Minimax-Rate Adaptive Nonparametric Regression with Unknown Correlations of Errors

Yang Guowu and Yang Yuhong

Citation: SCIENCE CHINA Mathematics ; doi: 10.1007/s11425-018-9394-x

View online: http://engine.scichina.com/doi/10.1007/s11425-018-9394-x

Published by the Science China Press

SCIENCE CHINA
Mathematics

• ARTICLES • doi:

Minimax-Rate Adaptive Nonparametric Regression with Unknown Correlations of Errors

Guowu Yang[1] & Yuhong Yang[2,*]

[1] *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu* 611731, *China;*
[2]*School of Statistics, Universit of Minnesota, Minneapolis* 55455, *U.S.A;*
Email:guowu@uestc.edu.cn,yangx374@umn.edu

Received ; accepted

Abstract Minimax-rate adaptive nonparametric regression has been intensively studied under the assumption of independent or uncorrelated errors in the literature. In many applications, however, the errors are dependent, including both short- and long-range dependent situations. In such a case, adaptation with respect to the unknown dependence is important. We present a general result in this direction under Gaussian errors. It is assumed that the covariance matrix of the errors is known to be in a list of specifications possibly including independence, short-range dependence and long-range dependence as well. The regression function is known to be in one of a countable (or uncountable but well structured) collection of function classes. Adaptive estimators are constructed to attain the minimax rate of convergence automatically for each function class under each correlation specification in the corresponding lists.

Keywords nonparametric regression, adaptive estimation, long-range dependence, rate of convergence

MSC(2010) 62G08, 62C20

Citation: Guowu Yang, Yuhong Yang. SCIENCE CHINA Mathematics journal sample. Sci China Math, 2017, 60, doi: 10.1007/s11425-000-0000-0

1 Introduction

Adaptive function estimation has been extensively studied since the pioneering work of Efroimovish and Pinsker [15]. Minimax-rate adaptive estimators have been constructed for familiar function classes in various important and interesting settings. While achieving the minimax rate of convergence under a local loss is proven to be impossible (e.g., Lepskii [25], Brown and Low [8]), minimax-rate adaptation under a global loss (such as squared L_2 loss) has been obtained using a variety of tools for various function classes. The strategies that have been successfully proposed to construct adaptive estimators include wavelet thresholding (e.g., Donoho and Johnstone [12]), kernel estimators with automated bandwidth selection (e.g., Härdle and Marron [18]), and model selection based on finite-dimensional approximating models for smooth function classes (see, e.g., Barron [4], Barron, Birgé and Massart [5], Yang [38], Baraud [3], Wegkamp [33] for some early references). Positive results on adaptive estimation over general function classes (smooth or not) under independent errors are in Yang [35]. See Wang et al [32] for references and interesting results on adaptive estimation based on the flexible and powerful approach of combining or aggregating estimators.

*Corresponding author

For nonparametric regression, rates of convergence of various kinds of estimators have been well-established (see, e.g., Zhao [39] for convergence of nearest neighbor estimators). Minimax-rate adaptive estimation has also been studied, but the investigation has almost exclusively focused on the case with independent errors. However, research in recent years indicates that regression with dependent errors is both theoretically interesting and practically important (see Opsomer et al [37] for a review of the topic). Indeed, in various disciplines (e.g., hydrology and finance), long-range dependence phenomena are well-known (see, e.g., [28], [6], [23]). Despite much increased technical difficulties in various ways for handling dependent errors (especially the long-range dependent cases), there has been a substantial progress in the literature.

The results include identification of minimax rate of convergence for infinite dimensional classes by Hall and Hart [16], Wang [31], Johnstone and Silverman [21], Efromovich [13] and Yang [36]. The first three papers show that under a one-dimensional fixed design, long-range dependence always damages the rate of convergence under the global L_2 type of loss for some classical function classes. Very interestingly, with a random design, however, under the assumption that the random errors are independent of the design variables, the latter two papers show that long-range dependence does not necessarily affect the rate of convergence, and even if it does, it damages the rate of convergence to a lesser degree. At first glance, this is rather surprising because for the estimation of the regression function, in the familiar nonparametric regression contexts, a fixed uniform design and a random design usually do not have any difference in terms of rate of convergence, which is apparently not true here. The key aspect is that for a fixed design, the correlations among the *adjacent* observations *in terms of the X value* (also in terms of observation order) are highest; in contrast, for the random design, the correlations among the *adjacent* observations *in terms of the observation order* (not in terms of the the X value) are highest. This subtle difference leads to drastically different rates of convergence.

More generally, regardless of the input dimension and smoothness nature of the regression function class, Yang [36] shows that under mild conditions, the minimax rate of convergence under the square L_2 loss is either the minimax rate for the same function class but under independent errors or the rate of convergence of the sample mean of the errors, whichever is worse.

Besides the derivation of the rates of convergence above, Hall et al [17], Efromovich [13], Johnstone and Silverman [21] constructed adaptive regression estimators for specific smoothness classes and under specific correlation structures. These results are very useful because the true smoothness parameters are unknowable in practice.

The present work continues in the direction of adaptive nonparametric function estimation under correlated random errors. We address the theoretical matter in a much greater generality in that the function classes and the possible correlation structures are general. Therefore, unlike previous work, different types of function classes and different kinds of correlation structures are allowed to be considered at the same time for more flexibility. Under mild conditions, we show that minimax rate adaptive estimators can be constructed to achieve the rate of convergence without knowing the class membership nor the correlations of the errors. Note that our construction of the adaptive estimators in this paper is mainly for theoretical understanding, and they may not be practical for implementation in real applications at this time.

1.1 Notations and setup

Consider the regression model

$$Y_i = u(X_i) + \varepsilon_i,\ i \geqslant 1.$$

Suppose the Gaussian errors ε_i, $i \geqslant 1$ have zero means and an unknown covariance matrix Ω. Note that we do not require that the errors are stationary. The explanatory variables X_i, $i \geqslant 1$, are defined on a measurable space $\mathcal{X}$ and are assumed to be i.i.d. with a known design density function $h(x)$ with respect to a σ-finite measure μ. We further assume that $\{X_i, i \geqslant 1\}$ are independent of the errors $\{\varepsilon_i, i \geqslant 1\}$. This is likely the case, for instance, when the random errors are due to the intrinsic nature of the measurement device that may induce a serial correlation in consecutive uses. It is readily seen that the above setup of

the regression function and the dependence of the errors is appropriate in terms of identifiability. Our goal is to estimate the regression function u based on data $(X_i, Y_i)_{i=1}^n$.

Let $\mathcal{U}$ be a class of regression functions. Let $\|\cdot\|$ denote the L_2 norm with respect to the distribution of X, i.e., $\| u \| = \left(\int u^2(x)h(x)\mu(dx)\right)^{1/2}$. Given $\mathcal{U}$ and Ω, the minimax risk for estimating $u \in \mathcal{U}$ under the squared L_2 loss is

$$R(\mathcal{U}; \Omega; n) = \inf_{\widehat{u}} \sup_{u \in \mathcal{U}} E \| \widehat{u} - u \|^2,$$

where the minimization is over all estimators of u based on $\{X_i, Y_i\}_{i=1}^n$ with the true regression in $\mathcal{U}$ and under the dependence Ω. This risk describes how well the regression function can be estimated uniformly over $\mathcal{U}$ when the random errors in the observations have the covariance matrix Ω. In our setting, rate of convergence of the minimax risk is generally identified in Yang [36] under minor additional conditions. Note that it may also be of interest to study the minimax risk over both the function class $\mathcal{U}$ and Ω in a class of covariance matrix, which is not handled in this work.

For the purpose of estimating u, we consider a list of function classes $\mathcal{L}_{class} = \{\mathcal{U}_\xi : \xi \in \Upsilon\}$. Here Υ can be either countable (possibly finite) or uncountable. The function classes are general and are not necessarily restricted to certain smoothness types. Some of them may be of the same type with different hyper-parameters and others may be drastically different.

Since the dependence of the errors is also typically unknown, it is natural to consider a covariance structure for the errors with parameters to be estimated based on the data. Both parametric and nonparametric forms can be considered. Here we consider a general list of classes of covariance matrices $\mathcal{L}_{depend} = \{\Xi_j : j \geqslant 1\}$. It may include both short and long range dependent cases as will be seen. The class $\Xi_j = \{\Omega(j, \theta) : \theta \in \Theta_j\}$ is typically a parametric family of covariance matrix (with Θ_j being a finite-dimensional parameter space), but it does not have to be.

Now the true unknown regression function is assumed to be in one of the classes in $\mathcal{L}_{class}$ and the dependence of the errors is from any one in $\mathcal{L}_{depend}$.

1.2 Example

Suppose we are estimating a d-dimensional regression function $f(x_1, ..., x_d)$ defined on $[0, 1]^d$. Here we assume that d is not a small integer, say $d = 10$ and thus we may face the well-known "curse of dimensionality". Since it is very hard to visually inspect the data in a moderate or high dimension, in general, it seems very unlikely that we can propose, with much confidence, a reasonable parametric family or a nonparametric class for the estimation of f. Thus it seems very natural to consider various candidates. In the same vein, different dependence structures can be considered to better capture the true dependence of the random errors.

We choose $\mathcal{L}_{class}$ to include:

1. Besov classes of full-dimension d;
2. Besov classes with different interaction orders;
3. A neural network class.

We choose $\mathcal{L}_{depend}$ to include:

1. Independence;
2. Short-range dependence;
3. Long-range dependence;
4. Alternating dependence.

Note that in the above situation, determining which function class in $\mathcal{L}_{class}$ and which dependence structure in $\mathcal{L}_{depend}$ together best characterize the data at hand can be very difficult. It is desirable to have an adaptive estimator that will work well for all the plausible scenarios (i.e., any combination of choices of the function class and the correlation structure). There are two natural and closely related approaches that can be considered for deriving adaptive estimators in our situation, one based on a selection rule that adaptively chooses a proper function class and a correlation structure, and the other based on combining non-adaptive procedures. We will focus on the latter approach with an appropriate weighting of the candidate regression procedures.

The example will be provided with more details and treated for adapatation in Section 5.

1.3 Question of interest

We pose the following general question. Can one construct an adaptive estimator that achieves the minimax rate of convergence without knowing the true $\mathcal{U}$ and Ω in the corresponding lists? Specifically, can one construct an estimator $\widehat{u}_n$ based on the data $Z^n = (X_i, Y_i)_{i=1}^n$ only so that for each $\mathcal{U} \in \mathcal{L}_{class}$ and Ω as a member of any class in $\mathcal{L}_{depend}$, there exists a constant C (allowed to depend on $\mathcal{U}$ and Ω) so that

$$\sup_n \frac{\sup_{u\in\mathcal{U}} E \parallel \widehat{u}_n - u \parallel^2}{R(\mathcal{U};\Omega;n)} \leqslant C(\mathcal{U};\Omega)?$$

Such an estimator will be said to be minimax-rate adaptive in terms of both dependence (correlation) of the errors and the function class.

We will provide a positive answer to the above question under some mild conditions.

The rest of the paper is organized as follows. In Section 2, we provide some preliminaries for our main results. In Section 3, we consider minimax-rate adaptation over a countable collection of function classes and a list of dependence of the errors. In Section 4, an extension to the case of an uncountable collection of function classes is given. In Section 5, we give an example to illustrate the application of the main result. Conclusion/discussion follows in Section 6. The proofs of the technical results are in Section 7.

2 Some preliminaries

2.1 Metric entropy and its role in determining the rate of convergence

It is known that metric entropy (Kolmogorov and Tihomirov [22]) determines the minimax rate of convergence for nonparametric function estimation (see Le Cam [24], Birgé [7], Yang and Barron [38]). A finite subset N_ϵ is called an ϵ-packing set in $\mathcal{U}$ under a distance d if $d(u,v) > \epsilon$ for any $u, v \in N_\epsilon$ with $u \neq v$. Let $M_2(\epsilon;\mathcal{U})$ be the maximal logarithm of the cardinality of any ϵ-packing set under d_2, the distance induced by the $L_2(h)$-norm $\parallel \cdot \parallel$ defined earlier. The asymptotic behavior of $M_2(\epsilon;\mathcal{U})$ when $\epsilon \to 0$ reflects how massive the class $\mathcal{U}$ is under the given distance. We call $M_2(\epsilon;\mathcal{U})$ the packing ϵ-entropy or simply the metric entropy of $\mathcal{U}$ and ϵ is called the packing radius.

Assume $M_2(\epsilon;\mathcal{U}) < \infty$ holds for every $\epsilon > 0$ and for every $\mathcal{U}$ in $\mathcal{L}_{class}$ (which necessarily requires that $\mathcal{U}$ to be bounded in $L_2(h)$ norm). Assume also $M_2(\epsilon;\mathcal{U}) \to \infty$ as $\epsilon \to 0$ (which excludes the trivial case when $\mathcal{U}$ is finite). These conditions are satisfied if $\mathcal{U}$ is not finite, separable, and compact in the $L_2(h)$ norm. For the target function class $\mathcal{U}$, we assume that $\sup_{u\in\mathcal{U}} \parallel u \parallel_\infty \leqslant L < \infty$ for all $\mathcal{U}$ in $\mathcal{L}_{class}$, i.e., the function classes are uniformly bounded throughout the paper.

For most function classes, the metric entropies are known only up to orders. For this reason, we assume that $M(\epsilon;\mathcal{U})$ is an available non-increasing function known to be of order $M_2(\epsilon;\mathcal{U})$. As in [38], we call a class $\mathcal{U}$ rich if for some constant $0 < \tau < 1$,

$$\liminf_{\epsilon\to 0} M(\tau\epsilon;\mathcal{U})/M(\epsilon;\mathcal{U}) > 1. \tag{2.1}$$

This condition is met for typical nonparametric classes (see [38] for more discussions), for which the metric entropy is usually of order $\epsilon^{-\alpha}\log(1/\epsilon)^{\beta}$ for some $\alpha > 0$ and $\beta \in R$. Through out the paper, we assume all the function classes being considered are rich.

In the rest of the paper, when comparing two sequences of positive numbers, the notations $\asymp$, $\preceq$ and $\succeq$ mean that the ratio of the right hand side and the left hand side is bounded both above (from infinity) and below (away from zero), bounded above (from infinity), and bounded below (away from zero), respectively.

2.2 A distance between covariance matrices and the covering entropy of a class of dependence

To address the additional difficulty in regression estimation due to the unknown dependence of the errors, we consider a distance on covariance matrices. Let A and B be two $n \times n$ symmetric matrices for $n \geqslant 1$.

Definition 1: The largest absolute value of the eigenvalues of $A - B$, denoted by $\zeta(A, B)$, is said to be the eigenvalue distance between A and B.

Clearly, the distance ζ is a metric on the space of all $n \times n$ matrices for any $n \geqslant 1$, and it is induced by the matrix spectral norm.

For a family of $n \times n$ symmetric matrices, for measuring the largeness of the collection, consider the covering entropy under the eigenvalue distance.

Let $\{\Omega(j,\theta) : \theta \in \Theta_j\}$ be the j-th family of dependence (covariance matrix) of the errors, and let $\Omega_n(j,\theta)$ be the finite section of $\Omega(j,\theta)$ of size n. Let $M_\zeta(\epsilon; \Xi_j; n)$, or simply $M(\epsilon; \Theta_j)$ for simplicity, denote the covering entropy of the class of $n \times n$ matrices $\{\Omega_n(j,\theta) : \theta \in \Theta_j\}$ under the distance ζ. We assume that $M(\epsilon; \Theta_j)$ is finite for each $\epsilon > 0$. As will be seen, as long as the classes of dependence are not too large compared to the regression function classes, not knowing the covariance matrix of the random errors does not hurt the rate of convergence for adaptive estimation of the regression function. In particular, if the correlation families are parametric, under mild conditions, minimax-rate adaptation for regression estimation can be achieved.

2.3 Some conditions on the correlations of the errors

We do not require stationarity of the errors. For a given Ω (the infinite dimensional covariance matrix), let Ω_n be its finite section of size n, i.e., the covariance matrix of $(\varepsilon_1, ..., \varepsilon_n)$. Let $\sigma_i^2(j,\theta)$ denote $Var(\varepsilon_i)$, $i \geqslant 1$ under the covariance matrix $\Omega(j,\theta)$. Let $\widetilde{\Omega}(j,\theta) = \Omega(j,\theta)/\sigma_1^2(j,\theta)$. Note that the first diagnal element of $\widetilde{\Omega}(j,\theta)$ is always 1.

We assume the following conditions on $\mathcal{L}_{depend}$ hold.

Assumption 1:

1. For each $j \geqslant 1$ and $\theta \in \Theta_j$, $\sup_i \sigma_i^2(j,\theta) \leqslant \overline{\sigma}^2$ for some known $0 < \overline{\sigma}^2 < \infty$.

2. For each $j \geqslant 1$, there exists a positive constant $\underline{\lambda}_j$ such that the smallest eigenvalue of $\widetilde{\Omega}_n(j,\theta)$ is lower bounded by $\underline{\lambda}_j$ for all $\theta \in \Theta_j$ and all $n \geqslant 1$.

When the errors are stationary, the condition $\sup_i \sigma_i^2(j,\theta) \leqslant \overline{\sigma}^2$ obviously simplifies to $\sigma^2(j,\theta) \leqslant \overline{\sigma}^2$, where $\sigma^2(j,\theta)$ is the common variance of the errors. The second condition prevents the errors to be increasingly colinear. Under stationarity, a sufficient condition to ensure this requirement is that the spectral density of the error series exists and is uniformly bounded away from zero. In particular, it is satisfied if $\Omega_n(j,\theta)$ can be expressed as the sum of two components $\Omega_n(j,\theta) = \Omega_n^{(1)} + \Omega_n^{(2)}(j,\theta)$, where $\Omega_n^{(1)} = \text{diag}\,(\omega_{1,n}, ..., \omega_{n,n})$ with $\min_{1 \leqslant i \leqslant n} \omega_{i,n} \geqslant c > 0$ for some constant $c > 0$ independent of n and θ, and $\Omega_n^{(2)}(j,\theta)$ is nonnegative definite. For non-stationary errors, the checking of Assumption 1 may demand case-by-case more technically involved analysis.

3 Adaptation with respect to countable lists of function classes and dependence

We consider in this section the case that $\mathcal{L}_{class}$ is countable, i.e., $\mathcal{L}_{class} = \{\mathcal{U}_i, i \geqslant 1\}$. Choose $\epsilon_{n,i}$ such that

$$M(\epsilon_{n,i}; \mathcal{U}_i) = n\epsilon_{n,i}^2. \tag{3.1}$$

As shown in Yang [36], $\epsilon_{n,i}^2$ characterizes the part of difficulty in estimating u from the massiveness of $\mathcal{U}_i$. For each class in $\mathcal{L}_{depend}$, let $\varsigma_{n,j}$ be chosen such that $M(\varsigma_{n,j}; \Theta_j) = n\varsigma_{n,j}$. For a parametric family of dependence, $M(\varsigma; \Theta_j)$ is usually of order $\log(1/\varsigma)$. Then $\varsigma_{n,j}$ defined above is of order $(\log n)/n$. Let Ω_n be the true covariance matrix of the random errors ε_i, $1 \leqslant i \leqslant n$, i.e., the finite section of the true infinite-dimensional covariance matrix Ω. Let $\mathbf{1}$ denote the column vector with all entries 1.

THEOREM 1**:** *Suppose Assumption 1 is satisfied.* Then we can construct an estimator $\widehat{u}_n$ based on $(X_i, Y_i)_{i=1}^n$ such that under each Ω in the list $\mathcal{L}_{depend}$, for every $\mathcal{U}_i \in \mathcal{L}_{class}$, we have

$$\sup_{u \in \mathcal{U}_i} E \parallel \widehat{u}_n - u \parallel^2 \preceq \frac{\log n}{n} + \varsigma_{n,j} + \max\left(\left(\mathbf{1}'\mathbf{\Omega}_n\mathbf{1}\right)/n^2, \epsilon_{n,i}^2\right).$$

REMARK: From the proof of Theorem 1, it can be seen that the dimension d of the covariate vector (i.e., the numbers of explanatory variables) is not playing any directly important role beyond its influence on the orders of the metric entropy of the regression function classes. Thus similar results for high-dimensional regression can be stated in terms of the metric entropy orders of the high-dimensional function classes.

REMARK: As already mentioned before Theorem 1, for a parametric family of dependence, $\varsigma_{n,j}$ in the upper bound is typically of order $(\log n)/n$, which usually does not affect the minimax rate of convergence for a nonparametric function class $\mathcal{U}_i$. See [36] for examples of the rate $\mathbf{1}'\mathbf{\Omega}_n\mathbf{1}/n^2$. Yang and Barron [38] give a number of examples of the order of $\epsilon_{n,i}^2$.

Under some additional mild conditions, the upper bound in Theorem 1 is in fact the minimax rate of convergence of the class $\mathcal{U}_i$ under the dependence Ω (see [36]). For instance, assume further that

$$\left(\mathbf{1}'\Omega_n^{-1}\mathbf{1}\right)\left(\mathbf{1}'\Omega_n\mathbf{1}\right) \asymp n^2 \text{ and } \mathbf{1}'\Omega_n\mathbf{1} \succeq n, \tag{3.2}$$

then given $\mathcal{U}$ and Ω, the minimax rate of convergence is

$$R(\mathcal{U}; \Omega; n) \asymp \max\left(\left(\mathbf{1}'\Omega_n\mathbf{1}\right)/n^2, \epsilon_{n,i}^2\right).$$

The condition is satisfied by familiar short and long-range dependences (see [36]). In particular, for the long range dependence (which has autocovariance $Cov(\varepsilon_i, \varepsilon_{i+j}) \sim c|j|^{-\tau}$ for some $c > 0$ and $0 < \tau < 1$), the condition $\left(\mathbf{1}'\Omega_n^{-1}\mathbf{1}\right)\left(\mathbf{1}'\Omega_n\mathbf{1}\right) \asymp n^2$ is satisfied and $\mathbf{1}'\Omega_n\mathbf{1}/n \asymp n^{-\tau}$ (see, e.g., [2], [16]). We have the following corollary.

COROLLARY 1**:** Under the assumptions for Theorem 1, if a function class $\mathcal{U}_i$ in $\mathcal{L}_{class}$ is rich and the true covariance matrix $\Omega \in \Xi_j$ satisfies $Tr\left(\Omega_n^{-1}\right) \asymp n$ and (3.2), and additionally $M(\varsigma; \Theta_j)$ is of order $\log(1/\varsigma)$, then the estimator $\widehat{u}_n$ achieves the minimax rate of convergence $R(\mathcal{U}_i; \Omega; n)$.

For Corollary 1, we have assumed that $Tr(\Omega_n^{-1})$ is of order n. For stationary situations, this condition may not be needed. Suppose the errors $\{\varepsilon_j, -\infty < j < \infty\}$ follow an infinite order Gaussian auto-regression $\sum_{k=-\infty}^{\infty} b_k\varepsilon_{k+j} = \zeta_j$, where ζ_j, $-\infty < j < \infty$ are i.i.d. from a standard normal distribution. The coefficients b_j's are assumed to be absolutely summable and satisfy $\int_0^\pi |b(\lambda)|^{-2}d\lambda < \infty$, where $b(\lambda) = \sum_{j=-\infty}^{\infty} b_j e^{ij\lambda}$. These conditions ensure invertibility of the auto-regression process to a moving average process. Let $r(j) = \sum_{k=-\infty}^{\infty} b_k b_{k+j}$, $-\infty < j < \infty$. Then the (k, j)-element of Ω is $r(j-k)$. Under these conditions, the minimax rate of convergence is showed to be max $\left(\left(\mathbf{1}'\Omega_n\mathbf{1}\right)/n^2, \epsilon_{n,i}^2\right)$ (see [36]). Thus the estimator in Theorem 1 converges at the minimax rate $R(\mathcal{U}_i; \Omega; n)$ adaptively without knowing which class contains u nor Ω.

We next give a result on adaptive regression estimation under more specific conditions on the correlation families. Here the parameter spaces Θ_j may or may not be compact.

Assumption 2:

For each $j \geqslant 1$, suppose Θ_j is a subset of R^{d_j}. Let $\widetilde{\Omega}_n(j,\theta) = (w_{il}(\theta))$ for $\theta \in \Theta_j$. Assume that there exist positive constants c_j and A_j such that $\max_{1\leqslant i\leqslant n, 1\leqslant l\leqslant n}|w_{il}(\theta) - w_{il}(\theta^{'})| \leqslant c_j n^{A_j} d(\theta,\theta^{'})$ holds for all $n \geqslant 1$ and $\theta, \theta^{'} \in \Theta_j$ where d denotes the Euclidean distance.

Assumption 2 can be directly verified for parametric families of correlations with the covariance matrix explicitly given. For example, consider a fractional Gaussian noise model with autocorrelation $r(k) = \frac{\sigma^2}{2}\left((k+1)^{2-\gamma} - 2k^{2-\gamma} + (k-1)^{2-\gamma}\right)$ for $0 < \gamma < 2$ (see [26]). When $0 < \gamma < 1$, the errors have short-range dependence; when $\gamma = 1$, the errors are uncorrelated; when $1 < \gamma < 2$, the errors have long-range dependence. It can be easily verified that this family of dependence satisfies Assumption 2.

THEOREM 2**:** *Assume Assumptions 1 and 2 are satisfied and (3.2) holds.* If the function classes $\mathcal{U}_i$ in $\mathcal{L}_{class}$ are rich, then a properly constructed combined estimator $\widehat{u}_n$ adaptively achieves the minimax rate of convergence for all combinations of the regression function class and the correlation structure in $\mathcal{L}_{class}$ and $\mathcal{L}_{depend}$.

For stationary error series, when the dependence is given in terms of the spectral density, the following condition is useful. We focus on long-range dependence here.

Let $f_\theta(\omega)$ be the spectral density of the error series, where $\theta \in \Theta \subset R^k$ $(1 \leqslant k < \infty)$ is unknown. We assume that there exists a continuous function $0 < \gamma(\theta) < 1$ such that $f_\theta(\omega) \sim |\omega|^{-\gamma(\theta)}$.

Assumption 3:

There exists a constant C such that

$$|f_\theta(\omega) - f_{\theta'}(\omega)| \leqslant C d(\theta,\theta^{'}) f_{\theta'}(\omega)$$

holds for all ω and all θ and $\theta^{'}$ with $\gamma(\theta) \leqslant \gamma(\theta^{'})$.

The condition was used in [10] for studying the maximum likelihood estimator for the parameters for a long-range dependence Gaussian process. It is satisfied for the case when $f_\gamma(\omega) = |1 - e^{i\omega}|^{-\gamma} f^*(\omega)$, where $0 < \gamma < 1$ and $f^*(\omega)$ is continuous and bounded away from zero. It includes fractional ARIMA cases, e.g., $f_\theta(\omega) = \frac{1}{2\pi}|1 - e^{i\omega}|^{-(1-\gamma)}$ (see, e.g., [14], [20]).

It can be easily verified that Assumption 3 implies Assumption 2.

4 Adaptation over an uncountable collection of function classes

We extend the results in the previous section to the case of an uncountable collection of regression function classes that have some mild structural properties.

Consider a collection of function classes $\{\mathcal{U}_\xi : \xi \in \Upsilon\}$, where Υ is a subset in a finite-dimensional Euclidean space R^m, $1 \leqslant m < \infty$. Assume that each function class in $\mathcal{L}_{class}$ is rich. Let $\| \xi \|_2 = \sqrt{\xi_1^2 + ... + \xi_m^2}$ denote the Euclidean norm of $\xi = (\xi_1, ..., \xi_m) \in R^m$. Assume that there is a partial order on the hyper-parameter space Υ and that the order of the hyper-parameters is in accordance with the order of the corresponding function classes in the sense that if $\xi_1 \lhd \xi_2$ then $\mathcal{U}_{\xi_1} \subset \mathcal{U}_{\xi_2}$.

Let $M(\epsilon;\xi)$ be a continuous upper bound (of the same order) on the metric entropy of the class $\mathcal{U}_\xi$ under the $L_2(h)$ distance. Let $\epsilon_{n,\xi}$ be determined by the equation

$$M(\epsilon_{n,\xi};\xi) = n\epsilon_{n,\xi}^2.$$

Consider the following discretization of R^m. For each $j \geqslant 1$, consider the dyadic grid $\{i2^{-j} : i \in \mathcal{Z}\}$ for each coordinate, where $\mathcal{Z}$ denotes the set of all integers. Let N_j denote the corresponding discrete set in R^m. Then let $Q = \cup_{j\geqslant 1} N_j$ be the overall discrete set of all the dyadic rational numbers.

Assumption 4: For each fixed Ω in each class in $\mathcal{L}_{depend}$, for each $\xi_0 \in \Upsilon$, there exist a sequence $\xi_n \in \Upsilon \cap N_{j_n}$ with $\xi_0 \lhd \xi_n$ for some j_n of order $\log n$ such that $\frac{\epsilon_{n,\xi_n}}{\epsilon_{n,\xi_0}}$ stays upper bounded.

Note that Assumption 4 is automatically satisfied for the case of a countable collection of functional classes. For function classes indexed by continuous hyperparameters, this condition is also satisfied for familiar smoothness classes (see [35], [37]).

THEOREM 3: *Under the assumptions for Theorem 2 and Assumption 4, we can construct an estimator such that it achieves the minimax rate of convergence adaptively over all function classes and under all the correlation structures considered.*

Example Consider one-dimensional Lipschitz classes on $[0,1]$ as follows. For positive constants C, $C_1, ..., C_r$, $r \geqslant 0$ being an integer, and $\rho \in (0,1]$, with $\alpha = r + \rho$, define

$$U(\alpha, C) = \{f : |f^{(k)}(x)| \leqslant C_k \text{ for } k = 0, 1, ..., r, \ |f^{(r)}(x) - f^{(r)}(y)| \leqslant C|x-y|^{\rho} \text{ for all } x, y \in [0,1]\}.$$

From [22], the metric entropy of $U(\alpha, C)$ is upper bounded by $A\left(\frac{1}{\epsilon}\right)^{1/\alpha}$ with A locally bounded in terms of the hyper-parameters. It is then straightforward to verify Assumption 4. Therefore, for these Lipschitz regression function classes, if the errors satisfy Assumptions 1-2 and (3.2), we have minimax-rate adaptive estimators over all the above Lipschitz regression function classes and all the correlation structures considered.

5 An example application

In this demonstration, the unknown regression function u is assumed to be uniformly bounded, and the explanatory variable X takes values in $[0,1]^d$ with a known design density (with respect to Lebesgue measure) bounded above and away from zero. In practical situations, especially when d is large, it is usually difficult to know the form of u. For high dimensional function estimation, suitable reduction of dimensionality may significantly improve estimation accuracy. We here consider two different ways of dimension reduction, namely, additive or low-order interaction modeling and neural network modeling. Coupled with the difficulty in modeling the regression function is the modeling of the correlations of the errors: it is generally difficult to know how the errors are related to each other.

We accordingly consider different scenarios in hope that some of them properly capture the characteristics of the data and give good regression estimation. We consider several function classes and a few correlation structures for the errors below.

5.1 Function classes

1. *Besov classes of full-dimension.* For $1 \leqslant \sigma, q \leqslant \infty$ and $\alpha/d > 1/q$, let $B_{q,\sigma}^{\alpha,d}(C)$ be the collections of all functions $g \in L_q[0,1]^d$ such that the Besov norm satisfies $\| g \|_{B_{q,\sigma}^{\alpha,d}} \leqslant C$ (see, e.g., [30], [11]). Besov classes are rich, providing a lot of flexibility (e.g., spatial inhomogeneity) for statistical function estimation. The minimax rate of convergence for estimating a function $u \in B_{q,\sigma}^{\alpha,d}(C)$ under the squared L_2 loss is known to be $n^{-2\alpha/(2\alpha+d)}$ (see, e.g., [12], [38]). For $d = 1$, Donoho and Johnstone [12] show that wavelet thresholding based estimators adaptively achieve the minimax rate of convergence.

2. *Besov classes with different interaction order.* Suppose d is relatively large. When the smoothness parameter α is small or moderate, the convergence rate $n^{-2\alpha/(2\alpha+d)}$ for the Besov class $B_{q,\sigma}^{\alpha,d}(C)$ is slow, which reflects the well-known "curse of dimensionality". To improve the rate of convergence, one may entertain various dimension reduction features. Here we consider Besov classes of different interaction orders as follows:

$$S_{q,\sigma}^{\alpha,1}(C) = \{\textstyle\sum_{i=1}^{d} g_i(x_i) : g_i \in B_{q,\sigma}^{\alpha,1}(C), 1 \leqslant i \leqslant d\}$$

$$S_{q,\sigma}^{\alpha,2}(C) = \{\textstyle\sum_{1 \leqslant i < j \leqslant d} g_{i,j}(x_i, x_j) : g_{i,j} \in B_{q,\sigma}^{\alpha,2}(C), 1 \leqslant i < j \leqslant d\}$$

$$\cdots$$

$$S_{q,\sigma}^{\alpha,d}(C) = B_{q,\sigma}^{\alpha,d}(C).$$

From above, it is clear that the simplest function class $S_{q,\sigma}^{\alpha,1}(C)$ contains additive functions (no interaction between the variables) and these classes have different effective input dimensions between 1 and d. From

[38], via simple metric entropy arguments, the minimax rate of convergence under the squared L_2 loss for estimating u in $S^{\alpha,r}_{q,\sigma}(C)$ is readily seen to be $n^{-2\alpha/(2\alpha+r)}$ for $1 \leqslant r \leqslant d$, which is suggested by the heuristic dimensionality reduction principle of Stone [29]. When r is small relative to d, the convergence rate is much faster compared to $n^{-2\alpha/(2\alpha+d)}$. For each r, one can consider tensor-product wavelets of different interaction orders and use thresholding to obtain an estimator that adaptively converges optimally for every $S^{\alpha,r}_{q,\sigma}(C)$ without knowing the hyper-parameters.

3. *A neural network class.* Let $N(C)$ be the closure in $L_2[0,1]^d$ of the set of all functions $g : R^d \to R$ of the form $g(x) = c_0 + \sum_i c_i m(v_i' x + b_i)$ (where the prime denotes the transpose), with $|c_0| + \sum_i |c_i| \leqslant C$, and $\|v_i\| = 1$, where m is the step function $m(t) = 1$ for $t \geqslant 0$, and $m(t) = 0$ for $t < 0$. The minimax rate for estimating $u \in N(C)$ under the squared L_2 loss is shown to be bounded between

$$n^{-(1+2/d)/(2+1/d)} (\log n)^{-(1+1/d)(1+2/d)/(2+1/d)} \text{ and } (n/\log n)^{-(1+1/d)/(2+1/d)} \tag{5.1}$$

(see [38]). When d is large, the rate is slightly better than $n^{-1/2}$ (independent of d), which avoids the "curse of dimensionality". Estimators at rate $O(\log n/n^{1/2})$ using finite-dimensional neural network models are in e.g., Barron [4].

5.2 Correlation structures

We consider stationary errors. Due to lack of knowledge on dependence, the collection of covariance matrix $\mathcal{L}_{depend}$ is chosen to include those corresponding to i.i.d., short-range dependence and long-range dependence as follows.

1. *Independence.*

2. *Short range dependence.* Consider exponentially decaying autocovariance $r(j) = \sigma^2\theta^{|j|}$ for an integer j, where $\sigma^2 > 0$ and $-1 < \theta < 1$ are unknown parameters.

3. *Long range dependence.* Suppose the spectral density of the error series satisfies that $f_\gamma(\omega) = |1 - e^{i\omega}|^{-\gamma} f^*(\omega)$, where $0 < \gamma < 1$ and $f^*(\omega)$ is a given continuous function bounded away from zero.

4. *Alternating dependence.* For the above long-range dependence, the errors are eventually positively correlated. Here consider $r(j) = \frac{(-1)^j\sigma^2}{2}\left((j+1)^{2-\gamma} - 2j^{2-\gamma} + (j-1)^{2-\gamma}\right)$ with unknown $\gamma \in (0,2)$ and $\sigma^2 > 0$. Note that the correlations are positive and negative in an alternating way. Because the covariances essentially cancel out even when $0 < \gamma < 1$, the rate of convergence for regression estimation is the same as under independent errors.

5.3 Adaptation over the function classes and the correlation structures

Applying the results in Section 4, we construct an adaptive estimator over the classes of regression functions and the specified dependence structures described above. If the errors are independent or short-range dependent (Cases 1, 2 and 4 above), then when u is in $B^{\alpha,d}_{q,\sigma}(C)$ with α relatively large compared to d, the risk converges at a good rate $O\left(n^{-2\alpha/(2\alpha+d)}\right)$; when u is in $S^{\alpha,r}_{q,\sigma}(C)$ for some small r, then the risk converges at a faster rate $O\left(n^{-2\alpha/(2\alpha+r)}\right)$; when u is not in any of these cases, but fortunately has the neural net representation, then the risk also converges at a good rate $O(\log n/n^{1/2})$. If the errors are long-range dependent (Case 3 above), then our adaptive estimator converges at the aforementioned rates for the function classes respectively, or the rate $\left(\mathbf{1}'\Omega_n\mathbf{1}\right)/n^2 \asymp n^{-\gamma}$, whichever is slower.

The key point here is that by our strategy of mixing estimators, the combined procedure automatically adapts to different scenarios for a good rate of convergence. That is, without knowing the covariance structure, nor which type of dimension reduction is appropriate for the underlying function and the smoothness parameters of the function class that contains the true regression function, we can do as well as if we knew them in advance.

6 Conclusion and discussion

In this work, adaptive regression with respect to function classes and correlation structures of the errors are considered with the emphasis that the errors are dependent and unknown to certain degree. Under the assumption that the errors are Gaussian and independent of the explanatory variables, we derived rate-optimal adaptation risk bounds and showed that minimax-rate adaptive estimation of the regression function is achievable in spite of unknown possibly long-range dependent errors. To highlight the roles of the regression function class and the correlations of the errors, we have not handled the issue of unknown design distribution of the covariates and adaptation with respect to it. Complete adaptation over regression function classes, error dependences and design distributions simultaneously is an interesting open problem. In addition, adaptation with respect to unknown relationship between the covariates and the random errors is also of interest for future research.

7 Proofs of the results

Before we prove Theorem 1, we give more notations. Let $Z = (X, Y)$, $z = (x, y)$, $z^n = (z_1, .., z_n)^{'}$, and similarly define y^n and x^n. Let $U^n = (u(X_1), ..., u(X_n))^{'}$ and $u^n = (u(x_1), ..., u(x_n))^{'}$.

7.1 Proof of Theorem 1

Let us outline our scheme for the construction of an adaptive estimator. For each choice of $\mathcal{U}$ and Ω pair, we construct a joint density on the product space of the observations, and then mix these densities with different $\mathcal{U}$ and Ω. The mixture will be shown to be suitably close to the unknown joint density of the data uniformly over the function classes and the dependences in a proper sense. The mixture will be used to construct the adaptive estimator in a rather delicate way.

Here we prove the result with $\mathcal{L}_{class}$ being countable, i.e., $\mathcal{L}_{class} = \{\mathcal{U}_i, i \geqslant 1\}$. We divide the proof into several steps.

7.1.1 Constructing a cover set for each $\mathcal{U}$ and Ω

Fix $\mathcal{U}$ and Ω for a moment. Let G_{ϵ_n} be an ϵ_n-net for $\mathcal{U}$ under the L_2 distance. Let

$$p_{u,\Omega}(z^n) = (\Pi_{i=1}^n h(x_i)) (2\pi)^{-n/2} |\Omega_n|^{-1/2} \exp\left(-(1/2)(y^n - u^n)^{'} \Omega_n^{-1} (y^n - u^n)\right).$$

Let $P_{Z^n,u,\Omega}$ denote the corresponding distribution. Then from Lemma 1 later in this section, for $\Omega \in \Xi_j$, under Assumption 1, we have

$$D(P_{Z^n,u,\Omega} \parallel P_{Z^n,v,\Omega}) = \frac{1}{2} E (u^n - v^n)^{'} \Omega_n^{-1} (u^n - v^n) \leqslant \frac{n}{2\underline{\lambda}_j \sigma_1^2(j,\theta)} \parallel u - v \parallel_h^2 . \tag{7.1}$$

Thus for any $u \in \mathcal{U}$ and $\Omega \in \Xi_j$, there exists $\widetilde{u} \in G_{\epsilon_n}$ such that

$$D(P_{Z^n,u,\Omega} \parallel P_{Z^n,\widetilde{u},\Omega}(z^n)) \leqslant \frac{n\epsilon_n^2}{2\underline{\lambda}_j \sigma_1^2(j,\theta)}. \tag{7.2}$$

7.1.2 Discretization of the dependence for each Ξ

Now we consider discretizing Θ_j by an $\varsigma_{n,j}$-net $\Theta_j^{(n)}$ with an appropriate choice of $\varsigma_{n,j} > 0$ to be given later. Then for any $\theta_0 \in \Theta_j$, there exists $\theta_1 \in \Theta_j$ such that $\zeta(\widetilde{\Omega}_n(j,\theta_0), \widetilde{\Omega}_n(j,\theta_1)) \leqslant \varsigma_{n,j}$. It follows from Lemmas 2 and 3 that

$$|tr\left(\widetilde{\Omega}_n(j,\theta_0)\left(\widetilde{\Omega}_n(j,\theta_1)\right)^{-1}\right) - n| \leqslant n\varsigma_{n,j}/\underline{\lambda}_j.$$

$$|\log\det\left(\widetilde{\Omega}_n(j,\theta_0)\left(\widetilde{\Omega}_n(j,\theta_1)\right)^{-1}\right)| \leqslant n\varsigma_{n,j}/\underline{\lambda}_j,$$

where $\underline{\lambda}_j$ is a lower bound on the smallest eigenvalue of $\widetilde{\Omega}_n(j,\theta_0)$ and $\widetilde{\Omega}_n(j,\theta_1)$. Call the discretized set Θ_j^n.

We also discretize the parameter $\sigma_1^2(j,\theta)$ in $(0,\overline{\sigma}^2]$ with an equally spaced ϵ-net A_n of width $\frac{1}{n}$. Then for any $\sigma_1^2(j,\theta_0)\in(0,\overline{\sigma}^2]$, there exists $a\in A_n$ such that $a\geqslant\sigma_1^2(j,\theta_0)$ and $|a-\sigma_1^2(j,\theta_0)|\leqslant 1/n$. Let $\Omega_n(j,\theta_1)=a\widetilde{\Omega}_n(j,\theta_1)$. We then can bound $|tr(\Omega_n(j,\theta_0)\left(\Omega_n(j,\theta_1)\right)^{-1})-n|$ and $\log\det\left(\Omega_n(j,\theta_0)\left(\Omega_n(j,\theta_1)\right)^{-1}\right)$ as follows. Observe

$$\begin{aligned} & tr\left(\Omega_n(j,\theta_0)\left(\Omega_n(j,\theta_1)\right)^{-1}\right)-n \\ = & \frac{\sigma_1^2(j,\theta_0)}{a}tr\left(\widetilde{\Omega}_n(j,\theta_0)\left(\widetilde{\Omega}_n(j,\theta_1)\right)^{-1}\right)-n \\ = & \frac{\sigma_1^2(j,\theta_0)}{a}\left[tr\left(\widetilde{\Omega}_n(j,\theta_0)\left(\widetilde{\Omega}_n(j,\theta_1)\right)^{-1}\right)-n\right]+\frac{n(\sigma_1^2(j,\theta_0)-a)}{a} \end{aligned}$$

and

$$\log\det\left(\Omega_n(j,\theta_0)\left(\Omega_n(j,\theta_1)\right)^{-1}\right)=n\log\frac{\sigma_1^2(j,\theta_0)}{a}+\log\det\left(\widetilde{\Omega}_n(j,\theta_0)\left(\widetilde{\Omega}_n(j,\theta_1)\right)^{-1}\right).$$

It follows that

$$\begin{aligned} |tr\left(\Omega_n(j,\theta_0)\left(\Omega_n(j,\theta_1)\right)^{-1}\right)-n| & \leqslant |tr\left(\widetilde{\Omega}_n(j,\theta_0)\left(\widetilde{\Omega}_n(j,\theta_1)\right)^{-1}\right)-n|+\frac{1}{\sigma_1^2(j,\theta_0)} \\ & \leqslant n\varsigma_{n,j}/\underline{\lambda}_j+\frac{1}{\sigma_1^2(j,\theta_0)}, \end{aligned} \tag{7.3}$$

$$|\log\det\left(\Omega_n(j,\theta_0)\left(\Omega_n(j,\theta_1)\right)^{-1}\right)|\leqslant\frac{n(a-\sigma_1^2(j,\theta_0))}{\sigma_1^2(j,\theta)}+n\varsigma_{n,j}/\underline{\lambda}_j\leqslant n\varsigma_{n,j}/\underline{\lambda}_j+\frac{1}{\sigma_1^2(j,\theta_0)}. \tag{7.4}$$

7.1.3 Mixing over regression functions and dependences

Now we have a finite collection of functions in $\widetilde{G}_{\epsilon_{n,i}}(\mathcal{U}_i)$ for each $\mathcal{U}_i\in\mathcal{L}_{class}$ and a countable collection $\widetilde{\Omega}\in\{\widetilde{\Omega}(j,\theta):\theta\in\Theta_j^n\}$ for $j\geqslant 1$. Let $\{\pi_i:i\geqslant 1\}$ be a prior weight assignment on $\{\mathcal{U}_i,\ i\geqslant 1\}$, i.e., $\sum_{i\geqslant 1}\pi_i=1$ with all $\pi_i>0$, $i\geqslant 1$. For each $i\geqslant 1$, choose w_i to be the uniform weight on $\widetilde{G}_{\epsilon_{n,i}}(\mathcal{U}_i)$. Let $\{\psi_j:j\geqslant 1\}$ be a prior weight assignment on $\{\Xi_j,\ j\geqslant 1\}$ and let κ_j denote the uniform prior weight on Θ_j^n. Also let $\{\phi_k\}$ be the uniform weight on A_n. Then for $a\in A_n$ and $\theta\in\Theta_j^n$, let $\Omega(j,\theta)=a\cdot\widetilde{\Omega}(j,\theta)$. From the previous subsection, we know that for each θ_0, there exists $a^*\in A_n$ and $\theta_1\in\Theta_j^n$ such that (7.3) and (7.4) hold. Define

$$q(z^n)=\sum_{i\geqslant 1}\sum_{u\in\widetilde{G}_{\epsilon_{n,i}}(\mathcal{U}_i)}\sum_{j\geqslant 1}\sum_{\theta\in\Theta_j^n}\sum_{a\in A_n}\pi_i\psi_j w_i(u)\kappa_j(\theta)\phi_k(a)p_{u,\Omega(j,\theta)}(z^n).$$

Then $q(z^n)$, with the corresponding distribution denoted by Q_{Z^n}, is a density on $\mathcal{Z}^n$ which satisfies that for any given $\theta_0\in\Theta_j$, $u_0\in\mathcal{U}_i$, for any $\theta_1\in\Theta_j^n$, $u_1\in\widetilde{G}_{\epsilon_{n,i}}(\mathcal{U}_i)$,

$$\begin{aligned} & D\left(P_{Z^n,u,\Omega(j,\theta_0)}\parallel Q_{Z^n}\right) \\ = & E\log\frac{p_{u;\Omega(j,\theta_0)}(Z^n)}{\sum_{i\geqslant 1}\sum_{u\in\widetilde{G}_{\epsilon_{n,i}}(\mathcal{U}_i)}\sum_{j\geqslant 1}\sum_{\theta\in\Theta_j^n}\sum_{a\in A_n}\pi_i\psi_j w_i(u)\kappa_j(\theta)\phi_k(a)p_{u,\Omega(j,\theta)}(z^n)} \\ \leqslant & E\log\frac{p_{u;\Omega(j,\theta_0)}(Z^n)}{\pi_i\psi_j w_i(u_1)\kappa_j(\theta_1)\phi_k(a^*)p_{u_1,\Omega(j,\theta_1)}(z^n)} \\ = & -\log(\pi_i)-\log(\psi_j)-\log\phi_k+M(\epsilon_{n,i};\mathcal{U}_i)+M(\varsigma_n;\Theta_j)+D\left(P_{Z^n,u;\Omega(j,\theta_0)}\parallel P_{Z^n,u_1,\Omega(j,\theta_1)}\right). \end{aligned}$$

Let $\underline{u}_0=(u_0(x_1),...,u_0(x_n))'$ and $\underline{u}_1=(u_1(x_1),...,u_1(x_n))'$. Given X^n, by Lemma 1, the conditional K-L divergence between $P_{Z^n,u_0;\Omega(j,\theta_0)}$ and $P_{Z^n,u_1,\Omega(j,\theta_1)}$ is upper bounded as follows:

$$\frac{1}{2}\log\det\left(\left(\Omega(j,\theta_0)\right)^{-1}\Omega(j,\theta_1)\right)\quad+\quad\frac{1}{2}tr\left(\Omega(j,\theta_0)\left(\Omega(j,\theta_1)\right)^{-1}\right)-\frac{n}{2}+\frac{1}{2}(\underline{u}_0-\underline{u}_1)'\left(\Omega(j,\theta_1)\right)^{-1}(\underline{u}_0-\underline{u}_1)$$

$$\leqslant \quad n\varsigma_n/\underline{\lambda}_j + (1/2)\left(\underline{u}_0 - \underline{u}_1\right)'\left(\underline{u}_0 - \underline{u}_1\right)/(\underline{\lambda}_j\sigma_1^2(j,\theta_0)) + \frac{1}{\sigma_1^2(j,\theta_0)}.$$

Taking expectation with respect to X^n, together with the setup of the discretizations, we have

$$\begin{aligned} D\left(P_{Z^n,u,\Omega(j,\theta_0)} \parallel Q_{Z^n}\right) \quad &\leqslant \quad -\log(\pi_i) - \log(\psi_j) - \log\phi_k + M(\epsilon_{n,i};\mathcal{U}_i) + M(\varsigma_{n,j};\Theta_j) + \\ &\qquad n\varsigma_{n,j}/\underline{\lambda}_j + \frac{n\epsilon_{n,i}^2}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)} + \frac{1}{\sigma_1^2(j,\theta_0)}. \end{aligned}$$

Note that for each $i \geqslant 1$, $j \geqslant 1$, and $\theta_0 \in \Theta_j$, we have

$$\begin{aligned} \sup_{u\in\mathcal{U}_i} D\left(P_{Z^n,u,\Omega(j,\theta_0)} \parallel Q_{Z^n}\right) \quad &\leqslant \quad -\log(\pi_i) - \log(\psi_j) - \log\phi_k + \\ &\qquad M(\epsilon_{n,i};\mathcal{U}_i) + M(\varsigma_{n,j};\Theta_j) + n\varsigma_{n,j}/\underline{\lambda}_j + \frac{n\epsilon_{n,i}^2}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)} + \frac{1}{\sigma_1^2(j,\theta_0)}. \end{aligned}$$

We choose $\epsilon_{n,i}$ such that $M(\epsilon_{n,i};\mathcal{U}_i) = n\epsilon_{n,i}^2$ and $\varsigma_{n,j}$ such that $M(\varsigma_{n,j};\Theta_j) = n\varsigma_{n,j}$. Then

$$\begin{aligned} \sup_{u\in\mathcal{U}_i} D\left(P_{Z^n,u,\Omega(j,\theta_0)} \parallel Q_{Z^n}\right) \quad &\leqslant \quad -\log(\pi_i) - \log(\psi_j) - \log\phi_k + \\ &\qquad \left(1 + \frac{1}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)}\right) n\epsilon_{n,i}^2 + \left(1 + \frac{1}{\underline{\lambda}_j}\right) n\varsigma_{n,j} + \frac{1}{\sigma_1^2(j,\theta_0)}. \end{aligned}$$

7.1.4 Estimating the conditional distributions

Thus we have constructed a distribution Q_{Z^n} on the product space, which is uniformly appropriately close to $P_{Z^n,u;\Omega}$ for all $u \in \mathcal{U}_i$ and $\theta \in \Theta_j$ in terms of the Kullback-Leibler divergence. Let $\theta_0 \in \Theta_j$ and $u \in \mathcal{U}_i$ be the true dependence and the regression function. For simplicity, let Ω denote $\Omega(j,\theta_0)$.

The density $q(z^n)$ can be written as product of conditional densities, i.e., $q(z^n) = q_0(z_1)\cdots q_{n-1}(z_n|z^{n-1})$. Let $\widehat{p}_{i-1}(z_i) = q_{i-1}(z_i|z^{i-1})$. For $n \geqslant 1$, let $\Omega_n = \begin{pmatrix} \Omega_{n-1} & \beta_{n-1} \\ \beta_{n-1}' & \sigma_n^2 \end{pmatrix}$ be the partition of Ω_n. Under the Gaussian assumption, given $X_{i+1} = x$ and $(X_j, Y_j)_{j=1}^i$ under Ω, Y_{i+1} has a normal distribution with mean $m_{i,u;\Omega}(x|Z^i) = u(x) + \beta_i'\Omega_i^{-1}\left(Y^i - U^i\right)$ and variance $\sigma_{i+1}^2 - \beta_i'\Omega_i^{-1}\beta_i$. Let

$$\begin{aligned} p_{z_{i+1}|Z^i;u;\Omega}(x_{i+1},y_{i+1}) \quad &= \quad h(x_{i+1})\left(2\pi\left(\sigma_{i+1}^2 - \beta_i'\Omega_i^{-1}\beta_i\right)\right)^{-1/2} \times \\ &\qquad \times \exp\left(-1/\left(2\left(\sigma_{i+1}^2 - \beta_i'\Omega_i^{-1}\beta_i\right)\right)\left(y_{i+1} - m_{i,u}(x_{i+1}|Z^i)\right)^2\right). \end{aligned}$$

It is the conditional density of Z_{i+1} given Z^i under the regression function u and Ω. Then we have,

$$\begin{aligned} &\textstyle\sum_{i=0}^{n-1} E\log\frac{p_{z_{i+1}|Z^i;u;\Omega}(Z_{i+1})}{\widehat{p}_i(Z_{i+1})} \\ &= E\log\tfrac{p_{u,\Omega}(Z^n)}{q(Z^n)} \\ &= D\left(P_{Z^n,u;\Omega} \parallel Q_{Z^n}\right) \\ &\leqslant -\log(\pi_i) - \log(\psi_j) - \log\phi_k + \left(1 + \tfrac{1}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)}\right) n\epsilon_{n,i}^2 + \left(1 + \tfrac{1}{\underline{\lambda}_j}\right) n\varsigma_{n,j} + \tfrac{1}{\sigma_1^2(j,\theta_0)}. \end{aligned}$$

Since the squared Hellinger distance satisfies $d_H^2(p_1,p_2) = \int\left(p_1^{1/2} - p_2^{1/2}\right)^2 d\mu \leqslant D(p_1 \parallel p_2)$, we have

$$\begin{aligned} \max_{u\in\mathcal{U}_i}\sum_{i=0}^{n-1} E d_H^2(p_{z_{i+1}|Z^i;u;\Omega},\widehat{p}_i) \quad &\leqslant \quad -\log(\pi_i) - \log(\psi_j) - \log\phi_k + \\ &\qquad \left(1 + \frac{1}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)}\right) n\epsilon_{n,i}^2 + \left(1 + \frac{1}{\underline{\lambda}_j}\right) n\varsigma_{n,j} + \frac{1}{\sigma_1^2(j,\theta_0)}. \end{aligned}$$

This means that we can estimate well the conditional densities of Z_{i+1} given Z^i by $\widehat{p}_i$'s in terms of the cumulative squared Hellinger risk. We now construct estimators of u.

7.1.5 Estimating u up to a constant

Now, for each i, let $\widetilde{u}_i$, $\widetilde{j}_i$ and $\widetilde{\theta}_{\widetilde{j}_i}$ be the minimizer of $d_H^2(p_{z_{i+1}|Z^i;u;\Omega(j,\theta)}, \widehat{p}_i)$ over $u \in \cup_{i\geqslant 1}\mathcal{U}_i,\ j \geqslant 1$, and $\theta \in \Theta_j$ respectively. Then by the triangle inequality, we have

$$\begin{aligned}
&\max_{u\in\mathcal{U}_i} \sum_{i=0}^{n-1} E d_H^2(p_{z_{i+1}|Z^i;u;\Omega}, p_{z_{i+1}|Z^i;\widetilde{u}_i;\Omega(\widetilde{j}_i,\widetilde{\theta}_{\widetilde{j}_i})}) \\
\leqslant\ & \max_{u\in\mathcal{U}_i} \sum_{i=0}^{n-1} 2E\left(d_H^2(p_{z_{i+1}|Z^i;u;\Omega}, \widehat{p}_i) + d_H^2(p_{z_{i+1}|Z^i;\widetilde{u};\Omega(\widetilde{j}_i,\widetilde{\theta}_{\widetilde{j}_i})}, \widehat{p}_i)\right) \\
\leqslant\ & 4\max_{u\in\mathcal{U}_i} \sum_{i=0}^{n-1} E d_H^2(p_{z_{i+1}|Z^i;u;\Omega}, \widehat{p}_i) \\
\leqslant\ & -4\log(\pi_i) - 4\log(\psi_j) - 4\log\phi_k + 4\left(1 + \frac{1}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)}\right) n\epsilon_{n,i}^2 + 4\left(1+\frac{1}{\underline{\lambda}_j}\right) n\varsigma_{n,j} + \frac{4}{\sigma_1^2(j,\theta_0)}.
\end{aligned}$$

From Lemma 1 in [35], we have that if f_1 and f_2 are two densities with mean and variance μ_1, σ_1^2 and μ_2, σ_2^2 respectively, then

$$d_H^2(f_1, f_2) \geqslant \frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2) + (\mu_1-\mu_2)^2}.$$

Under the boundedness assumptions, we have that the squared Hellinger distance between $p_{z_{i+1}|Z^i;u;\Omega}$ and $p_{z_{i+1}|Z^i;\widetilde{u}_i;\Omega(\widetilde{j}_i,\widetilde{\theta}_{\widetilde{j}_i})}$ at a given x is lower bounded by $\frac{(\mu_1-\mu_2)^2}{4\overline{\sigma}^2+4L^2}$, where $\mu_1 = u(x) - \beta_i'\Omega_i^{-1}\left(Y^i - U^i\right)$ and $\mu_2 = \widetilde{u}_i(x) - \widetilde{\beta}_i'\left(\Omega(\widetilde{j}_i,\widetilde{\theta}_{\widetilde{j}_i})\right)^{-1}\left(Y^i - \widetilde{U}^i\right)$ is the mean of $p_{z_{i+1}|Z^i;\widetilde{u}_i;\Omega(\widetilde{j}_i,\widetilde{\theta}_{\widetilde{j}_i})}$ at x. It follows that

$$\begin{aligned}
&E d_H^2(p_{z_{i+1}|Z^i;u;\Omega}, p_{z_{i+1}|Z^i;\widetilde{u}_i;\Omega(\widetilde{j}_i,\widetilde{\theta}_{\widetilde{j}_i})}) \\
\geqslant\ & \frac{1}{4\overline{\sigma}^2+4L^2} E\int h(x)\left(u(x) - \widetilde{u}_i(x) - \widetilde{\beta}_i'\left(\Omega(\widetilde{j}_i,\widetilde{\theta}_{\widetilde{j}_i})\right)^{-1}\left(Y^i - \widetilde{U}^i\right) + \beta_i'\Omega_i^{-1}\left(Y^i - U^i\right)\right)^2 d\mu \\
\geqslant\ & \frac{1}{4\overline{\sigma}^2+4L^2} E\int h(x)\left(u(x) - \widetilde{u}_i(x) - \tau_i\right)^2 d\mu,
\end{aligned}$$

where $\tau_i = \int h(x)u(x)d\mu - \int h(x)\widetilde{u}_i(x)$. Thus for any $u \in \mathcal{U}_i$, $j \geqslant 1$, $\theta_0 \in \Theta_j$, we have

$$\begin{aligned}
&\sum_{i=0}^{n-1} E\int h(x)\left(u(x) - \widetilde{u}_i(x) - \tau_i\right)^2 d\mu \\
\leqslant\ & 4\left(\overline{\sigma}^2 + L^2\right)\sum_{i=0}^{n-1} E d_H^2(p_{z_{i+1}|Z^i;u;\Omega}, p_{z_{i+1}|Z^i;\widetilde{u}_i;\Omega(\widetilde{j}_i,\widetilde{\theta}_{\widetilde{j}_i})}) \\
\leqslant\ & 16\left(\overline{\sigma}^2 + L^2\right)\left(-\log(\pi_i) - \log(\psi_j) - \log\phi_k + \left(1 + \frac{1}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)}\right) n\epsilon_{n,i}^2 + \left(1+\frac{1}{\underline{\lambda}_j}\right) n\varsigma_{n,j} + \frac{1}{\sigma_1^2(j,\theta_0)}\right).
\end{aligned}$$

Thus we have obtained a sequence of estimators $\widetilde{u}_i$ of u with the variances

$$E\left(\int h(x)\left(u(x) - \widetilde{u}_i(x) - \tau_i\right)^2 d\mu\right)$$

of $u - \widetilde{u}_i$ well controlled on average. However, a possibly large bias remains and need to be handled.

7.1.6 Debias the estimators

To get a final estimator of u, we estimate the mean $\eta = \int h(x)u(x)d\mu$ based on the current data Z^i. For any $\widehat{\eta}$ based on Z^n, let $\widehat{\widetilde{u}}_i(x) = \widetilde{u}_i(x) - \int \widetilde{u}_i(x)h(x)d\mu + \widehat{\eta}$. Then the new estimator satisfies

$$\int h(x)\left(u(x) - \widehat{\widetilde{u}}_i(x)\right)^2 d\mu \quad = \quad \int h(x)\left(u(x) - \widetilde{u}_i(x) - \tau_i\right)^2 d\mu$$

$$+\left(\widehat{\eta}-\eta\right)^2.$$

It follows that

$$\begin{aligned}
&\sum_{i=0}^{n-1} E\int h(x)\left(u(x)-\widehat{\widetilde{u}}_i(x)\right)^2 d\mu \\
=\;& \sum_{i=0}^{n-1} E\int h(x)\left(u(x)-\widetilde{u}_i(x)-\tau_i\right)^2 d\mu + nE(\widehat{\eta}-\eta)^2 \\
\leqslant\;& 16\left(\overline{\sigma}^2+L^2\right)\left(-\log\pi_i-\log\psi_j-\log\phi_k+\left(1+\frac{1}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)}\right)n\epsilon_{n,i}^2+\left(1+\frac{2}{\underline{\lambda}_j}\right)n\varsigma_{n,j}+\frac{1}{\sigma_1^2(j,\theta_0)}\right) \\
&+nE(\widehat{\eta}-\eta)^2.
\end{aligned}$$

A simple estimator of η based on Z^n is the sample mean of Y^n. Let $\widehat{\eta}=(1/n)\sum_{j=1}^n Y_j$, then

$$\begin{aligned}
E(\widehat{\eta}-\eta)^2 &= E\left(\frac{1}{n}\sum_{i=1}^n\left(u(X_i)-\eta\right)+\frac{1}{n}\sum_{i=1}^n e_i\right)^2 \\
&= E\left(\frac{1}{n}\sum_{i=1}^n\left(u(X_i)-\eta\right)\right)^2+E\left(\frac{1}{n}\sum_{i=1}^n e_i\right)^2 \\
&= \frac{1}{n}\int\left(u(x)-\eta\right)^2 h(x)d\mu+\frac{\mathbf{1}'\Omega_n\mathbf{1}}{n^2} \\
&\leqslant \frac{4L^2}{n}+\frac{\mathbf{1}'\Omega_n\mathbf{1}}{n^2}.
\end{aligned}$$

From all above, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n E\parallel u-\widehat{\widetilde{u}}_i\parallel^2\leqslant\frac{4L^2}{n}+\frac{\mathbf{1}'\Omega_n\mathbf{1}}{n^2}+ \\
&16\left(\overline{\sigma}^2+L^2\right)\left(\frac{-\log\pi_i-\log\psi_j-\log\phi_k}{n}+\left(1+\frac{1}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)}\right)\epsilon_{n,i}^2+\left(1+\frac{2}{\underline{\lambda}_j}\right)\varsigma_{n,j}+\frac{1}{n\sigma_1^2(j,\theta_0)}\right).
\end{aligned}$$

Note that $\log\phi_k$ is of order $\log n$. Due to convexity of the squared L_2 norm, the upper bound above also holds for $E\parallel u-\frac{1}{n}\sum_{i=1}^n\widehat{\widetilde{u}}_i\parallel^2$. This completes the proof of Theorem 1.

7.2 Proof of Theorem 2

Under Assumption 2, by Lemma 4 in Section 7, we have that

$$\zeta\left(\Omega_n(j,\theta),\Omega_n(j,\theta')\right)\leqslant n\max_{1\leqslant i\leqslant n,1\leqslant l\leqslant n}|w_{il}(\theta)-w_{il}(\theta')|\leqslant c_j n^{A_j+1}d(\theta,\theta').$$

Thus in order to have $\varsigma_{n,j}$ of order $1/n$, $d(\theta,\theta')$ should be of order $n^{-(A_j+2)}$. Consider a rectangular grid on R^{d_j} for an ϵ-net with ϵ of order $n^{-(A_j+2)}$ (ignore those cubes that do not intersect with Θ_j and modify a grid point if it is not in Θ_j). We can assign a prior weight on the grid in such a way that for each $\theta\in\Theta_j$, there exists $\theta_{(n)}$ in the grid with $\zeta\left(\Omega_n(j,\theta),\Omega_n(j,\theta')\right)\leqslant 1/n$ and $\log w(\theta_{(n)})$ is of order $\log n$ (see [35] or [36]). Then the rate given in Theorem 1 is known to be the minimax rate of convergence (see [36]).

7.2.1 Proof of Theorem 3

Consider the following countable collection of classes:

$$\{\mathcal{U}_\xi:\xi\in Q\cap\Upsilon\}. \tag{7.5}$$

We assign prior weights $\{\pi_\xi, \xi \in Q \cap \Upsilon\}$ based on a description of the indices of the classes according to coding theory as follows. For every dyadic rational number q, it can be written as $q = i(q)+\sum_{j=1}^{l(q)} a_j(q)2^{-j}$ for some $l \geqslant 1$, a_j's being either 0 or 1, and i is the integer part of q. To describe such a q, we just need to describe the integers i, l, and the a_j's. To describe i, we first describe the sign of i using $\log_2 2 = 1$ bit, and then describe the absolute value of i using $\log^*(i) =: \log_2(i)+2\log_2(\log_2(i+1))$ bits (ignoring rounding). Then describe l using $\log^*(l)$ bits, and finally describe a_j's using l bits. By this way, we describe all the hyper-parameter components $\beta_1, ..., \beta_m$ for $\xi = (\beta_1, ..., \beta_m) \in Q \cap \Upsilon$. The total description length for ξ then is

$$\sum_{k=1}^{m} (1 + \log^*(i(\beta_k)) + \log^*(l(\beta_k)) + l(\beta_k)).$$

The prior weight of class $\mathcal{U}_\xi$ in the countable collection is π_ξ with $-\log_2 \pi_\xi$ equal the above expression. The coding interpretation guarantees $\{\pi_\xi : \xi \in Q \cap \Upsilon\}$ is a sub-probability (see, e.g., [9]), i.e., $\sum_{\xi \in Q\cap\Upsilon} \pi_\xi \leqslant 1$. One can either normalize π_ξ to be a probability or put the remaining probability on any chosen class without any effect on rates of convergence.

As in the proof of Theorem 1, for each $\xi \in Q \cap \Upsilon$, discretize $\mathcal{U}_\xi$ by a covering set of size $\epsilon_{n,\xi}$ as defined earlier in this section and discretize the dependence as before. Follow the same construction of the adaptive estimator. Let $\widehat{u}^*$ denote the final estimator. It remains to show it is minimax-rate optimal for all the classes $\{\mathcal{U}_\xi : \xi \in \Upsilon\}$.

Based on Assumption 4, for each $\xi_0 \in \Upsilon$, there exists a sequence $\xi_n \in \Upsilon \cap N_{m_n}$ with $\xi_0 \lhd \xi_n$ for some m_n of order $\log n$ such that $\frac{\epsilon_{n,\xi_n}}{\epsilon_{n,\xi_0}}$ stays upper bounded. Note that then π_i is bounded above by order $\log n$. By Theorem 1, for each $u \in \mathcal{U}_{\xi_0} \subset \mathcal{U}_{\xi_n}$, we have

$$E \parallel u - \widehat{u} \parallel^2 \leqslant \frac{4L^2}{n} + \frac{\mathbf{1}'\Omega_n\mathbf{1}}{n^2} +$$
$$16\left(\overline{\sigma}^2 + L^2\right)\left(-\frac{\log(\pi_i)}{n} - \frac{\log(\psi_j)}{n} - \frac{\log(\phi_i)}{n} + \left(1 + \frac{1}{2\underline{\lambda}_j\sigma_1^2(j,\theta_0)}\right)\epsilon_{n,\xi_n}^2 + \left(1 + \frac{1}{\underline{\lambda}_j}\right)\varsigma_{n,j} + \frac{1}{n\sigma_1^2(j,\theta_0)}\right).$$

Together with $\log(\pi_i) = O(\log n)$ and $\epsilon_{n,\xi_n}^2 = O(\epsilon_{n,\xi_0}^2)$, we obtain

$$E \parallel u - \widehat{u} \parallel^2 = O\left(\frac{\log n}{n} + \epsilon_{n,\xi}^2 + \varsigma_{n,j} + \frac{\mathbf{1}'\Omega_n\mathbf{1}}{n^2}\right) = O\left(\max\left(\epsilon_{n,\xi_0}^2, \frac{\mathbf{1}'\Omega_n\mathbf{1}}{n^2}\right)\right).$$

The conclusion follows. This completes the proof of Theorem 3.

7.3 Lemmas and their proofs

Lemma 1: Let P_{μ_1,Ω_1} and P_{μ_2,Ω_2} denote the n-dimensional normal distributions with means μ_1 and μ_2 and covariance matrices Ω_1 and Ω_2, respectively. Then

$$\begin{aligned} & D\left(P_{\mu_1,\Omega_1} \parallel P_{\mu_2,\Omega_2}\right) \\ = \; & (1/2)\log\det\left(\Omega_1^{-1}\Omega_2\right) + (1/2)\,tr(\Omega_1\Omega_2^{-1}) - (1/2)\,n + (1/2)\left(\mu_1 - \mu_2\right)'\Omega_2^{-1}(\mu_1 - \mu_2). \end{aligned}$$

Proof: Direct calculation gives the result.

Lemma 2: Let Ω_1 and Ω_2 be two $n \times n$ positive definite matrix. Then we have

$$n - tr(\Omega_1\Omega_2^{-1}) \leqslant \log\det\left(\Omega_1\Omega_2^{-1}\right) \leqslant tr(\Omega_1^{-1}\Omega_2) - n$$

and consequently

$$|\log\det\left(\Omega_1\Omega_2^{-1}\right)| \leqslant \max\left(|tr(\Omega_1\Omega_2^{-1}) - n|, |tr(\Omega_1^{-1}\Omega_2) - n|\right).$$

Proof: From Lemma 1, by taking $\mu_1 = \mu_2$, we have

$$\log\det\left(\Omega_1^{-1}\Omega_2\right) + tr(\Omega_1\Omega_2^{-1}) - n \geqslant 0,$$

and similarly

$$\log\det\left(\Omega_1\Omega_2^{-1}\right) \geqslant n - tr(\Omega_1^{-1}\Omega_2).$$

Observing that $\log\det\left(\Omega_1\Omega_2^{-1}\right) = -\log\det\left(\Omega_1^{-1}\Omega_2\right)$, the conclusion follows.

Lemma 3: Let Ω_1 and Ω_2 be two $n\times n$ positive definite matrix. Let ν^* denote the maximum of the absolute values of the eigenvalues of $\Omega_2-\Omega_1$ and let w_1 and w_2 denote the smallest eigenvalues of Ω_1 and Ω_2 respectively. Then

$$|tr(\Omega_1\Omega_2^{-1}) - n| \leqslant n\nu^*/w_2.$$

$$|\log\det(\Omega_1\Omega_2^{-1})| \leqslant n\nu^*/(\min(w_1,w_2)).$$

Proof: By positive definiteness, there exist two orthogonal matrice O and $\widetilde{O}$ such that

$$\Omega_2 = O^{'}diag(\nu_{1,1},...,\nu_{1,n})O$$

and

$$\Omega_2-\Omega_1 = \widetilde{O}^{'}diag(\nu_{2,1},...,\nu_{2,n})\widetilde{O},$$

where the two diagonal matrices are formed by the eigenvalues of Ω_2 and $\Omega_2-\Omega_1$, respectively. It follows that

$$\begin{aligned} tr(\Omega_1\Omega_2^{-1}) - n &= tr\left(\Omega_2^{-1}(\Omega_1-\Omega_2)\right) = tr\left(O^{'}diag(\nu_{1,1}^{-1},...,\nu_{1,n}^{-1})O\widetilde{O}^{'}diag(\nu_{2,1},...,\nu_{2,n})\widetilde{O}\right) \\ &= tr\left(diag(\nu_{1,1}^{-1},...,\nu_{1,n}^{-1})O\widetilde{O}^{'}diag(\nu_{2,1},...,\nu_{2,n})\widetilde{O}O^{'}\right). \end{aligned}$$

Based on the simple observation that for symmetric matrices A, B_1, B_2, if $A\geqslant 0$ and $B_1\leqslant B_2$, then $tr(AB_1)\leqslant tr(AB_2)$, together with that $O\widetilde{O}^{'}diag(\nu_{2,1},...,\nu_{2,n})\widetilde{O}O^{'} \leqslant |v^*|I_n$, where v^* is the largest eigenvalue of $\Omega_1-\Omega_2$ in absolute value and I_n denotes the identity matrix, we have

$$tr(\Omega_1\Omega_2^{-1}) - n \leqslant v^*(\nu_{11}^{-1}+...+\nu_{1n}^{-1}) \leqslant nv^*/\nu_{11}^{-1}.$$

With a similarly established lower bound, i.e.,

$$tr(\Omega_1\Omega_2^{-1}) - n \geqslant -v^*(\nu_{11}^{-1}+...+\nu_{1n}^{-1}) \geqslant -nv^*/\nu_{11}^{-1},$$

we know

$$|tr(\Omega_1\Omega_2^{-1}) - n| \leqslant n\nu^*/w_2.$$

Together with Lemma 2, the conclusion on the determinant follows.

Lemma 4: Let $A=(a_{i,j})$ be an $n\times n$ square matrix. Then for any eigenvalue λ of A, $|\lambda|$ is upper bounded by

$$\min\left(\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}|, \max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{ij}|\right).$$

For a proof of this simple result, see, e.g., [19].

Acknowledgements This paper is dedicated to Professor Lin-Cheng Zhao, in celebration of his 75th Birthday. Both authors were inspired by Dr. Zhao as a leading scholar in statistical sciences. The second author had the honor of being his MS student (co-advised by Dr. Baiqi Miao), and received the solid initial training of minimax philosophy and tools. This research was partially supported by National Natural Science Foundation of China (Grant No. 61572109). We greatly appreciate the helpful comments by two referees on improving our work.

References

1 Guo Y Q, Shyr H J, Thierrin G. F-disjunctive languages. Intern J Comput Math, 1986, 18: 219–237
2 R.K. Adensdedt. On large sample estimation for the mean of a stationary sequence. Ann. Statist., 1974, 2: 1095-1107
3 Y. Baraud. Model selection for regression on a random design. ESAIM Probab. Statist., 2002, 6: 127-146.
4 A.R. Barron. Approximation and estimation bounds for artificial neural networks, Machine Learning, 1994, 14: 115-133
5 A.R. Barron, L. Birgé and P. Massart. Risk bounds for model selection via penalization. Probability Theory and Related Fields, 1999, 113: 301-413
6 J. Beran. Statistics for Long-Memory Processes. Technometrics,1994, 39: 105-106
7 L. Birgé. Approximation dans les espaces metriques et theorie de l'estimation. Z. Wahrscheinlichkeitstheor. Verw. Geb., 1983, 65: 181-237
8 L.D. Brown and M. G. Low. A constrained risk inequality with applications to non-parametric functional estimation, Ann. Statist, 1996, 24: 2524-2535
9 T.M. Cover and J.A. Thomas. Elements of Information Theory. Tsinghua University Press, 1991
10 R. Dahlhaus. Efficient parameter estimation for self-similar processes. Ann. Statistics, 1989, 17: 1749-1766
11 R.A. DeVore and G.G. Lorentz. Constructive Approximation. Springer-Verlag: 1993
12 D.L. Donoho and I.M. Johnstone. Minimax estimation via wavelet shrinkage. Annals of Statistics, 1988, 26: 879-921
13 S. Efromovich. How to overcome the curse of long-memory errors. IEEE Trans. Inform.Theory, 1999, 45: 1735-1741
14 Granger, C.W.J. and Joyeux, R. An introduction to long-range time series models and fractional differencing. J. Time Ser. Anal., 1980, 1: 15-30
15 S.Yu. Efroimovich and M.S. Pinsker. A self-educating nonparametric filtration algorithm. Automation and Remote Control, 1984, 45: 58-65
16 P. Hall and J.D. Hart. Nonparametric regression with long-range dependence. Stochastic Process. Appli., 1990, 36: 339-351
17 Hall, P. Lahiri, S. N. and Polzehl, J.. On bandwidth choice in nonparametric regression with both short-and long-range dependent errors. Ann. Statist., 1994, 23: 1921-1936
18 W. Härdle and J.S. Marron. Optimal bandwidth selection in nonparametric regression function estimation. Ann. Statist., 1985, 13: 1465-1481
19 R. Horn and C.R. Johnson. Matrix Analysis. Cambridge University Press, 1985
20 Hosking, J.R.M..Fractional differencing. Biometrika, 1981, 68: 165-176
21 I.M. Johnstone and B.W. Silverman. Wavelet threshold estimators for data with correlated noise. J. Roy. Statist. Soc. Ser. B, 1997, 59: 319-351
22 Kolmogorov, A.N. and Tihomirov, V.M.. ϵ-entropy and ϵ-capacity of sets in functional spaces. Uspekhi Mat. Nauk, 1959, 14: 3-86
23 H. Künsch, J. Beran and F. Hampel. Contrasts under long-range correlations. Ann. Statist., 1993, 21: 943-964
24 L.M. Le Cam. Convergence of estimates under dimensionality restrictions. Ann. Statist., 1973, 1: 38-53
25 O.V. Lepskii. Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. Theory probab. Appl., 1991, 36: 682-697
26 Mandelbrot, B.B. and Van Ness, J.W.. Fractional Brownian motions, fractional noises and applications. SIAM Rev., 1968, 10: 422-437
27 J. Opsomer, Y. Wang, and Y. Yang. Nonparametric regression with correlated errors. Statistical Science, 2001, 16: 134-153
28 A. Samorov and M.S. Taqqu. On the efficiency of sample mean in long memory noise. J. Time Ser. Anal., 1988, 9: 191-200
29 C.J. Stone. Additive regression and other nonparametric models. Annals of Statistics, 1985, 13: 689-705
30 H. Triebel. Interpolation properties of ϵ-entropy and diameters. Geometric characteristics of imbedding for function spaces of Sobolev-Besov type. Mat. Sbornik, 1975, 98: 27-41
31 Y. Wang. Function estimation via wavelet shrinkage for long-memory data. Ann. Statist., 1996, 24: 466-484
32 Z. Wang, S. Paterlini, F. Gao, and Y. Yang. Adaptive minimax regression estimation over sparse ℓ_q-hulls. Journal of Machine Learning Research, 2014, 15: 1675-1711
33 M. Wegkamp. Model selection in nonparametric regression. Annals of Statistics, 2003, 31: 252-273
34 Y. Yang. Model Selection for Nonparametric Regression, Statistica Sinica, 1999, 9: 475-499
35 Y. Yang. Combining Different Procedures for Adaptive Regression. Journal of Multivariate Analysis 2000, 74: 135-161
36 Y. Yang. Nonparametric regression and prediction with dependent errors. Bernoulli., 2001, 7: 633-655
37 Y. Yang. Minimax rate adaptive estimation over continuous hyper-parameters, IEEE Transactions on Information Theory, 2001, 47: 2081-2085
38 Y. Yang and A.R. Barron. Information-theoretic determination of minimax rates of convergence. Ann. Statistics, 1999, 27: 1564-1599
39 L.C. Zhao. Exponential bounds of mean error for the nearest neighbor estimates of regression functions. J. Multivariate

Analysis, 1987, 21: 168-178