Sequential Nonparametric Estimation of Controlled Multivariate Regression

SAM EFROMOVICH

Department of Mathematical Sciences, University of Texas at Dallas

Richardson, Texas, USA

Abstract: The paper considers an adaptive sequential nonparametric estimation of a multivariate regression with assigned mean integrated squared error (MISE) and minimax mean stopping time when the estimator matches performance of an oracle knowing all nuisance parameters and functions. It is known that the problem has no solution if regression belongs to a Sobolev class of differentiable functions. What if an underlying regression is smoother, say analytic? It is shown that in this case it is possible to match performance of the oracle. Furthermore, similarly to the classical Stein's solution for a parameter estimation, a two-stage sequential procedure solves the problem. The proposed regression estimator for the first stage, based on a sample with fixed sample size, is of interest on its own, and a thought-provoking environmental example of reducing potent greenhouse gas emission by an anaerobic digestion system is used to discuss a number of important topics for small samples.

Keywords: Adaptation; Minimax; MISE; Oracle approach; Minimal stopping time; Greengas.

Subject Classifications: 62G05; 62N01.

Address correspondence to Professor Sam Efromovich, Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75083-0688, USA; E-mail: efrom@utdallas.edu

1 Introduction

Nonparametric curve estimation is devoted to estimation of functions whose shape is unknown. A classical statistical setting is when a sample of size n is available, the problem is to propose a feasible estimator with a minimal mean integrated squared error (MISE), and oracle approach is used to find a benchmark for an adaptive estimator. The oracle knows some information about an underlying estimated function, including its smoothness, and everything about nuisance functions. Then a sharp lower bound for the MISE of oracle-estimators is established. The notion "sharp" means that both constant and rate of the MISE convergence are established. Then a good data-driven estimator should match performance of the oracle, and if the latter is possible then the estimator is called adaptive because it adapts to smoothness of an underlying estimated function and all nuisance functions. It is well known that adaptive nonparametric estimation is possible for a wide variety of statistical models and function classes of interest, see a discussion in Efromovich (1999,2018) and Wassermann (2006). This is a good news for nonparametric estimation with deterministic sample size.

Situation changes rather dramatically if we are interested in the Wald problem of sequential estimation with assigned value of a risk and a minimal mean stopping time. No adaptive sequential estimator, matching performance of the oracle, exists for the case of differentiable functions. More about the Wald problem, sequential estimation and the lack of adaptation can be found in Wald (1947), Stein and Wald (1947), Anscombe (1947,1953), Ghosh and Sen (1991), Ghosh, Mukhopadhyay and Sen (1997), Mukhopadhyay (1997), and Efromovich (1995,2007, 2018). While there is no way to change this outcome for estimation of differentiable functions, the paper shows that this is possible for smoother functions like analytic ones. Further, sequential estimation can use the simplest two-stage strategy whose roots go back to

Stein (1945) and Wald (1947), see also an interesting discussion in Aoshima and Yata (2011), Mukhopadhyay and Zacks (2018), and Mukhopadhyay (2019).

Let us describe a considered regression model and review relevant known results beginning with the case of a fixed sample size. We observe a sample $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ of size n from $(\mathbf{X}, Y)$ where $\mathbf{X} := (X_1, \dots, X_k)$ is a vector of continuous covariates (predictors) and Y is a response. The regression is controlled implying that the distribution of $\mathbf{X}$ is known, and in what follows it is supposed that the joint density $f^{\mathbf{X}}$ of the vector-predictor is supported and positive on k-dimensional cube $R := [0,1]^k$. The underlying regression model is

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\xi, \tag{1.1}$$

where $m(\mathbf{x}) := \mathbb{E}\{Y|\mathbf{X} = \mathbf{x}\}$ is the regression function of interest, ξ is a zero-mean regression error independent of $\mathbf{X}$, and a positive function $\sigma(\mathbf{x})$ is called a scale function. Let us formulate one of the main known theoretical results due to Hoffmann and Lepskii (2002). Consider a cosine tensor-product basis $\varphi_{\mathbf{i}}(\mathbf{x}) := \prod_{r=1}^k \varphi_{i_r}(x_r)$ on R where $\varphi_0(x) = 1$, $\varphi_i = 2^{1/2}\cos(\pi i x)$, $i = 1, 2, \dots$, $\mathbf{i} := (i_1, \dots, i_k)$, set $\theta_{\mathbf{i}} := \int_R m(\mathbf{x})\varphi_{\mathbf{i}}(\mathbf{x})d\mathbf{x}$ for Fourier coefficients of $m(\mathbf{x})$, and introduce an anisotropic Sobolev class $\mathcal{S}(\vec{\alpha}, Q) := \{m(\mathbf{x}) : m(\mathbf{x}) = \sum_{\mathbf{i}=0}^\infty \theta_{\mathbf{i}}\varphi_{\mathbf{i}}(\mathbf{x}), \mathbf{x} \in R;\ \sum_{\mathbf{i}=0}^\infty [1 + \sum_{r=1}^k (1 + \pi i_r)^{2\alpha_r}]\theta_{\mathbf{i}}^2 \leq Q\}$ of differentiable functions. Note that we use notation $\sum_{\mathbf{i}=0}^\infty := \sum_{i_1, \dots, i_k = 0}^\infty$. Then it is established that the optimal (oracle's) minimax rate of the MISE convergence is $n^{-2\alpha/(2\alpha+1)}$ where $\alpha := [\sum_{r=1}^k \alpha_r^{-1}]^{-1}$ is the effective smoothness. For univariate case $k = 1$ not only the rate but a sharp constant is known that is achieved by a data-driven estimator that matches performance of the oracle that knows parameters of the Sobolev class and the nuisance functions $f^X(x)$ and $\sigma(x)$, see Efromovich (1999). Oracle's lower bounds, used as benchmarks for data-driven estimators, are discussed in Barron, Birge and Massart (1999), Galtchouk and Pergamenshchikov (2009ab), and Efromovich (2018) where further references may be found. In short, the theory and methodology of regression estimation for $k = 1$ and a fixed sample size is well developed. For sequential estimation it is known that neither the constant nor the rate can be improved by a sequential plan with stopping time T satisfying $\mathbb{E}\{T\} \le n$. Further, if we restrict our attention to sequential estimators with an assigned MISE and minimal expected stopping time (the Wald problem), then no data-driven estimator can match performance of the oracle, see Efromovich (2007,2018).

As we will see shortly, that negative outcome for the Wald problem changes if we consider an analytic class of regression functions on R with faster decreasing Fourier coefficients,

$$\mathcal{A} := \mathcal{A}(\mathbf{b}, \mathbf{c}, Q)$$

$$:= \{m(\mathbf{x}) : \ m(\mathbf{x}) = \sum_{\mathbf{i}=0}^{\infty} \theta_{\mathbf{i}} \varphi_{\mathbf{i}}(\mathbf{x}), \ \mathbf{x} \in R; \ \sum_{\mathbf{i}=0}^{\infty} [1 + \sum_{r=1}^{k} (1 + i_r)^{2b_r} e^{c_r i_r}] \theta_{\mathbf{i}}^2 \le Q\}. \qquad (1.2)$$

Here $\theta_{\mathbf{i}} := \int_R m(\mathbf{x}) \varphi_{\mathbf{i}}(\mathbf{x}) d\mathbf{x}$, $\mathbf{b} := (b_1, \ldots, b_k)$ and $\mathbf{c} := (c_1, \ldots, c_k)$ are vectors of constants and $\min(c_1, \ldots, c_r) > 0$. Analytic function classes are familiar in statistical literature and well suited for many practical applications, see a discussion in Ibragimov (2001) and Efromovich (1999,2018). In what follows parameters $(\mathbf{b}, \mathbf{c}, Q)$ of the class $\mathcal{A}$ are known to the oracle and unknown to the statistician.

Now let us formulate our main aim. We are interested in estimation of the regression function $m(\mathbf{x})$ in model (1.1) by a sequential estimator $\mathcal{E} := \mathcal{E}(\{\check{m}_r(\mathbf{x}, \mathbf{Z}_1^r), r = 1, 2, \ldots\}, T)$. Here $\check{m}_r(\mathbf{x}, \mathbf{Z}_1^r)$ is a regression estimate based on a sample $\mathbf{Z}_1^r := \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_r, Y_r)\}$ with fixed sample size r, and T is a stopping time. Accordingly when the stopping time is defined then the regression estimate is $\check{m}_T(\mathbf{x}, \mathbf{Z}_1^T)$. Stopping time is a positive integer-valued random variable such that after observing $\mathbf{Z}_1^r$ we make a decision as to whether or not $T = r$. If the decision is $T = r$, then we stop observations and use the regression estimate $\check{m}_r(\mathbf{x}, \mathbf{Z}_1^r)$, otherwise we continue the sampling. More rigorously, let $(\Omega, \mathcal{F}, P)$ be an underlying probability space, $\{\mathbf{Z}_1^r, r = 1, 2, \dots\}$ be a sequence of multivariate random variables on $\{\Omega, \mathcal{F}, P\}$, and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$ be an increasing sequence of sub sigma-fields of $\mathcal{F}$ such that $\mathbf{Z}_1^r$ is $\mathcal{F}_r$-measurable, then the stopping time is a map $T : \Omega \to \{1, 2, \dots\}$ such that $\{T \leq r\} \in \mathcal{F}_r$. Suppose that the oracle knows the underlying class (1.2) of regression functions, the design density $f^{\mathbf{X}}$ and the scale function σ. Then we are considering two classical sequential regression problems when a data-driven estimator tries to match performance of the oracle. (1) Value of the mean stopping time is assigned, and then MISE of a sequential estimator matches the oracle's MISE; (ii) Value of the MISE is assigned, and then the mean stopping time matches the oracle's mean stopping time (the Wald problem). As we will see shortly, for the former problem sequential estimation does not dominate estimation based on a fixed sample size, but for the Wald problem sequential estimation is superior and can match the oracle.

The content of the paper is as follows. Asymptotic theory for oracle-estimators is presented in Section 2. Oracle's lower bounds serve as a benchmark for an estimator, and oracle-estimators inspire data-driven estimators. Sequential estimation with assigned MISE (the Wald problem) is considered in Section 3 where a sharp-minimax two-stage estimator matching performance of the oracle, is introduced. The used methodology mimics the classical Stein's approach proposed for parametric models. Proofs are deferred to Section 4. Conclusions and topics for future research can be found in Section 5. The online Supplementary Materials contain an important environmental example devoted to a new civil engineering technology for reducing greenhouse gas emission. This is a thought-provoking controlled regression with 5 covariates and only $n = 86$ observations. The example and its discussion shed a new light on the first stage of proposed estimator.

In the paper the following notations are used. For the first above-formulated problem of estimation with minimal MISE, given the mean stopping time is bounded by a positive integer n, we are interested in the asymptotic when $n \to \infty$. Set $q := q_n := \lceil 2 + \ln(n+1) \rceil$ and $q' := q'_n := \lceil 2 + \ln(\ln(n+3)) \rceil$ where $\lceil c \rceil$ denotes the smallest integer larger or equal to c. It is assumed that $\sum_{l=n+1}^{n} := 0$, $\sum_{\mathbf{i}=0}^{J} := \sum_{i_1,\dots,i_k=0}^{J}$, $0/0 := 0$, $\vee \mathbf{i} := \max(i_1, \dots, i_r)$, $\wedge \mathbf{i} := \min(i_1, \dots, i_k)$, $\mathbf{ic} := (i_1 c_1, \dots, i_k c_k)$, sup is the supremum over considered function classes, $o_n(1)$'s are generic vanishing sequences in n. $I(\cdot)$ is the indicator, w's are generic positive constants. $\mathbf{Z}_{r+1}^{r+j} := \{(\mathbf{X}_{r+1}, Y_{r+1}), (\mathbf{X}_{r+2}, Y_{r+2}), \dots, (\mathbf{X}_{r+j}, Y_{r+j})\}$ denotes a sequence of iid observations (sample of size j) from $(\mathbf{X}, Y)$. For the Wald problem, ε denotes a given positive real number that is used to bound the MISE, and then we are interested in the asymptotic as $\varepsilon \to 0$. Because notation n for a given sample size is no longer used, with some obvious abuse of notation it is convenient to consider $q := q(\varepsilon) := \lceil 2 + \ln(\varepsilon^{-1} + 1) \rceil$, $q' := q'(\varepsilon) := \lceil 2 + \ln(\varepsilon^{-1} + 1) \rceil$, and $o_\varepsilon(1)$ denotes a generic function that vanishes as $\varepsilon \to 0$.

2 Asymptotic theory of oracle-estimators

Consider regression model (1.1) and the oracle who knows the model, the design density $f^{\mathbf{X}}$ and the scale function σ (the two so-called nuisance functions), and parameters of the function class (1.2). The problem is to understand how well the oracle can solve the two sequential problems formulated in the Introduction, namely minimization of the MISE given assigned expected stopping time and the Wald problem of minimizing the expected stopping time given assigned value of the MISE. While the oracle knows the design density and the scale function, it is of interest to understand how the oracle can deal with a class of these nuisance functions.

Introduce a class of positive and differentiable on $R := [0,1]^k$ k-variate functions

$$\mathcal{N} := \mathcal{N}_1(w_1, w_2, w_3) := \{g: \quad w_1 \leq g(\mathbf{x}) \leq w_2, \ |\frac{\partial^k(g(\mathbf{x}))}{\partial x_1 \dots \partial x_k}| \leq w_3, \ \mathbf{x} \in R\Big\}, \tag{2.1}$$

where w_1, w_2 and w_3 are positive constants whose specific values play no role in the presented below theoretical results. Accordingly, to simplify formulas we may write that both f^X and σ belong to $\mathcal{N}$ keeping in mind that the constants can be different.

Assumption 1. *The design density f^X and the scale function σ are from class (2.1) with possibly different constants (w_1, w_2, w_3), and $\int_R f^{\mathbf{X}}(\mathbf{x})d\mathbf{x} = 1$. Regression error ξ in (1.1) is independent of $\mathbf{X}$ and its distribution F^ξ belongs to a class Ξ of distributions with zero mean and variance bounded by 1.*

The assumption is mild, and let us note that even in a classical univariate regression theory it is traditionally assumed that the nuisance functions are positive and differentiable, see Wassermann (2006) and Efromovich (2018).

The following theorem considers the two above-presented sequential problems and presents sharp lower bounds for oracle-estimators. Recall notation $\mathcal{E}(\{\check{m}_r(\mathbf{x}, \mathbf{Z}_1^r), r = 1, 2, \dots\}, T)$ for a sequential estimator introduced in the Introduction. In the theorem considered sequential regression estimates $\check{m}_r$ and stopping times T are constructed by the oracle, and accordingly the estimators are called oracle-estimators. To highlight that a statistic is constructed by the oracle we use the asterisk, for instance $\check{m}_r^*$, T^* and $\mathcal{E}^*$.

Theorem 1 (Lower Bounds for the Oracle). *Let Assumption 1 hold. The two sequential problems for oracle-estimators with assigned mean stopping time and assigned MISE are explored in turn:*

(i) Introduce a class $\mathcal{S}(n)$ of sequential oracle-estimators $\mathcal{E}^*(\{\tilde{m}_r^*(\mathbf{x}, \mathbf{Z}_1^r), r = 1, 2, \dots\}, T^*)$ with

stopping time T^* satisfying

$$\sup_{f^X\in\mathcal{N},\sigma\in\mathcal{N},F^\xi\in\Xi,m\in\mathcal{A}} \mathbb{E}\{T^*/n\} \le 1. \tag{2.2}$$

Then the following lower bound for minimax MISE of oracle-estimators holds,

$$\inf_{\mathcal{E}^*\in\mathcal{S}(n)} \sup_{f^X\in\mathcal{N},\sigma\in\mathcal{N},F^\xi\in\Xi,m\in\mathcal{A}} \Big[\mathbb{E}\{\int_R (\tilde{m}^*_{T^*}(\mathbf{x},\mathbf{Z}_1^{T^*})-m(\mathbf{x}))^2 d\mathbf{x}\}/\int_R \frac{\sigma^2(\mathbf{x})}{f^{\mathbf{X}}(\mathbf{x})}d\mathbf{x}\Big]$$

$$\ge n^{-1}[\ln(n)]^k[\prod_{r=1}^k c_r^{-1}](1+o_n(1)). \tag{2.3}$$

(ii) Consider the Wald problem of minimizing the expected stopping time given an assigned value of the MISE. Introduce a class $\mathcal{S}'(\varepsilon)$ of sequential oracle-estimators $\mathcal{E}'(\{\hat{m}^*_r(\mathbf{x},\mathbf{Z}_1^r), r=1,2,\ldots\},T^*)$ whose MISE satisfies the upper bound

$$\sup_{f^X\in\mathcal{N},\sigma\in\mathcal{N},F^\xi\in\Xi,m\in\mathcal{A}} \mathbb{E}\{\int_R (\hat{m}^*_{T^*}(\mathbf{x},\mathbf{Z}_1^{T^*})-m(\mathbf{x}))^2 d\mathbf{x}\} \le \epsilon. \tag{2.4}$$

Denote by $n^*(\epsilon)$ a minimal integer n such that $n^{-1}[\ln(n)]^k[\prod_{r=1}^k c_r^{-1}][\int_R \sigma^2(\mathbf{x})/f^{\mathbf{X}}(\mathbf{x})d\mathbf{x}] \le \epsilon$. Then

$$\inf_{\mathcal{E}'\in\mathcal{S}'(\varepsilon)} \sup_{f^X\in\mathcal{N},\sigma\in\mathcal{N},F^\xi\in\Xi,m\in\mathcal{A}} \mathbb{E}\{T^*/n^*(\epsilon)\} \ge (1+o_\epsilon(1)). \tag{2.5}$$

Let us comment on the lower bounds, and then proceed to presenting oracle-estimators that establish sharpness (attainability) of the lower bounds. We begin with the first problem and lower bound (2.3). Note that the MISE is proportional to the integral $\int_R[\sigma^2(\mathbf{x})/f^{\mathbf{X}}(\mathbf{x})]dx$ which shows how the two nuisance functions affect the MISE. This dependence is known for non-sequential regression estimators. Next, let us rewrite the right side of (2.3) as $n^{-1}P^*$, then the constant P^* is traditionally referred to as Pinsker constant to honor the pioneering result of Pinsker (1980) devoted to filtering signals from white Gaussian noise. Pinsker constant describes the effect of an underlying function class $\mathcal{A}$ on the MISE convergence. If we return to definition (1.2) of the class $\mathcal{A} := \mathcal{A}(\mathbf{b}, \mathbf{c}, Q)$, then we can conclude that only vector $\mathbf{c}$ affects the first order of the MISE convergence while $\mathbf{b}$ and Q do not. This is an interesting specific of analytic regression functions because for Sobolev function classes, considered in Pinsker (1980), all constants defining the class affect the Pinsker constant. Now let us look at the lower bound for the Wald problem. It points upon a conjecture that the oracle may use an estimator with a priori chosen fixed sample size $n^*(\epsilon)$ to solve the Wald problem. At first glance this conjecture looks strange, but let us stress that we are dealing with oracle-estimators that know nuisance functions and an underlying class of regression functions. Because the statistician does not have that information, sequential estimation is the only option to solve the classical Wald problem. Several more remarks about Theorem 1 are as follows. The class of distributions Ξ of the regression errors is large and includes both continuous and discrete random variables. Proof of the lower bound (2.3) uses a standard gaussian regression error. It is well known in point estimation theory that gaussian distribution is the least favorable for estimation of the mean, and here we have a similar property for the multivariate regression. Recall that for a gaussian variable its Fisher information is reciprocal of the variance, and this sheds light on factor $\sigma^2(\mathbf{x})$ in the integral on the left side of (2.3). It is reasonable to conjecture that for a fixed distribution of ξ we will see a corresponding Fisher information in the integral. Another thought-provoking comment is as follows. Suppose that the regression is additive, that is $m(\mathbf{x}) = \sum_{r=1}^{k} m_r(x_r)$, and we are interested in estimation of the univariate function $m_1(x_1)$. Additive models are often recommended to remedy the curse of multidimensionality, see Wassermann (2006). Then sharp minimax estimation of a component in additive model becomes dramatically more complicated according to Efromovich (2013,2018). In other words, minimax lower bounds for and adaptive minimax estimators of a multivariate regression and components in an additive regression are different, and this is an interesting specific of a multivariate regression.

Now we are in a position to introduce oracle-estimators that attain the lower bounds of Theorem 1. As a result, we will be able to conclude that the lower bounds are sharp-minimax and can be used as benchmarks for data-driven estimators. We begin with part (i) of Theorem 1 and introduce a minimax oracle-estimator based on a sample $\mathbf{Z}_1^n$. Below we define more general statistics and sequences than needed because later they will be used in Section 3 for construction of sharp-minimax sequential estimators. Also we are using notations and sequences introduced at the end of the Introduction.

Consider a sample $\mathbf{Z}_1^{n_1}$ with shortly defined deterministic $n_1 < n$, and introduce a low-frequency regression estimate

$$\tilde{m}_0(\mathbf{x}, n_1) := \sum_{\vee \mathbf{i} \leq q/(q')^4} \tilde{\theta}_{\mathbf{i}}(n_1)\varphi_{\mathbf{i}}(\mathbf{x}), \quad \tilde{\theta}_{\mathbf{i}}(n_1) := n_1^{-1} \sum_{l=1}^{n_1} Y_l [f^{\mathbf{X}}(\mathbf{X}_l)]^{-1} \varphi_{\mathbf{i}}(\mathbf{X}_l). \tag{2.6}$$

Then this estimate is used to construct a Fourier estimate

$$\hat{\theta}_{\mathbf{i}}(n_1) := [n - n_1]^{-1} \sum_{l=n_1+1}^{n} [Y_l - \tilde{m}_0(\mathbf{X}_l, n_1)][f^{\mathbf{X}}(\mathbf{X}_l)]^{-1} \varphi_{\mathbf{i}}(\mathbf{X}_l). \tag{2.7}$$

The proposed regression estimator is

$$\tilde{m}^*(\mathbf{x}) := \tilde{m}^*(\mathbf{x}, \mathbf{Z}_1^n) := \tilde{m}_0(\mathbf{x}, n_1) + \sum_{\{\mathbf{i}:\, q/(q')^4 < \vee \mathbf{i},\, \mathbf{i} \leq q(1+1/q')/\mathbf{c}\}} \hat{\theta}_{\mathbf{i}}(n_1)\varphi_{\mathbf{i}}(\mathbf{x}). \tag{2.8}$$

Here the sum complements the low-frequency Fourier components of $\tilde{m}_0$ by high-frequency components with indices $i_r \leq q(1 + 1/q')/c_r$, $r = 1, 2, \ldots, k$. Note that $\tilde{m}_0(\mathbf{x}, n_1)$ is based on $\mathbf{Z}_1^{n_1}$ while $\tilde{m}^*(\mathbf{x})$ on $\mathbf{Z}_{n_1+1}^n$ and $\tilde{m}_0(\mathbf{x}, n_1)$. Further, the two terms on the right side of (2.8) are low and high frequency components of the regression estimate, respectively. Finally, note that the estimator is data-driven and based solely on data. Also recall that for the Wald problem the oracle's sample size $n^*(\varepsilon)$ was defined in Theorem 1.

Theorem 2 (Oracle-estimator). *Let Assumption 1 hold and* $n_1 := n_1(n) := \lceil n/(q')^{k+2} \rceil$.
(i) Consider a deterministic stopping time $T = n$ *and a corresponding data-driven estimate* $\tilde{m}(\mathbf{x}, \mathbf{Z}_1^n) := \tilde{m}^*(\mathbf{x}, \mathbf{Z}_1^n)$. *The MISE of this estimate attains the lower bound (2.3) and*

$$\sup_{f^{\mathbf{X}} \in \mathcal{N}, \sigma \in \mathcal{N}, F^{\xi} \in \Xi, m \in \mathcal{A}} \mathbb{E}\{\int_R (\tilde{m}(\mathbf{x}, \mathbf{Z}_1^n) - m(\mathbf{x}))^2 d\mathbf{x}\} / \int_R \frac{\sigma^2(\mathbf{x})}{f^{\mathbf{X}}(\mathbf{x})} d\mathbf{x}$$

$$= n^{-1}[\ln(n)]^k [\prod_{r=1}^{k} c_r^{-1}](1 + o_n(1)). \tag{2.9}$$

(ii) For the Wald problem, consider an oracle-estimator $\tilde{m}^*(\mathbf{x}, \mathbf{Z}_1^{n^*(\varepsilon)})$ *defined in (2.8) and based on a sample with the oracle's deterministic sample size* $n^*(\varepsilon)$. *Then*

$$\sup_{f^{X} \in \mathcal{N}, \sigma \in \mathcal{N}, F^{\xi} \in \Xi, m \in \mathcal{A}} \mathbb{E}\{\int_R (\tilde{m}(\mathbf{x}, \mathbf{Z}_1^{n^*(\varepsilon)}) - m(\mathbf{x}))^2 d\mathbf{x}\} \leq \epsilon(1 + o_{\epsilon}(1)). \tag{2.10}$$

Theorem 2 implies two important conclusions. First, the lower bounds of Theorem 1 are sharp. Second, the oracle can solve the two classical sequential problems without invoking stochastic stopping times. These outcomes shed an interesting light on the two problems. For the first one (bounded mean stopping time and minimal MISE), the oracle suggest to use a data-driven estimator. For the Wald problem the oracle suggests the same regression estimator only with a deterministic sample size defined by nuisance functions and the underlying function class. This is an interesting conclusion for the theory of sequential nonparametric regression. Accordingly, the oracle tells the statistician that only sequential estimation can solve the Wald problem, and then simplicity of a proposed solution becomes paramount. As we will see shortly in Section 3, a two-stage Stein's methodology allows us to solve the problem. Finally, let us make several technical comments. As we will see in the proof of Theorem 2, there is a large choice of sequences $n_1(n) = o_n(1)n$ for which (2.9) holds. Also note while $\tilde{m}^*(\mathbf{x}, \mathbf{Z}_1^{n^*(\varepsilon)})$ is an oracle-estimator, statistics $\tilde{\theta}_{\mathbf{i}}$, $\hat{\theta}_{\mathbf{i}}$ and $\tilde{m}_0(\mathbf{x}, n_1)$ based solely on data. We may conclude that mimicking $n^*(\varepsilon)$ by a data-driven stopping time will be the main issue in the next section devoted to solving the Wald problem.

3 Two-stage sequential estimation with assigned MISE

The aim is to solve the Wald problem and suggest a data-driven sequential estimator that matches performance of the sharp-minimax oracle-estimator of Theorem 2. As we will see shortly, the renown Stein methodology of two-stage sequential estimation is applicable for the considered multivariate heteroscedastic regression.

We continue to use notations and statistics introduced in the previous sections, and let us make a specific remark about regression estimate $\tilde{m}_0(\mathbf{x}, n_1)$ and Fourier estimates $\tilde{\theta}_{\mathbf{i}}(n_1)$ and $\hat{\theta}_{\mathbf{i}}(n_1)$ defined in (2.6) and (2.7). In the proposed two-stage sequential regression estimator these statistics are used twice by both stages but using different observations collected by the corresponding stages. It is convenient to utilize the same notation for these statistics and keep in mind that they are based on different observations.

Now let us describe two stages of sequential estimation. The first one is based on n_0 observations $\mathbf{Z}_1^{n_0}$ from $(\mathbf{X}, Y)$ where $n_0 := n_0(\varepsilon, k)$ is the smallest integer such that $n_0 > \varepsilon^{-1}q^k/[q']^{k+1}$ and $q := q(\varepsilon)$, $q' := q'(\varepsilon)$ are defined at the end of the Introduction. Note that $\sup n_0/n^*(\varepsilon) = o_\varepsilon(1)$, where the supremum is over the same classes as in (2.10) and $n^*(\varepsilon)$ is the oracle's benchmark for the mean stopping time, see part (ii) of Theorem 2. Observations of the first stage are used to calculate size $\tilde{n}$ of an extra sample for the second stage,

$$\tilde{n} := \tilde{n}(\varepsilon, \mathbf{Z}_1^{n_0}) := \lceil \varepsilon^{-1} \tilde{d} \prod_{r=1}^{k} \tilde{J}_r \rceil. \tag{3.1}$$

Here

$$\tilde{J}_r := \min\Big\{J : \ \tilde{F}_r(J) \le \varepsilon/q', \quad J \in \{\lceil q/q' \rceil, \lceil q/q' \rceil + 1, \dots, qq'\}\Big\}, \tag{3.2}$$

$$\tilde{F}_r(J) := \sum_{\mathbf{i}_{-r}=0}^{qq'} \sum_{i_r=J+1}^{qq'} \Big[2[n_0(n_0-1)]^{-1} \sum_{1 \le l_1 < l_2 \le n_0} \frac{Y_{l_1} Y_{l_2} \varphi_{\mathbf{i}}(\mathbf{X}_{l_1}) \varphi_{\mathbf{i}}(\mathbf{X}_{l_2})}{f^{\mathbf{X}}(\mathbf{X}_{l_1}) f^{\mathbf{X}}(\mathbf{X}_{l_2})}\Big] \tag{3.3}$$

is U-statistic used to estimate Sobolev functional $F_r(J) := \sum_{\mathbf{i}_{-r}=0}^{qq'} \sum_{i_r=J+1}^{qq'} \theta_{\mathbf{i}}^2$, and $(k-1)$-dimension vector $\mathbf{i}_{-r}$ is obtained from vector $\mathbf{i}$ by removing its rth element, for instance $\mathbf{i}_{-2} := (i_1, i_3, \dots, i_k)$. Used in (3.1) statistic

$$\tilde{d} := \max(1/q', \min(q', (n_0 - n_1) q^{-2} \sum_{j=qq'+1}^{qq'+q^2} [\hat{\theta}_{j,0,\dots,0}(n_1)]^2)) \tag{3.4}$$

evaluates integral $\int_R [\sigma^2(\mathbf{x})/f^{\mathbf{X}}(\mathbf{x})] d\mathbf{x}$ that we have seen in the lower bound (2.3) of Theorem 1. Recall that this integral describes the effect of two nuisance functions on the MISE. Let us also stress that statistics $\tilde{\theta}_{\mathbf{i}}(n_1)$, $\hat{\theta}_{\mathbf{i}}(n_1)$, $\tilde{m}_0(\mathbf{x}, n_1)$ are based on the sample $\mathbf{Z}_1^{n_0}$, $n_1 = n_1(n_0) = \lceil n_0/(q')^{k+2} \rceil$, $q = q_\varepsilon$, $q' = q'_\varepsilon$, and ε is as small as desired.

The second stage is defined as follows. The stopping time is $T := n_0 + \tilde{n}$, where $\tilde{n}$ is defined in (3.1), and accordingly we get an extra sample $\mathbf{Z}_{n_0+1}^T$ from $(\mathbf{X}, Y)$. In what follows, to use notations of Section 2, we formally set $n := \tilde{n}$, $n_1 := n_1(\varepsilon) := \lceil \varepsilon^{-1} q^k/(q')^{k+2} \rceil$ (note that n_1 is not random), and using the extra sample $\mathbf{Z}_{n_0+1}^T$ and formulas (2.6) and (2.7) calculate Fourier estimates $\tilde{\theta}_{\mathbf{i}}(n_1)$, $\hat{\theta}_{\mathbf{i}}(n_1)$ and the low-frequency regression estimate

$$\check{m}_0(\mathbf{x}, n_1) := \sum_{\vee \mathbf{i} \le q/(q')^4} \tilde{\theta}_{\mathbf{i}}(n_1) \varphi_{\mathbf{i}}(\mathbf{x}). \tag{3.5}$$

The proposed sequential regression estimator, mimicking (2.8), is

$$\hat{m}_T(\mathbf{x}, \mathbf{Z}_1^T) := \check{m}_0(\mathbf{x}, n_1) + \sum_{\{\mathbf{i}:\, q/(q')^4 < \vee \mathbf{i},\, \mathbf{i} \le \tilde{\mathbf{J}}\}} \hat{\theta}_{\mathbf{i}}(n_1) \varphi_{\mathbf{i}}(\mathbf{x}), \tag{3.6}$$

where $\tilde{\mathbf{J}} := (\tilde{J}_1, \dots, \tilde{J}_k)$ and cutoffs $\tilde{J}_r$ are defined in (3.2). Note that the regression estimator (3.6) uses observations $\mathbf{Z}_1^{n_0}$ to calculate $\tilde{n}$ and $\tilde{\mathbf{J}}$, while all other statistics are calculated using the extra observations $\mathbf{Z}_{n_0+1}^T$.

Theorem 3 (Sequential Estimator for the Wald problem). *Let Assumption 1 hold and* $\sup_{F^\xi \in \Xi} \mathbb{E}\{\xi^4\} < \infty$. *Then the two-stage sequential regression estimator* $\hat{m}_T(\mathbf{x}, \mathbf{Z}_1^T)$, *defined in (3.6), is sharp-minimax. Namely, its MISE satisfies*

$$\sup_{f^{\mathbf{X}} \in \mathcal{N}, \sigma \in \mathcal{N}, F^\xi \in \Xi, m \in \mathcal{A}} \mathbb{E}_f\{\int_R (\hat{m}_T(\mathbf{x}, \mathbf{Z}_1^T) - m(\mathbf{x}))^2 d\mathbf{x}\} \leq \varepsilon(1 + o_\varepsilon(1)), \tag{3.7}$$

and its mean stopping time matches the oracle's one,

$$\sup_{f^{\mathbf{X}} \in \mathcal{N}, \sigma \in \mathcal{N}, F^\xi \in \Xi, m \in \mathcal{A}} \mathbb{E}\{T\}/n^*(\varepsilon) = 1 + o_\varepsilon(1). \tag{3.8}$$

This is an interesting theoretical outcome for a multivariate heteroscedastic regression which states that it is possible to suggest an adaptive sequential estimator that solves the Wald problem and matches performance of the oracle. Moreover, the two-stage sequential approach is motivated by and resembles classical pioneering methods of Stein (1945), Wald (1947) and Anscombe (1949,1953) for sequential estimation of parameters. Several possible extensions of the result are discussed in the Conclusion.

4 Proofs

Recall that main notations were introduced at the end of the Introduction. In the proofs we are interested in asymptotic in $n \to \infty$ for Theorems 1 and 2, and in $\varepsilon \to 0$ for Theorem 3.

Proof of Theorem 1. We begin with proving the oracle's lower bound (2.3) given the restriction (2.2) on the mean stopping time. To make the proof shorter we convert it to Efromovich (1989,2000). The left side of (2.3) does not increase if we consider: (i) Specific nuisance functions $f^{\mathbf{X}}(\mathbf{x})$ and $\sigma(\mathbf{x})$ from the class $\mathcal{N}$; (ii) Specific distribution of the regression error ξ from the class Ξ; (iii) A subclass of considered regression functions. Let us consider these suggestions in turn and explain the motivation. (i) We choose $f^{\mathbf{X}}(\mathbf{x}) = I(\mathbf{x} \in R)$ and $\sigma^2(\mathbf{x}) = d^*$. This functions belongs to $\mathcal{N}$ and they are the only constant functions on R that maximize the integral in (2.3); (ii) We choose a standard normal distribution for the regression error ξ. Recall that Fisher information for the mean of a gaussian distribution is reciprocal of the variance, and that for all distributions with bounded variance a gaussian one is the least favorable for estimating the mean; (iii) A subclass of regression functions is chosen in such a way that each Fourier coefficient can be treated independently, and accordingly the subclass should be a parallelepiped in place of ellipsoid $\mathcal{A}$. To define the parallelepiped set $J_{*r} := \lceil (1/c_r)q(1-1/q') \rceil$, $\mathbf{J}_* := (J_{*1}, \ldots, J_{*k})$, and note that for any $b > 0$

$$q^{-b}e^{-q(1-1/q')} = n^{-1}[q^{-b}n^{1/q'}](1+o_n(1)) \quad \text{and} \quad q^b = o_n(1)n^{1/q'}. \tag{4.1}$$

These relations allow us to introduce a parallelepiped

$$\mathcal{R}_n := \{m(\mathbf{x}) : \; m(\mathbf{x}) = \sum_{\mathbf{i}=0}^{\infty} I(\wedge \mathbf{i} \geq q/q', \; \vee(\mathbf{i}/\mathbf{J}_*) \leq 1)\theta_{\mathbf{i}}\varphi_{\mathbf{i}}(\mathbf{x}), \theta_{\mathbf{i}}^2 \leq n^{-1+1/(2q')}\}, \tag{4.2}$$

and according to (4.1) for all sufficiently large n we have $\mathcal{R}_n \subset \mathcal{A}$.

Using the above-described steps (i)-(iii) and the Bessel inequality we get for all sufficiently large n,

$$\inf_{\tilde{m}_T^*} \sup_{(f^X,\sigma)\in\mathcal{N}(d^*), F^{\xi}\in\Xi, m\in\mathcal{A}} \Big[\mathbb{E}\{\int_R (\tilde{m}_T^*(\mathbf{x}, \mathbf{Z}^T) - m(\mathbf{x}))^2 d\mathbf{x}\} / \int_R \frac{\sigma^2(\mathbf{x})}{f^{\mathbf{X}}(\mathbf{x})} d\mathbf{x} \Big]$$

$$\geq \inf_{\tilde{m}_T^*} \sup_{m \in \mathcal{R}_n} E_*\{\int_R (\tilde{m}_T^*(\mathbf{x}, \mathbf{Z}^T) - m(\mathbf{x}))^2 d\mathbf{x}\}/d^*$$

$$\geq \sup_{m \in \mathcal{R}_n} \sum_{\mathbf{i}=0}^{\infty} I(\min(\mathbf{i}) \geq q/q', \vee(\mathbf{i}/\mathbf{J}_* \leq 1) \inf_{\tilde{\theta}_{\mathbf{i}}^*(T)} \mathbb{E}_*\{(\tilde{\theta}_{\mathbf{i}}^*(T) - \theta_{\mathbf{i}})^2\}/d^*. \quad (4.3)$$

Here the expectation $\mathbb{E}_*$ stresses that distribution of the regression function ξ is standard normal, $f^X(\mathbf{x}) = I(\mathbf{x} \in R)$, and $\sigma^2(\mathbf{x}) = d^*$. In other words, on the right side of (4.3) the underlying model is $Y = m(\mathbf{x}) + \sqrt{d^*}\xi_0$ where ξ_0 is standard normal and independent of uniformly distributed on R predictor $\mathbf{X}$, and $m(\mathbf{x})$ belongs to the parallelepiped (4.2). We converted the setting into one considered in Efromovich (1989,2000), recall that parametric Fisher information for $Y' := \theta + \sqrt{d^*}\xi_0$ is $1/d^*$, and then validity of the lower bound (2.3) is established.

Lower bound (2.5) given (2.4) follows from (2.2) and (2.3) using proof by contradiction. Assume that (2.5) does not hold and instead its right side is $(1 - \gamma + o_{\epsilon^{-1}}(1))$ for some positive constant γ. Then this contradicts (2.2)-(2.3). Theorem 1 is proved.

Let us present a technical lemma that will be used shortly in proofs of Theorems 2 and 3.

Lemma 1. *(i) The following relation holds for function class $\mathcal{A} = \mathcal{A}(\mathbf{b}, \mathbf{c}, Q)$ defined in (1.2),*

$$\mathcal{A} \subset \mathcal{A}^* := \{m(\mathbf{x}) : m(\mathbf{x}) = \sum_{\mathbf{i}=0}^{\infty} \theta_{\mathbf{i}}\varphi_{\mathbf{i}}(\mathbf{x}), \ \theta_{\mathbf{i}}^2 \leq [Q/k][\prod_{r=1}^{k}(i_r + 1)^{-b_r} e^{-c_r i_r}]^{1/k}, \ \mathbf{x} \in R\}. \quad (4.4)$$

(ii) Let function $g(\mathbf{x})$ be square integrable on R and $\int_R [\partial^k g(\mathbf{x})/\partial x_1 \dots \partial x_k]^2 d\mathbf{x} < \infty$. Then the following relation is valid for Fourier coefficients of $g(\mathbf{x})$,

$$\sum_{i_1,\dots,i_k=1}^{\infty} \prod_{r=1}^{k} i_r^2 [\int_R g(\mathbf{x})\varphi_{\mathbf{i}}(\mathbf{x}) d\mathbf{x}]^2 = \int_R [\partial^k g(\mathbf{x})/\partial x_1 \dots \partial x_k]^2 d\mathbf{x}/\pi^{2k}. \quad (4.5)$$

Remark 1. There are two useful corollaries of Lemma 1. The first one is that for functions from $\mathcal{A}$ their Fourier coefficients are absolutely summable and the functions are uniformly bounded. The second one is that the same can be said about the ratio $\sigma^2(\mathbf{x})/f^{\mathbf{X}}(\mathbf{x})$ from the class $\mathcal{N}$ of nuisance functions. To see that Fourier coefficients are absolutely summable note that for any set $\mathcal{K} \subset \{0, 1, \dots\}^k$ equality (4.5) and the Cauchy-Schwarz inequality imply

$$\sum_{\mathbf{i}\in\mathcal{K}} |\kappa_{\mathbf{i}}| \leq \Big[\sum_{\mathbf{i}\in\mathcal{K}} [\prod_{r=1}^{k} (1+i_r^2)]^{-1} \sum_{i)\in\mathcal{K}} [\prod_{r=1}^{k} (1+i_r^2)] \kappa_{\mathbf{i}}^2 \Big]^{1/2}. \tag{4.6}$$

Proof of Lemma 1. Inequality (4.4) follows from the classical inequality between geometric and arithmetic means. Verification of (4.5) is more involved. To simplify formulas, set $g(\mathbf{x}) := \sigma^2(\mathbf{x})/f^{\mathbf{X}}(\mathbf{x})$ and note that g is square-integrable on R and the derivative $g'(\mathbf{x}) := \partial^k g(\mathbf{x})/\partial x_1 \dots \partial x_k$ exists and square-integrable on R. Then in place of the cosine tensor-product we use the sine tensor-product, and write using Parseval's identity and integration by parts,

$$\int_R [\partial^k g(\mathbf{x})/\partial x_1 \dots \partial x_k]^2 d\mathbf{x}$$

$$= \sum_{i_1,\dots,i_k=1}^{\infty} \Big[\int_R [\partial^k g(\mathbf{x})/\partial x_1 \dots \partial x_k] \prod_{r=1}^{k} 2^{1/2} \sin(\pi i_r x_r) d\mathbf{x} \Big]^2$$

$$= 2^k \sum_{i_1,\dots,i_k=1}^{\infty} \Big[\int_{[0,1]^{k-1}} \Big((\partial^{k-1} g(\mathbf{x})/\partial x_1 \dots \partial x_{k-1}) \sin(\pi i_k x_k) \Big|_{x_k=0}^{1}$$

$$- \int_0^1 (\pi i_k) \cos(\pi i_k x_k)(\partial^{k-1} g(\mathbf{x})/\partial x_1 \dots \partial x_{k-1}) dx_k \Big) \prod_{r=1}^{k-1} \sin(\pi i_r x_r) dx_t \Big]^2.$$

Using $\sin(0) = \sin(\pi i_k) = 0$ and then repeating the above-made step for $x_{k-1}, \dots, x_1$ we conclude that

$$\sum_{i_1,\dots,i_k=1}^{\infty} \Big[\pi^k \prod_{r=1}^{k} i_r \int_R g(\mathbf{x}) \prod_{r=1}^{k} \varphi_{i_r}(x_r) d\mathbf{x} \Big]^2 = \int_R [\partial^k g(\mathbf{x})/\partial x_1 \dots \partial x_k]^2 d\mathbf{x}. \tag{4.7}$$

Lemma 1 is proved.

Proof of Theorem 2. Some parts of the proof will be used later in the proof of Theorem 3. This explains why several more general relations than needed are presented.

First of all let us check that Fourier estimates introduced in (2.6) and (2.7) are unbiased. Using Assumption 1 we can write,

$$\mathbb{E}\{\tilde{\theta}_{\mathbf{i}}(n_1)\} = \mathbb{E}\{Y[f^X(X)]^{-1}\varphi_{\mathbf{i}}(\mathbf{X})\} = \mathbb{E}\{\frac{[m(\mathbf{X})+\sigma(\mathbf{X})\xi]\varphi_{\mathbf{i}}(\mathbf{X})}{f^{\mathbf{X}}(\mathbf{X})}\}$$
$$= \int_R \frac{f^{\mathbf{X}}(\mathbf{x})m(\mathbf{x})\varphi_{\mathbf{i}}(\mathbf{x})}{f^{\mathbf{X}}(\mathbf{x})}d\mathbf{x} = \int_R m(\mathbf{x})\varphi_{\mathbf{i}}(\mathbf{x})d\mathbf{x} = \theta_{\mathbf{i}}. \tag{4.8}$$

Now we are considering Fourier estimate $\hat{\theta}_{\mathbf{i}}(n_2)$. Using Assumption 1 and that $\mathbf{Z}_{n_1+1}^n$ and $\tilde{m}_0(\mathbf{x}, n_1)$ are independent, we can write for $\vee\mathbf{i} > q/(q')^4$,

$$\mathbb{E}\{\hat{\theta}_{\mathbf{i}}(n_1)\} = \mathbb{E}\{\frac{[m(\mathbf{X}_n)+\sigma(\mathbf{X}_n)\xi_n - \tilde{m}_0(\mathbf{X}_n, n_1)]\varphi_{\mathbf{i}}(\mathbf{X}_n)}{f^{\mathbf{X}}(\mathbf{X}_l)}\}$$
$$= \mathbb{E}\{\frac{m(\mathbf{X}_n)\varphi_{\mathbf{i}}(\mathbf{X}_n)}{f^{\mathbf{X}}(\mathbf{X}_n)}\} - \mathbb{E}\{\int_R \tilde{m}_0(\mathbf{x}, n_1)\varphi_{\mathbf{i}}(\mathbf{x})d\mathbf{x}\} = \theta_{\mathbf{i}}. \tag{4.9}$$

In the last equality we used (2.6) and $\int_R \varphi_{\mathbf{j}}(\mathbf{x})\varphi_{\mathbf{i}}(\mathbf{x})d\mathbf{x} = I(\mathbf{j} = \mathbf{i})$.

Unbiasedness of the two Fourier estimates is established, and now we are exploring their variances (mean squared errors). Using Remark 2 it is plain to realize that

$$\sup \mathbb{E}\{(\tilde{\theta}_{\mathbf{i}}(n_1) - \theta_{\mathbf{i}})^2\} \le wn_1^{-1}. \tag{4.10}$$

Here and in what follows the supremum is over the same function classes as in (2.9), and recall that w's are generic positive constants.

Using this result, the Parseval identity and definition of n_1 we conclude that

$$\sup \mathbb{E}\{\int_R (\tilde{m}_0(\mathbf{x}, n_1) - \sum_{\vee\mathbf{i}\le q/(q')^4} \theta_{\mathbf{i}}\varphi_{\mathbf{i}}(\mathbf{x}))^2\} \le wn_1^{-1}(1 + q/(q')^4)^k = o_n(1)n^{-1}q^k. \tag{4.11}$$

For $\hat{\theta}_{\mathbf{i}}(n_1)$ we need to establish a more accurate upper bound than (4.10), namely we need to get $\int_R[\sigma^2(\mathbf{x})/f^{\mathbf{X}}(\mathbf{x})]dx$ in place of a generic constant w. Write using (4.9), the Cauchy inequality, and a constant $\gamma \in (0, 1)$,

$$(n - n_1)^2\mathbb{E}\{(\hat{\theta}_{\mathbf{i}}(n_1) - \theta_{\mathbf{i}})^2\} = \mathbb{E}\Big\{\Big[\sum_{l=n_1+1}^{n} [\frac{[m(\mathbf{X}_l)+\sigma(\mathbf{X}_l)\xi_l - \tilde{m}_0(\mathbf{X}_l, n_1)]\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} - \theta_{\mathbf{i}}]\Big]^2\Big\}$$

$$= \mathbb{E}\Big\{\Big[\sum_{l=n_1+1}^{n}[\frac{(m(\mathbf{X}_l)+\sigma(\mathbf{X}_l)\xi_l - m(\mathbf{X}_l))\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} + \frac{(m(\mathbf{X}_l)-\tilde{m}_0(\mathbf{X}_l,n_1)\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} - \theta_i]\Big]^2\Big\}$$

$$\le (1+\gamma)(n-n_1)\int_R \frac{\sigma^2(\mathbf{x})\varphi^2_{\mathbf{i}}(\mathbf{x})}{[f^{\mathbf{X}}(\mathbf{x})]^2}d\mathbf{x} + (1+\gamma^{-1})\mathbb{E}\Big\{\Big[\sum_{l=n_1+1}^{n}[\frac{(m(\mathbf{X}_l)-\tilde{m}_0(\mathbf{X}_l,n_1)\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} - \theta_i]\Big]^2\Big\}$$

$$=: (1+\gamma)(n-n_1)A_{1\mathbf{i}} + (1+\gamma^{-1})A_{2\mathbf{i}}. \tag{4.12}$$

To analyze A_1 we note that $\varphi_i^2(x) = 1 + 2^{-1/2}\varphi_{2i}(x)$, and this together with Lemma 1 and Remark 2 allow us to conclude that

$$A_{1\mathbf{i}} = \int_R [\sigma^2(\mathbf{x})/f^{\mathbf{X}}(\mathbf{x})]d\mathbf{x}(1+\rho_{\mathbf{i}}), \text{ where } \sup\sum_{\mathbf{i}=0}^{\infty}|\rho_{\mathbf{i}}| < w < \infty. \tag{4.13}$$

For the term A_2 on the right side of (4.12) we can write for any considered $\mathbf{i}$ satisfying $\vee\mathbf{i} > q/(q')^4$,

$$A_{2\mathbf{i}} = \sum_{l=n_1+1}^{n}\mathbb{E}\{[\frac{(m(\mathbf{X}_l)-\tilde{m}_0(\mathbf{X}_l,n_1)\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} - \theta_i]^2\}$$

$$+2\sum_{n_1+1\le l<r\le n}\mathbb{E}\{[\frac{(m(\mathbf{X}_l)-\tilde{m}_0(\mathbf{X}_l,n_1)\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} - \theta_i][\frac{(m(\mathbf{X}_r)-\tilde{m}_0(\mathbf{X}_r,n_1)\varphi_{\mathbf{i}}(\mathbf{X}_r)}{f^{\mathbf{X}}(\mathbf{X}_r)} - \theta_i]\}$$

$$= (n-n_1)[\mathbb{E}\{[\frac{(m(\mathbf{X}_n)-\tilde{m}_0(\mathbf{X}_n,n_1)\varphi_{\mathbf{i}}(\mathbf{X}_n)}{f^{\mathbf{X}}(\mathbf{X}_n)}]^2\} - \theta_{\mathbf{i}}^2]. \tag{4.14}$$

In the last equality we used $\int_R[m(\mathbf{x}) - \tilde{m}_0(\mathbf{x},n_1)]\varphi_{\mathbf{i}}(\mathbf{x})d\mathbf{x} = \theta_{\mathbf{i}}$. With the help of (4.11) we conclude that $\sup A_{2\mathbf{i}} = o_n(1)nq^k$. Combining the results we get

$$\sup\frac{\mathbb{E}\{(\hat{\theta}_{\mathbf{i}}(n_1)-\theta_{\mathbf{i}})^2\}}{\int_R \frac{\sigma^2(\mathbf{x})}{f^{\mathbf{X}}(\mathbf{x})}d\mathbf{x}} \le n^{-1}(1+o_n(1)+\rho'_{\mathbf{i}}), \quad \text{where} \quad \sum_{\mathbf{i}=0}^{\infty}|\rho'_{\mathbf{i}}| < w < \infty. \tag{4.15}$$

Using (4.15) we conclude that

$$\sup\frac{\sum_{\{\mathbf{i}:\, q/(q')^4<\vee\mathbf{i},\, \vee(\mathbf{ic})\le q(1+1/q')\}}\mathbb{E}\{(\hat{\theta}_{\mathbf{i}}(n_1)-\theta_{\mathbf{i}})^2\}}{\int_R[\sigma^2(\mathbf{x})/f^{\mathbf{X}}(\mathbf{x})]d\mathbf{x}} \le n^{-1}q^k[\prod_{r=1}^{k}c_r^{-1}](1+o_n(1)).$$

This relation, together with $\sup\sum_{\vee(\mathbf{ic})>q(1+1/q')}\theta_{\mathbf{i}}^2 = o_n(1)n^{-1}q^k$, (4.11), (4.15) and the Parseval identity, verify Theorem 2.

Proof of Theorem 3. In what follows we are considering functions and distributions from the classes considered in the supremum of (3.7), and the sup means the supremum over that classes. Recall that general notations and specific sequences are introduced at the end of the Introduction, and in particular $q = q(\varepsilon)$, $q = q'(\varepsilon)$, and $o_\varepsilon(1) \to 0$ as $\varepsilon \to 0$. Also, to simplify formulas it is convenient to introduce a generic function $o^*_\varepsilon(1)$ such that $\sup |o_\varepsilon(1)| = o_\varepsilon(1)$.

We begin with verification of the following upper bound on the mean sample size of the second stage,

$$\sup \mathbb{E}\{\tilde{n}\} \leq n^*(\varepsilon)(1 + o_\varepsilon(1)). \tag{4.16}$$

Set $J^*_r := \lceil (1/c_r) q(1 + 1/q') \rceil$, $r = 1, \ldots, k$ and recall from Section 2 that these are optimal cutoffs of the sharp-minimax oracle-estimator (2.8). Write,

$$\mathbb{E}\{\tilde{n}\} = \mathbb{E}_f\{I(\bigcap_{r=1}^{k}\{\tilde{J}_r/J^*_r \leq 1\})\lceil \varepsilon^{-1}\tilde{d}\prod_{r=1}^{k}\tilde{J}_r\rceil\} + \mathbb{E}_f\{I(\bigcup_{r=1}^{k}\{\tilde{J}_r/J^*_r > 1\})\lceil \varepsilon^{-1}\tilde{d}\prod_{r=1}^{k}\tilde{J}_r\rceil\}$$

$$\leq 1 + \varepsilon^{-1}\mathbb{E}\{\tilde{d}\}\prod_{r=1}^{k}J^*_r + \varepsilon^{-1}\mathbb{E}_f\{I(\bigcup_{r=1}^{k}\{\tilde{J}_r/J^*_r > 1\})\tilde{d}\prod_{r=1}^{k}\tilde{J}_r\} =: 1 + A_1 + \varepsilon^{-1}A_2. \tag{4.17}$$

To continue we recall one familiar inequality for moments of U-statistics due to Lemma 4.1 and Remark 4.1 in Efromovich (2000), and note that this is the place where the extra assumption of a bounded fourth moment of the regression error ξ is used. The inequality is

$$\sup \mathbb{E}\{(\hat{F} - F)^4/(F + Ln_0^{-1})^2\} \leq w n_0^{-2}, \quad w < \infty. \tag{4.18}$$

Here $\hat{F} := \sum_{\mathbf{i}\in\mathcal{K}} 2[n_0(n_0-1)]^{-1}\sum_{1\leq l_1<l_2\leq n_0}\frac{Y_{l_1}Y_{l_2}\varphi_{\mathbf{i}}(\mathbf{X}_{l_1})\varphi_{\mathbf{i}}(\mathbf{X}_{l_2})}{f^{\mathbf{X}}(\mathbf{X}_{l_1})f^{\mathbf{X}}(\mathbf{X}_{l_2})}$, $F := \sum_{\mathbf{i}\in\mathcal{K}}\theta^2_{\mathbf{i}}$, $\mathcal{K} \subset \{0, 1, \ldots\}^k$, and L is cardinality of set $\mathcal{K}$. Further note that used in (3.4) statistic $\hat{\theta}_{j,0,\ldots,0}(n_1)$ estimates a univariate Fourier coefficient $\int_0^1[\int_{[0,1]^{k-1}} m(\mathbf{x})dx_2\ldots dx_k]\varphi_j(x_1)dx_1$. Then using Lemma A.1 in Efromovich (2013) we get a rough inequality

$$\sup \mathbb{E}\{(\tilde{d} - d)^4\} \leq wq^{-4}. \tag{4.19}$$

Now we can return to considering terms A_1 and A_2 on the right side of (4.17). For A_1 we write using (4.19),

$$\mathbb{E}\{\tilde{d}\} = d + \mathbb{E}\{\tilde{d} - d\} = d + o_\varepsilon^*(1).$$

This relation yields that

$$1 + A_1 = n^*(\varepsilon)(1 + o_\varepsilon^*(1)). \tag{4.20}$$

For term A_2 we can write using $\tilde{d} \le q'$ and $\tilde{J}_r \le qq'$,

$$A_2 = \mathbb{E}_f\{I(\bigcup_{r=1}^{k}\{\tilde{J}_r/J_r^* > 1\})\tilde{d}\prod_{r=1}^{k}\tilde{J}_r\} \le q'\sum_{r=1}^{k}\mathbb{E}\{I(\tilde{J}_r > J_r^*)\prod_{s=1}^{k}\tilde{J}_s\}$$

$$\le q'\sum_{r=1}^{k}(qq')^{k-1}\mathbb{E}\{I(\tilde{J}_r > J_r^*)\tilde{J}_r\}. \tag{4.21}$$

Recall that $F_r(J) = \sum_{\mathbf{i}_{-r}=0}^{qq'}\sum_{i_r=J+1}^{qq'}\theta_{\mathbf{i}}^2$, and if $J > J_r^*$ then

$$F_r(J) \le w(1+J)^{|b_r|}e^{-c_r q(1+1/q')/c_r} \le w\epsilon/(q')^2, \quad \text{and} \quad \tilde{F}_r(\tilde{J}_r - 1) > \varepsilon/q'.$$

Accordingly, for all sufficiently small ε and uniformly over the considered function classes we get

$$\mathbb{E}\{I(\tilde{J}_r > J_r^*)\tilde{J}_r\} = \sum_{j=J_r^*+1}^{qq'}\mathbb{E}\{I(\tilde{J}_r = J)J\} \le \sum_{J=J_r^*}^{qq'}(J+1)\mathbb{P}(\tilde{F}_r(J) - F(J) > \varepsilon/(2q')).$$

Using (4.18) and the Chebyshev inequality we continue

$$\mathbb{E}\{I(\tilde{J}_r > J_r^*)\tilde{J}_r\} \le \sum_{j=J_r^*}^{qq'}(J+1)\frac{\mathbb{E}\{(\tilde{F}_r(J) - F(J))^4\}}{(\varepsilon/2q')^4}$$

$$\le w(qq')^2(\varepsilon/2q')^{-4}[q^{|b_r|}\varepsilon^{1+1/q'} + (qq')^k n_0^{-1}]^2 n_0^{-2}$$

$$\le wq^2(q')^6\varepsilon^{-4}[q^{|b_r|}\varepsilon^{1+1/q'} + (qq')^k\varepsilon q^{-k}(q')^{k+1}]^2\varepsilon^2 q^{-2k}(q')^{2k+2}$$

$$= o_\varepsilon^*(1)q^{2-2k+1/2}. \tag{4.22}$$

We conclude that $\varepsilon^{-1}A_2 = o^*_\varepsilon(1)n^*(\varepsilon)$. This, (4.20) and (4.17) verify (4.16).

Now we are verifying that MISE of the proposed sequential regression estimator is at most $\varepsilon(1+o^*_\varepsilon(1))$. For an underlying regression function $m := m(\mathbf{x})$, introduce its specific vector $\mathbf{J}_m := (J_{m1}, \ldots, J_{mr})$ of oracle's cutoffs implying sharp-minimax estimation,

$$J_{mr} := \min\Big(J: \ F_r(J) = \sum_{\mathbf{i}_{-r}=0}^{qq'} \sum_{j=J+1}^{qq'} \theta^2_{\mathbf{i}} \le 2\varepsilon/q', \ J = 0, 1, \ldots, J^*_r\Big), \quad r = 1, \ldots, k. \quad (4.23)$$

The subscript m in J_{mr} emphasizes that this is a special oracle's cutoff based on the underlying regression function $m(\mathbf{x})$ from the class $\mathcal{A}$. Previously introduced cutoffs J^*_r are minimax and use only information about function class $\mathcal{A}$. The latter explains the upper bound J^*_r in (4.23).

We are analyzing the MISE using two vectors of oracle-cutoffs $(J_{m1}, \ldots, J_{mk})$ and $(J^*_1, \ldots, J^*_k)$. For an underlying regression function $m \in \mathcal{A}$ introduce a set of indexes $\mathcal{D}_m := \cap_{r=1}^k \{J_r : J_{mr} \le J_r \le J^*_r\}$, and note that the complementary set is

$$\mathcal{D}^c_m := \cup_{r=1}^k \{J_r : \{J_{mr} \le J_r \le J^*_r\}^c\} = \cup_{r=1}^k \{J_r : \{J_r < J_{mr}\} \cup \{J_r > J^*_r\}\}.$$

We also introduce notation $\tilde{\mathbf{J}} := (\tilde{J}_1, \ldots, \tilde{J}_k)$ for the vector of estimated cutoffs. Using these notations we can write,

$$\mathbb{E}\{\int_R (\hat{m}_T(\mathbf{x}, \mathbf{Z}_1^T) - m(\mathbf{x}))^2 d\mathbf{x}\}$$
$$= \mathbb{E}\{I(\tilde{\mathbf{J}} \in \mathcal{D}_m) \int_R (\hat{m}_T(\mathbf{x}, \mathbf{Z}_1^T) - m(\mathbf{x}))^2 d\mathbf{x}\} + \mathbb{E}\{I(\tilde{\mathbf{J}} \in \mathcal{D}^c_m) \int_R (\hat{m}_T(\mathbf{x}, \mathbf{Z}_1^T) - m(\mathbf{x}))^2 d\mathbf{x}\}$$
$$=: U_1 + U_2. \quad (4.24)$$

Term U_1 is the oracle's MISE and $\sup U_1 \le \varepsilon(1+o_\varepsilon(1))$. Accordingly, we need to show that

$$\sup U_2 = \varepsilon o_\varepsilon(1). \quad (4.25)$$

Using the above-presented formula for $\mathcal{D}^c_m$ we get,

$$U_2 \le \sum_{r=1}^k \mathbb{E}\{I(\tilde{J}_r < J_{mr}) \int_R (\hat{m}_T(\mathbf{x}, \mathbf{Z}_1^T) - m(\mathbf{x}))^2 d\mathbf{x}\}$$

$$+\sum_{r=1}^{k}\mathbb{E}\{I(\tilde{J}_r > J_r^*)\int_R(\hat{m}_T(\mathbf{x},\mathbf{Z}_1^T)-m(\mathbf{x}))^2d\mathbf{x}\} =: \sum_{r=1}^{k}U_{21r}+\sum_{r=1}^{k}U_{22r}. \tag{4.26}$$

Note that terms in the first sum use cutoffs smaller than recommended by the oracle for an underlying regression function m, and this may lead to larger bias. Terms in the second sum use cutoffs larger than suggested by the minimax oracle, and this may lead to larger variance. We are considering these two cases in turn. Write using the Parseval identity,

$$U_{21r} = E\{I(\tilde{J}_r < J_{mr})\int_R(\hat{m}_T(\mathbf{x},\mathbf{Z}_1^T)-m(\mathbf{x}))^2d\mathbf{x}\}$$

$$\le \mathbb{E}\{I(\tilde{J}_r < J_{mr})\sum_{\mathbf{i}_{-r}=0}^{qq'}\sum_{i_r=0}^{J_{mr}-1}(\check{\theta}_{\mathbf{i}}-\theta_{\mathbf{i}})^2\}+\mathbb{E}\{I(\tilde{J}_r < J_{mr})\sum_{\mathbf{i}_{-r}=0}^{qq'}\sum_{i_r\ge\tilde{J}_r}\theta_{\mathbf{i}}^2\}+\sum_{\vee\mathbf{i}>qq'}\theta_{\mathbf{i}}^2$$

$$=: V_1+V_2+V_3. \tag{4.27}$$

Here $\check{\theta}_{\mathbf{i}} := \tilde{\theta}_{\mathbf{i}}I(\vee\mathbf{i} \le q/(q')^4)+\hat{\theta}_{\mathbf{i}}I(\vee\mathbf{i} > q/(q')^4)$. For $m \in \mathcal{A}$ we have $\sup V_3 = o_\varepsilon(1)\varepsilon$. Accordingly, V_3 is sufficiently small, and we are evaluating V_1 and V_2 in turn. For V_1 we use Cauchy-Schwarz inequality and get

$$V_1 \le \Big[\mathbb{P}(\tilde{J}_r < J_{mr})\Big]^{1/2}\Big[\mathbb{E}\Big\{[\sum_{\mathbf{i}_{-r}=0}^{qq'}\sum_{i_r=0}^{J_{mr}-1}(\check{\theta}_{\mathbf{i}}-\theta_{\mathbf{i}})^2]^2\Big\}\Big]^{1/2}. \tag{4.28}$$

For the probability term we note that if $\tilde{J}_r < J_{mr}$ then $\tilde{F}_r(\tilde{J}) \le \varepsilon/q'$ and $F_r(\tilde{J}) > 2\varepsilon/q'$. This remark, Chebyshev inequality, $\tilde{J}_r \ge \lceil q/q' \rceil$ almost sure, and (4.18) yield

$$\mathbb{P}(\tilde{J}_r < J_{mr}) = \sum_{J=\lceil q/q'\rceil}^{J_{mr}-1}\mathbb{P}(\tilde{J}_r = J) \le \sum_{J=\lceil q/q'\rceil}^{J_{mr}-1}\mathbb{P}(\tilde{F}_r(J)-F_r(J) > (1/2)F_r(J), F_r(J) > 2\varepsilon/q')$$

$$\le w\sum_{J=\lceil q/q'_\varepsilon\rceil}^{J_{mr}-1}\frac{n_0^{-2}(F_r(J)+(qq')^kn_0^{-1})^2}{F_r^4(J)}I(F_r(J) > 2\varepsilon/q') \le o_\varepsilon^*(1)q^{-2k}(q')^w. \tag{4.29}$$

Next step is to evaluate the expectation on the right side of (4.28). Recall that Fourier estimate $\tilde{\theta}_{\mathbf{i}}(n_1)$ is not sequential and based on a sample of size n_1. This yields $\mathbb{E}\{(\tilde{\theta}_{\mathbf{i}}(n_1)-\theta_{\mathbf{i}})^4\} < wn_1^{-2} = o_\varepsilon^*(1)\varepsilon^2q^{-2k}(q')^{2k+5}$. To evaluate moments of $\hat{\theta}_{\mathbf{i}}$ we note that this Fourier estimator is based on a sample with random size $\tilde{n}-n_1$ defined by the first stage. Accordingly, set $n_* := \lceil \varepsilon^{-1}q^k/(q')^{k+1}\rceil$, $n^* := \lceil \varepsilon^{-1}q^k(q')^{k+1}\rceil$, note that for all sufficiently small ε we have $n_* \leq \tilde{n} \leq n^*$ almost sure, and recall that we are considering asymptotic in $\varepsilon \to 0$ and hence $n_1 = o^*_\varepsilon(1)n_*$ allows us to assume that $n_1 < n_*$. Using these remarks we can write,

$$\mathbb{E}\{(\hat{\theta}_{\mathbf{i}}-\theta_{\mathbf{i}})^4\} = \mathbb{E}\{[(\tilde{n}-n_1)^{-1}\sum_{l=n_0+n_1+1}^{n_0+\tilde{n}} \frac{(Y_l-\check{m}_0(\mathbf{X}_l,n_1))\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} - \theta_{\mathbf{i}}]^4\}$$

$$= \sum_{s=n_*}^{n^*} \mathbb{E}\{I(\tilde{n}=s)[(s-n_1)^{-1}\sum_{l=n_0+n_1+1}^{n_0+s} \frac{(Y_l-\check{m}_0(\mathbf{X}_l,n_1))\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} - \theta_{\mathbf{i}}]^4\}$$

$$= \sum_{s=n_*}^{n^*} \mathbb{P}(\tilde{n}=s)\mathbb{E}\{[(s-n_1)^{-1}\sum_{l=n_0+n_1+1}^{n_0+s} \frac{(Y_l-\check{m}_0(\mathbf{X}_l,n_1))\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} - \theta_{\mathbf{i}}]^4\}. \tag{4.30}$$

For the expectation on the right side of (4.30) we can use inequality (1.3.50) in Efromovich (2018) and get for $\vee\mathbf{i} > q/(q')^4$ (also recall a similar calculation in the proof of Theorem 2)

$$\mathbb{E}\{[(s-n_1)^{-1}\sum_{l=n_0+n_1+1}^{n_0+s} \frac{(Y_l-\check{m}_0(\mathbf{X}_l,n_1))\varphi_{\mathbf{i}}(\mathbf{X}_l)}{f^{\mathbf{X}}(\mathbf{X}_l)} - \theta_{\mathbf{i}}]^4\} \leq w(s-n_1)^{-2}. \tag{4.31}$$

Here we used the above-mentioned $n_1 = o^*_\varepsilon(1)n_*$ and that in (4.31) statistic $\check{m}_0(\mathbf{x},n_1)$ is independent of $(\mathbf{X}_l,Y_l)$. Using this inequality in (4.30) we conclude that

$$\mathbb{E}\{(\hat{\theta}_{\mathbf{i}}-\theta_{\mathbf{i}})^4\} \leq wn_*^{-2} = o^*_\varepsilon(1)\varepsilon^2 q^{-2k}(q')^{2k+3}. \tag{4.32}$$

Using (4.32) in (4.28) we conclude that $V_1 = o^*_\varepsilon(1)\varepsilon$.

To evaluate V_2 on the right side of (4.27) we note that this is the main term when $\tilde{J}_r < J_{mr}$ because a smaller cutoff increases squared bias and decreases variance of a regression estimate. To evaluate the increased squared bias and to simplify formulas, set $Z_r := \sum_{\mathbf{i}_{-r}=0}^{qq'}\sum_{j\geq\tilde{J}_r}\theta^2_{\mathbf{i}}/(\varepsilon/q')$. Note that Z_r is a random variable (function of $\tilde{J}_r$), and according to definition of $\tilde{J}_r$ we have $\tilde{F}_r(\tilde{J}_r) \leq \varepsilon/q'$, and also for $m \in \mathcal{A}$ and any positive constant c_* we have $\sum_{\vee\mathbf{i}>qq'}\theta_{\mathbf{i}}^2 = o_\varepsilon^*(1)\varepsilon q^{-c_*}$. Using this remark and the Chebyshev inequality we conclude that for a constant $z \geq 2$ and all sufficiently small ε we have

$$\mathbb{P}(Z_r \geq z, \tilde{J}_r < J_{mr}) \leq \sum_{J=\lceil q/q'\rceil}^{J_{mr}-1} \mathbb{P}(F_r(J) - \tilde{F}_r(J) > (z/3)F(J), F(J) > (z/2)\varepsilon/q')$$

$$\leq w \sum_{J=\lceil q/q'\rceil}^{J_{mr}-1} \frac{\mathbb{E}\{(F_r(J) - \tilde{F}_r(J))^4\}}{z^4[F_r(J)]^4}\} I(F(J) > (z/2)\varepsilon/q') = o_\varepsilon^*(1)q^{-1/2}z^{-2}. \tag{4.33}$$

Using this inequality and a classical inequality $\mathbb{E}\{\eta I(\eta \geq 2)\} \leq \sum_{r=1}^{\infty}\mathbb{P}(\eta \geq r)$ we can finish evaluation of V_2,

$$V_2 = \mathbb{E}_f\{I(\tilde{J}_r < J_{mr}) \sum_{\mathbf{i}_{-r}=0}^{qq'} \sum_{i_r \geq \tilde{J}_r} \theta_{\mathbf{i}}^2\} = (\varepsilon/q')\mathbb{E}\{I(\tilde{J}_r < J_{mr})Z_r\}$$

$$\leq 2\varepsilon/q' + (\varepsilon/q')\sum_{z=1}^{\infty}\mathbb{P}(Z_r \geq z, \tilde{J}_r < J_{mr}) \leq 2\varepsilon/q' + o_\varepsilon^*(1)\varepsilon\sum_{z=1}^{\infty} z^{-2} = o_\varepsilon^*(1)\varepsilon. \tag{4.34}$$

Using the already evaluated terms $V_1 + V_3 = o_\varepsilon^*(1)\varepsilon$ together with (4.34) in (4.27) we conclude that $U_{21r} = o_\varepsilon^*(1)\varepsilon$. Accordingly, $\sum_{r=1}^{k} U_{21r} = o_\varepsilon^*(1)\varepsilon$.

Now we are evaluating a term U_{22r} in (4.26). Recall that J_r^* is the oracle's minimax cutoff. Accordingly, the case $\tilde{J}_r > J_{r^*}$ increases the variance part of the MISE while its squared bias part remains sufficiently small. To realize that, we may write using the Parseval's identity (compare with (4.27))

$$U_{22r} = \mathbb{E}\{I(\tilde{J}_r > J_r^*)\int_R (\hat{m}_T(\mathbf{x}, \mathbf{Z}_1^T) - m(\mathbf{x}))^2 d\mathbf{x}$$

$$\leq \mathbb{E}\{I(\tilde{J}_r > J_r^*)\sum_{\mathbf{i}_{-r}=0}^{qq'}\sum_{i_r=0}^{\tilde{J}_r}(\check{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}})^2\} + \mathbb{E}\{I(\tilde{J}_r > J_r^*)\sum_{\mathbf{i}_{-r}=0}^{qq'}\sum_{i_r \geq J_r^*}\theta_{\mathbf{i}}^2\} + \sum_{\vee\mathbf{i}>qq'}\theta_{\mathbf{i}}^2$$

$$=: V_1' + V_2' + V_3'. \tag{4.35}$$

Here, similarly to (4.27), we use notation $\check{\theta}_{\mathbf{i}} := \tilde{\theta}_{\mathbf{i}} I(\vee\mathbf{i} \leq q/(q')^4) + \hat{\theta}_{\mathbf{i}} I(\vee\mathbf{i} > q/(q')^4)$. Definition of J_r^* implies that for $m \in \mathcal{A}$ we have $V_2' + V_3' = o_\varepsilon^*(1)\varepsilon$. The term V_1' is evaluated similarly to the term V_1 in (4.27) and we get $V_1' = o_\varepsilon^*(1)\varepsilon$. We have shown that all considered terms are $o_\varepsilon^*(1)\varepsilon$, and there are only a finite number of these terms.

Theorem 3 is proved,

5 Conclusion

The developed theory shows that asymptotically a sequential estimation of analytic multivariate regression functions with assigned MISE and minimax mean stopping time is possible. The proposed data-driven sequential estimator matches performance of the oracles that knows smoothness of an estimated multivariate regression and all nuisance functions, and accordingly the estimator can be referred to as adaptive. The asymptotic theory sheds a new light on the potential of sequential estimation because the only theoretical result known so far has been that no minimax adaptive sequential estimation is possible for differentiable regression functions. Another important result is that, similarly to classical parametric models, a two-stage sequential methodology solves the problem.

Let us mention several interesting open problems for future research. First, it is of interest to understand the effect of missing data on sequential estimation. Missing data are typical in regression problems. Some theoretical results are known for fixed sample sizes, see Efromovich (2018). In particular, it is known that different remedies should be used for missing predictors and missing responses. It is understood that for sequential estimation an underlying missing mechanism must be evaluated and taken into account, and the latter is a challenging problem on its own. Second, it is of interested to apply the developed sequential methodology to other familiar sequential problems like change point discussed in Baron (2001) and Schmegner and Baron (2004), confidence bands and hypotheses testing discussed in Wald (1947). Third, sequential estimation of the scale function $\sigma(\mathbf{x})$ and distribution of regression error ξ is another practically important and theoretically challenging problem. Fourth, as it follows from the discussion in Section 2, it is of interest to develop theory of sequential estimation for additive regression models. Fifth, in Galtchouk and Pergamenshchikov (2009ab) a practically important type of heteroscedastic regression $Y = m(x) + \sigma(x, m)\xi$ is considered where the scale may depend on both x and the regression function $m(x)$. It will be of interest to consider a multivariate regression of this type and then explore sequential estimation of the regression and scale.

Finally, a challenging and urgent open problem is the practically important case of a small sample for a first stage. First stage is based on a priori chosen sample size, and this creates a possibility of no feasible estimation due to curse of multidimensionality and a sample size which may be too small for an underlying regression (of course, using multiple stages is a possible remedy but still simplicity of just two stages is appealing). Supplementary Materials shed light on challenges of first stage and point upon possible modifications of proposed asymptotically optimal estimates via exploring a thought-provoking environmental example of a multivariate regression with 5 covariates and sample size $n = 86$. The practical example also highlights a connection between studied nonparametric multivariate regression and nonparametric functional regression where predictor is a process. Accordingly, sequential functional regression is another important topic for future research.

Acknowledgements

A discussion with Nitis Mukhopadhyay and valuable comments of the Associate Editor and reviewers are gratefully appreciated. The research is supported in part by NSF Grant DMS-1915845 and Grants from CAS and BIFAR.

References

Anscombe, F. J. (1949). Large-Sample Theory of Sequential Estimation, *Biometrika* 36: 455-458.

Anscombe, F. (1953). Sequential estimation, *Journal of Royal Statistical Society* 15: 1-29.

Aoshima, M and Yata K. (2011). Two-Stage Procedures for High-Dimensional Data. *Sequential Analysis* 30: 356-399.

Barron, A., Birge, L. and Massart, P. (1999). Risk Bounds for Model Selection via Penalization, *Probability Theory and Related Fields* 113: 301-413.

Baron, M. (2001). Bayes Stopping Rules in a Change-Point Model with a Random Hazard Rate, *Sequential Analysis* 20: 147-163.

Efromovich, S. (1989). On Sequential Nonparametric Estimation of Density, *Theory of Probability and Applications* 34: 263-276.

Efromovich, S. (1995). Sequential Nonparametric Estimation with Assigned Risk, *Annals of Statistics* 23: 1376-1392.

Efromovich, S. (1999). *Nonparametric Curve Estimation*, New York: Springer.

Efromovic, S. (2000). On Sharp Adaptive Estimation of Multivariate Curves, *Mathematical Methods of Statistics* 9: 117-139.

Efromovich, S. (2007). Sequential Design and Estimation in Heteroscedastic Nonparametric Regression, *Sequential Analysis* 26: 3-25.

Efromovich, S. (2013). Nonparametric Regression with the Scale Depending on Auxiliary Variable, *Annals of Statistics* 41: 1542-1568.

Efromovich, S. (2018). *Missing and Modified Data in Nonparameteric Estimation*, Boca Raton: Chapman & Hall.

Hoffmann, M. and Lepski, O. (2002). Random Rates in Anisotropic Regression, *Annals of Statistics*

30: 325-396.

Ibragimov, I. (2001). Estimation of Analytic Functions, *Lecture Notes-Monograph Series* 36: 359-383.

Ghosh, M., Mukhopadhyay, N., and Sen, P. K. (1997). *Sequential Estimation*, New York: Wiley.

Ghosh, B. K. and Sen, P. K., eds. (1991). *Handbook of Sequential Analysis*, New York: Marcel Dekker.

Galtchouk, L.I. and Pergamenshchikov, S. M. (2009a). Sharp Non Asymptotic Oracle Inequalities for Nonparametric Heteroscedastic Regression Models, *Journal of Nonparametric Statistics* 21: 1 - 16.

Galtchouk, L. and Pergamenshchikov, S.(2009b). Adaptive Asymptotically Efficient Estimation in Heteroscedastic Nonparametric Regression, *Journal of Korean Statistical Society* 38: 305-322.

Mukhopadhyay, N. (1997). An Overview of Sequential Nonparametric Density Estimation. *Nonlinear Analysis, Theory, Methods and Applications* 30: 4395–4402.

Mukhopadhyay, N. and Zacks, S. (2018). Modified Two-Stage and Purely Sequential Estimation of the Variance in a Normal Distribution with Illustrations Using Horticultural Data, *Journal of Statistical Theory and Practice* 12: 111-135.

Mukhopadhyay, N. (2019). Two-Stage and Multi-Stage Estimation, *The Exponential Distribution*, 429-452, New York: Taylor,

Pinsker, M.S. (1980). Optimal Filtering Square Integrable Signals in White Gaussian Noise, *Problems Information Transmission* 16: 52-68.

Schmegner, C. and Baron, M. (2004). Principles of Optimal Sequential Planning, *Sequential Analysis* 23: 11-32.

Stein, C. (1945). Two-Sample Test for a Linear Hypothesis Whose Power is Independent of the Variance, *Annals of Mathematical Statistics* 16: 243-258.

Stein, C. and Wald, A. (1947). Sequential Confidence Intervals for the Mean of a Normal Distribution with Known Variance, *Annals of Mathematical Statistics* 18: 427-433.

Wald, A. (1947). *Sequential Analysis*, New York: John Wiley.

Wassermann, L. (2006). *All of Nonparametric Statistic*, New York: Springer.