Sequential Model Selection Method for Nonparametric Autoregression

Ouerdia Arkoun, Jean-Yves Brua, Serguei Pergamenshchikov

Sequential Model Selection Method for Nonparametric Autoregression *

Ouerdia Arkoun †, Jean-Yves Brua ‡and
Serguei Pergamenshchikov§

Abstract

In this paper for the first time the nonparametric autoregression estimation problem for the quadratic risks is considered. To this end we develop a new adaptive sequential model selection method based on the efficient sequential kernel estimators proposed by Arkoun and Pergamenshchikov (2016). Moreover, we develop a new analytical tool for general regression models to obtain the non asymptotic sharp oracle inequalities for both usual quadratic and robust quadratic risks. Then, we show that the constructed sequential model selection procedure is optimal in the sense of oracle inequalities.

MSC: primary 62G08, secondary 62G05

Keywords: Non-parametric estimation; Non parametric autoregresion; Non-asymptotic estimation; Robust risk; Model selection; Sharp oracle inequalities.

*This work was supported by the RSF grant 17-11-01049 (National Research Tomsk State University).

†Sup'Biotech, Laboratoire BIRL, 66 Rue Guy Moquet, 94800 Villejuif, France and Laboratoire de Mathématiques Raphael Salem, Normandie Université, UMR 6085 CNRS-Université de Rouen, France, e-mail: ouerdia.arkoun@gmail.com

‡Laboratoire de Mathématiques Raphael Salem, Normandie Université, UMR 6085 CNRS- Université de Rouen, France, e-mail : jean-yves.brua@univ-rouen.fr

§Laboratoire de Mathématiques Raphael Salem, UMR 6085 CNRS- Université de Rouen Normandie, France and International Laboratory of Statistics of Stochastic Processes and Quantitative Finance, National Research Tomsk State University, e-mail: Serge.Pergamenshchikov@univ-rouen.fr

1 Introduction

One of the standard linear models in general theory of time series is the autoregressive model (see, for example, [1] and the references therein). Natural extensions for such models are nonparametric autoregressive models which are defined by

$$y_k = S(x_k)y_{k-1} + \xi_k \quad \text{and} \quad x_k = a + \frac{k(b-a)}{n}\,, \tag{1.1}$$

where $S(\cdot) \in \mathbf{L}_2[a,b]$ is unknown function, $a < b$ are fixed known constants, $1 \le k \le n$, the initial value y_0 is a constant and the noise $(\xi_k)_{k\ge1}$ is i.i.d. sequence of unobservable random variables with $\mathbf{E}\xi_1 = 0$ and $\mathbf{E}\xi_1^2 = 1$.

The problem is to estimate the function $S(\cdot)$ on the basis of the observations $(y_k)_{1\le k\le n}$ under the condition that the noise distribution is unknown.

It should be noted that the varying coefficient principle is well known in the regression analysis. It permits the use of more complex forms for regression coefficients and, therefore, the models constructed via this method are more adequate for applications (see, for example, [9], [23]). In this paper we consider the varying coefficient autoregressive models (1.1). There is a number of papers which consider these models such as [7], [8] and [6]. In all these papers, the authors propose some asymptotic (as $n \to \infty$) methods for different identification studies without considering optimal estimation issues. To our knowledge, for the first time, the minimax estimation problem for the model (1.1) has been treated in [3] and [24] in the nonadaptive case, i.e. for the known regularity of the function S. Then, in [2] it is proposed to use the sequential analysis method for the adaptive pointwise estimation problem in the case where the unknown Hölder regularity is less than one, i.e when the function S is not differentiable. Also it should be noted (see, [2]) that for the model (1.1), the adaptive pointwise estimation is possible only in the sequential analysis framework. That is why we study sequential estimation methods for the smooth function S. In this paper we consider the quadratic risk defined as

$$\mathcal{R}_p(\widehat{S}_n, S) \;=\; \mathbf{E}_{p,S}\|\widehat{S}_n - S\|^2\,, \quad \|S\|^2 = \int_a^b S^2(x)\mathrm{d}x\,, \tag{1.2}$$

where $\widehat{S}_n$ is an estimator of S based on observations $(y_k)_{1\le k\le n}$ and $\mathbf{E}_{p,S}$ is the expectation with respect to the distribution law $\mathbf{P}_{p,S}$ of the process $(y_k)_{1\le k\le n}$ given the distribution density p and the coefficient S. Moreover, taking into account that the distribution p is unknown, we use the robust nonparametric estimation approach proposed in [12]. To this end we set the robust risk as

$$\mathcal{R}^*(\widehat{S}_n, S) \;=\; \sup_{p\in\mathcal{P}} \mathcal{R}_p(\widehat{S}_n, S)\,, \tag{1.3}$$

where $\mathcal{P}$ is a family of the distributions defined in Section 2.

In order to estimate the function S in model (1.6) we make use of the estimator family $(\widehat{S}_\lambda\,, \lambda \in \Lambda)$, where $\widehat{S}_\lambda$ is a weighted least square estimator with the Pinsker weights. For this family, similarly to [14], we construct a special selection rule, i.e. a random variable $\widehat{\lambda}$ with values in Λ, for which we define the selection estimator as $\widehat{S}_* = \widehat{S}_{\widehat{\lambda}}$. Our goal in this paper is to show the non asymptotic sharp oracle inequality for the robust risks (1.3), i.e. to show that for any $\check{\varrho} > 0$ and $n \geq 1$

$$\mathcal{R}^*(\widehat{S}_*, S) \;\leq\; (1+\check{\varrho}) \min_{\lambda\in\Lambda} \mathcal{R}^*(\widehat{S}_\lambda, S) + \frac{\mathcal{B}_n}{\check{\varrho} n}\,, \tag{1.4}$$

where $\mathcal{B}_n$ is a rest term such that for any $\check{\delta} > 0$,

$$\lim_{n\to\infty} \frac{\mathcal{B}_n}{n^{\check{\delta}}} = 0\,. \tag{1.5}$$

In this case the estimator $\widehat{S}_*$ is called optimal in the oracle inequality sense. In this paper, in order to obtain this inequality for model (1.1) we develop a new model selection method based on the truncated sequential procedures developed in [4] for the pointwise efficient estimation. Then we use the non asymptotic analysis tool proposed in [14] based on the non-asymptotic studies from [5] for a family of least-squares estimators and extended in [10] to some other estimator families. To this end we use the approach proposed in [16], i.e. we pass to a discrete time regression model by making use of the truncated sequential procedure introduced in [4]. To this end, at any point $(z_l)_{1\leq l\leq d}$ of a partition of the interval $[a,b]$, we define a sequential procedure (τ_l, S_l^*) with a stopping rule τ_l and an estimator S_l^*. For $Y_l = S_l^*$ with $1 \leq l \leq d$, we come to the regression equation on some set $\Gamma \subseteq \Omega$:

$$Y_l = S(z_l) + \zeta_l\,, \quad 1 \leq l \leq d\,. \tag{1.6}$$

Here, in contrast with the classical regression model, the noise sequence $(\zeta_l)_{1\leq l\leq d}$ has a complex structure, namely,

$$\zeta_l = \xi_l^* + \varpi_l\,, \tag{1.7}$$

where $(\xi_l^*)_{1\leq l\leq d}$ is a "main noise" sequence of uncorrelated random variables and $(\varpi_l)_{1\leq l\leq n}$ is a sequence of bounded random variables.

We will use the oracle inequality (1.4) to prove the asymptotic efficiency for the proposed procedure, using the same method as it is been used in [15]. The asymptotic efficiency means that the procedure provides the optimal convergence rate and the asymptotically minimal rate normalized risk which

coincides with the Pinsker constant. It should be emphasized that only sharp oracle inequalities of type (1.4) allow to synthesis efficiency property in the adaptive setting.

The paper is organized as follows: In Section 2 we state the main conditions for the model (1.1). In Section 3 we describe the passage to the regression scheme. In Section 4 we describe the sequential model selection procedure. In Section 5 we announce the main results. In Section 6 we study the properties of the obtained regression model (1.6). In Section 7 we prove all basic results. In Appendix A we give all the auxiliary and technical tools.

2 Main Conditions

As in [4] we assume that in the model (1.1) the i.i.d. random variables $(\xi_k)_{k\geq 1}$ have a density p (with respect to Lebesgue measure) from the functional class $\mathcal{P}$ defined as

$$\mathcal{P} := \Big\{ p \geq 0 \, : \, \int_{-\infty}^{+\infty} p(x)\,\mathrm{d}x = 1\,, \quad \int_{-\infty}^{+\infty} x\,p(x)\,\mathrm{d}x = 0\,, \\ \int_{-\infty}^{+\infty} x^2\,p(x)\,\mathrm{d}x = 1 \quad \text{and} \quad \sup_{k\geq 1} \frac{\int_{-\infty}^{+\infty} |x|^{2k}\,p(x)\,\mathrm{d}x}{\varsigma^k (2k-1)!!} \leq 1 \Big\}\,, \tag{2.1}$$

where $\varsigma \geq 1$ is some fixed parameter, which may be a function of the number observation n, i.e. $\varsigma = \varsigma(n)$, such that for any $\check{\delta} > 0$

$$\lim_{n\to\infty} \frac{\varsigma(n)}{n^{\check{\delta}}} = 0\,. \tag{2.2}$$

Note that the $(0,1)$-Gaussian density belongs to $\mathcal{P}$. In the sequel we denote this density by p_0. It is clear that for any $q > 0$

$$\mathbf{m}_q^* = \sup_{p\in\mathcal{P}} \mathbf{E}_p\,|\xi_1|^q < \infty\,, \tag{2.3}$$

where $\mathbf{E}_p$ is the expectation with respect to the density p from $\mathcal{P}$. To obtain the stable (uniformly with respect to the function S) model (1.1), we assume that for some fixed $0 < \epsilon < 1$ and $L > 0$ the unknown function S belongs to the ε - *stability set* introduced in [4] as

$$\Theta_{\epsilon,L} = \Big\{ S \in \mathbf{C}_1([a,b],\mathbb{R}) : |S|_* \leq 1-\epsilon \quad \text{and} \quad |\dot{S}|_* \leq L \Big\}\,, \tag{2.4}$$

where $\mathbf{C}_1([a,b],\mathbb{R})$ is the Banach space of continuously differentiable $[a,b] \to \mathbb{R}$ functions and $|S|_* = \sup_{a\leq x\leq b} |S(x)|$.

3 Passage to a discrete time regression model

We will use as a basic procedure the pointwise procedure from [4] at the points $(z_l)_{1\le l\le d}$ defined as

$$z_l = a + \frac{l}{d}(b-a) \quad \text{and} \quad d = [\sqrt{n}]\,, \tag{3.1}$$

where $[a]$ is the integer part of a number a. So we propose to use the first ι_l observations for the auxiliary estimation of $S(z_l)$. We set

$$\widehat{S}_{\iota_l} = \frac{1}{A_{\iota_l}} \sum_{j=1}^{\iota_l} Q_{l,j}\, y_{j-1}\, y_j\,, \quad A_{\iota_l} = \sum_{j=1}^{\iota_l} Q_{l,j}\, y_{j-1}^2\,, \tag{3.2}$$

where $Q_{l,j} = Q(u_{l,j})$ and the kernel $Q(\cdot)$ is the indicator function of the interval $[-1;1]$, i.e. $Q(u) = \mathbf{1}_{[-1,1]}(u)$. The points $(u_{l,j})$ are defined as

$$u_{l,j} = \frac{x_j - z_l}{h}\,. \tag{3.3}$$

Note that to estimate $S(z_l)$ on the basis of kernel estimator with the kernel Q we use only the observations $(y_j)_{k_{1,l}\le j\le k_{2,l}}$ from the h - neighborhood of the point z_l, i.e.

$$k_{1,l} = [n\widetilde{z}_l - n\widetilde{h}] + 1 \quad \text{and} \quad k_{2,l} = [n\widetilde{z}_l + n\widetilde{h}] \wedge n\,, \tag{3.4}$$

where $\widetilde{z}_l = (z_l - a)/(b-a)$ and $\widetilde{h} = h/(b-a)$. Note that, only for the last point $z_d = b$, we take $k_{2,d} = n$. We chose ι_l in (3.2) as

$$\iota_l = k_{1,l} + \mathbf{q} \quad \text{and} \quad \mathbf{q} = \mathbf{q}_n = [(n\widetilde{h})^{\mu_0}] \tag{3.5}$$

for some $0 < \mu_0 < 1$. In the sequel for any $0 \le k < m \le n$ we set

$$A_{k,m} = \sum_{j=k+1}^{m} Q_{l,j}\, y_{j-1}^2 \quad \text{and} \quad A_m = A_{0,m}\,. \tag{3.6}$$

Next, similarly to [2], we use a kernel sequential procedure based on the observations $(y_j)_{\iota_l\le j\le n}$. To transform the kernel estimator in a linear function of observations and we replace the number of observations n by the following stopping time

$$\tau_l = \inf\{\iota_l + 1 \le k \le k_{2,l} \,:\, A_{\iota_l,k} \ge H_l\}\,, \tag{3.7}$$

where $\inf\{\emptyset\} = k_{2,l}$ and the positive threshold H_l will be chosen as a positive random variable which is measurable with respect to the σ-field $\{y_1, \ldots, y_{\iota_l}\}$. Now we define the sequential estimator as

$$S_l^* = \frac{1}{H_l} \left(\sum_{j=\iota_l+1}^{\tau_l-1} Q_{l,j}\, y_{j-1}\, y_j \; + \; \varkappa_l\, Q_{l,\tau_l}\, y_{\tau_l-1}\, y_{\tau_l} \right) \mathbf{1}_{\Gamma_l}\,, \tag{3.8}$$

where $\Gamma_l = \{A_{\iota_l,k_{2,l}-1} \geq H_l\}$ and the correcting coefficient $0 < \varkappa_l \leq 1$ on this set is defined as

$$A_{\iota_l,\tau_l-1} + \varkappa_l^2\, Q_{l,\tau_l} y_{\tau_l-1}^2 = H_l\,. \tag{3.9}$$

Note that, to obtain the efficient kernel estimator of $S(z_l)$ we need to use all $k_{2,l} - \iota_l - 1$ observations. Similarly to [19], one can show that $\tau_l \approx \gamma_l\, H_l$ as $H_l \to \infty$, where

$$\gamma_l = 1 - S^2(z_l)\,. \tag{3.10}$$

Therefore, one needs to chose H_l as $(k_{2,l} - \iota_l - 1)/\gamma_l$. Taking into account that the coefficients γ_l are unknown, we define the threshold H_l as

$$H_l = \frac{1-\widetilde{\epsilon}}{\widetilde{\gamma}_l}(k_{2,l} - \iota_l - 1) \quad \text{and} \quad \widetilde{\epsilon} = \frac{1}{2 + \ln n}\,, \tag{3.11}$$

where $\widetilde{\gamma}_l = 1 - \widetilde{S}^2_{\iota_l}$ and $\widetilde{S}_{\iota_l}$ is the projection of the estimator $\widehat{S}_{\iota_l}$ in the interval $]-1+\widetilde{\epsilon}, 1-\widetilde{\epsilon}[$, i.e.

$$\widetilde{S}_{\iota_l} = \min(\max(\widehat{S}_{\iota_l}, -1+\widetilde{\epsilon}), 1-\widetilde{\epsilon})\,. \tag{3.12}$$

To obtain the uncorrelated stochastic terms in kernel estimator for $S(z_l)$ we chose the bandwidth h as

$$h = \frac{b-a}{2d}\,. \tag{3.13}$$

As to the estimator $\widehat{S}_{\iota_l}$, we can show the following property.

Proposition 3.1. *The convergence rate in probability of the estimator* (3.12) *is more rapid than any power function, i.e. for any* $\mathbf{b} > 0$

$$\lim_{n\to\infty} n^{\mathbf{b}} \max_{1\leq l\leq d} \sup_{S\in\Theta_{\epsilon,L}} \sup_{p\in\mathcal{P}} \mathbf{P}_{p,S}\left(|\widetilde{S}_{\iota_l} - S(z_l)| > \epsilon_0 \right) = 0\,, \tag{3.14}$$

where $\epsilon_0 = \epsilon_0(n) \to 0$ *as* $n \to \infty$ *such that* $\lim_{n\to\infty} n^{\check{\delta}}\epsilon_0 = \infty$ *for any* $\check{\delta} > 0$.

Now we set

$$Y_l = S_l^*(z_l)\mathbf{1}_\Gamma \quad \text{and} \quad \Gamma = \cap_{l=1}^d \Gamma_l\,. \tag{3.15}$$

Using the convergence (3.14), we study the probability properties of the set Γ in the following proposition.

Proposition 3.2. *For any* $\mathbf{b} > 0$, *the probability of the set* Γ *satisfies the following asymptotic equality*

$$\lim_{n\to\infty} n^{\mathbf{b}} \sup_{S\in\Theta_{\epsilon,L}} \mathbf{P}_{p,S}\left(\Gamma^c\right) = 0\,. \tag{3.16}$$

In view of this proposition we can shrink the set Γ^c. So, using the estimators (3.15) on the set Γ we obtain the discrete time regression model (1.6) in which

$$\xi_l^* = \frac{\sum_{j=\iota_l+1}^{\tau_l-1} Q_{l,j}\, y_{j-1}\, \xi_j + \varkappa_l\, Q(u_{l,\tau_l})\, y_{\tau_l-1}\, \xi_{\tau_l}}{H_l} \tag{3.17}$$

and $\varpi_l = \varpi_{1,l} + \varpi_{2,l}$, where

$$\varpi_{1,l} = \frac{\sum_{j=\iota_l+1}^{\tau_l-1} Q_{l,j}\, y_{j-1}^2 \check{s}_{l,j} + \varkappa_l^2\, Q(u_{l,\tau_l})\, y_{\tau_l-1}^2 \check{s}_{l,\tau_l}}{H_l}\,, \quad \check{s}_{l,j} = S(x_j) - S(z_l)$$

and

$$\varpi_{2,l} = \frac{(\varkappa_l - \varkappa_l^2)\, Q(u_{l,\tau_l})\, y_{\tau_l-1}^2\, S(x_{\tau_l})}{H_l}\,.$$

Note that in the model (1.6) the random variables $(\xi_j^*)_{1\le j\le d}$ are defined only on the set Γ. For technical reasons we need to define these variables on the set Γ^c as well. To this end, for any $j \ge 1$ we set

$$\check{Q}_{l,j} = Q_{l,j}\, y_{j-1}\, \mathbf{1}_{\{j<k_{2,l}\}} + \sqrt{H_l}\, Q_{l,j}\, \mathbf{1}_{\{j=k_{2,l}\}} \tag{3.18}$$

and $\check{A}_{\iota_l,m} = \sum_{j=\iota_l+1}^{m} \check{Q}_{l,j}^2$. Note, that for any $j \ge 1$ and $l \ne m$

$$\check{Q}_{l,j}\, \check{Q}_{m,j} = 0\,. \tag{3.19}$$

and $\check{A}_{\iota_l,k_{2,l}} \ge H_l$. So now we can modify the stopping time (3.7) as

$$\check{\tau}_l = \inf\{k \ge \iota_l + 1 : \check{A}_{\iota_l,k} \ge H_l\}\,. \tag{3.20}$$

Obviously, $\check{\tau}_l \le k_{2,l}$ and $\check{\tau}_l = \tau_l$ on the set Γ for any $1 \le l \le d$. Now similarly to (3.9) we define the correction coefficient as

$$\check{A}_{\iota_l,\check{\tau}_l-1} + \check{\varkappa}_l^2\, \check{Q}_{l,\check{\tau}_l}^2 = H_l\,. \tag{3.21}$$

It is clear that $0 < \check{\varkappa}_l \le 1$ and $\check{\varkappa}_l = \varkappa_l$ on the set Γ for $1 \le l \le d$. Using this coefficient we set

$$\eta_l = \frac{\sum_{j=\iota_l+1}^{\check{\tau}_l-1} \check{Q}_{l,j}\, \xi_j + \check{\varkappa}_l\, \check{Q}_{l,\check{\tau}_l}\, \xi_{\check{\tau}_l}}{H_l}\,. \tag{3.22}$$

Note that on the set Γ, for any $1 \leq l \leq d$, the random variables $\eta_l = \xi_l^*$. Moreover (see Lemma A.2), for any $1 \leq l \leq d$ and $p \in \mathcal{P}$

$$\mathbf{E}_{p,S}\left(\eta_l \,|\mathcal{G}_l\right) = 0\,, \quad \mathbf{E}_{p,S}\left(\eta_l^2 \,|\mathcal{G}_l\right) = \sigma_l^2 \quad \text{and} \quad \mathbf{E}_{p,S}\left(\eta_l^4 \,|\mathcal{G}_l\right) \leq \breve{\mathbf{m}}\sigma_l^4\,, \tag{3.23}$$

where $\sigma_l = H_l^{-1/2}$, $\mathcal{G}_l = \sigma\{\eta_1, \ldots, \eta_{l-1}, \sigma_l\}$ and $\breve{\mathbf{m}} = 4(144/\sqrt{3})^4\,\mathbf{m}_4^*$. It is clear that

$$\sigma_{0,*} \leq \min_{1\leq l\leq d} \sigma_l^2 \leq \max_{1\leq l\leq d} \sigma_l^2 \leq \sigma_{1,*}\,, \tag{3.24}$$

where

$$\sigma_{0,*} = \frac{1-\epsilon^2}{2(1-\widetilde{\epsilon})nh} \quad \text{and} \quad \sigma_{1,*} = \frac{1}{(1-\widetilde{\epsilon})(2nh-\mathbf{q}-3)}\,.$$

Now, taking into account that $|\varpi_{1,l}| \leq Lh$, for any $S \in \Theta_{\epsilon,L}$ we obtain that

$$\sup_{S\in\Theta_{\epsilon,L}} \mathbf{E}_{p,S}\mathbf{1}_\Gamma \varpi_l^2 \leq \left(L^2h^2 + \frac{\breve{\upsilon}_n}{(nh)^2}\right)\,, \tag{3.25}$$

where $\breve{\upsilon}_n = \sup_{p\in\mathcal{P}} \sup_{S\in\Theta_{\epsilon,L}} \mathbf{E}_{p,S} \max_{1\leq j\leq n} y_j^4$. The behavior of this coefficient is studied in the following proposition.

Proposition 3.3. *For any $\mathbf{b} > 0$ the sequence $(\breve{\upsilon}_n)_{n\geq 1}$ satisfies the following limiting equality*

$$\lim_{n\to\infty} n^{-\mathbf{b}}\,\breve{\upsilon}_n = 0\,. \tag{3.26}$$

Remark 3.1. *It should be noted that the property (3.26) means that the asymptotic behavior of the upper bound (3.25) is approximately as h^{-2} when $n \to \infty$. We will use this in the oracle inequalities below.*

Remark 3.2. *It should be emphasized that to estimate the function S in (1.1) we use the approach developed in [16] for the sequential drift estimation problem in the stochastic differential equation. On the basis of the efficient sequential kernel procedure developped in [11, 13, 18] with the kernel-indicator, the stochastic differential equation is replaced by regression model. It should be noted that to obtain the efficient estimator one needs to take the kernel-indicator estimator. By this reason, in this paper, we use the kernel-indicator in the sequential estimator (3.8). It also should be noted that the sequential estimator (3.8) which has the same form as in [4], except the last term, in which the correction coefficient is replaced by the square root of the coefficient used in [22]. We modify this procedure to calculate the variance of the stochastic term (3.17).*

4 Model selection

In this section we consider the nonparametric estimation problem in the non asymptotic setting for the regression model (1.6) for some set $\Gamma \subseteq \Omega$. The design points $(z_l)_{1\le l\le d}$ are defined in (3.1). The function $S(\cdot)$ is unknown and has to be estimated from observations the $Y_1, \ldots, Y_d$. Moreover, we assume that the unobserved random variables $(\eta_l)_{1\le l\le d}$ satisfy the properties (3.23) with some nonrandom constant $\breve{\mathbf{m}} > 1$ and the known random positive coefficients $(\sigma_l)_{1\le l\le d}$ satisfy the inequality (3.24) for some nonrandom positive constants $\sigma_{0,*}$ and $\sigma_{1,*}$ Concerning the random sequence $\varpi = (\varpi_l)_{1\le l\le n}$ we suppose that

$$\mathbf{u}_d^* = \mathbf{E}_{p,S}\mathbf{1}_\Gamma \|\varpi\|_d^2 < \infty\,. \tag{4.1}$$

The performance of any estimator $\widehat{S}$ will be measured by the empirical squared error

$$\|\widehat{S} - S\|_d^2 = (\widehat{S} - S, \widehat{S} - S)_d = \frac{b-a}{d}\sum_{l=1}^{d}(\widehat{S}(z_l) - S(z_l))^2\,.$$

Now we fix a basis $(\phi_j)_{1\le j\le n}$ which is orthonormal for the empirical inner product:

$$(\phi_i\,,\,\phi_j)_d = \frac{b-a}{d}\sum_{l=1}^{d}\phi_i(z_l)\phi_j(z_l) = \mathbf{1}_{\{i=j\}}\,. \tag{4.2}$$

For example, we can take the trigonometric basis $(\phi_j)_{j\ge 1}$ in $\mathbf{L}_2[a,b]$ defined as

$$\phi_1 = 1\,, \quad \phi_j(x) = \sqrt{\frac{2}{b-a}}\mathrm{Tr}_j\left(2\pi[j/2]\mathbf{l}_0(x)\right)\,, \quad j \ge 2\,, \tag{4.3}$$

where the function $\mathrm{Tr}_j(x) = \cos(x)$ for even j and $\mathrm{Tr}_j(x) = \sin(x)$ for odd j, $[x]$ denotes the integer part of x. and $\mathbf{l}_0(x) = (x-a)/(b-a)$. Note that, using the orthonormality property (4.2) we can represent for any $1 \le l \le d$ the function S as

$$S(z_l) = \sum_{j=1}^{d}\theta_{j,d}\,\phi_j(z_l) \quad \text{and} \quad \theta_{j,d} = \left(S, \phi_j\right)_d\,. \tag{4.4}$$

So, to estimate the function S we have to estimate the Fourier coefficients $(\theta_{j,d})_{1\le j\le d}$. To this end, we replace the function S by these observations, i.e.

$$\widehat{\theta}_{j,d} = \frac{b-a}{d}\sum_{l=1}^{d}Y_l\phi_j(z_l)\,. \tag{4.5}$$

From (1.6) we obtain immediately the following regression scheme

$$\widehat{\theta}_{j,d} \,=\, \theta_{j,d} \,+\, \zeta_{j,d} \quad \text{with} \quad \zeta_{j,d} = \sqrt{\frac{b-a}{d}}\eta_{j,d} + \varpi_{j,d}\,, \tag{4.6}$$

where

$$\eta_{j,d} = \sqrt{\frac{b-a}{d}} \sum_{l=1}^{d} \eta_l \phi_j(z_l) \quad \text{and} \quad \varpi_{j,d} = \frac{b-a}{d} \sum_{l=1}^{d} \varpi_l\, \phi_j(z_l)\,.$$

Note that the upper bound (3.24) and the Bounyakovskii-Cauchy-Schwarz inequality imply that

$$|\varpi_{j,d}| \le \|\varpi\|_d \, \|\phi_j\|_d = \|\varpi\|_d\,. \tag{4.7}$$

We estimate the function S on the grid (3.1) by the weighted least-squares estimator

$$\widehat{S}_\lambda(z_l) \,=\, \sum_{j=1}^{d} \lambda(j)\, \widehat{\theta}_{j,d}\, \phi_j(z_l)\, \mathbf{1}_\Gamma\,, \quad 1 \le l \le d\,, \tag{4.8}$$

where the weight vector $\lambda = (\lambda(1), \ldots, \lambda(d))'$ belongs to some finite set $\Lambda \subset [0,1]^d$, the prime denotes the transposition. We set for any $a \le t \le b$

$$\widehat{S}_\lambda(t) = \widehat{S}_\lambda(z_1)\mathbf{1}_{\{a \le t \le z_1\}} + \sum_{l=2}^{d} \widehat{S}_\lambda(z_l)\mathbf{1}_{\{z_{l-1} < t \le z_l\}}\,. \tag{4.9}$$

Moreover, denoting $\lambda^2 = (\lambda^2(1), \ldots, \lambda^2(n))'$ we define the following sets

$$\Lambda_1 = \{\lambda^2\,,\, \lambda \in \Lambda\} \quad \text{and} \quad \Lambda_2 = \Lambda \cup \Lambda_1\,. \tag{4.10}$$

Denote by ν the cardinal number of the set Λ and

$$\nu^* = \max_{\lambda \in \Lambda} \sum_{j=1}^{d} \mathbf{1}_{\{\lambda(j) > 0\}}\,.$$

In order to obtain a good estimator, we have to write a rule to choose a weight vector $\lambda \in \Lambda$ in (4.8). We define the empirical squared risk as

$$\mathrm{Err}_d(\lambda) = \|\widehat{S}_\lambda - S\|_d^2\,.$$

Using (4.4) and (4.8) we can rewrite this risk as

$$\mathrm{Err}_d(\lambda) = \sum_{j=1}^{d} \lambda^2(j)\widehat{\theta}_{j,d}^2 \,-\, 2\sum_{j=1}^{d} \lambda(j)\widehat{\theta}_{j,d}\,\theta_{j,d} \,+\, \sum_{j=1}^{d} \theta_{j,d}^2\,. \tag{4.11}$$

Since the coefficient $\theta_{j,d}$ is unknown, we need to replace the term $\widehat{\theta}_{j,d}\,\theta_{j,d}$ by some of its estimators which we choose as

$$\widetilde{\theta}_{j,d} = \widehat{\theta}^2_{j,d} - \frac{b-a}{d}\,\mathbf{s}_{j,d} \quad \text{with} \quad \mathbf{s}_{j,d} = \frac{b-a}{d}\sum_{l=1}^{d}\sigma_l^2\,\phi_j^2(z_l)\,. \tag{4.12}$$

Note that from (3.24) - (4.2) it follows that

$$\mathbf{s}_{j,d} \le \sigma_{1,*}\,. \tag{4.13}$$

Finally, we define the cost function of the form

$$J_d(\lambda) \;=\; \sum_{j=1}^{d}\lambda^2(j)\widehat{\theta}^2_{j,d} \;-\; 2\sum_{j=1}^{d}\lambda(j)\,\widetilde{\theta}_{j,d} \;+\; \delta P_d(\lambda)\,, \tag{4.14}$$

where the penalty term is defined as

$$P_d(\lambda) = \frac{b-a}{d}\sum_{j=1}^{d}\lambda^2(j)\mathbf{s}_{j,d} \tag{4.15}$$

and $0 < \delta < 1$ is some positive constant which will be chosen later. We set

$$\widehat{\lambda} = \operatorname{argmin}_{\lambda\in\Lambda} J_d(\lambda) \tag{4.16}$$

and define an estimator of $S(t)$ of the form (4.9):

$$\widehat{S}_*(t) = \widehat{S}_{\widehat{\lambda}}(t) \quad \text{for} \quad a \le t \le b\,. \tag{4.17}$$

Remark 4.1. *We use the procedure* (4.17) *to estimate the function* S *in the autoregressive model* (1.1) *through the regression scheme* (1.6) *generated by the sequential procedures* (3.15).

5 Main results

In this section we formulate all main results. First we obtain the sharp oracle inequality for the selection model procedure (4.17) for the general regression model (1.6).

Theorem 5.1. *There exists some constant* $\mathbf{l}^* > 0$ *such that for any weight vectors set* Λ*, any* $p \in \mathcal{P}$*, any* $n \ge 1$ *and* $0 < \delta \le 1/12$*, the procedure* (4.17)*, satisfies the following oracle inequality*

$$\mathbf{E}_{p,S}\|\widehat{S}_* - S\|_d^2 \le \frac{1+4\delta}{1-6\delta}\min_{\lambda\in\Lambda}\mathbf{E}_{p,S}\|\widehat{S}_\lambda - S\|_d^2 + \mathbf{l}^*\frac{\nu_\varsigma^2}{\delta}\left(\frac{\sigma_{1,*}^2}{\sigma_{0,*}d} + \mathbf{u}_d^* + \delta^2\sqrt{\mathbf{P}_S(\Gamma^c)}\right)\,. \tag{5.1}$$

Using now Lemma A.7 we obtain the oracle inequality for the quadratic risks (1.2).

Theorem 5.2. *There exists some constant* $\mathbf{l}^* > 0$ *such that for any weight vectors set* Λ*, any continuously differentiable function* S*, any* $p \in \mathcal{P}$*, any* $n \geq 1$ *and* $0 < \delta \leq 1/12$*, the procedure* (4.17) *satisfies the following oracle inequality*

$$\mathcal{R}_p(\widehat{S}_*, S) \leq \frac{(1+4\delta)(1+\delta)^2}{1-6\delta} \min_{\lambda\in\Lambda} \mathcal{R}_p(\widehat{S}_\lambda, S) + \mathbf{l}^* \frac{\varsigma^2 \nu}{\delta} \left(\frac{\|\dot{S}\|^2}{d^2} + \frac{\sigma_{1,*}^2}{\sigma_{0,*} d} + \mathbf{u}_d^* + \delta^2 \sqrt{\mathbf{P}_S(\Gamma^c)} \right) . \tag{5.2}$$

Now we assume that the cardinal ν of Λ and the parameter ς in the density family (2.1) are functions of the number observations n, i.e. $\nu = \nu(n)$ and $\varsigma = \varsigma(n)$ such that for any $\check{\delta} > 0$

$$\lim_{n\to\infty} \frac{\nu(n)}{n^{\check{\delta}}} = 0 . \tag{5.3}$$

Using Propositions 3.2 – 3.3 and the bounds (3.24) - (3.25) we obtain the oracle inequality for the estimation problem for the model (1.1).

Theorem 5.3. *Assume that the conditions* (2.2) *and* (5.3) *hold. Then for any* $p \in \mathcal{P}$*,* $S \in \Theta_{\epsilon,L}$*,* $n \geq 3$ *and* $0 < \delta \leq 1/12$*, the procedure* (4.17) *satisfies the following oracle inequality*

$$\mathcal{R}_p(\widehat{S}_*, S) \leq \frac{(1+4\delta)(1+\delta)^2}{1-6\delta} \min_{\lambda\in\Lambda} \mathcal{R}_p(\widehat{S}_\lambda, S) + \frac{\check{\mathbf{B}}_n(p)}{\delta n} , \tag{5.4}$$

where the term $\check{\mathbf{B}}_n(p)$ *is such that for any* $\check{\delta} > 0$

$$\lim_{n\to\infty} \frac{\check{\mathbf{B}}_n(p)}{n^{\check{\delta}}} = 0 .$$

We obtain the same inequality for the robust risks

Theorem 5.4. *Assume that the conditions* (2.2) *and* (5.3) *hold. Then for any* $n \geq 3$*, any* $S \in \Theta_{\epsilon,L}$ *and any* $0 < \delta \leq 1/12$*, the procedure* (4.17) *satisfies the following oracle inequality*

$$\mathcal{R}^*(\widehat{S}_*, S) \leq \frac{(1+4\delta)(1+\delta)^2}{1-6\delta} \min_{\lambda\in\Lambda} \mathcal{R}^*(\widehat{S}_\lambda, S) + \frac{\mathbf{B}_n^*}{\delta n} , \tag{5.5}$$

where the term $\check{\mathbf{B}}_n$ *is such that for any* $\check{\delta} > 0$

$$\lim_{n\to\infty} \frac{\mathbf{B}_n^*}{n^{\check{\delta}}} = 0 .$$

It is well known that to obtain the efficiency property we need to specify the weight coefficients $(\lambda(j))_{1\le j\le n}$ (see, for example, [15]). Consider for some fixed $0<\varepsilon<1$ a numerical grid of the form

$$\mathcal{A}=\{1,\dots,k^*\}\times\{\varepsilon,\dots,m\varepsilon\}\,, \tag{5.6}$$

where $m=[1/\varepsilon^2]$. We assume that both parameters $k^*\ge 1$ and ε are functions of n, i.e. $k^*=k^*(n)$ and $\varepsilon=\varepsilon(n)$, such that

$$\begin{cases} \lim_{n\to\infty} k^*(n)=+\infty\,, \quad \lim_{n\to\infty}\dfrac{k^*(n)}{\ln n}=0\,, \\ \\ \lim_{n\to\infty}\varepsilon(n)=0 \quad \text{and} \quad \lim_{n\to\infty} n^{\check{\delta}}\varepsilon(n)=+\infty \end{cases} \tag{5.7}$$

for any $\check{\delta}>0$. One can take, for example, for $n\ge 2$

$$\varepsilon(n)=\frac{1}{\ln n} \quad \text{and} \quad k^*(n)=k_0^*+\sqrt{\ln n}\,, \tag{5.8}$$

where $k_0^*\ge 0$ is some fixed constant. For each $\alpha=(\beta,\mathbf{l})\in\mathcal{A}$, we introduce the weight sequence

$$\lambda_\alpha=(\lambda_\alpha(j))_{1\le j\le n}$$

with the elements

$$\lambda_\alpha(j)=\mathbf{1}_{\{1\le j<j_*\}}+\bigl(1-(j/\omega_\alpha)^\beta\bigr)\,\mathbf{1}_{\{j_*\le j\le\omega_\alpha\}}, \tag{5.9}$$

where $j_*=1+[\ln n]$, $\omega_\alpha=(\mathrm{d}_\beta\,\mathbf{l}\,n)^{1/(2\beta+1)}$ and

$$\mathrm{d}_\beta=\frac{(\beta+1)(2\beta+1)}{\pi^{2\beta}\beta}\,.$$

Now we define the set Λ as

$$\Lambda\;=\;\{\lambda_\alpha\,,\;\alpha\in\mathcal{A}\}\,. \tag{5.10}$$

Note that these weight coefficients are used in [20, 21] for continuous time regression models to show the asymptotic efficiency. It will be noted that in this case the cardinal of the set Λ is $\nu=k^*m$. It is clear that properties (5.7) imply condition (5.3).

6 Properties of the regression model (1.6)

In order to prove the oracle inequalities we need to study the conditions introduced in [20] for the general semi-martingale model. To this end we set for any $\lambda \in \mathbb{R}^d$ the functions

$$\mathbf{B}(\lambda) = \frac{b-a}{\sqrt{d}} \sum_{j=1}^{d} \lambda(j)\, \widetilde{\eta}_{j,d}\,, \quad \widetilde{\eta}_{j,d} = \eta_{j,d}^2 - \mathbf{E}_{p,S}\, \eta_{j,d}^2\,. \tag{6.1}$$

Proposition 6.1. *For any $d \geq 1$ and any $\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$*

$$\mathbf{E}_{p,S}\, \mathbf{1}_{\Gamma}\, \mathbf{B}^2(\lambda) \leq 10(b-a)\sigma_{1,*}\check{\mathbf{m}}\, \mathbf{E}_{p,S}\, P_d(\lambda)\,. \tag{6.2}$$

where $\check{\mathbf{m}}$ is defined in (3.23).

Proof. First note that the random variable $\widetilde{\eta}_{j,d}$ can be represented as

$$\widetilde{\eta}_{j,d} = \frac{b-a}{d} \sum_{l=1}^{d} \left(\phi_j^2(z_l)\, \check{\eta}_l + 2\mathbf{1}_{\{l \geq 2\}}\, \check{\upsilon}_{j,l} \eta_l \right)\,,$$

where $\check{\eta}_l = \eta_l^2 - \sigma_l^2$ and $\check{\upsilon}_{j,l} = \phi_j(z_l) \sum_{r=1}^{l-1} \phi_j(z_r)\eta_r$. Therefore, we can rewrite the term $\mathbf{B}(\lambda)$ as

$$\mathbf{B}(\lambda) = \mathbf{B}_1(\lambda) + 2\mathbf{B}_2(\lambda)\,.$$

The terms $\mathbf{B}_1(\lambda)$ and $\mathbf{B}_2(\lambda)$ are defined as

$$\mathbf{B}_1(\lambda) = \frac{(b-a)^2}{d\sqrt{d}} \sum_{l=1}^{d} \psi_{1,l}(\lambda)\check{\eta}_l \quad \text{and} \quad \mathbf{B}_2(\lambda) = \frac{(b-a)^2}{d\sqrt{d}} \sum_{l=2}^{d} \psi_{2,l}(\lambda)\eta_l\,,$$

where

$$\psi_{1,l}(\lambda) = \sum_{j=1}^{d} \lambda(j)\phi_j^2(z_l) \quad \text{and} \quad \psi_{2,l}(\lambda) = \sum_{j=1}^{d} \lambda(j)\check{\upsilon}_{j,l}\,.$$

So,

$$\mathbf{E}_{p,S}\, \mathbf{B}^2(\lambda) \leq 2\mathbf{E}_{p,S}\, \mathbf{B}_1^2(\lambda) + 8\mathbf{E}_{p,S}\, \mathbf{B}_2^2(\lambda)\,. \tag{6.3}$$

Taking into account property (4.2) and Bounyakovskii - Cauchy - Schwarz inequality we get

$$\psi_{1,l}^2(\lambda) \leq \sum_{j=1}^{d} \lambda^2(j)\phi_j^2(z_l) \sum_{j=1}^{d} \phi_j^2(z_l) = \frac{d}{b-a} \sum_{j=1}^{d} \lambda^2(j)\phi_j^2(z_l)\,.$$

In view of properties (3.23) we obtain that

$$\mathbf{E}_{p,S}\,\mathbf{B}_1^2(\lambda) = \frac{(b-a)^4}{d^3}\sum_{l=1}^{d}\psi_{1,l}^2(\lambda)\,\mathbf{E}_{p,S}\,\check{\eta}_l^2 \le \frac{(b-a)^4}{d^3}\sum_{l=1}^{d}\psi_{1,l}^2(\lambda)\,\mathbf{E}_{p,S}\,\eta_l^4$$

$$\le \frac{\sigma_{1,*}\check{\mathbf{m}}(b-a)^3}{d^2}\mathbf{E}_{p,S}\sum_{j=1}^{d}\lambda^2(j)\sum_{l=1}^{d}\sigma_l^2\,\phi_j^2(z_l)$$

$$= \sigma_{1,*}\,(b-a)\check{\mathbf{m}}\,\mathbf{E}_{p,S}\,P_d(\lambda)\,.$$

To estimate the last term in the right hand side of the inequality (6.3), noting that the term $\psi_{2,l}$ can be represented as

$$\psi_{2,l}(\lambda) = \sum_{r=1}^{l-1} g_{l,r}\,\eta_r\,,$$

where $g_{l,r} = \sum_{j=1}^{d}\lambda(j)\,\phi_j(z_l)\phi_j(z_r)$, we use properties (3.23) to obtain

$$\mathbf{E}_{p,S}\,\mathbf{B}_2^2(\lambda) = \frac{(b-a)^4}{d^3}\sum_{l=2}^{d}\mathbf{E}_{p,S}\,\psi_{2,l}^2(\lambda)\,\eta_l^2 \le \frac{\sigma_{1,*}(b-a)^4}{d^3}\sum_{l=2}^{d}\mathbf{E}_{p,S}\psi_{2,l}^2(\lambda)$$

$$= \frac{\sigma_{1,*}(b-a)^4}{d^3}\sum_{l=2}^{d}\sum_{r=1}^{l-1} g_{l,r}^2\,\mathbf{E}_{p,S}\sigma_r^2$$

$$\le \frac{\sigma_{1,*}(b-a)^4}{d^3}\sum_{r=1}^{d}\mathbf{E}_{p,S}\sigma_r^2\sum_{l=1}^{d} g_{l,r}^2 = \sigma_{1,*}\,(b-a)\,\mathbf{E}_{p,S}\,P_d(\lambda)\,.$$

Hence Proposition 6.1. □

Now we need the following moment bound.

Proposition 6.2. *For any non random* $v_1,\dots,v_d$

$$\mathbf{E}\left(\sum_{j=1}^{d} v_j\,\eta_{j,d}\right)^2 \le \sigma_{1,*}\sum_{j=1}^{d} v_j^2\,. \tag{6.4}$$

Proof. Note that

$$\mathbf{E}\left(\sum_{j=1}^{d} v_j\,\eta_{j,d}\right)^2 = \frac{b-a}{d}\,\mathbf{E}\sum_{l=1}^{d}\sigma_l^2\left(\sum_{j=1}^{d} v_j\phi_j(z_l)\right)^2$$

$$\le \frac{\sigma_{1,*}(b-a)}{d}\sum_{l=1}^{d}\left(\sum_{j=1}^{d} v_j\phi_j(z_l)\right)^2.$$

By applying the orthonormal property (4.2) we obtain the desired inequality. Hence Proposition 6.2. □

7 Proofs

7.1 Proof of Proposition 3.1

First recall that

$$\widehat{S}_{\iota_l} = \frac{1}{A_{\iota_l}} \sum_{j=1}^{\iota_l} Q_{l,j} y_{j-1} y_j \quad \text{and} \quad \widetilde{S}_{\iota_l} = \min(\max(\widehat{S}_{\iota_l}, -1+\tilde{\epsilon}), 1-\tilde{\epsilon}),$$

where $\tilde{\epsilon} = 1/(2+\ln n)$. Note that for sufficiently large n, for which we have $\tilde{\epsilon} < \epsilon$ and then $S(z_l) \in [-1+\tilde{\epsilon}; 1-\tilde{\epsilon}]$. We can write

$$|\widetilde{S}_{\iota_l} - S(z_l)| \le |\widehat{S}_{\iota_l} - S(z_l)| \le \frac{\sum_{j=k_{1,l}}^{\iota_l} y_{j-1}^2 |S(x_j) - S(z_l)|}{\sum_{j=k_{1,l}}^{\iota_l} y_{j-1}^2} + |I_n|,$$

where $I_n = \sum_{j=k_{1,l}}^{\iota_l} y_{j-1}\xi_j / \sum_{j=k_{1,l}}^{\iota_l} y_{j-1}^2$. Taking into account that $|x_j - z_l| \le h$ for $k_{1,l} \le j \le k_{2,l}$, we obtain that for any $S \in \Theta_{\epsilon,L}$,

$$|\widehat{S}_{\iota_l} - S(z_l)| \le Lh + |I_n|.$$

So, for sufficiently large n

$$\begin{aligned} \mathbf{P}_{p,S}\left(|\widetilde{S}_{\iota_l} - S(z_l)| > \epsilon_0\right) &\le \mathbf{P}_{p,S}\left(I_n > \frac{\epsilon_0}{2}\right) \\ &\le \mathbf{P}_{p,S}\left(I_n > \frac{\epsilon_0}{2}, \Xi\right) + \mathbf{P}_{p,S}\left(\Xi^c\right), \end{aligned} \tag{7.1}$$

where $\Xi = \left\{\left|\Upsilon_{m_0,m_1}(z_l)\right| \le 1/2\right\}$, $m_0 = k_{1,l} - 2$, $m_1 = \iota_l - 1$ and $\Upsilon_{m_0,m_1}(z_l)$ is defined in (A.7). Hence we obtain the following inequality on the set Ξ:

$$\sum_{j=k_{1,l}}^{\iota_l} y_{j-1}^2 = (\iota_l - k_{1,l} + 1)\left(\frac{1}{1-S^2(z_l)} + \Upsilon_{m_0,m_1}(z_l)\right) \ge \frac{\mathbf{q}}{2}.$$

Therefore, for any $\check{p} > 2$,

$$\mathbf{P}_{p,S}\left(I_n > \frac{\epsilon_0}{2}\,,\,\Xi\right) \le \mathbf{P}_{p,S}\left(\left|\sum_{j=k_{1,l}}^{\iota_l} y_{j-1}\xi_j\right| > \frac{\mathbf{q}}{2}\right)$$

$$\le \frac{2^{\check{p}}}{\mathbf{q}^{\check{p}}}\,\mathbf{E}_{p,S}\left|\sum_{j=k_{1,l}}^{\iota_l} y_{j-1}\xi_j\right|^{\check{p}}\,. \tag{7.2}$$

Using here the correlation inequality (A.2) and the bound (A.6), we obtain that

$$\max_{1\le l\le d}\sup_{S\in\Theta_{\epsilon,L}}\sup_{p\in\mathcal{P}}\mathbf{E}_{p,S}\left|\sum_{j=k_{1,l}}^{\iota_l} y_{j-1}\xi_j\right|^{\check{p}} \le c_{\check{p}}\,\mathbf{q}^{\check{p}/2}\,.$$

Applying this bound in (7.1) and using Lemma A.6 we obtain Proposition 3.1. □

7.2 Proof of Proposition 3.2

First note, that

$$\mathbf{P}_{p,S}(\Gamma^c) \le \sum_{l=1}^{d}\mathbf{P}_{p,S}\left(A_{\iota_l,k_{2,l}-1} < H_l\right)\,.$$

Moreover, note that in view of definition (A.7) the term $A_{\iota_l,k_{2,l}-1}$ can be represented as

$$A_{\iota_l,k_{2,l}-1} = (m_{1,l}-m_{0,l})\left(\frac{1}{\gamma_l} + \Upsilon_{m_{0,l},m_{1,l}}(z_l)\right)\,,$$

where $m_{0,l} = \iota_l - 1$ and $m_{1,l} = k_{2,l} - 2$. Taking into account the definition of H_l in (3.11) and the fact that $0 < \widetilde{\gamma}_l, \gamma_l \le 1$ and that $|\widetilde{\gamma}_l - \gamma_l| \le 2|\widetilde{S}_{\iota_l} - S(z_l)|$, we obtain

$$\mathbf{P}_{p,S}\left(A_{\iota_l,k_{2,l}-1} < H_l\right) = \mathbf{P}_{p,S}\left(\frac{1}{\gamma_l} + \Upsilon_{m_{0,l},m_{1,l}}(z_l) < \frac{1-\widetilde{\epsilon}}{\widetilde{\gamma}_l}\right)$$

$$\le \mathbf{P}_{p,S}\left(\left|\frac{1}{\gamma_l} - \frac{1}{\widetilde{\gamma}_l}\right| > \frac{\widetilde{\epsilon}}{2}\right) + \mathbf{P}_{p,S}\left(\left|\Upsilon_{m_{0,l},m_{1,l}}(z_l)\right| > \frac{\widetilde{\epsilon}}{2}\right)$$

$$\le \mathbf{P}_{p,S}\left(\left|\widetilde{S}_{\iota_l} - S(z_l)\right| > \frac{\widetilde{\epsilon}^3}{4}\right) + \mathbf{P}_{p,S}\left(\left|\Upsilon_{m_{0,l},m_{1,l}}(z_l)\right| > \frac{\widetilde{\epsilon}}{2}\right)\,.$$

Applying here Proposition 3.1 and Lemma A.6 we obtain Proposition 3.2. □

7.3 Proof of Proposition 3.3

Note that, for any $m \geq 1$

$$\begin{aligned}\mathbf{E}_{p,S} \max_{1\leq j\leq n} y_j^4 &\leq n^{\mathbf{b}/2} + \sum_{j=1}^{n} \int_{n^{\mathbf{b}/2}}^{+\infty} \mathbf{P}_{p,S}\left(y_j^4 \geq z\right) \mathrm{d}z \\ &\leq n^{\mathbf{b}/2} + n \max_{1\leq j\leq n} \mathbf{E}_{p,S}\,|y_j|^{4m} \int_{n^{\mathbf{b}/2}}^{+\infty} z^{-m}\mathrm{d}z \\ &= n^{\mathbf{b}/2} + \max_{1\leq j\leq n} \mathbf{E}_{p,S}\,|y_j|^{4m}\, \frac{n^{1-\mathbf{b}(m-1)/2}}{m-1}\,.\end{aligned}$$

Choosing here $m > 1+2/\mathbf{b}$ and using the bound (A.6) we obtain the property (3.26). Hence Proposition 3.3. □

7.4 Proof of Theorem 5.1

First of all, note that on the set Γ we can represent the empirical squared error $\mathrm{Err}_d(\lambda)$ in the form

$$\mathrm{Err}_d(\lambda) = J_d(\lambda) + 2\sum_{j=1}^{d} \lambda(j)\theta'_{j,d} + \|S\|_d^2 - \delta\, P_d(\lambda) \tag{7.3}$$

with $\theta'_{j,d} = \widetilde{\theta}_{j,d} - \theta_{j,d}\widehat{\theta}_{j,d}$. From (4.6) we find that

$$\theta'_{j,d} = \theta_{j,d}\zeta_{j,d} + \frac{b-a}{d}\widetilde{\eta}_{j,d} + 2\sqrt{\frac{b-a}{d}}\eta_{j,d}\varpi_{j,d} + \varpi_{j,d}^2\,,$$

where $\widetilde{\eta}_{j,d} = \eta_{j,d}^2 - \mathbf{s}_{j,d}$. Now putting

$$M(\lambda) = \sum_{j=1}^{d} \lambda(j)\,\theta_{j,d}\,\zeta_{j,d}\,, \tag{7.4}$$

we rewrite (7.3) as follows

$$\begin{aligned}\mathrm{Err}_d(\lambda) = J_d(\lambda) + 2M(\lambda) + 2\frac{1}{\sqrt{d}}\,\mathbf{B}(\lambda) \\ + 2\Delta(\lambda) + \|S\|_d^2 - \delta\, P_d(\lambda)\,,\end{aligned} \tag{7.5}$$

where $\mathbf{B}(\lambda)$ is given in (6.1), $\Delta(\lambda) = \Delta_1(\lambda) + \Delta_2(\lambda)$,

$$\Delta_1(\lambda) = \sum_{j=1}^{d} \lambda(j)\, \varpi_{j,d}^2 \quad \text{and} \quad \Delta_2(\lambda) = 2\sqrt{\frac{b-a}{d}} \sum_{j=1}^{d} \lambda(j)\, \eta_{j,d} \varpi_{j,d}\,.$$

In view of Proposition 6.1, for any $\lambda \in \mathbb{R}^d$,

$$\mathbf{E}_{p,S}\, \mathbf{1}_\Gamma \mathbf{B}^2(\lambda) \le 10\sigma_{1,*}\,(b-a)\check{\mathbf{m}}\, \mathbf{E}_{p,S}\, P_d(\lambda)\,. \tag{7.6}$$

Note that the inequalities (3.24) imply that

$$P_{0,d}(\lambda) \le P_d(\lambda) \le P_{1,d}(\lambda)\,, \tag{7.7}$$

where

$$P_{0,d}(\lambda) = \frac{\sigma_{0,*}(b-a)|\lambda|^2}{d} \quad \text{and} \quad P_{1,d}(\lambda) = \frac{\sigma_{1,*}(b-a)|\lambda|^2}{d}\,.$$

For $\Delta_1(\lambda)$, taking into account the properties of Fourier coefficients we obtain that

$$\sup_{\lambda\in[0,1]^d} |\Delta_1(\lambda)| \le \sum_{j=1}^{d} \varpi_{j,d}^2 = \|\varpi\|_d^2\,. \tag{7.8}$$

To estimate the term $\Delta_2(\lambda)$ we recall that, for any $\varepsilon > 0$ and any $x, y \in \mathbb{R}$

$$2xy \le \varepsilon x^2 + \varepsilon^{-1} y^2\,. \tag{7.9}$$

Therefore, for some $0 < \varepsilon < 1$,

$$|\Delta_2(\lambda)| \le \varepsilon\, \frac{b-a}{d} \sum_{j=1}^{d} \lambda^2(j)\, \eta_{j,d}^2 + \frac{\|\varpi\|_d^2}{\varepsilon} = \varepsilon P_d(\lambda) + \varepsilon \frac{|\mathbf{B}(\lambda^2)|}{\sqrt{d}} + \frac{\|\varpi\|_d^2}{\varepsilon}\,,$$

where the vector $\lambda^2 \in \Lambda_1$ as in (4.10). Thus, for any $\lambda \in [0,1]^d$,

$$|\Delta(\lambda)| \le \varepsilon P_d(\lambda) + \varepsilon \frac{\mathbf{B}(\lambda^2)|}{\sqrt{d}} + 2\varepsilon^{-1}\|\varpi\|_d^2\,. \tag{7.10}$$

Putting

$$M_1(\lambda) = 2\frac{\mathbf{B}(\lambda)}{\sqrt{d}} + 2\Delta(\lambda)\,,$$

we can rewrite the empirical risk (7.5) as

$$\mathrm{Err}_d(\lambda) = J_d(\lambda) + 2M(\lambda) + M_1(\lambda) + \|S\|_d^2 - \delta\, P_d(\lambda)\,. \tag{7.11}$$

From (7.10) we obtain

$$|M_1(\lambda)| \le 2\frac{|\mathbf{B}(\lambda)|}{\sqrt{d}} + 2\frac{|\mathbf{B}(\lambda^2)|}{\sqrt{d}} + 2\varepsilon P_d(\lambda) + 4\varepsilon^{-1}\|\varpi\|_d^2 .$$

Moreover, setting

$$\mathbf{B}^* = \sup_{\lambda\in\Lambda}\left(\frac{\mathbf{B}^2(\lambda)}{P_d(\lambda)} + \frac{\mathbf{B}^2(\lambda^2)}{P_d(\lambda^2)}\right)$$

and taking into account that $P_d(\lambda^2) \le P_d(\lambda)$ for any $\lambda \in \Lambda$, we get

$$2\frac{|\mathbf{B}(\lambda)|}{\sqrt{d}} + 2\,\frac{|\mathbf{B}(\lambda^2)|}{\sqrt{d}} \le 2\varepsilon P_d(\lambda) + \varepsilon^{-1}\frac{\mathbf{B}^*}{d} .$$

By choosing $\varepsilon = \delta/4$ we find

$$|M_1(\lambda)| \le \delta P_d(\lambda) + \frac{16}{\delta}\Upsilon_d , \quad \Upsilon_d = \frac{\mathbf{B}^*}{4d} + \|\varpi\|_d^2 . \tag{7.12}$$

Now from (7.11) we obtain that, for some fixed λ_0 from Λ,

$$\begin{aligned}\mathrm{Err}_d(\widehat{\lambda}) - \mathrm{Err}_d(\lambda_0) \;&=\; J_d(\widehat{\lambda}) \;-\; J_d(\lambda_0) + 2\,M(\widehat{\mu}) \\ &\quad + M_1(\widehat{\lambda}) - \delta P_d(\widehat{\lambda}) - M_1(\lambda_0) + \delta P_d(\lambda_0) ,\end{aligned}$$

where $\widehat{\mu} = \widehat{\lambda} - \lambda_0$. By the definition of $\widehat{\lambda}$ in (4.16) we obtain on the set Γ

$$\mathrm{Err}_d(\widehat{\lambda}) \le \mathrm{Err}_d(\lambda_0) \;+\; 2\,M(\widehat{\mu}) + 32\frac{\Upsilon_d}{\delta} + 2\delta P_d(\lambda_0) . \tag{7.13}$$

From (7.6) and (7.7) it follows that

$$\begin{aligned}\mathbf{E}_{p,S}\,\mathbf{1}_\Gamma \mathbf{B}^* &\le \sum_{\lambda\in\Lambda}\mathbf{E}_{p,S}\,\mathbf{1}_\Gamma\left(\frac{\mathbf{B}^2(\lambda)}{P_d(\lambda)} + \frac{\mathbf{B}^2(\lambda^2)}{P_d(\lambda^2)}\right) \\ &\le 10\sigma_{1,*}(b-a)\check{\mathbf{m}}\sum_{\lambda\in\Lambda}\left(\frac{P_{1,d}(\lambda)}{P_{0,d}(\lambda)} + \frac{P_{1,d}(\lambda^2)}{P_{0,d}(\lambda^2)}\right) \\ &= 20\check{\mathbf{m}}(b-a)\nu\,\overline{\sigma}_* .\end{aligned}$$

and $\overline{\sigma}_* = \dfrac{\sigma_{1,*}^2}{\sigma_{0,*}}$. Therefore, for $0 < \delta < 1$ this inequality allows to bound Υ_d as

$$\mathbf{E}_{p,S}\mathbf{1}_\Gamma\Upsilon_d \le \frac{5\check{\mathbf{m}}(b-a)\overline{\sigma}_*\nu}{d} + \mathbf{u}_d^* , \tag{7.14}$$

where $\mathbf{u}_d^*$ is given by (4.1).
Now we study the second term on the right-hand side of inequality (7.13). For any weight vector $\lambda \in \Lambda$ we set $\mu = \lambda - \lambda_0$. Then we decompose this term as

$$M(\mu) = Z(\mu) + V(\mu) \tag{7.15}$$

with

$$Z(\mu) = \sqrt{\frac{b-a}{d}} \sum_{j=1}^{d} \mu(j)\, \theta_{j,d} \eta_{j,d} \quad \text{and} \quad V(\mu) = \sum_{j=1}^{d} \mu(j)\, \theta_{j,d} \varpi_{j,d}\,.$$

We define now the weighted discrete Fourier transformation of S, i.e. we set

$$\check{S}_\mu = \sum_{j=1}^{d} \mu(j)\, \theta_{j,d} \phi_j\,. \tag{7.16}$$

Now by using Propositioin 6.2 we can estimate the term $Z(\mu)$ as

$$\mathbf{E}_{p,S} \mathbf{1}_\Gamma\, Z^2(\mu) \;\le\; \frac{\sigma_{1,*}(b-a)}{d}\, \|\check{S}_\mu\|_d^2 := \sigma_{1,*}(b-a)\mathbf{D}(\mu)\,. \tag{7.17}$$

Moreover, by the inequalities (7.9) with $\varepsilon = \delta$ and (7.8) we can estimate $V(\mu)$ as follows

$$2V(\mu) = 2\sum_{j=1}^{d} \mu(j)\, \theta_{j,d} \varpi_{j,d} \le \delta\, \|\check{S}_\mu\|_d^2 + \frac{\|\varpi\|_d^2}{\delta}\,. \tag{7.18}$$

Setting

$$Z^* = \sup_{\mu \in \Lambda - \lambda_0} \frac{Z^2(\mu)}{\mathbf{D}(\mu)}\,,$$

we obtain on the set Γ

$$2M(\mu) \le\; 2\delta \|\check{S}_\mu\|_d^2 + \frac{Z^*}{d\delta} + \frac{\|\varpi\|_d^2}{\delta}\,. \tag{7.19}$$

Note now that from (7.17) it follows that

$$\mathbf{E}_{p,S}\, \mathbf{1}_\Gamma\, Z^* \le \sum_{\mu \in \Lambda - \lambda_0} \frac{\mathbf{E}_{p,S}\, \mathbf{1}_\Gamma\, Z^2(\mu)}{\mathbf{D}(\mu)} \le \nu \sigma_{1,*}(b-a)\,. \tag{7.20}$$

Now we estimate the first term on the right-hand side of the inequality (7.19). On the set Γ we have

$$\|\check{S}_\mu\|_d^2 - \|\widehat{S}_\mu\|_d^2 = \sum_{j=1}^d \mu^2(j)(\theta_{j,d}^2 - \widehat{\theta}_{j,d}^2) \le -2\sum_{j=1}^d \mu^2(j)\,\theta_{j,d}\,\zeta_{j,d}$$

$$= -2Z_1(\mu) - 2V_1(\mu)\,, \tag{7.21}$$

where

$$Z_1(\mu) = \sqrt{\frac{b-a}{d}}\sum_{j=1}^d \mu^2(j)\theta_{j,d}\eta_{j,d} \quad \text{and} \quad V_1(\mu) = \sum_{j=1}^d \mu^2(j)\,\theta_{j,d}\varpi_{j,d}\,.$$

Taking into account that $|\mu(j)| \le 1$, similarly to inequality (7.17), we find

$$\mathbf{E}_{p,S}\,\mathbf{1}_\Gamma\, Z_1^2(\mu) \le \sigma_{1,*}\,\mathbf{D}(\mu)\,.$$

Moreover, for the random variable

$$Z_1^* = \sup_{\mu\in\Lambda-\lambda_0}\frac{Z_1^2(\mu)}{\mathbf{D}(\mu)}\,,$$

we obtain the same upper bound as in (7.20), i.e.

$$\mathbf{E}_{p,S}\,Z_1^*\,\mathbf{1}_\Gamma \le \nu\sigma_{1,*}(b-a)\,. \tag{7.22}$$

Furthermore, similarly to (7.18) we estimate the second term in (7.21) as

$$2|V_1(\mu)| \le \delta\|\check{S}_\mu\|_d^2 + \frac{\|\varpi\|_d^2}{\delta}\,.$$

Therefore, on the set Γ

$$\|\check{S}_\mu\|_d^2 \le \|\widehat{S}_\mu\|_d^2 + 2\delta\|\check{S}_\mu\|_d^2 + \frac{Z_1^*}{d\delta} + \frac{\|\varpi\|_d^2}{\delta}\,,$$

i.e.

$$\|\check{S}_\mu\|_d^2 \le \frac{1}{1-2\delta}\,\|\widehat{S}_\mu\|_d^2 + \frac{1}{(1-2\delta)\delta}\left(\frac{Z_1^*}{d} + \|\varpi\|_d^2\right)\,. \tag{7.23}$$

Using this inequality in (7.19) and putting $Z_2^* = Z^* + Z_1^*$ yield on the set Γ

$$2M(\widehat{\mu}) \le \frac{2\delta}{1-2\delta}\|\widehat{S}_{\widehat{\mu}}\|_d^2 + \frac{1}{\delta(1-2\delta)}\left(\frac{Z_2^*}{d} + \|\varpi\|_d^2\right)$$

$$\le \frac{4\delta(\mathrm{Err}_d(\widehat{\lambda}) + \mathrm{Err}_d(\lambda_0))}{1-2\delta} + \frac{1}{\delta(1-2\delta)}\left(\frac{Z_2^*}{d} + \|\varpi\|_d^2\right)\,.$$

Therefore from the preceding inequality and (7.13) we obtain

$$\mathrm{Err}_d(\widehat{\lambda})\mathbf{1}_\Gamma \le \frac{1+2\delta}{1-6\delta}\mathrm{Err}_d(\lambda_0)\mathbf{1}_\Gamma \ + \ \frac{32(1-2\delta)}{\delta(1-6\delta)}\Upsilon_d\mathbf{1}_\Gamma$$
$$+ \frac{1}{\delta(1-6\delta)}\left(\frac{Z_2^*}{d} + \|\varpi\|_d^2\right)\mathbf{1}_\Gamma + \frac{2\delta(1-2\delta)}{1-6\delta}\,P_d(\lambda_0)\mathbf{1}_\Gamma$$

and through the inequalities (7.14), (7.20) and (7.22) we estimate the empirical risk as

$$\mathbf{E}_{p,S}\mathrm{Err}_d(\widehat{\lambda})\mathbf{1}_\Gamma \le \frac{1+2\delta}{1-6\delta}\mathbf{E}_{p,S}\mathrm{Err}_d(\lambda_0)\mathbf{1}_\Gamma \ + \ \frac{32(1-2\delta)}{\delta(1-6\delta)}\left(\frac{5\check{\mathbf{m}}\overline{\sigma}_*\nu(b-a)}{d} + \mathbf{u}_d^*\right)$$
$$+ \frac{1}{\delta(1-6\delta)}\left(\frac{2\nu\sigma_{1,*}(b-a)}{d} + \mathbf{u}_d^*\right) + \frac{2\delta(1-2\delta)}{1-6\delta}\,\mathbf{E}_{p,S}\mathbf{1}_\Gamma P_d(\lambda_0)\,.$$

Taking into account that $\sigma_{*,1} \le \overline{\sigma}_*$ and that $1-6\delta > 1/2$ for $0 < \delta < 1/12$, we get

$$\mathbf{E}_{p,S}\mathrm{Err}_d(\widehat{\lambda})\mathbf{1}_\Gamma \le \frac{1+2\delta}{1-6\delta}\mathbf{E}_{p,S}\mathrm{Err}_d(\lambda_0)\mathbf{1}_\Gamma + \frac{320}{\delta}\left(\frac{(\check{\mathbf{m}}+1)\overline{\sigma}_*\nu(b-a)}{d} + \mathbf{u}_d^*\right)$$
$$+ \frac{2\delta(1-2\delta)}{1-6\delta}\,\mathbf{E}_{p,S}\mathbf{1}_\Gamma P_d(\lambda_0)\,.$$

By applying Lemma A.4 with $\varepsilon = 2\delta$ we get that

$$\mathbf{E}_{p,S}\mathrm{Err}_d(\widehat{\lambda})\mathbf{1}_\Gamma \le \frac{1+4\delta}{1-6\delta}\mathbf{E}_{p,S}\mathrm{Err}_d(\lambda_0)\mathbf{1}_\Gamma + \frac{320}{\delta}\left(\frac{(\check{\mathbf{m}}+1)\overline{\sigma}_*\nu(b-a)}{d} + 3\mathbf{u}_d^*\right)$$
$$+ 10\delta\sqrt{\sigma_{1,*}\check{\mathbf{m}}\mathbf{P}_{p,S}(\Gamma^c)}\,.$$

Taking into account the definition of $\check{\mathbf{m}}$ in (3.23) and that $\mathbf{m}_4^* \le 3\varsigma^2$, then by replacing

$$\mathbf{E}_{p,S}\mathrm{Err}_d(\widehat{\lambda})\mathbf{1}_\Gamma \quad \text{and} \quad \mathbf{E}_{p,S}\mathrm{Err}_d(\lambda_0)\mathbf{1}_\Gamma$$

by

$$\mathbf{E}_{p,S}\|\widehat{S}_* - S\|_d^2 - \|S\|_d^2\mathbf{P}_{p,S}(\Gamma^c) \quad \text{and} \quad \mathbf{E}_{p,S}\|\widehat{S}_{\lambda_0} - S\|_d^2 - \|S\|_d^2\mathbf{P}_{p,S}(\Gamma^c)$$

respectively, we come to the inequality (5.1). Hence Theorem 5.1. □

Acknowledgements. The last author was partially supported by by the Russian Federal Professor program (project no. 1.472.2016/1.4, the Ministry of Education and Science of the Russian Federation), the research project no. 2.3208.2017/4.6 (the Ministry of Education and Science of the Russian Federation), by RFBR Grant 16-01-00121 A and by "The Tomsk State University competitiveness improvement program" grant 8.1.18.2018.

A Appendix

A.1 Burkhölder inequality

We need the following from [25].

Proposition A.1. *Let $(M_k)_{1\le k\le n}$ be a martingale. Then for any $q>1$*

$$\mathbf{E}\,|M_n|^q \le \mathbf{b}_q^*\mathbf{E}\left(\sum_{j=1}^n (M_j - M_{j-1})^2\right)^{q/2}, \tag{A.1}$$

where the coefficient $\mathbf{b}_q^* = 18(q)^{3/2}/(q-1)^{1/2}$.

A.2 Properties of the sequential procedures

Lemma A.2. *The properties (3.23) hold for the random variables $(\eta_l)_{1\le l\le d}$ defined in (3.22).*

Proof. First, we set $\mathcal{F}_j = \sigma\{\xi_1,\dots,\xi_j\}$ for $1\le j\le n$ and as usual $\mathcal{F}_0 = \{\Omega,\emptyset\}$. Moreover, note that

$$\eta_l = \sum_{j=1}^n \check{t}_{l,j}\,\xi_j \quad\text{and}\quad \check{t}_{l,j} = \sigma_l^2\left(\mathbf{1}_{\{\iota_l\le j<\check{\tau}_l\}}\check{Q}_{l,j} + \mathbf{1}_{\{j=\check{\tau}_l\}}\check{\varkappa}_l\,\check{Q}_{l,\check{\tau}_l}\right).$$

Taking into account that $\check{t}_{l,j}$ is $\mathcal{F}_{j-1}$ - measurable for any $1\le j\le n$ and

$$\sum_{j=1}^n \check{t}_{l,j}^2 = \sigma_l^2\,.$$

Note also that $\mathcal{G}_l = \sigma\{\eta_1,\dots,\eta_{l-1},\sigma_l,\} \subset \mathcal{F}_{\iota_l}$. Noting that

$$\mathbf{E}\left(\eta_l|\mathcal{F}_{\iota_l}\right) = 0 \quad\text{and}\quad \mathbf{E}\left(\eta_l^2|\mathcal{F}_{\iota_l}\right) = 1\,,$$

we obtain the first two equalities in (3.23). As to the last inequality, note that through (A.1) we can write

$$\mathbf{E}_{p,S}\left(\left(\sum_{j=1}^n \check{t}_{l,j}\,\xi_j\right)^4 |\mathcal{F}_{\iota_l}\right) \le \mathbf{b}_4^*\,\mathbf{E}_{p,S}\left(\left(\sum_{j=1}^n \check{t}_{l,j}^2\,\xi_j^2\right)^2 |\mathcal{F}_{\iota_l}\right).$$

Now, note that

$$\left(\sum_{j=1}^{n} \check{t}_{l,j}^2\, \xi_j^2\right)^2 \le 2\sigma_l^4 + 2\left(\sum_{j=1}^{n} \check{t}_{l,j}^2\, \widetilde{\xi}_j\right)^2$$

where $\widetilde{\xi}_j = \xi_j^2 - 1$. Taking into account that

$$\mathbf{E}_{p,S}\left(\left(\sum_{j=1}^{n} \check{t}_{l,j}^2\, \widetilde{\xi}_j\right)^2 |\mathcal{F}_{\iota_l}\right) = \mathbf{E}_p\, \widetilde{\xi}_1^2 \sum_{j=1}^{n} \check{t}_{l,j}^4 \le \sigma_l^4\, \mathbf{E}_p\, \widetilde{\xi}_1^2\,,$$

we obtain the last bound in (3.23). Hence Lemma A.2.

□

A.3 Correlation inequality

Now we give the correlation inequality from [17].

Proposition A.3. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_j)_{1\le j\le n}, \mathbf{P})$ be a filtered probability space and $(X_j, \mathcal{F}_j)_{1\le j\le n}$ be a sequence of random variables such that for some $p \ge 2$*

$$\max_{1\le j\le n} \mathbf{E}\,|X_j|^p < \infty\,.$$

Define

$$b_{j,n}(p) = \left(\mathbf{E}\,(|X_j| \sum_{k=j}^{n} |\mathbf{E}\,(X_k|\mathcal{F}_j)|)^{p/2}\right)^{2/p}.$$

Then

$$\mathbf{E}\,|\sum_{j=1}^{n} X_j\,|^p \le (2p)^{p/2}\left(\sum_{j=1}^{n} b_{j,n}(p)\right)^{p/2}. \tag{A.2}$$

A.4 Upper bound for the penalty term

Lemma A.4. *For sufficiently large n and $0 < \varepsilon < 1$,*

$$\mathbf{E}_{p,S} P_d(\lambda) \le \frac{1}{1-\varepsilon}\mathbf{E}_{p,S}\, Err_d(\lambda)\mathbf{1}_\Gamma + \frac{\mathbf{u}_d^*}{(1-\varepsilon)\varepsilon} + \frac{10}{1-\varepsilon}\sqrt{\sigma_{1,*}\check{\mathbf{m}}\mathbf{P}_{p,S}(\Gamma^c)}\,.$$

Proof. Indeed, by the definition of $\mathrm{Err}_d(\lambda)$ on the set Γ we have

$$\mathrm{Err}_d(\lambda) \;=\; \sum_{j=1}^{d} \big((\lambda(j)-1)\theta_{j,d} + \lambda(j)\zeta_{j,d}\big)^2$$

$$= \sum_{j=1}^{d} \left((\lambda(j)-1)\theta_{j,d} + \lambda(j)\varpi_{j,d} + \lambda(j)\sqrt{\frac{b-a}{d}}\eta_{j,d}\right)^2 .$$

Therefore, putting

$$I_1 = \sum_{j=1}^{d} \lambda(j)(\lambda(j)-1)\theta_{j,d}\eta_{j,d} \quad \text{and} \quad I_2 = \sum_{j=1}^{d} \lambda^2(j)\varpi_{j,d}\eta_{j,d}\,,$$

we get on the set Γ the following lower bound for the empirical risk

$$\mathrm{Err}_d(\lambda) \ge \frac{b-a}{d}\sum_{j=1}^{d} \lambda^2(j)\,\eta_{j,d}^2 + 2\sqrt{\frac{b-a}{d}}\,I_1 + 2\sqrt{\frac{b-a}{d}}\,I_2\,.$$

Taking into account that for $0<\varepsilon<1$,

$$2\sqrt{\frac{b-a}{d}}|I_2| \le \varepsilon\,\frac{b-a}{d}\sum_{j=1}^{d}\lambda^2(j)\eta_{j,d}^2 + \frac{\|\varpi\|_d^2}{\varepsilon}\,,$$

we get

$$\mathrm{Err}_d(\lambda) \;\ge\; (1-\varepsilon)\,\frac{b-a}{d}\sum_{j=1}^{d}\lambda^2(j)\,\eta_{j,d}^2 + 2\sqrt{\frac{b-a}{d}}\,I_1 - \frac{\|\varpi\|_d^2}{\varepsilon}\,. \tag{A.3}$$

Let us consider the first term in (A.3), then we have

$$\mathbf{E}_{p,S}\mathbf{1}_\Gamma \sum_{j=1}^{d}\lambda^2(j)\,\eta_{j,d}^2 = \mathbf{E}_{p,S}\sum_{j=1}^{d}\lambda^2(j)\,\mathbf{s}_{j,d} - \mathbf{E}_{p,S}\mathbf{1}_{\Gamma^c}\sum_{j=1}^{d}\lambda^2(j)\,\eta_{j,d}^2\,.$$

Using the correlation inequality (A.2) and the upper bound for the fourth moment in (3.23) we obtain

$$\mathbf{E}_{p,S}\eta_{j,d}^4 \le 64\breve{\mathbf{m}}\,\sigma_{1,*}^2\,. \tag{A.4}$$

This implies

$$\mathbf{E}_{p,S}\mathbf{1}_\Gamma \sum_{j=1}^{d}\lambda^2(j)\,\eta_{j,d}^2 \ge \mathbf{E}_{p,S}\sum_{j=1}^{d}\lambda^2(j)\,\mathbf{s}_{j,d} - 8\sigma_{1,*}\,d\,\sqrt{\breve{\mathbf{m}}\,\mathbf{P}_{p,S}(\Gamma^c)}\,. \tag{A.5}$$

Therefore, we obtain

$$\frac{b-a}{d}\mathbf{E}_{p,S}\mathbf{1}_{\Gamma}\sum_{j=1}^{d}\lambda^2(j)\,\eta_{j,d}^2 \geq \mathbf{E}_{p,S}\,P_d(\lambda) - 8(b-a)\sigma_{1,*}\sqrt{\breve{\mathbf{m}}\mathbf{P}_{p,S}(\Gamma^c)}\,.$$

Moreover, taking into account that $\mathbf{E}_{p,S}I_1 = 0$ and in view of Proposition (6.2)

$$\mathbf{E}_{p,S}I_1^2 \leq \sigma_{1,*}\,\|S\|_d^2\,.$$

So, recalling that that $\|S\|_d \leq b-a$, we estimate $\mathbf{E}_{p,S}I_1\mathbf{1}_{\Gamma}$ as

$$\left|\mathbf{E}_{p,S}\,I_1\mathbf{1}_{\Gamma}\right| = \left|\mathbf{E}_{p,S}\,I_1\mathbf{1}_{\Gamma^c}\right| \leq \sqrt{\sigma_{1,*}}\sqrt{\mathbf{P}_{p,S}(\Gamma^c)}\,.$$

Hence Lemma A.4. □

A.5 Properties of the model (1.1)

Lemma A.5. *For all $t \in \mathbb{N}^*$ and $0 < \epsilon < 1$, the random variables y_k in (1.1) satisfy the following :*

$$\sup_{n\geq 1}\sup_{0\leq k\leq n}\sup_{S\in\Theta_{\epsilon,L}}\mathbf{E}_{p,S}\,y_k^{2t} < \infty. \tag{A.6}$$

Proof. This lemma is shown in [2] (Lemma A.1). □

We set

$$\Upsilon_{m_0,m_1}(z_l) = \frac{1}{m_1-m_0}\sum_{j=m_0+1}^{m_1} y_j^2 - \frac{1}{\gamma_l}\,, \tag{A.7}$$

where $(k_{1,l}-2)_+ \leq m_0 < m_1 \leq k_{2,l}$, $(a)_+$ is positive part of a and γ_l is defined in (3.10).

Lemma A.6. *Assume that the bounds m_0 and m_1 in (A.7) are such that for some $0 < \epsilon_1 < 1/2$*

$$\liminf_{n\to\infty} n^{\epsilon_1}(m_1-m_0) > 0\,.$$

Then, for any $\mathbf{b} > 0$

$$\lim_{n\to\infty} n^{\mathbf{b}}\max_{1\leq l\leq d}\sup_{S\in\Theta_{\epsilon,L}}\sup_{p\in\mathcal{P}}\mathbf{P}_{p,S}\left(\left|\Upsilon_{m_0,m_1}(z_l)\right| > \epsilon_0\right) = 0\,, \tag{A.8}$$

where $\epsilon_0 = \epsilon_0(n) \to 0$ as $n \to \infty$ is such that $\lim_{n\to\infty} n^{\check{\delta}}\epsilon_0 = \infty$ for any $\check{\delta} > 0$.

Proof. This lemma is shown in [2] (Lemma A.2). □

A.6 Properties of the norms

Lemma A.7. *Let f be an absolutely continuous $[a,b] \to \mathbb{R}$ function with $\|\dot{f}\| < \infty$ and g be a simple $[a,b] \to \mathbb{R}$ function of the form*

$$g(t) = \sum_{j=1}^{p} c_j \, \chi_{(t_{j-1},t_j]}(t),$$

where c_j are some constants. Then for all $\widetilde{\varepsilon} > 0$, the function $\Delta = f - g$ satisfies the following inequalities

$$\|\Delta\|^2 \leq (1+\widetilde{\varepsilon})\|\Delta\|_d^2 + \left(1 + \frac{1}{\widetilde{\varepsilon}}\right) \frac{\|\dot{f}\|^2}{d^2}(b-a)^2 \,,$$

and

$$\|\Delta\|_d^2 \leq (1+\widetilde{\varepsilon})\|\Delta\|^2 + \left(1 + \frac{1}{\widetilde{\varepsilon}}\right) \frac{\|\dot{f}\|^2}{d^2}(b-a)^2 \,.$$

Proof. Lemma A.7 is proven in [21]. (Lemma A.2.) □

References

[1] Anderson, T.W. : *The Statistical Analysis of Time Series,* New York : *Wiley.*, 1994.

[2] Arkoun, O (2011): Sequential Adaptive Estimators in Nonparametric Autoregressive Models, *Sequential Analysis* 30: 228-246.

[3] Arkoun, O. and Pergamenchtchikov, S. (2008) : Nonparametric Estimation for an Autoregressive Model, *Journal of Mathematics and Mechanics of Tomsk State University* 2: 20-30.

[4] Arkoun, O. and Pergamenchtchikov, S. (2016) : Sequential Robust Estimation for Nonparametric Autoregressive Models. - *Sequential Analysis*, 2016, **35** (4), 489 – 515

[5] Baron, A., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, **113**, 301-413.

[6] Belitser, E. (2000) : Recursive Estimation of a Drifted Autoregressive Parameter, *The Annals of Statistics* 26: 860-870.

[7] Dahlhaus, R. (1996) : Maximum Likelihood Estimation and Model Selection for Locally Stationary Processes, *Journal of Nonparametric Statistics* 6: 171-191.

[8] Dahlhaus, R. (1996) : On the Kullback-Leibler Information Divergence of Locally Stationary Processes, *Stochastic Processes and their Applications* 62: 139-168.

[9] Fan, J. and Zhang, W. (2008) : Statistical Methods with Varying Coefficient Models, *Statistics and Its Interface* 1: 179-195.

[10] Fourdrinier, D. and Pergamenshchikov, S. M. (2007) Improved model selection method for a regression function with dependent noise. - *Annals of the Institute of Statistical Mathematics*, 59, p. 435-464

[11] Galtchouk, L. and Pergamenshchikov, S. (2005) : Nonparametric Sequential Minimax Estimation of the Drift Coefficient in Diffusion Processes, *Sequential Analysis* 24: 303-330.

[12] Galtchouk, L. and Pergamenshchikov, S. (2006) : Asymptotically Efficient Estimates for Nonparametric Regression Models, *Statistics and Probability Letters* 76 : 852-860.

[13] Galtchouk, L. and Pergamenshchikov, S. (2006) : Asymptotically Efficient Sequential Kernel Estimates of the Drift Coefficient in Ergodic Diffusion Processes, *Statistical Inference for Stochastic Processes* 9: 1-16

[14] Galtchouk, L.I. and Pergamenshchikov, S.M. (2009) Sharp nonasymptotic oracle inequalities for nonparametric heteroscedastic regression models. *Journal of Nonparametric Statistics*, **21**, 1-16.

[15] Galtchouk, L.I. and Pergamenshchikov, S.M. (2009) Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression. *Journal of the Korean Statistical Society*, **38** (4), 305 - 322.

[16] Galtchouk, L. and Pergamenshchikov, S. (2011) : Adaptive Sequential Estimation for Ergodic Diffusion Processes in Quadratic Metric, *Journal of Nonparametric Statistics* 23: 255-285.

[17] Galtchouk, L. and Pergamenshchikov, S. (2013) Uniform concentration inequality for ergodic diffusion processes observe at discrete times, *Stochastic processes and their applications*, **123**, 91 - 109.

[18] Galtchouk, L. and Pergamenshchikov, S. (2015) Efficient pointwise estimation based on discrete data in ergodic nonparametric diffusions. *Bernoulli*, **21** (4), 2569 - 2594.

[19] Konev, V. and Pergamenshchikov, S. (1984): Estimate of the Number of Observations in Sequential Identification of the Parameters of Dynamical Systems, *Avtomat. i Telemekh* 12: 56-63.

[20] V.V. Konev and S.M. Pergamenshchikov. Efficient robust nonparametric estimation in a semimartingale regression model. *Ann. Inst. Henri Poincaré Probab. Stat.*, **48** (4), 2012, 1217–1244.

[21] V.V. Konev and S.M. Pergamenshchikov. Robust model selection for a semimartingale continuous time regression from discrete data. *Stochastic processes and their applications*, **125**, 2015, 294 – 326.

[22] Konev V.V. On One Property of Martingales with Conditionally Gaussian Increments and Its Application in the Theory of Nonasymptotic Inference. *Doklady Mathematics*, 2016. Vol. 94, No 3. P. 1-5.

[23] Luo, X. H., Yang, Z. H. and Zhou, Y. (2009) : Nonparametric Estimation of the Production Function with Time-Varying Elasticity Coefficients, *Systems Engineering-Theory & Practice* 29: 144-149.

[24] Moulines, E., Priouret, P. and Roueff, F. (2005): On Recursive Estimation for Time Varying Autoregressive Processes, *The Annals of Statistics* 33: 2610-2654.

[25] Shiryaev, A. N. : *Probability, New York: Springer-Verlag, 2004.*