.* Consider the following multivariate version of (2):

$$Y_i = f(\boldsymbol{X}_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\{\boldsymbol{X}_i\}_{i=1}^n = \{(X_{i,1}, \ldots, X_{i,d})^\top\}_{i=1}^n$ are i.i.d. copies of $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ in $\mathbb{R}^d$ for some fixed positive integer d, $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. copies of ε with zero mean and unit variance and are independent of $\{\boldsymbol{X}_i\}_{i=1}^n$, and $f : \mathbb{R}^d \to \mathbb{R}$ belongs to a d-dimensional anisotropic Hölder class with smoothness index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\top$ defined below. The goal is to estimate σ^2 with $f(\cdot)$ and the distribution of $\boldsymbol{X}$ as nuisance parameters. This problem has been studied in Spokoiny (2002), Munk et al. (2005), Cai, Levine and Wang (2009), to name a few, again with a focus on the fixed design setting.

Let $I_1, \ldots, I_d$ be d fixed (possibly infinite) intervals on $\mathbb{R}$ and let $\boldsymbol{I}$ be their Cartesian product $I_1 \times \cdots \times I_d \subset \mathbb{R}^d$. Following Barron, Birgé and Massart (1999) and Bhattacharya, Pati and Dunson (2014), we define an anisotropic Hölder class $\Lambda_{\boldsymbol{\alpha},\boldsymbol{I}}(C_{\mathcal{F}})$ on $\boldsymbol{I}$ as follows. For any $\boldsymbol{x} \in \boldsymbol{I}$ and $k \in [d]$, let $f_k(\cdot \mid \boldsymbol{x}_{-k})$ denote the univariate function $y \mapsto f(x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_d)$, with $\boldsymbol{x}_{-k}$ defined as $\boldsymbol{x}$ without the kth component. Then $\Lambda_{\boldsymbol{\alpha},\boldsymbol{I}}(C_{\mathcal{F}})$ is defined as all $f : \boldsymbol{I} \mapsto \mathbb{R}$ such that

$$\max_{1\le k\le d} \max_{0\le j\le \lfloor\alpha_k\rfloor} \sup_{\boldsymbol{x}\in\boldsymbol{I}} \big\| f_k^{(j)}(\cdot \mid \boldsymbol{x}_{-k}) \big\|_\infty \le C_{\mathcal{F}}$$

and

$$\max_{1\le k\le d} \sup_{x\in\boldsymbol{I}} \sup_{y_1,y_2\in I_k} \frac{|f_k^{(\lfloor\alpha_k\rfloor)}(y_1 \mid \boldsymbol{x}_{-k}) - f_k^{(\lfloor\alpha_k\rfloor)}(y_2 \mid \boldsymbol{x}_{-k})|}{|y_1 - y_2|^{\alpha_k'}} \le C_{\mathcal{F}},$$

where again $\lfloor\alpha_k\rfloor$ is the largest integer strictly smaller than α_k and $\alpha_k' := \alpha_k - \lfloor\alpha_k\rfloor$. Let $\operatorname{supp}(\boldsymbol{X})$ be the support of $\boldsymbol{X}$.

Define $\mathcal{P}_{\mathrm{mcv},(\boldsymbol{X},\varepsilon)}$ (where "mcv" stands for "multivariate constant variance") as the multivariate counterpart of $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$:

(a) $\boldsymbol{X}$ satisfies $\operatorname{supp}(\boldsymbol{X}) \subset \boldsymbol{I}$.

(b) $\boldsymbol{X}$ has density $p_{\boldsymbol{X}}(\cdot)$ and there exists a fixed positive constant C_0 such that

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d} p_{\boldsymbol{X}}(\boldsymbol{u}) \le C_0.$$

(c) There exist two fixed constants $\delta_0 > 0$ and $c_0 > 0$ such that for any $\boldsymbol{\delta} \in \mathbb{R}^d$ that satisfies $\|\boldsymbol{\delta}\|_\infty < \delta_0$, there exists a set $\mathcal{U} := \mathcal{U}_{\boldsymbol{\delta}} \subset [-1, 1]^d$ such that

$$\boldsymbol{\lambda}(\mathcal{U}_{\boldsymbol{\delta}}) \ge c_0 \quad \text{and} \quad \inf_{\boldsymbol{u}\in\mathcal{U}_{\boldsymbol{\delta}}} p_{\widetilde{\boldsymbol{X}}_{ij}}(u_1\delta_1, \ldots, u_d\delta_d) \ge c_0,$$

where $\boldsymbol{\lambda}(\cdot)$ represents the Lebesgue measure on $\mathbb{R}^d$.

(d) $\mathbb{E}\varepsilon^4 \le C_\varepsilon$ for some fixed positive constant C_ε.

For an upper bound on the minimax risk, we propose the following multivariate extension of (8) via a product kernel (again with convention $0/0 = 0$):

$$\widehat{\sigma}_d^2 := \frac{\binom{n}{2}^{-1} \sum_{i<j} \big(\prod_{k=1}^d K_{h_k}(X_{i,k} - X_{j,k})\big)(Y_i - Y_j)^2/2}{\binom{n}{2}^{-1} \sum_{i<j} \big(\prod_{k=1}^d K_{h_k}(X_{i,k} - X_{j,k})\big)}, \tag{19}$$

where $K(\cdot)$ is a kernel chosen to satisfy (9), and $\{h_k\}_{k=1}^d$ is a kernel bandwidth sequence.

In the following results, we will use $\underline{\alpha}$ to denote the harmonic mean of the d-dimensional smoothness index $\boldsymbol{\alpha}$, that is, $\underline{\alpha} := d/(\sum_{k=1}^{d} 1/\alpha_k)$. This quantity is known as the *effective smoothness* in classical problems such as anisotropic density estimation (Ibragimov and Khas'minskiĭ (1981), Birgé (1986)) and anisotropic function estimation (Nussbaum (1986), Hoffmann and Lepski (2002)).

PROPOSITION 1. *Suppose $0 < \alpha_k \le 1$, $k \in [d]$. Suppose the kernel $K(\cdot)$ in $\widehat{\sigma}_d^2$ is chosen such that (9) is satisfied with constants $\overline{M}_K$ and $\underline{M}_K$, and the bandwidth sequence is chosen as $h_k \asymp n^{-2\underline{\alpha}/(\alpha_k(4\underline{\alpha}+d))}$ for all $k \in [d]$. Then, under (4) with random design, it holds that*

$$\sup_{f\in\Lambda_{\boldsymbol{\alpha},\boldsymbol{I}}(C_{\mathcal{F}})}\sup_{\sigma^2\le C_\sigma}\sup_{\mathbb{P}_{(\boldsymbol{X},\varepsilon)}\in\mathcal{P}_{\mathrm{mcv},(\boldsymbol{X},\varepsilon)}}\mathbb{E}\big(\widehat{\sigma}_d^2-\sigma^2\big)^2 \le C\big(n^{-8\underline{\alpha}/(4\underline{\alpha}+d)}\vee n^{-1}\big),$$

where C is some fixed positive constant that only depends on $\overline{M}_K$, $\underline{M}_K$, $\boldsymbol{\alpha}$, $C_{\mathcal{F}}$, C_σ and C_0, c_0, C_ε in $\mathcal{P}_{\mathrm{mcv},(\boldsymbol{X},\varepsilon)}$.

PROPOSITION 2. *Under (4) with random design, it holds that*

$$\inf_{\widetilde{\sigma}^2}\sup_{f\in\Lambda_{\boldsymbol{\alpha},\boldsymbol{I}}(C_{\mathcal{F}})}\sup_{\sigma^2\le C_\sigma}\sup_{\mathbb{P}_{(\boldsymbol{X},\varepsilon)}\in\mathcal{P}_{\mathrm{mcv},(\boldsymbol{X},\varepsilon)}}\mathbb{E}\big(\widetilde{\sigma}^2-\sigma^2\big)^2 \ge c\big(n^{-8\underline{\alpha}/(4\underline{\alpha}+d)}\vee n^{-1}\big),$$

where c is some fixed positive constant that only depends on $\boldsymbol{\alpha}$, $C_{\mathcal{F}}$, C_σ and C_0, c_0, C_ε in $\mathcal{P}_{\mathrm{mcv},(\boldsymbol{X},\varepsilon)}$, and $\widetilde{\sigma}^2$ ranges over all estimators of σ^2.

We note that Proposition 1 is only proved for $\alpha_k \in (0, 1]$, $k \in [d]$. The general case when α_k is possibly larger than 1 is much more involved due to the difficulty in the random design analysis. Propositions 1 and 2, combined, imply that the minimax rate is $n^{-8\underline{\alpha}/(4\underline{\alpha}+d)} \vee n^{-1}$ for $\alpha_k \in (0, 1]$, $k \in [d]$. In particular, when f is in an isotropic α-Hölder class ($0 < \alpha \le 1$), this rate becomes $n^{-8\alpha/(4\alpha+d)} \vee n^{-1}$. We also remark that a different estimator achieving the rate $n^{-8\alpha/(4\alpha+d)} \vee n^{-1}$ over an isotropic α-Hölder class has been briefly sketched in Robins et al. (2008).

For completeness, we also state without proof some results for model (4) in the fixed design setting. In particular, we consider the following two types of fixed designs in the d-dimensional unit cube $[0, 1]^d$, namely, the grid design (GD):

$$\begin{aligned}(X_{(i_1,\ldots,i_d),1},\ldots,X_{(i_1,\ldots,i_d),d}) &= \big(i_1/n^{1/d},\ldots,i_d/n^{1/d}\big),\\ (i_1,\ldots,i_d) &\in \big[n^{1/d}\big]\times\cdots\times\big[n^{1/d}\big]\end{aligned} \tag{20}$$

assuming $n^{1/d}$ is an integer, and the diagonal design (DD):

$$(X_{i,1},\ldots,X_{i,d}) = (i/n,\ldots,i/n), \quad i \in [n]. \tag{21}$$

Here, for any positive integer n, $[n]$ denotes the set $\{1, 2, \ldots, n\}$. Let $\alpha_{\max} := \max_{k\in[d]}\alpha_k$ and $\alpha_{\min} := \min_{k\in[d]}\alpha_k$. The first result for (GD) is a simple modification of the isotropic result in Cai, Levine and Wang (2009) by taking differences along the smoothest direction with index $\alpha_{\max}$. The second result can be readily deduced from the fact that $Y_i = \widetilde{f}(i/n) + \sigma\varepsilon_i$, $i \in [n]$, where $\widetilde{f}(x) := f(x,\ldots,x)$ is $\alpha_{\min}$-Hölder smooth.

PROPOSITION 3. *Under (4) with fixed design (GD), it holds that*

$$\inf_{\widetilde{\sigma}^2}\sup_{f\in\Lambda_{\boldsymbol{\alpha},[0,1]^d}(C_{\mathcal{F}})}\sup_{\sigma^2\le C_\sigma}\sup_{\mathbb{E}\varepsilon^4\le C_\varepsilon}\mathbb{E}\big(\widetilde{\sigma}^2-\sigma^2\big)^2 \asymp n^{-4\alpha_{\max}/d}\vee n^{-1}$$

up to some fixed positive constant that only depends on $\boldsymbol{\alpha}$, $C_{\mathcal{F}}$, C_σ, C_ε, where $\widetilde{\sigma}^2$ ranges over all estimators of σ^2.

PROPOSITION 4. *Under* (4) *with fixed design* (*DD*), *it holds that*

$$\inf_{\widetilde{\sigma}^2} \sup_{f\in\Lambda_{\boldsymbol{\alpha},[0,1]^d}(C_{\mathcal{F}})} \sup_{\sigma^2\le C_\sigma} \sup_{\mathbb{E}\varepsilon^4\le C_\varepsilon} \mathbb{E}\big(\widetilde{\sigma}^2-\sigma^2\big)^2 \asymp n^{-4\alpha_{\min}} \vee n^{-1}$$

up to some fixed positive constant that only depends on $\boldsymbol{\alpha}$, $C_{\mathcal{F}}$, C_σ, C_ε, *where* $\widetilde{\sigma}^2$ *ranges over all estimators of* σ^2.

When $f(\cdot)$ belongs to an isotropic α-Hölder class, Proposition 3 implies the minimax rate $n^{-4\alpha/d}\vee n^{-1}$ derived in Cai, Levine and Wang (2009). Comparison with the random design rate $n^{-8\alpha/(4\alpha+d)}\vee n^{-1}$ thus shows that, for $0<\alpha\le 1$, a faster rate is again achievable in the random design setting for $\alpha < d/4$.

4.2. *Nonparametric additive model.* Consider variance estimation in the additive model (5):

$$Y_i = \sum_{k=1}^{d} f_k(X_{i,k}) + \sigma\varepsilon_i, \quad i=1,2,\ldots,n,$$

for some fixed integer $d\ge 2$, where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. with zero mean and unit variance and are independent from $\{\boldsymbol{X}_i\}_{i=1}^n = \{(X_{i,1},\ldots,X_{i,d})^\top\}_{i=1}^n$ in the random design setting. Unlike Section 4.1, we specify $d\ge 2$, since the minimax rate in the fixed design (GD) has completely different behavior for $d=1$ and $d\ge 2$ (see Proposition 5 below).

4.2.1. *Fixed design.* We first consider the two fixed designs (GD) and (DD) defined in (20) and (21). For both designs, we consider an error distribution class with only a finite fourth moment condition. We start with (GD), where by iteratively taking pairwise differences, one is able to estimate the variance at the parametric rate n^{-1} without any smoothness assumption on the additive components $\{f_k\}_{k=1}^d$. For simplicity, we illustrate this idea with $d=2$ with two additive components $f(\cdot)$ and $g(\cdot)$, and assume that $\sqrt{n}$ is an even number. In this case,

$$Y_{i,j} = f\left(\frac{i}{\sqrt{n}}\right) + g\left(\frac{j}{\sqrt{n}}\right) + \sigma\varepsilon_{i,j}, \quad (i,j)\in[\sqrt{n}]\times[\sqrt{n}],$$

where $\{\varepsilon_{i,j}\}_{i,j\in[\sqrt{n}]}$ are i.i.d. with zero mean and unit variance. By taking the pairwise difference in the first dimension, we have

$$Y_{(i_1,i_2),j} := Y_{i_1,j} - Y_{i_2,j} = f\left(\frac{i_1}{\sqrt{n}}\right) - f\left(\frac{i_2}{\sqrt{n}}\right) + \sigma(\varepsilon_{i_1,j}-\varepsilon_{i_2,j})$$

for all $j\in[\sqrt{n}]$ and $(i_1,i_2)\in[\sqrt{n}]\times[\sqrt{n}]$ such that $i_1\ne i_2$. Taking again the pairwise difference in the second dimension, we have

$$Y_{(i_1,i_2),(j_1,j_2)} := Y_{(i_1,i_2),j_1} - Y_{(i_1,i_2),j_2} = \sigma(\varepsilon_{i_1,j_1}-\varepsilon_{i_2,j_1}-\varepsilon_{i_1,j_2}+\varepsilon_{i_2,j_2})$$

for all $(i_1,i_2,j_1,j_2)\in[\sqrt{n}]\times[\sqrt{n}]\times[\sqrt{n}]\times[\sqrt{n}]$ such that $i_1\ne i_2$ and $j_1\ne j_2$. Clearly, we have $\mathbb{E}Y_{(i_1,i_2),(j_1,j_2)}=0$ and $\operatorname{Var}(Y_{(i_1,i_2),(j_1,j_2)})=4\sigma^2$. Let $m:=\sqrt{n}/2$ and define $\mathcal{I}:=\{(1,2),(3,4),\ldots,(2m-1,2m)\}$ with cardinality m. Then, for the set of data points $\{Y_{(i_1,i_2),(j_1,j_2)}\}_{(i_1,i_2),(j_1,j_2)\in\mathcal{I}}$ with cardinality $m^2=n/4$, it can be readily verified that they are i.i.d. with mean 0 and variance $4\sigma^2$. Therefore, with $\overline{Y}$ defined as the sample average of $\{Y_{(i_1,i_2),(j_1,j_2)}\}_{(i_1,i_2),(j_1,j_2)\in\mathcal{I}}$, the sample variance estimator,

$$\widehat{\sigma}^2_{\text{add,GD}} := \frac{1}{n}\sum_{(i_1,i_2),(j_1,j_2)\in\mathcal{I}} (Y_{(i_1,i_2),(j_1,j_2)}-\overline{Y})^2,$$

achieves the parametric rate n^{-1}. A similar derivation holds for general d.

PROPOSITION 5. *Suppose $d \geq 2$. Under (5) with fixed design (GD), it holds that*

$$\inf_{\widetilde{\sigma}^2} \sup_{f_k, k\in[d]} \sup_{\sigma^2 \leq C_\sigma} \sup_{\mathbb{E}\varepsilon^4 \leq C_\varepsilon} \mathbb{E}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2 \asymp n^{-1}$$

up to some fixed positive constant that only depends on C_σ and C_ε, where $\widetilde{\sigma}^2$ ranges over all estimators of σ^2, and the first supremum is taken over all functions defined on $[0, 1]$ for each $k \in [d]$.

Now we move on to the design (DD), where we assume each additive component f_k in (5) is α_k-Hölder smooth on $[0, 1]$ with some fixed constant $C_{\mathcal{F}}$. In this case, the model can equivalently be written as

$$Y_i = \widetilde{f}(i/n) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\widetilde{f} := \sum_{k=1}^d f_k$ is $\alpha_{\min}$-Hölder smooth. Therefore, the univariate estimator and lower bound in Wang et al. (2008) can be directly applied.

PROPOSITION 6. *Under (5) with fixed design (DD), it holds that*

$$\inf_{\widetilde{\sigma}^2} \sup_{f_k \in \Lambda_{\alpha_k,[0,1]}(C_{\mathcal{F}}), k\in[d]} \sup_{\sigma^2 \leq C_\sigma} \sup_{\mathbb{E}\varepsilon^4 \leq C_\varepsilon} \mathbb{E}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2 \asymp n^{-4\alpha_{\min}} \vee n^{-1}$$

up to some fixed positive constant that only depends on $C_{\mathcal{F}}$, C_σ, C_ε, where $\widetilde{\sigma}^2$ ranges over all estimators of σ^2.

Comparison of Propositions 6 and 4 shows that, in contrast to grid design (GD) and random design below, there is no gain from an additive structure in the mean function for the diagonal design (DD).

4.2.2. *Random design.* We now discuss (5) with a random design for $\{\boldsymbol{X}_i\}_{i=1}^n$ when f_k is α_k-Hölder smooth on some fixed set I_k for each $k \in [d]$. Since a shift in the mean does not affect the estimation of variance, we assume $\mathbb{E} f_k(X_{1,k}) = 0$ for each $k \in [d]$ for simplicity. Recall the definition of $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$ in the beginning of Section 2. Define the joint distribution class $\mathcal{P}_{\mathrm{add},(\boldsymbol{X},\varepsilon)}$ (where "add" stands for "additive") as:

For each $k \in [d]$, the joint distribution of (X_k, ε) belongs to $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$ and the components of $\boldsymbol{X}$ are mutually independent.

In view of Theorem 2, the following lower bound is immediate.

PROPOSITION 7. *Under (5) with random design, it holds that*

$$\inf_{\widetilde{\sigma}^2} \sup_{f_k \in \Lambda_{\alpha_k,I_k}(C_{\mathcal{F}}), k\in[d]} \sup_{\sigma^2 \leq C_\sigma} \sup_{\mathbb{P}_{(\boldsymbol{X},\varepsilon)} \in \mathcal{P}_{\mathrm{add},(\boldsymbol{X},\varepsilon)}} \mathbb{E}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2 \geq c\big(n^{-\frac{8\alpha_{\min}}{4\alpha_{\min}+1}} \vee n^{-1}\big),$$

where c is a fixed positive constant that only depends on $\boldsymbol{\alpha}$, $C_{\mathcal{F}}$, C_σ and C_0, c_0, C_ε in $\mathcal{P}_{\mathrm{add},(\boldsymbol{X},\varepsilon)}$, and $\widetilde{\sigma}^2$ ranges over all estimators of σ^2.

We now describe a procedure that matches the lower bound in Proposition 7, but depends crucially on mutual independence. For illustrative purposes, we again consider the case of only two additive components $f(\cdot)$ and $g(\cdot)$, which are α- and β-Hölder smooth, respectively. Let X and W denote the two covariates. For each $i \in [n]$, define

$$\varepsilon_i^X := f(X_i) + \sigma\varepsilon_i \quad \text{and} \quad \varepsilon_i^W := g(W_i) + \sigma\varepsilon_i,$$

and their corresponding variances

$$\sigma_X^2 := \mathbb{E}f^2(X) + \sigma^2 \quad \text{and} \quad \sigma_W^2 := \mathbb{E}g^2(W) + \sigma^2.$$

Clearly, we have $\mathbb{E}\varepsilon_i^X = 0$ and $\mathbb{E}\varepsilon_i^W = 0$, and ε_i^X and ε_i^W are independent of $g(W_i)$ and $f(X_i)$, respectively. Now, notice that the additive model in (5) can be equivalently viewed as $Y_i = f(X_i) + \varepsilon_i^W$. Thus by applying the univariate kernel estimator defined in (8) to $\{(Y_i, X_i)\}_{i=1}^n$, which we denote as $\widehat{\sigma}_W^2$, one obtains

$$\mathbb{E}\big(\widehat{\sigma}_W^2 - \sigma_W^2\big)^2 \le C\big(n^{-8\alpha/(4\alpha+1)} \vee n^{-1}\big)$$

for some fixed positive constant C. Similarly, defining $\widehat{\sigma}_X^2$ as $\widehat{\sigma}_W^2$, one has

$$\mathbb{E}\big(\widehat{\sigma}_X^2 - \sigma_X^2\big)^2 \le C\big(n^{-8\beta/(4\beta+1)} \vee n^{-1}\big).$$

Lastly, under a finite fourth moment assumption on ε, a sample variance estimator of $\{Y_i\}_{i=1}^n$, denoted as $\widehat{\sigma}_Y^2$, achieves the parametric rate n^{-1} in estimating the total variance $\mathrm{Var}(Y)$, which can be decomposed as $\mathbb{E}f^2(X) + \mathbb{E}g^2(W) + \sigma^2$. Consequently, we have shown that the method-of-moments estimator

$$\widehat{\sigma}_{\mathrm{moment},2}^2 := \widehat{\sigma}_X^2 + \widehat{\sigma}_W^2 - \widehat{\sigma}_Y^2 \tag{22}$$

achieves the optimal rate in Proposition 7. We summarize the above derivation for the natural extension $\widehat{\sigma}_{\mathrm{moment},d}^2$ to general d.

PROPOSITION 8. *Under (5) with random design, it holds that*

$$\sup_{f_k \in \Lambda_{\alpha_k, I_k}(C_{\mathcal{F}}), k\in[d]} \sup_{\sigma^2 \le C_\sigma} \sup_{\mathbb{P}_{(\boldsymbol{X},\varepsilon)} \in \mathcal{P}_{\mathrm{add},(\boldsymbol{X},\varepsilon)}} \mathbb{E}\big(\widehat{\sigma}_{\mathrm{moment},d}^2 - \sigma^2\big)^2 \le C\big(n^{-\frac{8\alpha_{\min}}{4\alpha_{\min}+1}} \vee n^{-1}\big),$$

where C is some fixed positive constant that only depends on $\boldsymbol{\alpha}$, $C_{\mathcal{F}}$, C_σ and C_0, c_0, C_ε in $\mathcal{P}_{\mathrm{add},(\boldsymbol{X},\varepsilon)}$.

Propositions 7 and 8 together imply the minimax rate over $\mathcal{P}_{\mathrm{add},(\boldsymbol{X},\varepsilon)}$, which further illustrates the fact that an additive structure in the mean function could possibly avoid the "curse of dimensionality" in variance estimation. However, we note that our results crucially rely on the mutual independence condition. It is still largely unclear if the same minimax rate could apply to the general case without this condition, though a discussion of an interesting connection to variance estimation under linear models shall be made in Section 4.4.

4.3. *Connection to quadratic functional estimation.* We now formally state the connection between quadratic functional estimation and variance estimation in (2), the first of which has been studied in, for example, Doksum and Samarov (1995), Ruppert, Sheather and Wand (1995), Huang and Fan (1999) and Robins et al. (2009).

Recall the definition of Q in (3) with some nonnegative weight function $w(\cdot)$. Squaring both sides of (2), multiplying by $w(X_i)$, and then taking the expectation, one has

$$\mathbb{E}\big(Y_i^2 w(X_i)\big) = \mathbb{E}\big(f^2(X_i)w(X_i)\big) + \sigma^2\mathbb{E}\big(w(X_i)\varepsilon_i^2\big) = Q + \sigma^2\mathbb{E}w(X_i).$$

Under a finite fourth moment assumption on ε, both $\mathbb{E}(Y_i^2 w(X_i))$ and $\mathbb{E}w(X_i)$ can be estimated at the parametric rate via the sample mean estimator, and σ^2 can be estimated via $\widehat{\sigma}^2$ in (8) with rate $n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$ under the quadratic risk. Therefore, the estimator

$$\widehat{Q} := \frac{1}{n}\sum_{i=1}^n Y_i^2 w(X_i) - \left(\frac{1}{n}\sum_{i=1}^n w(X_i)\right) \cdot \widehat{\sigma}^2$$

achieves the same rate $n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$. In fact, it is not possible to improve upon this rate since if there exists an estimator $\widetilde{Q}$ with a faster convergence rate, then the "conjugate" estimator of σ^2 defined as

$$\widetilde{\sigma}^2 := \max\left\{\frac{\frac{1}{n}\sum_{i=1}^n Y_i^2 w(X_i) - \widetilde{Q}}{\frac{1}{n}\sum_{i=1}^n w(X_i)}, 0\right\} \cdot \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^n w(X_i) > 0\right\}$$

will also converge to σ^2 at a faster rate, violating the lower bound in Theorem 2.

The following result summarizes the derivation. Recall the definition of $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$ in the beginning of Section 2.

PROPOSITION 9. *Suppose the weight function $w(\cdot)$ in the definition of Q is uniformly bounded on $\mathbb{R}$. Then it holds that*

$$\inf_{\widetilde{Q}} \sup_{f\in\Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{\sigma^2\le C_\sigma} \sup_{\mathbb{P}_{(X,\varepsilon)}\in\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}} \mathbb{E}(\widetilde{Q}-Q)^2 \asymp n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$$

up to some fixed positive constant that only depends on $w(\cdot)$, α, $C_{\mathcal{F}}$, C_σ and C_0, c_0, C_ε in $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$, where $\widetilde{Q}$ ranges over all estimators of Q.

4.4. *Connection to the linear model.* Throughout this paper, we have treated the distribution of $\boldsymbol{X}$ as a nuisance parameter. Interestingly, when we do know the distribution of $\boldsymbol{X}$, variance estimation in nonparametric regression with random design becomes substantially easier with the aid of parallel work in the high-dimensional linear model (Verzelen and Villers (2010), Dicker (2014), Kong and Valiant (2018), Verzelen and Gassiat (2018)). We first elaborate on this point using the simple model (2), and then formulate corresponding results for (4) and (5).

By applying the inverse of the distribution function F of X, (2) can be equivalently written as

$$Y_i = \overline{f}(U_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\{U_i\}_{i=1}^n = \{F(X_i)\}_{i=1}^n$ are i.i.d. uniform on $[0, 1]$, and $\overline{f}(\cdot) := f \circ F^{-1}(\cdot)$ is still α-Hölder smooth under Lipschitz continuity on F^{-1}. Then, using a wavelet expansion for Hölder classes (cf. Proposition 2.5 in Meyer (1990)), one has

$$Y_i = \overline{f}_1(U_i) + \sum_{j=1}^{2^J} \psi_j(U_i) + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{23}$$

where $\{\psi_j\}_{j=1}^\infty$ is an L_2-orthonormal wavelet basis under the Lebesgue measure on $[0, 1]$, and $\overline{f}_1(\cdot)$ is the remainder term after truncation at resolution $J = J_n$ which satisfies $\|\overline{f}_1\|_\infty = O(2^{-\alpha J_n})$. Let $\boldsymbol{\psi} := (\psi_1, \ldots, \psi_{2^J})$ and assume without loss of generality that $\mathbb{E}\boldsymbol{\psi} = \mathbf{0}_{2^J}$, since a mean shift does not affect the estimation of variance. Moreover, due to the orthonormality of $\{\psi_j\}_{j=1}^\infty$, we have $\mathrm{Cov}(\boldsymbol{\psi}) = \mathbb{E}(\boldsymbol{\psi}\boldsymbol{\psi}^\top) = \mathbf{I}_{2^J}$. Following Verzelen and Gassiat (2018) and Kong and Valiant (2018), the estimator

$$\widehat{\sigma}^2_{\mathrm{proj}} := \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2 - \binom{n}{2}^{-1} \sum_{i<j} Y_i Y_j \boldsymbol{\psi}^\top(U_i)\boldsymbol{\psi}(U_j)$$

has a variance term of the order $(2^{J_n} + n)/n^2$ and a bias term of the order $2^{-2\alpha J_n}$. Therefore, by choosing the optimal truncation level $2^{J_n} \asymp n^{2/(4\alpha+1)}$, $\widehat{\sigma}^2_{\mathrm{proj}}$ recovers the optimal rate $n^{-8\alpha/(4\alpha+1)} \vee n^{-1}$ in Theorem 1.

Define $\widehat{\sigma}^2_{\mathrm{proj},d}$ (with tensor wavelet basis) and $\widehat{\sigma}^2_{\mathrm{proj,add}}$ as the natural extensions of $\widehat{\sigma}^2_{\mathrm{proj}}$ under (4) and (5), respectively (see the proofs of Propositions 10 and 11 in the Supplementary Material (Shen et al. (2020)) for exact definitions). In the wavelet expansion, we will use J_k to denote the truncation level for the kth component of $f(\cdot)$ in (4) and f_k in (5), and we use F_k to denote the marginal distribution of $X_{1,k}$. Recall that $\underline{\alpha} = d/(\sum_{k=1}^d 1/\alpha_k)$ for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\top$.

PROPOSITION 10 (Multivariate nonparametric regression, design known). *Suppose the distribution of $\boldsymbol{X}$ is known with* $\mathrm{supp}(\boldsymbol{X}) \subset \boldsymbol{I}$ *for some fixed set $\boldsymbol{I} \subset \mathbb{R}^d$, and $F_k^{-1}(\cdot)$ is Lipschitz continuous for all $k \in [d]$ with some fixed positive constant. Then, when 2^{J_k} is chosen to be of the order $n^{2\underline{\alpha}/(\alpha_k(4\underline{\alpha}+d))}$ for $k \in [d]$ in $\widehat{\sigma}^2_{\mathrm{proj},d}$, it holds that*

$$\sup_{f \in \Lambda_{\boldsymbol{\alpha},\boldsymbol{I}}(C_{\mathcal{F}})} \sup_{\sigma^2 \le C_\sigma} \sup_{\mathbb{E}\varepsilon^4 \le C_\varepsilon} \mathbb{E}\big(\widehat{\sigma}^2_{\mathrm{proj},d} - \sigma^2\big)^2 \le C\big(n^{-8\underline{\alpha}/(4\underline{\alpha}+d)} \vee n^{-1}\big),$$

where C is some fixed positive constant that only depends on $\boldsymbol{\alpha}$, $C_{\mathcal{F}}$, C_σ, C_ε and the distribution of $\boldsymbol{X}$.

PROPOSITION 11 (Nonparametric additive model, design known). *Suppose the distribution of $\boldsymbol{X}$ is known with* $\mathrm{supp}(\boldsymbol{X}) \subset I_1 \times \cdots \times I_d$ *for some fixed intervals $I_1, \ldots, I_d$ on the real line, and $F_k^{-1}(\cdot)$ is Lipschitz continuous for all $k \in [d]$ with some fixed positive constant. Then, when 2^{J_k} is chosen to be of the order $n^{2\alpha_k/(4\alpha_k+1)}$ for $k \in [d]$ in $\widehat{\sigma}^2_{\mathrm{proj,add}}$, it holds that*

$$\sup_{f_k \in \Lambda_{\alpha_k, I_k}(C_{\mathcal{F}}), k \in [d]} \sup_{\sigma^2 \le C_\sigma} \sup_{\mathbb{E}\varepsilon^4 \le C_\varepsilon} \mathbb{E}\big(\widehat{\sigma}^2_{\mathrm{proj,add}} - \sigma^2\big)^2 \le C\big(n^{-\frac{8\alpha_{\min}}{4\alpha_{\min}+1}} \vee n^{-1}\big),$$

where C is some fixed positive constant that only depends on $\boldsymbol{\alpha}$, $C_{\mathcal{F}}$, C_σ, C_ε and the distribution of $\boldsymbol{X}$.

As in the classical setting of mean function estimation via orthogonal series, the difference of the rates in Propositions 10 and 11 is clearly explained by the number of wavelet bases used to approximate f in (4) and $\{f_k\}_{k=1}^d$ in (5). We also note that, quite interestingly, Proposition 10 gives results beyond the case $0 < \alpha_1, \ldots, \alpha_d \le 1$ considered in Proposition 1, and Proposition 11 does not rely on the mutual independence of the components of $\boldsymbol{X}$.

4.5. *Adaptive estimation of constant variance.* In this subsection, we consider adaptive estimation of the variance σ^2 in model (2). This is achieved by a Lepski-type procedure (Lepski (1991, 1992)). Let $\widehat{\sigma}^2(h)$ be the estimator in (8) with an explicit dependence on the bandwidth parameter h. For any given sample size n and fixed positive constant δ, define two positive integers m_1 and m_2 such that $2^{-m_1} \le n^{-1} \le 2^{-m_1+1}$ and $2^{-m_2-1} \le n^{-(2-\delta)} \le 2^{-m_2}$, and define the following dyadic grid:

$$\mathcal{H}_\delta := \big\{2^{-j} : m_1 \le j \le m_2,\ j \in \mathbb{Z}\big\}.$$

Then define the estimator $\widehat{\sigma}^2_{\mathrm{adapt}} := \widehat{\sigma}^2(\widehat{h}_\delta)$ with

$$\widehat{h}_\delta := \max\big\{h \in \mathcal{H}_\delta : \big|\widehat{\sigma}^2(h) - \widehat{\sigma}^2(h')\big| \le \tau(\log n)^{1/2} n^{-1} (h')^{-1/2}, \forall h' \in \mathcal{H}_\delta, h' < h\big\}$$

for some sufficiently large positive constant τ. If the set being maximized is empty, we will take $\widehat{h}_\delta = n^{-(2-\delta)}$.

To state the error bound of $\widehat{\sigma}^2_{\mathrm{adapt}}$, we need the following variant $\mathcal{P}^{\mathrm{adapt}}_{\mathrm{cv},(X,\varepsilon)}$ of the distribution class $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$ considered in Theorem 1, where we replace the finite fourth-moment assumption (d) therein by the stronger sub-Gaussian tail condition:

(d′) There exist some fixed positive constants $C_{1,\varepsilon}$ and $C_{2,\varepsilon}$ such that $\mathbb{E}\exp(t\varepsilon) \le C_{1,\varepsilon}\exp(C_{2,\varepsilon}t^2)$ for any $t \in \mathbb{R}$.

A similar exponential moment assumption has been made in the context of adaptive estimation under fixed design (cf. Theorems 1 and 2 in Cai and Wang (2008)).

PROPOSITION 12. *For any given sufficiently small fixed $\alpha_* > 0$, fix some $\delta_* \in (0, 8\alpha_*/(4\alpha_*+1))$. Suppose the kernel $K(\cdot)$ in $\widehat{\sigma}^2_{\text{adapt}} = \widehat{\sigma}^2(\widehat{h}_{\delta_*})$ is chosen such that (9) is satisfied with constants $\overline{M}_K$ and $\underline{M}_K$, and τ in $\widehat{h}_{\delta_*}$ is chosen to be sufficiently large (only depending on δ_*, $C_{1,\varepsilon}$, $C_{2,\varepsilon}$). Then, under (2) with random design, it holds uniformly over all $\alpha \ge \alpha_*$ that*

$$\sup_{f\in\Lambda_{\alpha,I}(C_{\mathcal{F}})}\sup_{\sigma^2\le C_\sigma}\sup_{\mathbb{P}_{(X,\varepsilon)}\in\mathcal{P}^{\text{adapt}}_{\text{cv},(X,\varepsilon)}} \mathbb{E}\big(\widehat{\sigma}^2_{\text{adapt}} - \sigma^2\big)^2 \le C\left\{\left(\frac{\log n}{n^2}\right)^{4\alpha/(4\alpha+1)} \vee n^{-1}\right\},$$

where C is some fixed positive constant that only depends on δ_*, $\overline{M}_K$, $\underline{M}_K$, $C_{\mathcal{F}}$, C_σ and C_0, c_0, $C_{1,\varepsilon}$, $C_{2,\varepsilon}$ in $\mathcal{P}^{\text{adapt}}_{\text{cv},(X,\varepsilon)}$.

The following proposition shows that the extra polylogarithmic term cannot be removed.

PROPOSITION 13. *Let $\phi_{n,\alpha} := (\log n/n^2)^{2\alpha/(4\alpha+1)} \vee n^{-1/2}$ for any $\alpha > 0$ and positive integer n. Consider any fixed positive α_* and $\alpha_* \le \alpha_1 < \alpha_2 < \infty$. Then, for any sufficiently large n and sufficiently small fixed positive constant c, any estimator $\widetilde{\sigma}^2$ will satisfy that, if*

$$\sup_{f\in\Lambda_{\alpha_2,I}(C_{\mathcal{F}})}\sup_{\sigma^2\le C_\sigma}\sup_{\mathbb{P}_{(X,\varepsilon)}\in\mathcal{P}^{\text{adapt}}_{\text{cv},(X,\varepsilon)}} \mathbb{E}\big((\widetilde{\sigma}^2 - \sigma^2)/\phi_{n,\alpha_2}\big)^2 \le c,$$

then

$$\sup_{f\in\Lambda_{\alpha_1,I}(C_{\mathcal{F}})}\sup_{\sigma^2\le C_\sigma}\sup_{\mathbb{P}_{(X,\varepsilon)}\in\mathcal{P}^{\text{adapt}}_{\text{cv},(X,\varepsilon)}} \mathbb{E}\big((\widetilde{\sigma}^2 - \sigma^2)/\phi_{n,\alpha_1}\big)^2 \ge c.$$

The above two results combined are in line with analogous adaptation results in quadratic functional estimation (Efromovich and Low (1996), Cai and Low (2006)).

5. Proof of Theorem 2.

PROOF. We will only prove the lower bound $n^{-8\alpha/(4\alpha+1)}$ in the regime $0 < \alpha < 1/4$ since for $\alpha \ge 1/4$, the rate reduces to the parametric rate n^{-1} and the proof is straightforward. Throughout the proof, C represents a generic sufficiently large positive constant and c represents a generic sufficiently small positive constant always taken to be smaller than $1/4$. Both C and c only depend on α, $C_{\mathcal{F}}$, C_σ, C_ε, C_0, c_0 and might have different values for each occurrence. By appropriately rescaling the parameters in the lower bound construction, without loss of generality, we assume that the sample size n and the constants $C_{\mathcal{F}}$, C_σ, C_ε, C_0 are sufficiently large, c_0 is sufficiently small, and $[0,1] \subset I$.

We will make use of Le Cam's two point method. Introduce the following constants:

$$\theta_n^2 := h_n^{2\alpha} := cn^{-4\alpha/(4\alpha+1)} \quad \text{and} \quad N := N_n := 1/(6h_n), \tag{24}$$

where we tune the constant c in h_n so that N is a positive integer. We now specify $f(\cdot)$, distribution of X and distribution of ε in the null and alternative hypotheses, H_0 and H_1, respectively.

Choice of σ^2: Under H_0, let $\sigma^2 = 1 + \theta_n^2$. Under H_1, let $\sigma^2 = 1$.

Choice of ε: Under both H_0 and H_1, let $\varepsilon \sim \mathcal{N}(0,1)$.

Choice of X: Under both H_0 and H_1, let X be uniformly distributed on the union of the intervals $[(6i-5)h_n, (6i-1)h_n]$ for $i \in [N]$.

Choice of $f(\cdot)$: Under H_0, let $f \equiv 0$. Under H_1, let f take the value $h_n^\alpha r_i$ on $[(6i-5)h_n, (6i-1)h_n]$, where $\{r_i\}_{i=1}^N$ are N i.i.d. symmetric and bounded random variables with distribution $\mathbb{G}$ satisfying

$$\int_{-\infty}^{\infty} x^j \mathbb{G}(dx) = \int_{-\infty}^{\infty} x^j \varphi(x)\,dx, \quad j = 1, \ldots, q, \tag{25}$$

where q is some fixed odd integer strictly larger than $1 + 1/(2\alpha)$. Let f be 0 at points $6(i-1)h_n$ for $i \in [N]$, and then linearly interpolate f for the rest of the unspecified points on $[0,1]$.

See Figure 1 for an illustration. In the definition of $f(\cdot)$ under H_1, the existence of the distribution $\mathbb{G}$ is guaranteed by Lemma 1, and the range of $\{r_i\}_{i=1}^N$, which we denote as B, only depends on α.

Clearly, $\sigma^2 \le C_\sigma$ under both H_0 and H_1. Moreover, $f(\cdot)$ under both H_0 and H_1 belongs to $\Lambda_{\alpha,[0,1]}(C_{\mathcal{F}})$ due to the boundedness of $\{r_i\}_{i=1}^N$ in H_1. Next, we show that the joint distribution of (X, ε) belongs to $\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}$. Condition (d) clearly holds and Condition (a) holds with $I = [0,1]$. Condition (b) holds as well by the fact that $p_X(u) = 3/2$ for $u \in [(6i-5)h_n, (6i-1)h_n]$ for $i \in [N]$ and $p_X(u) = 0$ otherwise. Lastly, for Condition (c), it holds by the convolution formula that for any $0 < u < 1/2$,

$$\begin{aligned} p_{\widetilde{X}_{ij}}(u) &= \int_u^1 p_X(t) p_X(t-u)\,dt \ge \sum_{i=\lceil u/(6h_n)\rceil+1}^{N} \int_{(6i-5)h_n}^{(6i-1)h_n} p_X(t) p_X(t-u)\,dt \\ &\ge \sum_{i=\lceil u/(6h_n)\rceil+1}^{N} \frac{3}{2}\cdot\frac{3}{2}\cdot 2h_n \ge \frac{3}{8} - 9h_n \ge \frac{1}{4} \end{aligned}$$

for sufficiently large n. Here, the second inequality follows from the fact that for any fixed $t \in [(6i-5)h_n, (6i-1)h_n]$, $p_X(t) = 3/2$ and $p_X(t-u) = 0$ on a subset with Lebesgue measure at most $2h_n$. By symmetry of $\widetilde{X}_{ij}$, Condition (c) also holds with $\delta_0 = 1/2$ and $\mathcal{U}_\delta \equiv [-1, 1]$.

Denote by σ_i^2, f_i, $\mathbb{P}_{i,(X,\varepsilon)}$, $i = 0, 1$, the choice of σ^2, f, and $\mathbb{P}_{(X,\varepsilon)}$ under H_0 and H_1, respectively. Let π be the distribution on $\Lambda_{\alpha,I}(C_{\mathcal{F}})$ such that $f_1 \sim \pi$. Moreover, let $\mathbb{E}_{\sigma^2, f, \mathbb{P}_{(X,\varepsilon)}}$ represent the expectation with respect to the model (2) with parameters σ^2, f, $\mathbb{P}_{(X,\varepsilon)}$. Then we have

$$\begin{aligned} &\inf_{\widetilde{\sigma}^2} \sup_{f\in\Lambda_{\alpha,I}(C_{\mathcal{F}})} \sup_{\sigma^2\le C_\sigma} \sup_{\mathbb{P}_{(X,\varepsilon)}\in\mathcal{P}_{\mathrm{cv},(X,\varepsilon)}} \mathbb{E}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2 \\ &\quad\ge \inf_{\widetilde{\sigma}^2}\left\{\frac{1}{2}\mathbb{E}_{\sigma_0^2, f_0, \mathbb{P}_{0,(X,\varepsilon)}}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2 + \frac{1}{2}\int \mathbb{E}_{\sigma_1^2, f, \mathbb{P}_{1,(X,\varepsilon)}}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2\,d\pi(f)\right\} \\ &\quad\ge \inf_{\widetilde{\sigma}^2}\left\{\frac{1}{2}\mathbb{E}_{\sigma_0^2, f_0, \mathbb{P}_{0,(X,\varepsilon)}}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2 + \frac{1}{2}\mathbb{E}_{\sigma_1^2, f_1, \mathbb{P}_{1,(X,\varepsilon)}}\big(\widetilde{\sigma}^2 - \sigma^2\big)^2\right\}, \end{aligned}$$

where the first inequality follows by lower bounding the maximum risk with Bayes risk with prior π. In what follows, we will use $\mathbb{P}_0$ and $\mathbb{P}_1$ to denote the joint distribution of $\{Y_i, X_i\}_{i=1}^n$ under H_0 and H_1, respectively. Note that the choice of θ_n^2 in (24) leads to the desired lower bound under the quadratic loss. Therefore, adopting the standard reduction scheme with Le Cam's two point method (cf. Theorem 2.2 in Tsybakov (2009)), it suffices to show that $\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) \le c < 1$. To show this, let $\{\widetilde{r}_i\}_{i=1}^N$ be N i.i.d. standard normal random variables,

and $\widetilde{\mathbb{P}}_1$ be the joint distributions of $\{X_i, Y_i\}_{i=1}^n$ under H_1 with $\{r_i\}_{i=1}^N$ replaced by $\{\widetilde{r}_i\}_{i=1}^N$. Then, by triangle inequality, we have

$$\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) \le \mathsf{TV}(\mathbb{P}_0, \widetilde{\mathbb{P}}_1) + \mathsf{TV}(\mathbb{P}_1, \widetilde{\mathbb{P}}_1).$$

We will show $\mathsf{TV}(\mathbb{P}_0, \widetilde{\mathbb{P}}_1) \le c$ and $\mathsf{TV}(\mathbb{P}_1, \widetilde{\mathbb{P}}_1) \le c$ separately.

For the first inequality, define $\boldsymbol{x} := (x_1, \ldots, x_n)$, $d\boldsymbol{x} := dx_1 \cdots dx_n$ and similarly for $\boldsymbol{y}$ and $d\boldsymbol{y}$. Denote p_0, p_1, and $\widetilde{p}_1$ as the densities of $\mathbb{P}_0$, $\mathbb{P}_1$, and $\widetilde{\mathbb{P}}_1$ with respect to the Lebesgue measure. Then we have

$$\begin{aligned}\mathsf{TV}(\mathbb{P}_0, \widetilde{\mathbb{P}}_1) &= \frac{1}{2}\int\int \bigl|p_0(\boldsymbol{x}, \boldsymbol{y}) - \widetilde{p}_1(\boldsymbol{x}, \boldsymbol{y})\bigr|\, d\boldsymbol{x}\, d\boldsymbol{y} \\ &= \int p(\boldsymbol{x})\, d\boldsymbol{x} \left\{\frac{1}{2}\int \bigl|p_0(\boldsymbol{y} \mid \boldsymbol{x}) - \widetilde{p}_1(\boldsymbol{y} \mid \boldsymbol{x})\bigr|\, d\boldsymbol{y}\right\} \\ &= \int p(\boldsymbol{x})\, d\boldsymbol{x}\, \mathsf{TV}\bigl(\mathbb{P}_0(\boldsymbol{y} \mid \boldsymbol{x}), \widetilde{\mathbb{P}}_1(\boldsymbol{y} \mid \boldsymbol{x})\bigr),\end{aligned} \tag{26}$$

where $p(\boldsymbol{x}) := \prod_{i=1}^n p_X(X_i)$ stands for the common density of $\{X_i\}_{i=1}^n$ under $\mathbb{P}_0$ and $\widetilde{\mathbb{P}}_1$. Note that under $\mathbb{P}_0$, $\boldsymbol{y} \mid \boldsymbol{x} \sim \mathcal{N}_n(0, \boldsymbol{\Sigma_0})$, with $\boldsymbol{\Sigma_0} = (1 + \theta_n^2)\mathbf{I}_n$. Define $\{b_i\}_{i=1}^n$ to be the location index sequence of $\{X_i\}_{i=1}^n$ taking values in $[N]$, that is,

$$b_i = j \quad \text{if } X_i \in \bigl[(6j-5)h_n, (6j-1)h_n\bigr].$$

Then, due to the symmetry of $\{r_i\}_{i=1}^N$ and design of the nonparametric component f, it holds that under $\widetilde{\mathbb{P}}_1$, $\boldsymbol{y} \mid \boldsymbol{x} \sim \mathcal{N}_n(0, \boldsymbol{\Sigma_1})$, with $(\boldsymbol{\Sigma_1})_{ii} = 1 + h_n^{2\alpha} = 1 + \theta_n^2$ and $(\boldsymbol{\Sigma_1})_{ij} = h_n^{2\alpha}\mathbb{1}\{b_i = b_j\}$ for $i \ne j$. Define $N_0 := \sum_{i \ne j} \mathbb{1}\{b_i = b_j\}$. Since $\boldsymbol{\Sigma_1}$ is positive definite (see Lemma A5 in the Supplementary Material (Shen et al. (2020))), we have by Lemma 2 that

$$\mathsf{TV}\bigl(\mathbb{P}_0(\boldsymbol{y} \mid \boldsymbol{x}), \widetilde{\mathbb{P}}_1(\boldsymbol{y} \mid \boldsymbol{x})\bigr) \le C\frac{\theta_n^2}{1 + \theta_n^2} N_0^{1/2} \le C\theta_n^2 N_0^{1/2}.$$

Note that N_0 is a random variable that depends on $\{X_i\}_{i=1}^n$, and by (26) and Jensen's inequality we have

$$\mathsf{TV}(\mathbb{P}_0, \widetilde{\mathbb{P}}_1) \le C\theta_n^2 \mathbb{E} N_0^{1/2} \le C\theta_n^2 (\mathbb{E} N_0)^{1/2}.$$

Some simple algebra shows that $\mathbb{E}N_0 \le Cn^2 h_n$, thus by choosing a sufficiently small c in the definition of h_n in (24), we have

$$\mathsf{TV}(\mathbb{P}_0, \widetilde{\mathbb{P}}_1) \le C\theta_n^2 n h_n^{1/2} \le c.$$

To complete the proof, we now show that $\mathsf{TV}(\mathbb{P}_1, \widetilde{\mathbb{P}}_1) \le c$. Consider an arbitrary realization of $\{X_i\}_{i=1}^n$, and assume that based on their location indices $\{b_i\}_{i=1}^n$, $\{X_i\}_{i=1}^n$ is partitioned into L clusters with corresponding cardinality s_ℓ so that the X_i's in the same cluster have the same value b_i. Apparently, we have the relations $1 \le L \le n$ and $\sum_{\ell=1}^L s_\ell = n$. Let $m_{\max}$ be the maximum cluster size, and define the "good event" $\Omega_n := \{m_{\max} \le K\}$, where $K := \lfloor 2/(1 - 4\alpha)\rfloor + 2$. Then it holds that

$$\begin{aligned}\mathsf{TV}(\mathbb{P}_1, \widetilde{\mathbb{P}}_1) &= \mathbb{E}\bigl(\mathbb{1}_{\Omega_n}\mathsf{TV}\bigl(\mathbb{P}_1(\boldsymbol{y} \mid \boldsymbol{x}), \widetilde{\mathbb{P}}_1(\boldsymbol{y} \mid \boldsymbol{x})\bigr)\bigr) + \mathbb{E}\bigl(\mathbb{1}_{\Omega_n^c}\mathsf{TV}\bigl(\mathbb{P}_1(\boldsymbol{y} \mid \boldsymbol{x}), \widetilde{\mathbb{P}}_1(\boldsymbol{y} \mid \boldsymbol{x})\bigr)\bigr) \\ &\le \mathbb{E}\bigl(\mathbb{1}_{\Omega_n}\mathsf{TV}\bigl(\mathbb{P}_1(\boldsymbol{y} \mid \boldsymbol{x}), \widetilde{\mathbb{P}}_1(\boldsymbol{y} \mid \boldsymbol{x})\bigr)\bigr) + \mathbb{P}\bigl(\Omega_n^c\bigr).\end{aligned}$$

Under the choice of h_n in (24), N is of the order $n^{2/(4\alpha+1)}$, and

$$\lambda_K := \lim_{n \to \infty} \frac{n^K}{K! N^{K-1}} = 0.$$

Thus by Lemma 3 (and continuity), it holds that Ω_n has asymptotic probability 1 under both $\mathbb{P}_1$ and $\widetilde{\mathbb{P}}_1$. As a result, it suffices to upper bound $\mathsf{TV}(\mathbb{P}_1(\boldsymbol{y} \mid \boldsymbol{x}), \widetilde{\mathbb{P}}_1(\boldsymbol{y} \mid \boldsymbol{x}))$ for each realization $\boldsymbol{x}$ in Ω_n, where the maximum cluster size $m_{\max}$ is bounded by a fixed constant.

Denoting p_{1,π_ℓ} and $\widetilde{p}_{1,\pi_\ell}$ for each $\ell \in [L]$ as the joint density of those y_i's in the ℓth cluster π_ℓ conditioning on the given realization of $\{X_i\}_{i=1}^n$ under $\mathbb{P}_1$ and $\widetilde{\mathbb{P}}_1$, we obtain that

$$p_1(\boldsymbol{y} \mid \boldsymbol{x}) - \widetilde{p}_1(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{\ell=1}^{L} p_{1,\pi_\ell} - \prod_{\ell=1}^{L} \widetilde{p}_{1,\pi_\ell}.$$

The above inequality further implies by telescoping that

$$\left|p_1(\boldsymbol{y} \mid \boldsymbol{x}) - \widetilde{p}_1(\boldsymbol{y} \mid \boldsymbol{x})\right| \le \sum_{\ell=1}^{L} |p_{1,\pi_\ell} - \widetilde{p}_{1,\pi_\ell}|.$$

For each $\ell \in [L]$, $|p_{1,\pi_\ell} - \widetilde{p}_{1,\pi_\ell}|$ only depends on the ℓth cluster through its cardinality, which we now control for a general cluster size $d \ge 1$. Without loss of generality, we assume that $\ell = 1$ and the y_i's in this cluster are $\{y_1, \ldots, y_d\}$ with common location index $b_i = 1$ for $i \in [d]$. Then, under the choice of θ_n^2 in (24), we clearly have $Y_i = \theta_n r_1 + \varepsilon_i$ under $\mathbb{P}_1$ and $Y_i = \theta_n \widetilde{r}_1 + \varepsilon_i$ under $\widetilde{\mathbb{P}}_1$ for $i \in [d]$, where the sequence $\{\varepsilon_i\}_{i=1}^d$ follows the standard normal distribution under both $\mathbb{P}_1$ and $\widetilde{\mathbb{P}}_1$. Therefore, it holds that

$$\begin{aligned} p_{1,\pi_1}(y_1, \ldots, y_d) &= \int_{-\infty}^{\infty} \varphi(y_1 - \theta_n v) \ldots \varphi(y_d - \theta_n v) \mathbb{G}(dv), \\ \widetilde{p}_{1,\pi_1}(y_1, \ldots, y_d) &= \int_{-\infty}^{\infty} \varphi(y_1 - \theta_n v) \ldots \varphi(y_d - \theta_n v) \varphi(v)\, dv, \end{aligned}$$

where $\mathbb{G}$ is the distribution of $\{r_i\}_{i=1}^N$ specified in (25). Using the well-known equality $\varphi(t - \theta_n v) = \varphi(t)(\sum_{k=0}^{\infty} v^k \theta_n^k H_k(t)/k!)$ for any t, v, where H_k is the kth order Hermite polynomial, it holds that

$$\begin{aligned} &\varphi(y_1 - \theta_n v) \cdots \varphi(y_d - \theta_n v) \\ &\quad = \varphi(y_1) \cdots \varphi(y_d) \sum_{k_1, \ldots, k_d = 0}^{\infty} v^{\sum_{i=1}^d k_i} \theta_n^{\sum_{i=1}^d k_i} \frac{H_{k_1}(y_1)}{k_1!} \cdots \frac{H_{k_d}(y_d)}{k_d!} \\ &\quad = \varphi(y_1) \cdots \varphi(y_d) \sum_{k=0}^{\infty} v^k \theta_n^k \sum_{k_1 + \cdots + k_d = k} \frac{H_{k_1}(y_1)}{k_1!} \cdots \frac{H_{k_d}(y_d)}{k_d!} \end{aligned}$$

and, therefore,

$$\begin{aligned} &p_{1,\pi_1}(y_1, \ldots, y_d) - \widetilde{p}_{1,\pi_1}(y_1, \ldots, y_d) \\ &\quad = \varphi(y_1) \cdots \varphi(y_d) \sum_{k=0}^{\infty} \theta_n^k \sum_{k_1 + \cdots + k_d = k} \frac{H_{k_1}(y_1)}{k_1!} \cdots \frac{H_{k_d}(y_d)}{k_d!} \int v^k (\mathbb{G} - \Phi)(dv) \\ &\quad = \varphi(y_1) \cdots \varphi(y_d) \sum_{k=p}^{\infty} \theta_n^{2k} \sum_{k_1 + \cdots + k_d = 2k} \frac{H_{k_1}(y_1)}{k_1!} \cdots \frac{H_{k_d}(y_d)}{k_d!} \int v^{2k} (\mathbb{G} - \Phi)(dv), \end{aligned}$$

where the second equality follows by the symmetry and moment matching property of $\mathbb{G}$ in (25) and $p := (q+1)/2$ is a positive integer. This further yields

$$\begin{aligned}
&\bigl|p_{1,\pi_1}(y_1,\ldots,y_d) - \widetilde{p}_{1,\pi_1}(y_1,\ldots,y_d)\bigr| \\
&\quad\le \varphi(y_1)\cdots\varphi(y_d)\sum_{k=p}^{\infty}\theta_n^{2k}\sum_{k_1+\cdots+k_d=2k}\frac{|H_{k_1}(y_1)|}{k_1!}\cdots\frac{|H_{k_d}(y_d)|}{k_d!}\int v^{2k}\mathbb{G}(dv) \\
&\qquad+\varphi(y_1)\cdots\varphi(y_d)\sum_{k=p}^{\infty}\theta_n^{2k}\sum_{k_1+\cdots+k_d=2k}\frac{|H_{k_1}(y_1)|}{k_1!}\cdots\frac{|H_{k_d}(y_d)|}{k_d!}\int v^{2k}\varphi(v)\,dv \\
&\quad:= I + II.
\end{aligned}$$

For term I, since $\mathbb{G}$ is compactly supported on $[-B, B]$, one clearly has

$$I \le \varphi(y_1)\cdots\varphi(y_d)\sum_{k=p}^{\infty}\theta_n^{2k}B^{2k}\sum_{k_1+\cdots+k_d=2k}\frac{|H_{k_1}(y_1)|}{k_1!}\cdots\frac{|H_{k_d}(y_d)|}{k_d!}.$$

For term II, using the equality $\int\varphi(v)v^{2k}\,dv = (2k-1)!!$, with $(2k-1)!! := (2k-1)(2k-3)\ldots 1$, we obtain

$$II = \varphi(y_1)\cdots\varphi(y_d)\sum_{k=p}^{\infty}\theta_n^{2k}(2k-1)!!\sum_{k_1+\cdots+k_d=2k}\frac{|H_{k_1}(y_1)|}{k_1!}\cdots\frac{|H_{k_d}(y_d)|}{k_d!}.$$

We now upper bound $\int_{-\infty}^{\infty}|H_k(t)|\varphi(t)\,dt$ for an arbitrary positive integer k. When k is even, as has been calculated in Wang et al. (2008) (cf. chain of inequality after equation (19) on page 662), $\int_{-\infty}^{\infty}|H_k(t)|\varphi(t)\,dt \le 2^{k/2}(k-1)!!$. When k is odd, set $k = 2\widetilde{k}+1$, then we have

$$\begin{aligned}
\int_{-\infty}^{\infty}\bigl|H_k(t)\bigr|\varphi(t)\,dt &= \int_{-\infty}^{\infty}\varphi(t)\left|(2\widetilde{k}+1)!\sum_{m=0}^{\widetilde{k}}\frac{(-1)^m t^{2\widetilde{k}+1-2m}}{m!(2\widetilde{k}+1-2m)!2^m}\right|dt \\
&\le \sum_{m=0}^{\widetilde{k}}\frac{(2\widetilde{k}+1)!}{m!(2\widetilde{k}+1-2m)!2^m}\int_{-\infty}^{\infty}|t|^{2\widetilde{k}+1-2m}\varphi(t)\,dt \\
&= \sqrt{\frac{2}{\pi}}\sum_{m=0}^{\widetilde{k}}\frac{(2\widetilde{k}+1)!(2\widetilde{k}-2m)!!}{m!(2\widetilde{k}+1-2m)!2^m} \\
&= \sqrt{\frac{2}{\pi}}\sum_{m=0}^{\widetilde{k}}\frac{(2\widetilde{k}+1)!(2m)!!}{(\widetilde{k}-m)!(2m+1)!2^{\widetilde{k}-m}} \\
&= \sqrt{\frac{2}{\pi}}(2\widetilde{k}+1)!!\sum_{m=0}^{\widetilde{k}}\frac{\widetilde{k}!}{m!(\widetilde{k}-m)!}\frac{(m!)^2 2^{2m}}{(2m+1)!} \\
&\le (2\widetilde{k}+1)!!\sum_{m=0}^{\widetilde{k}}\frac{\widetilde{k}!}{m!(\widetilde{k}-m)!} \\
&= (2\widetilde{k}+1)!!2^{\widetilde{k}},
\end{aligned}$$

where in the third line we use the fact that $\int_{-\infty}^{\infty}|t|^{2\ell+1}\varphi(t)\,dt = \sqrt{2/\pi}(2\ell)!!$ for any positive integer ℓ. Define for any positive integer k: $[k]_1 := k-1$ if k is even and k if k is odd, and $[k]_2 := k/2$ if k is even and $(k-1)/2$ if k is odd. Then the above calculation implies that

$\int_{-\infty}^{\infty} |H_k(t)|\varphi(t)\,dt \le ([k]_1)!!2^{[k]_2}$ for any k, and moreover, it can be readily checked that $([k]_1)!!/(k!) = 1/(2^{[k]_2}([k]_2)!)$. Therefore, for term $I = I(y_1, \ldots, y_d)$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} I(y_1, \ldots, y_d)\,dy_1 \cdots dy_d \\
&\quad \le \sum_{k=p}^{\infty} \theta_n^{2k} (B^2)^k \sum_{k_1+\cdots+k_d=2k} \frac{1}{(k_1)! \cdots (k_d)!} ([k_1]_1)!!2^{[k_1]_2} \cdots ([k_d]_1)!!2^{[k_d]_2} \\
&\quad = \sum_{k=p}^{\infty} \theta_n^{2k} (B^2)^k \sum_{k_1+\cdots+k_d=2k} \frac{1}{([k_1]_2)! \cdots ([k_d]_2)!}.
\end{aligned}$$

Now note that the number of d-tuple $(k_1, \ldots, k_d)$ such that $k_1 + \cdots + k_d = 2k$ is upper bounded by $(Ck)^d$, which is further bounded by C^k for every $k \ge 0$ with some sufficiently large C that only depends on d, and for each such tuple, it holds that

$$k - \frac{d}{2} = \sum_{i=1}^{d} \frac{k_i - 1}{2} \le \sum_{i=1}^{d} [k_i]_2 \le \sum_{i=1}^{d} \frac{k_i}{2} = k,$$

thus we have

$$\sum_{k_1+\cdots+k_d=2k} \{([k_1]_2)! \cdots ([k_d]_2)!\}^{-1} \le C^k \sum_{k-d/2 \le \bar{k}_1+\cdots+\bar{k}_d \le k} \{(\bar{k}_1)! \cdots (\bar{k}_d)!\}^{-1}.$$

For the latter quantity, we have by the multinomial identity

$$\sum_{x_1+\cdots+x_{d+1}=k} k!/(x_1! \cdots x_{d+1}!)(d+1)^{-k} = 1$$

that

$$\begin{aligned}
\frac{(d+1)^k}{k!} &= \sum_{\bar{k}_1+\cdots+\bar{k}_{d+1}=k} \frac{1}{(\bar{k}_1)! \cdots (\bar{k}_{d+1})!} \\
&= \sum_{\bar{k}_1+\cdots+\bar{k}_d \le k} \frac{1}{(\bar{k}_1)! \cdots (\bar{k}_d)!(k - (\bar{k}_1 + \cdots + \bar{k}_d))!} \\
&\ge \sum_{k-d/2 \le \bar{k}_1+\cdots+\bar{k}_d \le k} \frac{1}{(\bar{k}_1)! \cdots (\bar{k}_d)!(k - (\bar{k}_1 + \cdots + \bar{k}_d))!} \\
&\ge \left(\left(\frac{d}{2}\right)!\right)^{-1} \sum_{k-d/2 \le \bar{k}_1+\cdots+\bar{k}_d \le k} \frac{1}{(\bar{k}_1)! \cdots (\bar{k}_d)!}.
\end{aligned}$$

This concludes that

$$\int_{\mathbb{R}^d} I(y_1, \ldots, y_d)\,dy_1 \cdots dy_d \le \theta_n^{2p} \sum_{k=p}^{\infty} \frac{(CB^2)^k}{k!} \le \theta_n^{2p} e^{CB^2}.$$

Using a similar argument for $II = II(y_1, \ldots, y_d)$, we obtain

$$\begin{aligned}
\int_{\mathbb{R}^d} II(y_1, \ldots, y_d)\,dy_1 \cdots dy_d &\le \sum_{k=p}^{\infty} \frac{(2k-1)!!}{k!} \theta_n^{2k} C^k \\
&= \sum_{k=p}^{\infty} \frac{(2k-1)!!}{(2k)!!} \theta_n^{2k} (2C)^k \le \theta_n^{2p} C^p
\end{aligned} \tag{27}$$

since $\theta_n^2 < 1/C$ for sufficiently large n.

Putting together the pieces, we have for every realization $\boldsymbol{x}$ in Ω_n,

$$\int_{\mathbb{R}^n} |p_1(\boldsymbol{y} \mid \boldsymbol{x}) - \widetilde{p}_1(\boldsymbol{y} \mid \boldsymbol{x})| d\boldsymbol{y} \le \sum_{\ell=1}^{L} \int_{\mathbb{R}^{|\pi_\ell|}} |p_{1,\pi_\ell} - \widetilde{p}_{1,\pi_\ell}| \le L \max_{1\le d\le K} \theta_n^{2p}\big(e^{CB^2} + C^p\big)$$

$$\le n\theta_n^{2p}\big(e^{CB^2} + C^p\big) \le c.$$

Here, the second inequality follows since every $|p_{1,\pi_\ell} - \widetilde{p}_{1,\pi_\ell}|$ depends on the ℓth cluster only through its cardinality, the third inequality follows since $L \le n$ and K is a fixed absolute constant that only depends on α, and the last inequality follows due to the choice $\theta_n^2 = h_n^{2\alpha} = cn^{-4\alpha/(4\alpha+1)}$ and the value of p. This completes the proof. □

LEMMA 1 (Lemma 1, Wang et al. (2008)). *For any fixed positive integer q, there exist a $B < \infty$ and a symmetric distribution $\mathbb{G}$ on $[-B, B]$ such that $\mathbb{G}$ and the standard normal distribution have the same first q moments, that is,*

$$\int_{-B}^{B} x^j \mathbb{G}(dx) = \int_{-\infty}^{\infty} x^j \varphi(x)\, dx, \quad j = 1, \dots, q.$$

LEMMA 2 (Theorem 1.1, Devroye, Mehrabian and Reddad (2018)). *If $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive definite $d \times d$ matrices, then*

$$\frac{1}{100} \le \frac{\mathrm{TV}(\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1), \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2))}{\min\{1, \|\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2 - \mathbf{I}_d\|_F\}} \le \frac{3}{2}.$$

For the following lemma, we first introduce some terminology regarding the multinomial distribution. Let m, M be two positive integers, and the random vector $(f_1, \dots, f_M)$ be the multinomial count with total count m and equal probability $(1/M, 1/M, \dots, 1/M)$. Define $\rho := m/M$. For any positive integer $r \ge 2$, define $\lambda := \lambda_r := \lim_{m\to\infty} m^r/(r!M^{r-1})$. Following Kolchin, Sevast'yanov and Chistyakov (1978) (Chapter 2, equation (11)), we will call the domain of variation $m, M \to \infty$, in which

$$\rho \to 0, \qquad 0 < \lambda_r < \infty$$

the *left-hand r-domain*. The following lemma characterizes the asymptotic behavior of the maximum frequency $f_{\max}$ defined as $\max_{1\le j\le M} f_j$.

LEMMA 3 (Theorem 1 of Section 2.6, Kolchin, Sevast'yanov and Chistyakov (1978)). *Suppose the multinomial distribution with total count m and equal probability $(1/M, \dots, 1/M)$ is in the left-hand r-domain for some positive integer $r \ge 2$ with limit λ_r, then it holds that*

$$\mathbb{P}(f_{\max} = r - 1) \to e^{-\lambda_r} \quad \text{and} \quad \mathbb{P}(f_{\max} = r) \to 1 - e^{-\lambda_r},$$

that is, the maximum frequency converges asymptotically to a two-point distribution.

Acknowledgments. The research of Y. Shen was supported in part by NSF Grant DMS-1252624. The research of C. Gao was supported in part by NSF Grant DMS-1712957. The research of D. Witten was supported in part by NIH Grant DP5OD009145, NSF CAREER Award DMS-1252624, and a Simons Investigator Award in Mathematical Modeling of Living Systems (Award Number 560585). The research of F. Han was supported in part by NSF Grant DMS-1712536. The authors thank Eric J. Tchetgen Tchetgen and Cun-Hui Zhang for helpful discussions and constructive comments. The authors also thank the Co-Editor Ming Yuan, an anonymous Associate Editor and two anonymous referees for their helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal estimation of variance in nonparametric regression with random design" (DOI: 10.1214/20-AOS1944SUPP; .pdf). This supplement contains proofs of remaining results.

REFERENCES

BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028 https://doi.org/10.1007/s004400050210

BHATTACHARYA, A., PATI, D. and DUNSON, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Ann. Statist.* **42** 352–381. MR3189489 https://doi.org/10.1214/13-AOS1192

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393. MR1065550

BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **71** 271–291. MR0816706 https://doi.org/10.1007/BF00332312

BROWN, L. D. and LEVINE, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *Ann. Statist.* **35** 2219–2232. MR2363969 https://doi.org/10.1214/009053607000000145

CAI, T. T., LEVINE, M. and WANG, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivariate Anal.* **100** 126–136. MR2460482 https://doi.org/10.1016/j.jmva.2008.03.007

CAI, T. T. and LOW, M. G. (2006). Optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **34** 2298–2325. MR2291501 https://doi.org/10.1214/009053606000000849

CAI, T. T. and WANG, L. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *Ann. Statist.* **36** 2025–2054. MR2458178 https://doi.org/10.1214/07-AOS509

DEVROYE, L., MEHRABIAN, A. and REDDAD, T. (2018). The total variation distance between high-dimensional Gaussians. Preprint. Available at arXiv:1810.08693.

DICKER, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101** 269–284. MR3215347 https://doi.org/10.1093/biomet/ast065

DOKSUM, K. and SAMAROV, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Ann. Statist.* **23** 1443–1473. MR1370291 https://doi.org/10.1214/aos/1176324307

DONOHO, D. L. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323. MR1081043 https://doi.org/10.1016/0885-064X(90)90025-9

EFROMOVICH, S. and LOW, M. (1996). On optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **24** 1106–1125. MR1401840 https://doi.org/10.1214/aos/1032526959

FAN, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19** 1273–1294. MR1126325 https://doi.org/10.1214/aos/1176348249

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998–1004. MR1209561

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21** 196–216. MR1212173 https://doi.org/10.1214/aos/1176349022

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587

FAN, J. and YAO, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85** 645–660. MR1665822 https://doi.org/10.1093/biomet/85.3.645

GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633. MR0897854 https://doi.org/10.1093/biomet/73.3.625

GINÉ, E., LATAŁA, R. and ZINN, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability, II* (*Seattle, WA*, 1999). *Progress in Probability* **47** 13–38. Birkhäuser, Boston, MA. MR1857312

GINÉ, E. and NICKL, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli* **14** 47–61. MR2401653 https://doi.org/10.3150/07-BEJ110

HALL, P. and CARROLL, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. *J. Roy. Statist. Soc. Ser. B* **51** 3–14. MR0984989

HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528. MR1087842 https://doi.org/10.1093/biomet/77.3.521

HALL, P. and MARRON, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77** 415–419. MR1064818 https://doi.org/10.1093/biomet/77.2.415

HÄRDLE, W. and TSYBAKOV, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics* **81** 223–242. MR1484586 https://doi.org/10.1016/S0304-4076(97)00044-4

HOFFMANN, M. and LEPSKI, O. (2002). Random rates in anisotropic regression. *Ann. Statist.* **30** 325–396. MR1902892 https://doi.org/10.1214/aos/1021379858

HUANG, L.-S. and FAN, J. (1999). Nonparametric estimation of quadratic regression functionals. *Bernoulli* **5** 927–949. MR1715445 https://doi.org/10.2307/3318450

IBRAGIMOV, I. A. and KHAS'MINSKIĬ, R. Z. (1981). More on estimation of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **108** 72–88. MR0629401

KOLCHIN, V. F., SEVAST'YANOV, B. A. and CHISTYAKOV, V. P. (1978). *Random Allocations*. V. H. Winston & Sons, Washington, DC. MR0471016

KONG, W. and VALIANT, G. (2018). Estimating learnability in the sublinear data regime. Preprint. Available at arXiv:1805.01626.

LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24** 659–681. MR1394981 https://doi.org/10.1214/aos/1032894458

LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785 https://doi.org/10.1214/aos/1015957395

LEPSKI, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

LEPSKI, O. (1992). Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

MEYER, Y. (1990). *Ondelettes et Opérateurs. I: Ondelettes. Actualités Mathématiques*. [*Current Mathematical Topics*.] Hermann, Paris. MR1085487

MÜLLER, U. U., SCHICK, A. and WEFELMEYER, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. *Statistics* **37** 179–188. MR1986175 https://doi.org/10.1080/0233188031000078051

MÜLLER, H.-G. and STADTMÜLLER, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15** 610–625. MR0888429 https://doi.org/10.1214/aos/1176350364

MUNK, A. and RUYMGAART, F. (2002). Minimax rates for estimating the variance and its derivatives in nonparametric regression. *Aust. N. Z. J. Stat.* **44** 479–488. MR1934736 https://doi.org/10.1111/1467-842X.00249

MUNK, A., BISSANTZ, N., WAGNER, T. and FREITAG, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 19–41. MR2136637 https://doi.org/10.1111/j.1467-9868.2005.00486.x

NUSSBAUM, M. (1986). On nonparametric estimation of a regression function, being smooth on a domain in $\mathbb{R}^k$. *Theory Probab. Appl.* **31** 118–125.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230. MR0760684 https://doi.org/10.1214/aos/1176346788

ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman. Inst. Math. Stat.* (*IMS*) *Collect.* **2** 335–421. IMS, Beachwood, OH. MR2459958 https://doi.org/10.1214/193940307000000527

ROBINS, J., TCHETGEN, E. T., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270. MR1379468

RUPPERT, D., WAND, M. P., HOLST, U. and HÖSSJER, O. (1997). Local polynomial variance-function estimation. *Technometrics* **39** 262–273. MR1462587 https://doi.org/10.2307/1271131

SHEN, Y., GAO, C., WITTEN, D. and HAN, F. (2020). Supplement to "Optimal estimation of variance in nonparametric regression with random design." https://doi.org/10.1214/20-AOS1944SUPP

SPOKOINY, V. (2002). Variance estimation for high-dimensional regression models. *J. Multivariate Anal.* **82** 111–133. MR1918617 https://doi.org/10.1006/jmva.2001.2023

THOMPSON, A. M., KAY, J. W. and TITTERINGTON, D. M. (1991). Noise estimation in signal restoration using regularization. *Biometrika* **78** 475–488. MR1130916 https://doi.org/10.1093/biomet/78.3.475

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

VERZELEN, N. and GASSIAT, E. (2018). Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli* **24** 3683–3710. MR3788186 https://doi.org/10.3150/17-BEJ975

VERZELEN, N. and VILLERS, F. (2010). Goodness-of-fit tests for high-dimensional Gaussian linear models. *Ann. Statist.* **38** 704–752. MR2604699 https://doi.org/10.1214/08-AOS629

VON NEUMANN, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Stat.* **12** 367–395. MR0006656 https://doi.org/10.1214/aoms/1177731677

VON NEUMANN, J. (1942). A further remark concerning the distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Stat.* **13** 86–88. MR0006657 https://doi.org/10.1214/aoms/1177731645

WANG, L., BROWN, L. D., CAI, T. T. and LEVINE, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *Ann. Statist.* **36** 646–664. MR2396810 https://doi.org/10.1214/009053607000000901