Nonparametric Regression and Prediction with Dependent Errors (Running Title: Regression and Prediction under Dependence)

Yuhong Yang
Department of Statistics
Iowa State University
Ames, IA 50011, USA

Abstract

We study minimax rates of convergence for nonparametric regression and prediction under a random design with dependent errors. It is shown that dependence among errors in general does not hurt a prediction of the next response. For estimating the regression function, however, dependence may damage the minimax rate of convergence. Under the assumption that the errors are independent of the explanatory variables, we show that minimax rates of convergence are determined in terms of the massiveness (characterized by metric entropy) of the function class assumed to contain the underlying regression function, and behavior of the covariance matrix of the errors. It is shown that the minimax risk is at the worse rate between two quantities: the minimax risk of the same function class but under the assumption of i.i.d. errors, and the minimax risk of estimating the mean of the regression function. Examples of function classes under different covariance structures including both short and long range dependences are given.

Key words and phrases. Long range dependent errors, minimax rate of convergence, nonparametric regression, prediction.

1 Introduction

1.1 Problem of interest

Assume we observe random variables $(X_i, Y_i)_{i=1}^n$, where Y_i takes value in R and X_i takes value in $\mathcal{X}$, a subset in R^d for some $d \geq 1$. The relationship between response variables Y_i's and the explanatory or experimental variables X_i's is modeled as

$$Y_i = u(X_i) + \varepsilon_i,\ i \geq 1, \tag{1}$$

where u is an unknown regression function. The random errors $\{\varepsilon_i,\ i \geq 1\}$ are assumed to have a joint normal distribution conditioned on $\{X_i,\ i \geq 1\}$ with mean zero and a known covariance matrix. A goal is to estimate the regression function u, which is assumed a priori to be in a nonparametric function class $\mathcal{U}$ (e.g., monotone or Lipschitz). Another related goal is to predict the next response Y_{n+1} given the data and X_{n+1}. In this paper, we study how well one can estimate u and how well one can predict Y_{n+1} both under a minimax consideration over the function class $\mathcal{U}$. The focus is on determination of minimax rates of convergence for the estimation and prediction problems when the errors are dependent. We will characterize how dependence of the errors as well as the function class affects the minimax rates of convergence under appropriate conditions.

1.2 Some background

In recent years, there has been an increasing interest in statistical estimation based on long-range dependent data (the reader is referred to Beran (1994) for a survey of work in this area). Long-range dependence has been observed in many applied scientific disciplines. Künsch et al (1993) wrote: "Perhaps most unbelievable to many is the observation that high-quality measurement series from astronomy, physics, chemistry, generally regarded as prototypes of 'i.i.d.' observations, are not independent but long-range correlated". Based on the empirical evidences of long-range dependence in measurements and other applications, it becomes important to study how long-range dependence affects statistical estimation.

For parametric regression with fixed designs, asymptotic results for MLE and least square estimators under long-range dependence are established by Yajima (e.g., 1991). Künsch et al (1993) show that for certain analysis of variance models with random designs, contrasts can be estimated at the same rate as that under independent errors. Asymptotic results for the estimation of long-range dependence parameters under parametric models are in Beran (1986), Fox and Taqqu (1986), Dahlhaus (1989), Giraitis and Surgailis (1990), Robinson (1995) and others.

For nonparametric regression, effect of long-range dependence on minimax rates of convergence is studied in a pioneering work of Hall and Hart (1990a) for a differentiable function class, later by Wang (1996), and Johnstone and Silverman (1997) for Besov classes, all under a fixed equally spaced design. These results show that a certain long-range dependence of errors damages the minimax rate of convergence for regression estimation. The latter two papers propose adaptive wavelet estimators. In addition, Wang shows that for some inhomogeneous Besov classes, linear estimators can not achieve the minimax rate of convergence, and Johnstone and Silverman show that when an unknown dependence parameter is properly estimated, a wavelet threshold estimator is adaptive with respect to both the dependence parameter and the smoothness parameters. Robinson (1996) derives local asymptotic normality for kernel estimators under long-range dependence.

In this work, we study effects of a general dependence among the errors on regression estimation and on prediction for a general nonparametric function class, under a random design. The focus is on the theoretic determination of the minimax rate of convergence. We do not address issues of estimation of dependence and adaptive estimation in this paper.

We finally point out that independently of our work, Efromovich (1999) obtains minimax rates of convergence for regression estimation for Hölder classes under a long-range dependence and a random design. He proposes a series expansion estimator and shows it is adaptive with respect to a smoothness parameter. Our results on minimax rates of convergence apply to general classes of regression functions satisfying a mild richness assumption.

1.3 A summary of our findings

We summarize our results informally below. The conclusions are in terms of minimax rates of convergence under square L_2 type of loss under a random design. The errors are assumed to be independent of the explanatory variables for the results on estimating the regression function, but the independence is not required for prediction.

1. If the variances of the errors are uniformly upper bounded, then the regression function up to a constant can be estimated as well as under i.i.d. errors.

2. Under some mild conditions, the minimax risk for estimating the regression function in a class converges at a rate of the maximum of two quantities: the minimax rate of the same function class but under i.i.d. errors, and the minimax rate for estimating the mean value of the regression function.

3. Dependence among the errors and/or dependence among the explanatory variables (as long as they have the same marginal density) do not make prediction of next response harder.

From above, the effect of dependence of serially correlated errors on regression is sort of "parametric", in the sense that it does not affect the rate of convergence more than adding the risk for estimating a single parameter (the mean of the regression function). Similar phenomena have been observed earlier for some parametric models (e.g., Künsch et al (1993)) and density estimation (Hall and Hart (1990b)) both under long-range dependence.

The paper is organized as follows. Some preliminary considerations are given first in Section 2. The main results on both regression estimation and prediction are presented in Section 3. A key proposition on minimax risk bounds is presented in Section 4. The proofs of the main results as well as useful lemmas are given in Section 5.

2 Risks of interests and metric entropy

2.1 Risk for regression estimation

We assume that $\{X_i, i \geq 1\}$ are i.i.d. with density h with respect to a measure μ. For the nonparametric class $\mathcal{U}$ supposed to contain u, we assume that $\mathcal{U}$ is uniformly bounded throughout the paper.

For regression estimation, we obtain results when the errors are independent of X^n, i.e., the conditional covariance matrix Ω_n of $\{\varepsilon_i, 1 \leq i \leq n\}$ given $X^n = (X_1, ..., X_n)$ does not depend on X^n. We assume that Ω_n is known. Let $\| u - v \|_{L_2(h)} = \left(\int (u-v)^2 h d\mu\right)^{1/2}$ be the L_2 distance between two functions u and v with respect to the design density of X_1. Since $\mathcal{U}$ is uniformly bounded, the distance is well-defined within the function class.

The minimax risk we examine for estimating the regression function u is

$$R(\mathcal{U}; \Omega; n) = \min_{\hat{u}} \max_{u \in \mathcal{U}} E \| u - \hat{u} \|^2_{L_2(h)},$$

where $\hat{u}$ is over all estimators based on $(X_i, Y_i)_{i=1}^n$ and the expectation is taken under the true regression function u. The minimax risk measures how well one can possibly estimate u uniformly over the function class.

A condition, namely, $Tr(\Omega_n^{-1})$ is of order n ($Tr(\cdot)$ denotes the trace of a square matrix), will be used for identifying minimax rates. It is satisfied by short- and long-range dependent cases as given in Section 3.3. It also holds for stationary invertible autoregressive errors as studied in Hall and Hart (1990a). Let $\sigma_i^2 = Var(\varepsilon_i)$. A simple sufficient condition for $Tr(\Omega_n^{-1}) \asymp n$ is that $\sup \sigma_i^2 < \infty$ and there is a white noise component in the errors, i.e., $\varepsilon_i = \varepsilon_i^{(1)} + \varepsilon_i^{(2)}$, where $\{\varepsilon_i^{(1)}, i \geq 1\}$ are i.i.d. and independent of $\{\varepsilon_i^{(2)}\}$ (see Lemma 7 in Section 5). When the trace condition is not satisfied, rates better than that under i.i.d. errors are possible. For instance, assume that the errors are independent with decreasing variances σ_i^2 of order i^{-1}. Then it is intuitively clear that the rate of convergence can be faster compared with that under i.i.d. errors.

2.2 Estimation of the mean of the regression function

Related to the above problem of regression estimation is the problem of estimating the mean value of the regression function with respect to the design density. As will be seen, this "parametric" problem characterizes the influence of serial dependence of errors on regression estimation.

Let $\Delta = \{\eta(u) = \int uhd\mu : u \in \mathcal{U}\}$ be the set of all possible mean values of $u(X)$ for the class $\mathcal{U}$. Let

$$r_n = \min_{\widehat{\eta}} \max_{u \in \mathcal{U}} E(\widehat{\eta} - \eta(u))^2 \qquad (2)$$

be the minimax risk for estimating $\eta(u)$, where the minimization is over $\widehat{\eta}$ based on $(X_i, Y_i)_{i=1}^n$.

2.3 Estimation of the regression function up to a constant

Long-range dependence makes the estimation of the mean of the regression function harder and therefore may affect the rate for estimating the whole regression function. In some applications, it is the trend or change of the function that is of interest. Then it is appropriate to estimate the regression function up to a constant.

Let $u_0(x) = u(x) - \eta(u)$ be a centered version of the regression function (centered according to the design density). The minimax risk for the estimation of u_0 is

$$R_0(\mathcal{U}; \Omega; n) = \min_{\hat{u}_0} \max_{u \in \mathcal{U}} E \parallel u_0 - \hat{u}_0 \parallel^2_{L_2(h)},$$

where $\hat{u}_0$ is over all estimators based on $(X_i, Y_i)_{i=1}^n$.

2.4 Risk for prediction

For prediction, we assume that X_i, $i \geq 1$ have the same marginal density h with the joint distribution known. When the data are dependent, the problem of prediction of the next response Y_n based on X_n and the past observations $(X_j, Y_j)_{j=1}^{n-1}$ may be essentially different from the problem of estimating the regression function. For the purpose of prediction, we may first "estimate" the conditional mean function of Y_n given X_n and $(X_j, Y_j)_{j=1}^{n-1}$. From a standard calculation, the conditional density of Y_n given $X_n = x$ and the past data, is normal with mean $m_{n-1,u}(x) = u(x) + \beta'_{n-1}\Omega^{-1}_{n-1}\left(Y^{n-1} - U^{n-1}\right)$ and variance $\sigma^2_n - \beta'_{n-1}\Omega^{-1}_{n-1}\beta_{n-1}$, where β_{n-1} is from partition $\Omega_n = \begin{pmatrix} \Omega_{n-1} & \beta_{n-1} \\ \beta'_{n-1} & \sigma^2_n \end{pmatrix}$ and $U^{n-1} = (U_1, ..., U_{n-1})'$ with $U_i = u(X_i)$ for $1 \leq i \leq n-1$. Here the conditional covariance matrix Ω_n is allowed to depend on X^n in general and therefore is random. For a given "estimator" $\widehat{m}_{n-1}$ of $m_{n-1,u}$, the square risk is

$$E\left(m_{n-1,u}(X_n) - \widehat{m}_{n-1}(X_n)\right)^2 = E \parallel m_{n-1,u} - \widehat{m}_{n-1} \parallel^2_{L_2(\nu_{n-1})},$$

where $\parallel \cdot \parallel^2_{L_2(\nu_i)}$ $(i \geq 0)$ denotes the L_2 distance with respect to the conditional distribution of X_{i+1} given X^i.

Once we have an estimator $\widehat{m}_{n-1}$, we may predict Y_n by $\widehat{m}_{n-1}(X_n)$ at X_n. This predictor has square risk

$$\begin{aligned} E\left(Y_n - \widehat{m}_{n-1}(X_n)\right)^2 &= E\left(E\left[\left(Y_n - \widehat{m}_{n-1}(X_n)\right)^2 \middle| X^{n-1}, Y^{n-1}, X_n\right]\right) \\ &= E\left(\sigma^2_n - \beta'_{n-1}\Omega^{-1}_{n-1}\beta_{n-1}\right) + E\left(m_{n-1,u}(X_n) - \widehat{m}_{n-1}(X_n)\right)^2. \end{aligned}$$

Note that the first term in the above decomposition does not depend on the predictors, indicating the prediction problem is equivalent to the problem of "estimating" the conditional mean $m_{n-1,u}$ as expected. Now we define the minimax average cumulative prediction risk as

$$\min_{\widehat{m}_{i-1}, 1 \leq i \leq n} \max_{u \in \mathcal{U}} (1/n) \sum_{i=1}^{n} E \parallel m_{i-1,u} - \widehat{m}_{i-1} \parallel^2_{L_2(\nu_{i-1})},$$

where the minimization is over all $\widehat{m}_{i-1}$ based on $(X_j, Y_j)_{j=1}^{i-1}$ for $1 \leq i \leq n$ respectively (for $i = 1$, $\widehat{m}_0$ is any initial guess, which does not have any effect in terms of rate of convergence). The average cumulative risk is natural for consideration for a prediction problem where one is interested in performance of a prediction strategy not just once but averaged over time.

Because of a technical reason, we study a clipped version of the prediction risk. For a fixed constant $A > 0$, let $\| g \|_{L_2(\nu_i),A} = \left(\int \min\left(|g|, A\right)^2 d\nu_i \right)^{1/2}$ denote the clipped L_2 norm with respect to the conditional distribution ν_i. We now redefine the minimax average cumulative prediction risk as follows

$$R_{ACP}(\mathcal{U};\Omega;n;A) = \min_{\widehat{m}_{i-1}, 1 \leq i \leq n} \max_{u \in \mathcal{U}} (1/n) \sum_{i=1}^{n} E \parallel m_{i-1,u} - \widehat{m}_{i-1} \parallel^2_{L_2(\nu_{i-1}),A} . \tag{3}$$

2.5 Metric entropy as a measure of massiveness of a function class

It is clear that the bigger the function class $\mathcal{U}$ is, the larger (at least no smaller) the minimax risk. For nonparametric regression with independent errors, it is known that massiveness of a target function class affects the minimax rate of convergence in terms of metric entropy order of the function class (see, e.g., Ibragimov and Hasminskii (1977), Bretagnolle and Huber (1979), Birgé (1983, 1986), Le Cam (1986, Chapter 16), Yatracos (1988), and Yang and Barron (1999)). Metric entropy as a measure of massiveness of a function class was intensively studied in Kolmogorov and Tihomirov (1959) and since then results have been obtained on the orders of metric entropy for the classical function classes and some others under various norms (see, e.g., Lorentz, Golitschek, and Makovoz (1996)).

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 $M_2(\epsilon) = M_2(\epsilon;\mathcal{U})$ be the maximal logarithm of the cardinality of any ϵ-packing set under the $L_2(h)$ distance. Clearly $M_2(\epsilon)$ is nonincreasing in ϵ. The asymptotic behavior of $M_2(\epsilon)$ as $\epsilon \to 0$ reflects how massive the class $\mathcal{U}$ is under the given distance. We call $M_2(\epsilon)$ the packing ϵ-entropy or simply the metric entropy of $\mathcal{U}$.

Throughout the paper, we assume $M_2(\epsilon) < \infty$ for every $\epsilon > 0$ (which necessarily requires $\mathcal{U}$ to be bounded in $L_2(h)$ norm) and $M_2(\epsilon) \to \infty$ as $\epsilon \to 0$ (which excludes trivial cases when $\mathcal{U}$ is finite). These conditions are satisfied if $\mathcal{U}$ is not finite, separable, and compact in $L_2(h)$ norm.

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

$$\liminf_{\epsilon \to 0} M(\tau\epsilon)/M(\epsilon) > 1. \tag{4}$$

This condition is a characteristic of familiar nonparametric classes (except classes of analytic functions), for which the metric entropy is usually of order $\epsilon^{-\alpha} \log(1/\epsilon)^{\beta}$ for some $\alpha > 0$ and $\beta \in R$.

3 Main results

In this paper, the expression $a_n \preceq b_n$ means that $\limsup (a_n/b_n) < \infty$. If $a_n \preceq b_n$ and $b_n \preceq a_n$ (i.e., a_n and b_n are of the same order), we write $a_n \asymp b_n$.

3.1 Regression Estimation

For regression estimation, the explanatory variables $X_1, X_2, ...$ are assumed to be i.i.d. with known density h with respect to a measure μ. They are further assumed to be independent of the errors ε_i's in model (1). The following additional assumptions will be used for our results.

Assumption A1: The class $\mathcal{U}$ is uniformly bounded, i.e., there exists a known constant L such that $\sup_{u \in \mathcal{U}} \| u \|_\infty \leq L < \infty$.

Assumption A2: The class $\mathcal{U}$ is rich as defined in (4).

Assumption A3: The class $\mathcal{U}$ contains the constant functions $u \equiv c$ with $c \in \Delta$.

Assumption A4: The mean value set Δ contains an interval $[a, b]$ with $a < b$.

Assumption A5: $\sup_{i \geq 1} \sigma_i^2 < \infty$.

Assumption A6: $Tr(\Omega_n^{-1}) \asymp n$.

Assumption A4 excludes cases where the estimation of $\eta(u)$ is trivial.

Choose ϵ_n such that

$$M(\epsilon_n) \asymp n\epsilon_n^2. \tag{5}$$

Under the richness assumption in (4), any two sequences of solution to the equation are of the same order, and ϵ_n^2 gives the minimax rate of convergence for estimating the regression function under i.i.d. errors (see, e.g., Birgé (1983), Le Cam (1985) and Yang and Barron (1999)). An interpretation of the equation is that if we discretize the function class $\mathcal{U}$ using an ϵ-net, then ϵ_n balances the estimation error of order $M(\epsilon)/n$ (due to identifying a good representor in the ϵ-net based on data) and the approximation error (bias squared, due to discretization) ϵ^2. Throughout the paper, unless stated otherwise, ϵ_n is defined as above.

Theorem 1: *If Assumptions A1-A6 are satisfied, we have the following conclusions.*

1. *The minimax risk for estimating u_0 is of order ϵ_n^2, i.e.,*

$$R_0(\mathcal{U}; \Omega; n) \asymp \epsilon_n^2. \tag{6}$$

2. *The minimax risk for regression function estimation is at rate of the maximum (or equivalently, the sum) of two quantities: the minimax rate of the same class but under i.i.d. errors, and the rate for estimating the mean $\eta = Eu(X)$ of the regression function under the correlated errors. That is,*

$$R(\mathcal{U}; \Omega; n) \asymp r_n + \epsilon_n^2. \tag{7}$$

REMARKS: 1. Without assuming $Tr(\Omega_n^{-1}) \asymp n$ (Assumption A6), the above quantities ϵ_n^2 and $r_n + \epsilon_n^2$ give valid upper rates respectively (see the proof of Theorem 1 in Section 5), but they are not necessarily optimal in general (see Section 2.1).

2. A parametric analogue of (6) is in Künsch et al (1993), where it is shown that the rate of convergence for estimating a contrast (similar in spirit to u_0) remain unchanged for some ANOVA models.

From above, in particular, for stationary Gaussian errors independent of X^n, the regression function up to a constant can be estimated as well as under i.i.d. errors. For the estimation of the whole regression function, however, the minimax rate for estimating $\eta(u)$ may hurt. Roughly speaking, the difficulty in estimating u is determined by the maximum of that caused by largeness of the function class $\mathcal{U}$ and that caused by the dependence among the errors in estimating a constant. The separation of the roles of the function class and dependence is somewhat surprising. This separation may not hold when the random errors and the explanatory variables are not independent.

From Theorem 1, once we know the metric entropy order of a nonparametric class and the minimax rate for estimating $\eta(u)$, the minimax rate for regression is determined. The metric entropies for classical function classes are usually of order $M(\epsilon) \asymp \epsilon^{-d/\alpha} (\log(1/\epsilon))^{\beta}$, where d is the dimension of $\mathcal{X}$, α is a smoothness parameter of the class measured in some way (e.g., in terms of derivatives, or a modulus of continuity) and $\beta \in R$. Then solving $M(\epsilon_n) = n\epsilon_n^2$, we have ϵ_n^2 of order $n^{-2\alpha/(2\alpha+d)} (\log n)^{2\alpha\beta/(2\alpha+d)}$. If $r_n \asymp n^{-\gamma}$ for some $0 < \gamma < 1$ (as for the long-range dependence case in Section 3.2), then

$$R(\mathcal{U};\Omega;n) \asymp \begin{cases} n^{-2\alpha/(2\alpha+d)} (\log n)^{2\alpha\beta/(2\alpha+d)} & \text{if } \gamma > 2\alpha/(2\alpha+d), \text{ or } \gamma = 2\alpha/(2\alpha+d) \text{ and } \beta \geq 0 \\ n^{-\gamma} & \text{if } \gamma < 2\alpha/(2\alpha+d), \text{ or } \gamma = 2\alpha/(2\alpha+d) \text{ and } \beta < 0. \end{cases}$$

If for some reason $Eu(X) = 0$ for all $u \in \mathcal{U}$, i.e., $\Delta = \{0\}$, then there is no need to estimate $\eta(u)$. As a consequence of Theorem 1, the rate of convergence for estimating the regression function is of order ϵ_n^2 regardless of the dependence among the errors.

We now consider the rate of convergence of r_n. Under Assumption A3, the problem of estimating $\eta \in \Delta$ based on $Y_i = \eta + \varepsilon_i$, $1 \leq i \leq n$ (without X_i, $1 \leq i \leq n$) is an easier subproblem with smaller minimax risk than that of estimating $\eta(u) = Eu(X)$ based on $(X_i, Y_i)_{i=1}^n$ with $Y_i = u(X_i) + \varepsilon_i$, $1 \leq i \leq n$ (see Lemma 6 in Section 5). That is, $r_n \geq \tilde{r}_n$, where $\tilde{r}_n$ is the minimax mean square error of the easier problem. Since $\{X_i\}_{i=1}^n$ is not involved, $\tilde{r}_n$ is handled more easily. Some results on $\tilde{r}_n$ were given in Hall and Hart (1990b). The following lemma gives useful bounds on r_n and $\tilde{r}_n$. Let $\mathbf{1}' = (1, 1, ..., 1)$ of dimension n.

LEMMA 1: *Under Assumption A4, the minimax risk $\tilde{r}_n$ satisfies*

$$\left(\mathbf{1}'\Omega_n^{-1}\mathbf{1}\right)^{-1} \preceq \tilde{r}_n \preceq \left(\mathbf{1}'\Omega_n\mathbf{1}\right)/n^2.$$

If $\left(\mathbf{1}'\Omega_n^{-1}\mathbf{1}\right)\left(\mathbf{1}'\Omega_n\mathbf{1}\right) \asymp n^2$ *and* $\mathbf{1}'\boldsymbol{\Omega}_n\mathbf{1} \succeq n$, *then under Assumptions A3 and A4,*

$$r_n \asymp \widetilde{r}_n \asymp \left(\mathbf{1}'\boldsymbol{\Omega}_n\mathbf{1}\right)/n^2.$$

Remarks: 1. The quantity $\left(\mathbf{1}'\boldsymbol{\Omega}_n^{-1}\mathbf{1}\right)^{-1}$ is the variance of the best linear unbiased estimator (BLUE) of η based on $Y_1, ..., Y_n$ with $Y_i = \eta + \varepsilon_i$, where $\{\varepsilon_i, 1 \leq i \leq n\}$ have the covariance matrix Ω_n. Adenstedt (1974) showed that for a wide range of stationary error sequences having a spectral density, the minimum variance $\left(\mathbf{1}'\boldsymbol{\Omega}_n^{-1}\mathbf{1}\right)^{-1}$ depends asymptotically only on the behavior of the spectral density near the origin.

2. Note that $\mathbf{1}'\Omega_n\mathbf{1}/n^2$ is the variance of the average of the errors, which determines the rate of convergence of $\sum_{i=1}^n Y_i/n$ as a simple estimator of η. For the case of long-range dependence, it behaves as well as the BLUE in terms of rate of convergence (see Adenstedt (1974) and Samarov and Taqqu (1988)). The condition $\left(\mathbf{1}'\boldsymbol{\Omega}_n^{-1}\mathbf{1}\right)\left(\mathbf{1}'\Omega_n\mathbf{1}\right) \asymp n^2$ is to say that BLUE and the simple estimator converge at the same speed (as in the case for the short- and long-range dependent cases in Section 3.3). The condition $\mathbf{1}'\Omega_n\mathbf{1} \succeq n$ exclude unusual situations (e.g., independent errors with $\sigma_i^2 = i^{-1}$) where a better rate than ϵ_n^2 is possible for regression.

3. If $|\sum_{i=1,j=1}^n Cov(\varepsilon_i, \varepsilon_j)| \preceq n$, then the dependence is weak and $\mathbf{1}'\boldsymbol{\Omega}_n\mathbf{1}/n^2 \asymp 1/n$. As a result, $r_n \asymp 1/n$ and from Theorem 1, we have the same rate of convergence for regression estimation as in the case of i.i.d. errors. For another extreme with $\mathbf{1}'\Omega_n\mathbf{1} \asymp n^2$ (see Section 3.3, Case 5), the minimax risk for estimating the regression function does not converge to zero at all under Assumptions A3 and A4, though the rate remains to be ϵ_n^2 for estimating u_0.

Theorem 2: *Under Assumptions A1-A6, if* $\left(\mathbf{1}'\boldsymbol{\Omega}_n^{-1}\mathbf{1}\right)\left(\mathbf{1}'\Omega_n\mathbf{1}\right) \asymp n^2$ and $\mathbf{1}'\Omega_n\mathbf{1} \succeq n$ then

$$R(\mathcal{U}; \Omega; n) \asymp \left(\mathbf{1}'\Omega_n\mathbf{1}\right)/n^2 + \epsilon_n^2. \tag{8}$$

3.2 Rates under long-range dependence

3.2.1 Long-range dependence

Assume that the errors are stationary and that the spectral density, say $f(\lambda)$ of the serially correlated errors exists. Let $r(i)$ denote the correlation between ε_j and ε_{j+i}. The error process is said to be long-range dependent if for some $c > 0$ and $0 < \gamma < 1$,

$$f(\lambda) \sim c\lambda^{-(1-\gamma)} \text{ as } \lambda \to 0 \tag{9}$$

(see, e.g., Cox (1984)). Then $r(j)$ is of order $|j|^{-\gamma}$.

Corollary 1: *Assume that* $f(\lambda)$ *satisfies (9), is continuous except at the origin and is bounded away from* 0. *Under Assumptions A1-A4, we have* $\widetilde{r}_n \asymp r_n \asymp n^{-\gamma}$ *and the minimax rate of convergence for regression estimation is*

$$R(\mathcal{U}; \Omega; n) \asymp n^{-\gamma} + \epsilon_n^2.$$

3.2.2 An example with Besov classes

For $1 \leq \sigma \leq \infty$, $1 \leq q \leq \infty$, and $\alpha/d > 1/q - 1/2$, let $B^{\alpha}_{\sigma,q}(C)$ be the collections of all functions $g \in L_q[0,1]^d$ such that the Besov norm satisfy $\| g \|_{B^{\alpha}_{\sigma,q}} \leq C$ (see e.g., DeVore and Lorentz (1993) and Triebel (1975)). Then the L_2 metric entropy is of order $\epsilon^{-d/\alpha}$ (see, e.g., Triebel (1975) and Lorentz, Golitschek, and Makovoz (1996, Chapter 15)). Assume the design density $h(x)$ of X with respect to Lebesgue measure μ is bounded above and away from zero. Then the metric entropy of the Besov class under $L_2(h)$ distance is of order $\epsilon^{-d/\alpha}$. Application of Corollary 1 yields the minimax rate of convergence under the long-range dependence:

$$R(B^{\alpha}_{\sigma,q}(C); \Omega; n) \asymp n^{-\min(2\alpha/(2\alpha+d),\gamma)}. \tag{10}$$

3.2.3 A comparison with an equally spaced fixed design

Results on minimax rates are obtained for long-range dependent errors with a one-dimensional equally spaced fixed design in Hall and Hart (1990a), Wang (1996), and Johnstone and Silverman (1997) for some concrete smoothness function classes. The model being considered is

$$Y_i = u(i/n) + \varepsilon_i, 1 \leq i \leq n,$$

where $Cor(\varepsilon_i, \varepsilon_j) \sim c|i-j|^{-\gamma}$ for some $0 < \gamma < 1$, and u is in Besov class $B^{\alpha}_{\sigma,q}(C)$ (or a differentiable class in Hall and Hart (1990a)). The minimax rate of convergence for estimating u under squared L_2 loss is shown to be of order $n^{-2\alpha\gamma/(2\alpha+\gamma)}$.

Assume there are only measurement errors (independent of the sampling sites X_i's) in the responses and the errors are long-range dependent in the order of measurements. For this case, if one uses an equally spaced fixed design, and if the order of measurements corresponds to the order of the sites, the rate of convergence is $n^{-2\alpha\gamma/(2\alpha+\gamma)}$ from above. Alternatively, if one uses a random design, from (10), the rate of convergence is $n^{-\min(2\alpha/(2\alpha+1),\gamma)}$, which is faster compared to that with the fixed design. An explanation of the difference in rates is as follows. Under the fixed design, observations with x values close to each other are highly correlated. With the random design, however, the orders of the measurements of the observations at nearby x values are not necessarily adjacent but on average quite far away from each other, resulting in weaker correlations between observations that are close in terms of x values. Thus it is clear that the latter is preferred to the former design. A closer look suggests that the difference in rates is not due to the difference in random and fixed designs, but rather because the order of measurements are not randomized for the fixed design case. If one uses an equally spaced fixed design, one should randomize the order of measurements and we expect the same rate of convergence as under the random design. This example also illustrates importance of the randomization principle in statistical experimental design as well demonstrated earlier in Künsch et al (1993) under some parametric settings with long-range dependence.

3.3 Examples of dependence

For simplicity, we focus on stationary errors.

1. *Exponentially decaying correlation.* Let $r(j) = \sigma^2\theta^j, j \geq 0$ for some constants $\sigma^2 > 0$ and θ with $|\theta| < 1$. Then it can be shown that $Tr(\Omega_n^{-1}) \asymp n$ and $\mathbf{1}'\Omega_n\mathbf{1}/n^2$ is of order n^{-1}.

2. *Short-range dependence.* More generally than the above case, we assume that the errors are weakly correlated or short-range dependent in the sense $\sum_{k=0}^{m}|r(k)|$ converges as $m \to \infty$. Then $\mathbf{1}'\Omega_n\mathbf{1}/n^2$ is of order n^{-1}. A special case is finite memory dependence where the errors are correlated only when they are not far away from each other, i.e., $r(j) = 0$ when $j \geq j^*$ for some $j^* > 1$. Another example is $r(k) \asymp |k|^{-\gamma}$ with $\gamma > 1$.

3. *Long-range dependence.* Assume $f(\lambda) = f^*(\lambda)|1 - e^{i\lambda}|^{-(1-\gamma)}$ for some $0 < \gamma < 1$, where $f^*(\lambda)$ is a strictly positive continuous function. This includes the spectral density of a fractional Gaussian noise model (Mandelbrot and Van Ness (1968)) and a fractional ARIMA model (Granger and Joyeux (1980) and Hosking (1981)). For the first case, $r(j) = c/2\left(|j+1|^{2-\gamma} - 2|j|^{2-\gamma} + |j-1|^{2-\gamma}\right)$ (then $r(j) \sim c'j^{-\gamma}$ for some constant $c' > 0$). Fractional ARIMA(p, d, q) process has a spectral density $f(\lambda; d, \phi, \theta) = c|\theta\left(e^{i\lambda}\right)|^2/|\phi\left(e^{i\lambda}\right)(1 - e^{i\lambda})^d|^2$, where $\theta(z) = 1 - \sum_{j=1}^{q}\theta_j z^j$ and $\phi(z) = 1 - \sum_{j=1}^{p}\phi_j z^j$ are polynomials of order q and p respectively. From Corollary 1, $r_n \asymp n^{-\gamma}$ (see also Hall and Hart (1990b)).

4. *Alternating dependence.* For the above long-range dependence, the errors are eventually positively correlated, i.e., $r(j) > 0$ when j is large enough. Now suppose $r(j) \sim c(-1)^j|j|^{-\gamma}$ for some $\gamma > 0$ as $j \to \infty$. One can obtain such a dependence from long-range dependent errors $\{\varepsilon_i\}$ by considering $\{(-1)^i\varepsilon_i\}$. Then because the covariances essentially cancel out even when $0 < \gamma < 1$, the rate of convergence for estimating $\eta(u)$ under this correlation is still of order $1/n$.

5. *An excessively highly correlated case.* Let Ω_n have diagonal elements σ^2 and off-diagonal elements $\sigma^2\theta$. For $0 < \theta < 1$, Ω_n is positive definite for all n. For this case, $\left(\mathbf{1}'\mathbf{\Omega_n}\mathbf{1}\right)/n^2 \asymp 1$, and since $\mathbf{1}$ is an eigenvector of Ω_n, the product $\left(\mathbf{1}'\mathbf{\Omega_n^{-1}}\mathbf{1}\right)\left(\mathbf{1}'\mathbf{\Omega_n}\mathbf{1}\right)$ is easily seen to be of order n^2 as useful for applying Theorem 2.

For Cases 2-4, it is assumed that the spectral density of the errors is bounded away from 0. Then $Tr(\Omega_n^{-1}) \asymp n$ (see Lemma 8 in the appendix). Note that the trace condition is automatically satisfied for the other cases.

Take the Besov classes $B^{\alpha}_{\sigma,q}(C)$ for examples. Based on Theorem 2, the minimax rate of convergence for estimating u is $n^{-2\alpha/(2\alpha+d)}$ for Cases 1, 2 and 4, and is worsened to $n^{-\min(2\alpha/(2\alpha+d),\gamma)}$ for Case 3 (as seen in the previous subsection). For Case 5, by Theorem 1, the minimax rate for estimating u_0 is still $n^{-2\alpha/(2\alpha+d)}$. However, since $\left(\mathbf{1}'\mathbf{\Omega_n}\mathbf{1}\right)/n^2 \asymp 1$, the minimax risk for estimating u does not converge at all.

3.4 Rate of minimax risk for prediction

For prediction, different assumptions will be used for model (1).

Assumption A7: $X_i, i \geq 1$ have the same marginal density function h with known joint distribution.

Assumption A8: There is an i.i.d. component in the errors, i.e., $\varepsilon_i = \varepsilon_i^{(1)} + \varepsilon_i^{(2)}$ with $\{\varepsilon_i^{(1)}\}$ i.i.d. independent of $\{\varepsilon_i^{(2)}\}$.

Assumption A9: The conditional covariances of the errors are uniformly bounded, i.e.,

$$\sup_{n \geq 1} \sup_{1 \leq i \leq n} \sup_{x^n} Var(\varepsilon_i | X^n = x^n) < \infty. \tag{11}$$

THEOREM 3: *Under Assumptions A1-A2 and A7-A9, dependence among the errors does not hurt the rate of convergence for prediction, i.e., the average cumulative prediction risk in (3) satisfies*

$$R_{ACP}(\mathcal{U}; \Omega; n; A) \preceq \epsilon_n^2.$$

Note that when the errors are independent, the average cumulative prediction risk and the individual prediction risk are of the same order as ϵ_n^2 (see Yang and Barron (1999)). The above result shows that as long as the covariances are known, dependence of errors does not harm prediction in terms of rate of convergence as intuition also suggests. Dependence can result in faster rate of convergence for prediction. For example, as an extreme case, if the errors are identical, then under smoothness conditions on the regression function u, prediction risk can be as small as of order n^{-2} by a simple interpolation.

For the prediction result, the errors are not required to be independent of the explanatory variables. For example, consider the following dependence structure

$$Cov(\varepsilon_i, \varepsilon_j | X^n = x^n) = \sigma_W^2 \delta_{ij} + \sigma_S^2 \rho_S(x_i - x_j) + \sigma_T^2 \rho_T(i - j), \tag{12}$$

where $\sigma_W^2 > 0$, σ_S^2 and σ_T^2 are nonnegative constants, δ_{ij} equals 1 if $i = j$ and equals 0 otherwise, ρ_S is a correlation function defined on $\{x - x' : x, x' \in \mathcal{X}\}$ and ρ_T is a correlation function defined on integers. An interpretation of this dependence structure is that the total errors in the response come from three independent components: white noise $\{\varepsilon_{W,i}, i \geq 1\}$, "spatially" correlated errors $\{\varepsilon_{S,i}, i \geq 1\}$, and time (or order of observation or measurement) dependent errors $\{\varepsilon_{T,i}, i \geq 1\}$. The two correlation functions could be general (but known), including short-range and long-range situations. Spatial long-range dependent processes have been constructed (e.g., Whittle (1962), Gay and Heyde (1990) and Renshaw (1994)). The representation of the overall covariance in terms of sum of the three components is not necessarily unique. Since the covariance is assumed to be known, this is not a problem here. The condition in (11) becomes $\rho_S(0) < \infty$ and $\rho_T(0) < \infty$.

Corollary 2: *With dependence given in (12), under Assumptions A1, A2 and A7, and* $\rho_S(0) < \infty$ and $\rho_T(0) < \infty$, *we have*

$$R_{ACP}(\mathcal{U}; \Omega; n; A) \preceq \epsilon_n^2.$$

Example 1: For simplicity, consider a one-dimensional case. Assume $\{X_i, i \geq 1\}$ are i.i.d. from Cauchy distribution (i.e., $h(x) = \left(\pi(1+x^2)\right)^{-1}$ with respect to Lebesgue measure). Let $\mathcal{U}$ consist of all monotone nondecreasing functions on R bounded between two constants. The ϵ-metric entropy of $\mathcal{U}$ under $L_2(h)$ distance is of order ϵ^{-1} (which is seen using $\tan^{-1}$ transformation on x and the fact that the L_2 metric entropy of a uniformly bounded monotone function class with a compact support is of order $1/\epsilon$). Assume that the errors have a conditional covariance matrix as in (12). The spatial and/or serial correlations can be either short- or long-range dependent. For an example of a long-range spatial correlation, let $\{V(t), t \in R\}$ be a stationary process with spectral density

$$f_V(\omega) = \sigma^2 \pi^{-1} \sin^2(\omega/2) \omega^{-2-4\gamma_S},$$

where $0 < \gamma_S < 1/4$ (see Gay and Heyde (1990)). The asymptotic covariance is $Cov(V(t), V(t+\tau) \sim c\tau^{4\gamma_S - 1}$ as $\tau \to \infty$. From Corollary 2, solving equation $1/\epsilon = n\epsilon^2$, we know that the minimax rate for prediction is $O\left(n^{-2/3}\right)$.

4 A key proposition and its derivation

4.1 Minimax upper and lower bounds for regression

Assume that the errors are independent of X^n. Let $\rho_n = Tr(\Omega_n^{-1})$.

Choose $\widetilde{\epsilon}_n$ such that

$$M_2(\widetilde{\epsilon}_n) = (1/2)\, \rho_n \widetilde{\epsilon}_n^2. \tag{13}$$

Let

$$\psi_n = (11/2)\, \rho_n \widetilde{\epsilon}_n^2 + \log\left(8Ln^{1/2}/\widetilde{\epsilon}_n\right)$$

and let $\underline{\epsilon}_n$ be chosen to satisfy

$$M_2(\underline{\epsilon}_n) = 2\psi_n. \tag{14}$$

Let $\overline{\epsilon}_n$ satisfy

$$M_2(\overline{\epsilon}_n) = n\overline{\epsilon}_n^2/2, \tag{15}$$

and define

$$\overline{\psi}_n = (11/2)\, n\overline{\epsilon}_n^2 + \log\left(8Ln^{1/2}/\overline{\epsilon}_n\right),$$

$$\psi_n^* = \min\left(\psi_n, \overline{\psi}_n\right).$$

Typically, (e.g., when ρ_n is of a polynomial order in n), the component $\rho_n\tilde{\epsilon}_n^2$ (or $n\bar{\epsilon}_n^2$) dominates the other term in ψ_n (or $\overline{\psi}_n$). Then under the richness condition in (4), $\tilde{\epsilon}_n$ and $\underline{\epsilon}_n$ are of the same order. If $\rho_n \asymp n$, then $\tilde{\epsilon}_n$, $\underline{\epsilon}_n$, $\bar{\epsilon}_n$, ψ_n/n, and $\overline{\psi}_n/n$ are all of the same order. They are also of the same order as ϵ_n determined by $M(\epsilon_n) = n\epsilon_n^2$ in (5) with $M(\epsilon)$ of order $M_2(\epsilon)$ (see Yang and Barron (1999)). Let $\overline{\sigma}^2 = \sup_{i\geq 1}\sigma_i^2$.

Proposition 0: *Under Assumptions A1 and A5, the minimax squared $L_2(h)$ risk for regression function estimation is bounded as follows*:

$$\max\left(\underline{\epsilon}_n^2/8, r_n\right) \leq R(\mathcal{U};\Omega;n) \quad \leq r_n + C_{L,\overline{\sigma}^2}\psi_n^*/n,$$

where $C_{L,\overline{\sigma}^2}$ is a constant depending on L and $\overline{\sigma}^2$.

Remark: Without the richness assumption (4), even under $\rho_n \asymp n$, the upper and lower bounds in the above proposition may not be of the same order. For example, for classes of analytic functions, the metric entropies are of polynomial orders of $\log(1/\epsilon)$ (Kolmogorov and Tihomirov (1959)) and the upper and lower bounds differ in a logarithmic term unless r_n dominates. It seems that an use of local entropy (instead of global entropy) as pioneered by Le Cam (1975) and Birgé (1983) in the construction of the upper bound may overcome the gap.

4.2 Proof of Proposition 0

In Yang and Barron (1999), minimax rates of convergence for regression under independent Gaussian errors are derived using a connection between density estimation and data compression. The Cesaro average of the Bayes predictive density estimators of the joint distribution of (X, Y) based on the uniform prior on a suitably chosen ϵ-net in the regression function class $\mathcal{U}$ is used to produce an estimator of the regression function to obtain a minimax upper bound. For regression with dependent errors, however, due to correlations, the Bayes predictive density "estimators" are targeted at the conditional distributions of (X_i, Y_i), $i \geq 1$, given the past observations. They are no longer appropriate for estimating the distributions of (X_i, Y_i). It becomes much harder to derive a rate-optimal estimator under general conditions on $\mathcal{U}$ and Ω. The difficulty is overcome through rather delicate adjustments of the Bayes predictive estimators as will be seen.

We give more notations first. Let $Z = (X, Y)$, $z = (x, y)$, $z^n = (z_1, .., z_n)$. Let $U^n = (u(X_1), ..., u(X_n))$ and $u^n = (u(x_1), ..., u(x_n))$.

4.2.1 Lower bound

We prove $R(\mathcal{U};\Omega;n) \geq \underline{\epsilon}_n^2/8$ and $R(\mathcal{U};\Omega;n) \geq r_n$ separately. The second inequality follows basically from the observation that estimating the whole regression function is at least as difficult as estimating the mean of the regression function. The proof of the first one utilizes Fano's inequality together with a suitable upper bound on the involved mutual information.

Let $N_{\underline{\epsilon}_n}$ be an $\underline{\epsilon}_n$-packing set with the maximum cardinality in $\mathcal{U}$ and let $G_{\tilde{\epsilon}_n}$ be an $\tilde{\epsilon}_n$-net for $\mathcal{U}$ both under $L_2(h)$ distance. Since an ϵ-packing set with the maximum cardinality is automatically an ϵ-covering set, we can find a $G_{\tilde{\epsilon}_n}$ such that $\log|G_{\tilde{\epsilon}_n}| = M_2(\tilde{\epsilon}_n)$. Following now a standard argument using Fano's inequality (see, e.g., Birgé (1983, Proposition 2.8), Yu (1996, p. 427) and Yang and Barron (1999, pp. 1570-1571), we have

$$\min_{\widehat{u}} \max_{u\in\mathcal{U}} E_u \parallel u - \widehat{u} \parallel^2_{L_2(h)} \geq (\underline{\epsilon}_n^2/4)\left(1 - \frac{I\left(U;Z^n\right) + \log 2}{\log|N_{\underline{\epsilon}_n}|}\right)$$

where the Shannon's mutual information $I\left(U;Z^n\right)$ is equal to the average (with respect to the uniform prior w) of the Kullback-Leibler (K-L) divergence between $p_u(z^n)$ and $p^w(z^n) = \sum_{u\in N_{\underline{\epsilon}_n}} p_u(z^n)/|N_{\underline{\epsilon}_n}|$. Here

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

Since the Bayes mixture density $p^w(z^n)$ minimizes the average K-L divergence over all choices of joint density $q(z^n)$ on the sample space $\mathcal{Z}^n$, the mutual information is upper bounded by the maximum K-L divergence between $p_u(z^n)$ and any $q(z^n)$. That is,

$$I\left(U;Z^n\right) \leq \max_{u\in N_{\underline{\epsilon}_n}} D\left(P_{Z^n,u} \parallel Q_{Z^n}\right).$$

We will choose $q(z^n) = (1/|G|)\sum_{u\in G} p_u(z^n)$ for a certain appropriate covering set G.

Key to the analysis is the following expression for the K-L divergence between $P_{Z^n,u}$ and $P_{Z^n,v}$ (see Lemma 2 in Section 5):

$$D\left(P_{Z^n,u} \parallel P_{Z^n,v}\right) = (1/2)\,\rho_n \parallel u - v \parallel^2_{L_2(h)} + (1/2)\left(\sum_{i\neq j}\omega_{i,j}^{-1}\right)\left(Eu(X) - Ev(X)\right)^2, \qquad (16)$$

where $\omega_{i,j}^{-1}$ denotes the (i,j)-th element of Ω_n^{-1}. When the errors are i.i.d. the second term in the above expression is zero and one can simply take G to be $G_{\tilde{\epsilon}_n}$ and obtain the right order upper bound on $\max_{u\in N_{\underline{\epsilon}_n}} D\left(P_{Z^n,u} \parallel Q_{Z^n}\right)$ as shown in Yang and Barron (1999). For dependent errors, $\sum_{i\neq j}\omega_{i,j}^{-1}$ might be large compared to ρ_n and the choice of $G_{\tilde{\epsilon}_n}$ together with the familiar bound $\left(Eu(X) - Ev(X)\right)^2 \leq \parallel u - v \parallel^2_{L_2(h)}$ is not sufficient for the result. We instead construct a covering set carefully to handle this term $\left(\sum_{i\neq j}\omega_{i,j}^{-1}\right)\left(Eu(X) - Ev(X)\right)^2$. The idea is to slightly enlarge $G_{\tilde{\epsilon}_n}$ by adding constants so that for each $u \in \mathcal{U}$, we can find v in the enlarged covering set such that both terms in (16) are well behaved. Details are as follows.

Let $A_n = \{a_1, a_2, ..., a_m\}$, $a_j = -2L + j\delta\tilde{\epsilon}_n$ be equally spaced points in $[-2L, 2L]$ with width $\delta\tilde{\epsilon}_n$ and $m = \lfloor 4L/\left(\delta\tilde{\epsilon}_n\right)\rfloor$ (recall that L is an upper bound on the sup-norms of functions in $\mathcal{U}$). Let us consider an enlarged net $\widetilde{G}_{\tilde{\epsilon}_n} = \{v + a : v \in G_{\tilde{\epsilon}_n}$ and $a \in A_n\}$. Note that $\log\left(|\widetilde{G}_{\tilde{\epsilon}_n}|\right) \leq M_2(\tilde{\epsilon}_n) + \log\left(4L/\left(\delta\tilde{\epsilon}_n\right)\right)$. For any $u \in \mathcal{U}$, there exist $\widetilde{u} \in G_{\tilde{\epsilon}_n}$ and $a^* \in A_n$ such that $\parallel u - \widetilde{u} \parallel_{L_2(h)} \leq \tilde{\epsilon}_n$ and $|\int\left(\widetilde{u} - u\right)hd\mu - a^*| \leq \delta\tilde{\epsilon}_n$. Then $|a^*| \leq \delta\epsilon_n + |\int(u - \widetilde{u})hd\mu| \leq (1+\delta)\tilde{\epsilon}_n$.

Let $\tilde{\tilde{u}} = \tilde{u} - a^*$, then $| \int \left(u - \tilde{\tilde{u}}\right) hd\mu| \leq \delta\tilde{\epsilon}_n$, and $\| u - \tilde{\tilde{u}} \|_{L_2(h)} \leq \| u - \tilde{u} \|_{L_2(h)} + \| \tilde{\tilde{u}} - \tilde{u} \|_{L_2(h)} \leq (2+\delta)\tilde{\epsilon}_n$. Clearly we have $\tilde{\tilde{u}} \in \widetilde{G}_{\tilde{\epsilon}_n}$. From (16), we have $D(P_{Z^n,u} \parallel P_{Z^n,\tilde{\tilde{u}}}) \leq (1/2)(2+\delta)^2\rho_n\epsilon_n^2 + (1/2)\max(0, \varpi_n)\delta^2\tilde{\epsilon}_n^2$, where $\varpi_n = \sum_{i\neq j, 1\leq i,j\leq n} \omega_{i,j}^{-1}$. Now choose w_1 to be the uniform prior on $\widetilde{G}_{\tilde{\epsilon}_n}$ and let $q(z^n) = p^{w_1}(z^n) = \sum_{u\in\widetilde{G}_{\epsilon_n}} w_1(u)p_u(z^n)$ and Q_{Z^n} be the corresponding Bayes mixture density and distribution respectively. Let $\lambda_{(1),n} \leq \lambda_{(2),n} \leq \cdots \leq \lambda_{(n),n}$ be the eigenvalues of Ω_n. Then $\varpi_n \leq \mathbf{1}'\Omega_n^{-1}\mathbf{1} \leq n\lambda_{(1),n}^{-1}$. Since $\rho_n = \sum_{i=1}^n \lambda_{(i),n}^{-1} \geq \lambda_{(1),n}^{-1}$, we have $\varpi_n/\rho_n \leq n$. From above we have that for any $u \in \mathcal{U}$,

$$\begin{aligned} D\left(P_{Z^n,u} \parallel Q_{Z^n}\right) &= E\log \frac{p_u(z^n)}{(1/|\widetilde{G}_{\tilde{\epsilon}_n}|)\sum_{u'\in\widetilde{G}_{\tilde{\epsilon}_n}} p_{u'}(z^n)} \\ &\leq E\log \frac{p_u(z^n)}{(1/|\widetilde{G}_{\tilde{\epsilon}_n}|)p_{\tilde{\tilde{u}}}(z^n)} \\ &= \log|\widetilde{G}_{\tilde{\epsilon}_n}| + D\left(P_{Z^n,u} \parallel P_{Z^n,\tilde{\tilde{u}}}\right) \\ &\leq M_2(\tilde{\epsilon}_n) + \log\left(4L/\left(\delta\tilde{\epsilon}_n\right)\right) + (1/2)\left(2+\delta\right)^2\rho_n\tilde{\epsilon}_n^2 + (1/2)\, n\rho_n\delta^2\tilde{\epsilon}_n^2. \end{aligned} \tag{17}$$

Taking $\delta = n^{-1/2}$, together with our choice of $\tilde{\epsilon}_n$ in (13), we have

$$D\left(P_{Z^n,u} \parallel Q_{Z^n}\right) \leq \log\left(4Ln^{1/2}/\tilde{\epsilon}_n\right) + (11/2)\,\rho_n\tilde{\epsilon}_n^2. \tag{18}$$

Thus we have shown that $I\left(U; Z^n\right) \leq \log\left(4Ln^{1/2}/\tilde{\epsilon}_n\right) + (11/2)\,\rho_n\tilde{\epsilon}_n^2$. By our choice of $\underline{\epsilon}_n$ in (14), $\left(I\left(U; Z^n\right) + \log 2\right)/\log|N_{\underline{\epsilon}_n}| \leq \frac{1}{2}$. Thus $\min_{\hat{u}}\max_{u\in\mathcal{U}} E \parallel u - \hat{u} \parallel_{L_2(h)}^2 \geq \underline{\epsilon}_n^2/8$.

The inequality $R\left(\mathcal{U}; \Omega; n\right) \geq r_n$ follows from the simple fact that for any estimator $\widehat{u}$ based on Z^n, let $\widehat{\eta} = \int \widehat{u}hd\mu$, then

$$E(\widehat{\eta} - \eta)^2 = E\left(\int\left(\widehat{u} - u\right)hd\mu\right)^2 \leq E \parallel \widehat{u} - u \parallel_{L_2(h)}^2.$$

4.2.2 Upper bound

We divide the proof of the upper bound in several steps. In Step 1, as in the derivation of the lower bound, consider the covering set $\widetilde{G}_{\tilde{\epsilon}_n}$ with uniform prior. We show the resulting Bayes predictive densities (at different sample sizes) are good "estimators" of the conditional densities of the observations Z_i given the past Z^{i-1}. The Bayes predictive densities are mixtures of Gaussian densities. In Step 2, based on the Bayes predictive densities, we construct density estimators (of the same conditional densities) that have the form of a single Gaussian density (instead of a mixture) still with good risk bounds. Being a single Gaussian density is important in the later construction of the regression estimator. In Step 3, the risk bounds on the estimators in Step 2 are shown to imply that the regression function can be estimated well up to a constant. In Step 4, the estimation of the constant is shown to be determined by the correlations between the errors. Together with Step 3, we have a good estimator of the regression function. In Step 5, we consider the case when ρ_n is of higher order than n. A suitable modification improves the upper rate of convergence. This is why ψ_n^* is used instead of ψ_n in the upper bound in Proposition 0.

Step 1 As in the derivation of lower bounds, consider the covering set $\widetilde{G}_{\tilde{\epsilon}_n}$ with uniform prior w_1. Let the Bayes predictive density estimators be $\hat{p}_i(z) = p\left(Z_{i+1}|Z^i\right)$ evaluated at $Z_{i+1} = z$, which equal $p^{w_1}(Z^i, z)/p^{w_1}(Z^i)$ for $i > 0$ and $\hat{p}_i(z) = p^{w_1}(z) = \left(1/|\widetilde{G}_{\tilde{\epsilon}_n}|\right) \sum_{u \in \widetilde{G}_{\tilde{\epsilon}_n}} p_u(z)$ for $i = 0$. For $n \geq 1$, 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 $(X_j, Y_j)_{j=1}^i$, Y_{i+1} has a normal distribution with mean $m_{i,u}(x|Z^i) = u(x) + \beta_i' \Omega_i^{-1} \left(Y^i - U^i\right)$ and variance $\sigma_{i+1}^2 - \beta_i' \Omega_i^{-1} \beta_i$. Let

$$\begin{aligned} p_{z_{i+1}|Z^i;u}(x_{i+1}, y_{i+1}) \quad = \quad & h(x_{i+1}) \left(2\pi \left(\sigma_{i+1}^2 - \beta_i' \Omega_i^{-1} \beta_i\right)\right)^{-1/2} \times \\ & \times \exp\left(-1/\left(2\left(\sigma_{i+1}^2 - \beta_i' \Omega_i^{-1} \beta_i\right)\right) \left(y_{i+1} - m_{i,u}(x_{i+1}|Z^i)\right)^2\right). \end{aligned} \tag{19}$$

It is the conditional density of Z_{i+1} given Z^i under the regression function u. Then by the chain rule (e.g., Barron (1987)), for any $u \in \mathcal{U}$,

$$\sum_{i=0}^{n-1} E \log \frac{p_{z_{i+1}|Z^i;u}(Z_{i+1})}{\widehat{p}_i(Z_{i+1})} = E \log \frac{p_u(Z^n)}{p^{w_1}(Z^n)} = D\left(P_{Z^n,u} \parallel Q_{Z^n}\right) \leq \psi_n,$$

where the last inequality is as in (18). Thus

$$\max_{u \in \mathcal{U}} \sum_{i=0}^{n-1} ED(p_{z_{i+1}|Z^i;u} \parallel \widehat{p}_i) \leq \psi_n. \tag{20}$$

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

$$\max_{u \in \mathcal{U}} \sum_{i=0}^{n-1} E d_H^2(p_{z_{i+1}|Z^i;u}, \widehat{p}_i) \leq \psi_n.$$

This means that we can estimate (or predict) well the conditional densities of Z_{i+1} given Z^i by $\widehat{p}_i$'s in terms of the cumulative squared Hellinger risk.

Step 2 Note that $\hat{p}_i(x_{i+1}, y_{i+1})$ takes the form of $h(x_{i+1})\hat{g}_i(y_{i+1}|x_{i+1})$, where $\hat{g}_i(y_{i+1}|x_{i+1})$ is an estimator of the conditional density of Y_{i+1} given X_{i+1} and Z^i. It is a mixture of Gaussians using a posterior based on the uniform prior on the ϵ-net. We now construct an estimator taking the form of a single Gaussian density. The simplified form (instead of a mixture) is easier to work with in the next step. First fix $v^i \in R^i$. For given $(X_j, Y_j)_{j=1}^i$ and v^i, for each x, let $\widetilde{m}_i(x) = \widetilde{m}_i(x|v^i)$ be the minimizer of the Hellinger distance $d_H\left(\hat{g}_i(\cdot|x), \phi_b\right)$ between $\hat{g}_i(y \mid x)$ and the normal density $\phi_b(y)$ with mean b and the variance $\sigma_{i+1}^2 - \beta_i' \Omega_i^{-1} \beta_i$ over choices of b with $|b - \beta_i' \Omega_i^{-1} \left(Y^i - v^i\right)| \leq L$. Here $\widetilde{m}_i(x|v^i)$ and $\overline{u}_i(x) = \overline{u}_i(x|v^i) = \widetilde{m}_i(x) - \beta_i' \Omega_i^{-1} \left(Y^i - v^i\right)$ can be viewed as "estimators" of the conditional mean $m_{i,u}$ and of u respectively based on $(X_j, Y_j)_{j=1}^i$ except that v^i is used in place of U^i (unknown) in the second term of $u(x) + \beta_i' \Omega_i^{-1} \left(Y^i - U^i\right)$.

Denote by $p_{z_{i+1}|Z^i;s;v^i}$ the density function of (x_{i+1}, y_{i+1}):

$$\begin{aligned} & h(x_{i+1}) \left(2\pi \left(\sigma_{i+1}^2 - \beta_i' \Omega_i^{-1} \beta_i\right)\right)^{-1/2} \times \\ & \quad \exp\left(-1/\left(2\left(\sigma_{i+1}^2 - \beta_i'\Omega_i^{-1}\beta_i\right)\right)\left(y_{i+1} - \left(s(x_{i+1}) + \beta_i'\Omega_i^{-1}(Y^i - v^i)\right)^2\right), \end{aligned}$$

with given Z^i, function $s(x)$, and v^i. Let v_*^i be the minimizer of $d_H^2(\widehat{p}_i, p_{z_{i+1}|Z^i;\overline{u}_i;v^i})$ over $v^i \in R^i$ and denote the corresponding $\widetilde{m}_i$ and $\overline{u}_i$ by $\widetilde{m}_i^*$ and $\overline{u}_i^*$. Then using triangle inequality,

$$\begin{aligned} & d_H^2(p_{z_{i+1}|Z^i;u}, p_{z_{i+1}|Z^i;\overline{u}_i^*;v_*^i}) \\ \leq \;& 2d_H^2(p_{z_{i+1}|Z^i;u}, \widehat{p}_i) + 2d_H^2(p_{z_{i+1}|Z^i;\overline{u}_i^*;v_*^i}, \widehat{p}_i) \\ \leq \;& 2d_H^2(p_{z_{i+1}|Z^i;u}, \widehat{p}_i) + 2d_H^2(p_{z_{i+1}|Z^i;\overline{u}_i^0;U^i}, \widehat{p}_i) \\ \leq \;& 2d_H^2(p_{z_{i+1}|Z^i;u}, \widehat{p}_i) + 2d_H^2(p_{z_{i+1}|Z^i;u}, \widehat{p}_i) \\ = \;& 4d_H^2(p_{z_{i+1}|Z^i;u}, \widehat{p}_i), \end{aligned}$$

where in the second inequality, $\overline{u}_i^0$ is $\overline{u}_i(x|U^i)$ $(v^i = U^i)$ and for the third inequality, we use the fact that $d_H^2(p_{z_{i+1}|Z^i;\overline{u}_i^0;U^i}, \widehat{p}_i) = \int h(x_{i+1}) d_H^2\left(\hat{g}_i(\cdot|x_{i+1}), \phi_{\overline{u}_i^0 + \beta_i'\Omega_i^{-1}(Y^i - U^i)}\right) d\mu$ is upper bounded by $\int h(x_{i+1}) d_H^2\left(\hat{g}_i(\cdot|x_{i+1}), \phi_{m_{i,u}}\right) d\mu = d_H^2(\widehat{p}_i, p_{z_{i+1}|Z^i;u})$. It follows that

$$\max_{u\in\mathcal{U}} \sum_{i=0}^{n-1} E d_H^2(p_{z_{i+1}|Z^i;u}, p_{z_{i+1}|Z^i;\overline{u}_i^*;v_*^i}) \leq 4 \max_{u\in\mathcal{U}} \sum_{i=0}^{n-1} E d_H^2(p_{z_{i+1}|Z^i;u}, \widehat{p}_i) \leq 4\psi_n.$$

Thus the estimators $p_{z_{i+1}|Z^i;\overline{u}_i^*;v_*^i}$ of a simpler form continue to have a good bound on the cumulative Hellinger risk.

Step 3 Now note that

$$E d_H^2(p_{z_{i+1}|Z^i;u}, p_{z_{i+1}|Z^i;\overline{u}_i^*;v_*^i}) = 2E \int h(x) \left(1 - e^{-\left(\left(u(x) - \overline{u}_i^*(x)\right) - \beta_i'\Omega_i^{-1}(U^i - v_*^i)\right)^2 / \left(8\left(\sigma_{i+1}^2 - \beta_i'\Omega_i^{-1}\beta_i\right)\right)}\right) d\mu.$$

From Lemma 3 in Section 5,

$$\begin{aligned} & \int h(x) \left(1 - e^{-\left(u(x) - \overline{u}_i^*(x) - \beta_i'\Omega_i^{-1}(U^i - v_*^i)\right)^2 / \left(8\left(\sigma_{i+1}^2 - \beta_i'\Omega_i^{-1}\beta_i\right)\right)}\right) d\mu \\ \geq \;& c_{L,\overline{\sigma}^2} \int h(x) \left(u(x) - \overline{u}_i^*(x) - \tau_i\right)^2 d\mu, \end{aligned}$$

where $\tau_i = \int h(x) \left(u(x) - \overline{u}_i^*(x)\right) d\mu$ and $c_{L,\overline{\sigma}^2}$ is a constant depending only on L and $\overline{\sigma}^2$. Thus for any $u \in \mathcal{U}$,

$$\sum_{i=0}^{n-1} E \int h(x) \left(u(x) - \overline{u}_i^*(x) - \tau_i\right)^2 d\mu \tag{21}$$

$$
\begin{aligned}
&\leq \left(c_{L,\overline{\sigma}^2}\right)^{-1} \sum_{i=0}^{n-1} E \int h(x) \left(1 - e^{-\left(u(x)-\overline{u}_i^*(x)-\beta_i\Omega_i^{-1}(U^i-v_*^i)'\right)^2/\left(8\left(\sigma_{i+1}^2-\beta_i\Omega_i^{-1}\beta_i'\right)\right)}\right) d\mu \\
&= \left(c_{L,\overline{\sigma}^2}\right)^{-1} \sum_{i=0}^{n-1} 2^{-1} E d_H^2 (p_{z_{i+1}|Z^i;u}, p_{z_{i+1}|Z^i;\overline{u}_i^*;v_*^i}) \\
&\leq \left(c_{L,\overline{\sigma}^2}\right)^{-1} 2\psi_n.
\end{aligned}
$$

This means that we have obtained a sequence of estimators $\overline{u}_i^*$ of u with the variances $E\left(\int h(x)\left(u(x) - \overline{u}_i^*(x) - \tau_i\right)^2 d\mu\right)$ of $u - \overline{u}_i^*$ well controlled on average. However, a possibly large bias remains. To get a final estimator of u, we estimate the mean $\eta(u) = \int h u d\mu$ based on current data Z^i.

Step 4 For any $\widehat{\eta}_i$ based on Z^i, let $\widehat{\widehat{u}}_i(x) = \overline{u}_i^*(x) - \int \overline{u}_i^*(x)h(x)d\mu + \widehat{\eta}_i$. Then the new estimator satisfies

$$
\int h(x)\left(u(x) - \widehat{\widehat{u}}_i(x)\right)^2 d\mu = \int h(x)\left(u(x) - \overline{u}_i^*(x) - \tau_i\right)^2 d\mu \ + \left(\widehat{\eta}_i - \eta(u)\right)^2.
$$

It follows that

$$
\begin{aligned}
& \sum_{i=0}^{n-1} E \int h(x)\left(u(x) - \widehat{\widehat{u}}_i(x)\right)^2 d\mu \\
=& \sum_{i=0}^{n-1} E \int h(x)\left(u(x) - \overline{u}_i^*(x) - \tau_i\right)^2 d\mu + \sum_{i=0}^{n-1} E(\widehat{\eta}_i - \eta(u))^2 \\
\leq& \ 2\left(c_{L,\overline{\sigma}^2}\right)^{-1} \psi_n + \sum_{i=0}^{n-1} E(\widehat{\eta}_i - \eta(u))^2.
\end{aligned}
$$

Taking $\widehat{\eta}_i$ to be the minimax estimator of η based on Z^i, we have

$$
\sum_{i=0}^{n-1} E \int h(x)\left(u(x) - \widehat{\widehat{u}}_i(x)\right)^2 d\mu \leq 2\left(c_{L,\overline{\sigma}^2}\right)^{-1} \psi_n + \sum_{i=0}^{n-1} r_i.
$$

Here $r_0 = \min_{\eta'} \max_{u \in \mathcal{U}} \left(\eta' - \eta(u)\right)^2$. As a consequence, we have the following cumulative risk bound

$$
(1/n) \sum_{i=0}^{n-1} E \parallel u - \widehat{\widehat{u}}_i \parallel_{L_2(h)}^2 \leq 2\left(c_{L,\overline{\sigma}^2}\right)^{-1} \psi_n/n + \overline{r}_n,
$$

where $\overline{r}_n = (1/n)\sum_{i=0}^{n-1} r_i$. For the usual risk $R(\mathcal{U};\Omega;n)$, we do not need to require $\widehat{\eta}_i$ to depend only on Z^i. Then we set $\widehat{\eta}_i = \widehat{\eta}_n$ for all $1 \leq i < n$, where $\widehat{\eta}_n$ is the minimax estimator based on Z^n. Then the above risk bound becomes $2\left(c_{L,\overline{\sigma}^2}\right)^{-1} \psi_n/n + r_n$. From Lemma 4 in Section 5, we have an estimator $\widehat{u}_n$ based on Z^n such that

$$
\max_{u \in \mathcal{U}} E \parallel u - \widehat{u}_n \parallel_{L_2(h)}^2 \leq \max_{u \in \mathcal{U}} E \sum_{i=0}^{n-1} \parallel u - \widehat{\widehat{u}}_i \parallel_{L_2(h)}^2 \leq 2\left(c_{L,\overline{\sigma}^2}\right)^{-1} \psi_n/n + r_n. \qquad (22)
$$

Step 5 When ρ_n is of higher order than n, the upper bound above may be suboptimal. For instance, suppose we have independent errors with $\sigma_i^2 = i^{1-\delta}$ for some $1 < \delta < 2$, which implies that $\rho_n \asymp n^{\delta}$. Assume that $M_2(\epsilon) \asymp \epsilon^{-d/\alpha}$ for some $\alpha > 0$. Then the upper bound rate given in terms of ψ_n is $n^{\delta-1-2\alpha\delta/(2\alpha+d)}$, which is worse than the rate $n^{-2\alpha/(2\alpha+d)}$ obtained with i.i.d. errors. Clearly, this inferior rate is not because the problem is more difficult. It can be improved in general as follows. Let us generate i.i.d. random variables $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, ..., \tilde{\varepsilon}_n$ from a standard normal distribution. Let $\tilde{Y}_i = Y_i + \tilde{\varepsilon}_i$, $1 \leq i \leq n$. Then the random errors $\varepsilon_i + \tilde{\varepsilon}_i$ in $\tilde{Y}_i$ have covariance matrix $\tilde{\Omega}_n = I_n + \Omega_n$ (I_n is the $n \times n$ identity matrix). Then $\tilde{\rho}_n = Tr\left(\tilde{\Omega}_n^{-1}\right) \leq Tr\left(I_n^{-1}\right) = n$ because $I_n + \Omega_n \geq I_n$ implies $\left(I_n + \Omega_n\right)^{-1} \leq I_n^{-1}$ (here the symbol "$\geq$" for matrix comparison means the difference is nonnegative definite). Note also that the variances of the new errors $\varepsilon_i + \tilde{\varepsilon}_i$ are uniformly upper bounded by $\tilde{\overline{\sigma}}^2 = \overline{\sigma}^2 + 1$. Applying similar analysis to $(X_i, \tilde{Y}_i)$ replacing ρ_n by n yields

$$\sum_{i=0}^{n-1} E \int h(x)\left(u(x) - \overline{u}_i^*(x) - \tau_i\right)^2 d\mu \leq 2\left(c_{L,\tilde{\overline{\sigma}}^2}\right)^{-1} \overline{\psi}_n, \tag{23}$$

where $\overline{u}_i^*$'s are obtained with the new data $(X_j, \tilde{Y}_j)_{j=1}^{i}$. Estimating $\eta(u)$ the same way as before, we obtain a randomized estimator $\hat{u}_n$ with risk bounded by $2\left(c_{L,\tilde{\overline{\sigma}}^2}\right)^{-1} \overline{\psi}_n/n + r_n$. The estimator depends on both Z^n and the generated random variables $\tilde{\varepsilon}_i$, $1 \leq i \leq n$. One could average out the randomness in $\tilde{\varepsilon}_i$ to get a nonrandomized estimator with no bigger risk since the loss being considered is convex. Thus $R\left(\mathcal{U}; \Omega; n\right) \leq 2\left(c_{L,\tilde{\overline{\sigma}}^2}\right)^{-1} \overline{\psi}_n/n + r_n$. This completes the proof of Proposition 0.

5 Proofs of the main results

Proof of Lemma 1: For the upper rate on $\tilde{r}_n$, taking $\hat{\eta} = \sum_{i=1}^n Y_i/n$, we get $\tilde{r}_n \leq \left(\mathbf{1}'\mathbf{\Omega}_n\mathbf{1}\right)/n^2$. For lower bound, consider 2^m equally spaced points in $\Delta_n = [a_n, b_n] \subset \Delta$. Denote the set of these points by D_n and let Θ take values in D_n with equal probability. Let $\delta_n = (b_n - a_n)\, 2^{-m}$. Then as in the proof of Proposition 0, we have

$$\tilde{r}_n \geq \delta_n^2/4\left(1 - \frac{I\left(\Theta; Y^n\right) + \log 2}{m \log 2}\right).$$

Similarly to the analysis there, consider a rougher net in Δ_n. Let $D_n^{'}$ be the set of $2^{m^{'}}$ equally spaced points in Δ_n and let $\delta_n^{'} = (b_n - a_n)\, 2^{-m^{'}}$. Then it can be shown similarly that $I\left(\Theta; Y^n\right) \leq m^{'} \log 2 + (1/2)\left(\delta_n^{'}\right)^2 \left(\mathbf{1}'\mathbf{\Omega}_n^{-1}\mathbf{1}\right)$. Take $b_n - a_n$ of order $\left(\mathbf{1}'\mathbf{\Omega}_n^{-1}\mathbf{1}\right)^{-1/2}$ and $m^{'} = 1$ to have $I\left(\Theta; Y^n\right) \preceq 1$ (note that $\left(\mathbf{1}'\mathbf{\Omega}_n^{-1}\mathbf{1}\right)^{-1}$ is the variance of the best linear unbiased estimator and thus $\left(\mathbf{1}'\Omega_n^{-1}\mathbf{1}\right)^{-1} \preceq 1$). Thus there exists a constant C such that $I\left(\Theta; Y^n\right) \leq C$ for all n. Take m suitably large (independent of n) such that $\left(C + \log 2\right)/(m \log 2) \leq 1/2$. Then $\tilde{r}_n \geq \delta_n^2/8$. This establishes the lower bound rate $\left(\mathbf{1}'\mathbf{\Omega}_n^{-1}\mathbf{1}\right)^{-1}$.

For an upper bound on r_n in the second statement, consider $\widehat{\eta}_n = \overline{Y} = (1/n)\sum_{j=1}^{n} Y_j$. Then

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

Under the given conditions, together with Lemma 6 later in this section, we have

$$\widetilde{r}_n \preceq r_n \preceq \left(\mathbf{1}'\Omega_n\mathbf{1}\right)/n^2.$$

If $\left(\mathbf{1}'\Omega_n^{-1}\mathbf{1}\right)\left(\mathbf{1}'\Omega_n\mathbf{1}\right) \asymp n^2$, then clearly $\widetilde{r}_n \asymp r_n \asymp \left(\mathbf{1}'\Omega_n\mathbf{1}\right)/n^2$. This completes the proof of Lemma 1.

Proof of Theorem 1: The upper bound part for the first conclusion follows from (23) in the proof of Proposition 0 using $\widehat{u}_0 = (1/n)\sum_{i=0}^{n-1}\left(\overline{u}_i^*(x) - \int \overline{u}_i^*(x)h(x)d\mu\right)$ as an estimator of u_0. From (23) and using Lemma 4, we have that

$$\begin{aligned} E\int h(x)\left(u_0(x) - \widehat{u}_0(x)\right)^2 d\mu &\leq \frac{1}{n}\sum_{i=0}^{n-1} E\int h(x)\left(u_0(x) - \left(\overline{u}_i^*(x) - \int \overline{u}_i^*(x)h(x)d\mu\right)\right)^2 d\mu \\ &\leq 2\left(c_{L,\tilde{\overline{\sigma}}^2}\right)^{-1}\overline{\psi}_n \preceq \epsilon_n^2. \end{aligned}$$

Note that Assumption A6 is not needed for the above upper rate of convergence for estimating u_0.

To prove ϵ_n^2 is also a lower rate for $R_0(\mathcal{U};\Omega;n)$, consider distance d_0 defined as $d_0(u,v) = \int (u_0 - v_0)^2 h d\mu$, where $u_0 = u - \int uhd\mu$ and $v_0 = v - \int vhd\mu$. Replacing $L_2(h)$ distance by d_0 in the derivation of the lower bound in the proof of Proposition 0, we have

$$R_0(\mathcal{U};\Omega;n) \geq \eta_n^2/8,$$

where η_n is determined by $M_0(\eta_n) = 2\psi_n$ with $M_0(\epsilon)$ being the packing entropy of $\mathcal{U}$ under d_0. It is straightforward to show that $M_0(\epsilon)$ is of the same order as $M_2(\epsilon)$ for a uniformly bounded rich class. As a consequence, under Assumption A6, $\eta_n \asymp \epsilon_n$.

The second conclusion in (7) follows directly from Proposition 0 using that $\underline{\epsilon}_n^2$ and ψ_n^*/n are both of order ϵ_n^2 under the condition $Tr(\Omega_n^{-1}) \asymp n$. Note that the upper bound in Proposition 0 always satisfies $\psi_n^*/n \preceq \epsilon_n^2$, regardless of the trace condition. This completes the proof of Theorem 1.

Proof of Theorem 2: The conclusion follows directly from Theorem 1 and Lemma 1.

Proof of Corollary 1: Assumption A5 is obviously satisfied. From Lemma 8 later in this section, Assumption 6 is satisfied. It remains to verify $\left(\mathbf{1}'\Omega_n^{-1}\mathbf{1}\right)\left(\mathbf{1}'\Omega_n\mathbf{1}\right) \asymp n^2$, $\left(\mathbf{1}'\Omega_n\mathbf{1}\right) \succeq n$ and $\mathbf{1}'\Omega_n\mathbf{1}/n^2 \asymp n^{-\gamma}$. Since $r(j) \sim |j|^{-\gamma}$, it is straightforward to show that $\left(\mathbf{1}'\Omega_n\mathbf{1}\right) \asymp n^{2-\gamma}$. Under our assumptions on the spectral density, Adenstedt (1974, Theorem 5.2) shows that $\left(\mathbf{1}'\Omega_n^{-1}\mathbf{1}\right)^{-1}$ is of order $n^{-\gamma}$ (note that $\left(\mathbf{1}'\Omega_n^{-1}\mathbf{1}\right)^{-1}$ is the variance of the BLUE). This completes the proof of Corollary 1.

Proof of Theorem 3: We make some modifications to the derivation of the minimax upper bound in Proposition 0. Note now that the conditional covariance matrix Ω_n depends on X^n in general. Under the assumption that there is an i.i.d. component of errors, the eigenvalues of Ω_n is uniformly lower bounded, i.e., there exists a positive constant λ_1 (independent of X^n and n) such that $a'\Omega_n a \geq \lambda_1 \parallel a \parallel^2$ for all n-dimensional vector a (λ_1 is the variance of the i.i.d. portion of errors). Then from (24) in the proof of Lemma 2 later in this section, we have

$$\begin{aligned} 2D\left(P_{Z^n,u} \parallel P_{Z^n,v}\right) &= E\left(U^n - V^n\right)'\Omega_n^{-1}\left(U^n - V^n\right) \\ &\leq \lambda_1^{-1}\sum_{i=1}^{n} E\left(u(X_i) - v(X_i)\right)^2 \\ &= n\lambda_1^{-1} \parallel u - v \parallel_{L_2(h)}^2 . \end{aligned}$$

Similarly to (17), we have that the mutual information $I\left(U; Z^n\right)$ is upper bounded in order by $n\epsilon_n^2 + M_2(\epsilon_n)$. Without the i.i.d. assumption on $X_i, i \geq 1$, $h(x_{i+1})$ in (19) should be replaced by the conditional distribution of X_{i+1} given X^i. It will be denoted by $h(x_{i+1}|X^i)$ for convenience regardless of whether the conditional density with respect to μ exists or not. Proceed as before but replacing $h(x_{i+1})$ by $h(x_{i+1}|X^i)$, we have that

$$\max_{u\in\mathcal{U}} \sum_{i=0}^{n-1} Ed_H^2(p_{z_{i+1}|Z^i;u}, p_{z_{i+1}|Z^i;\overline{u}_i^*;v_*^i}) \preceq n\epsilon_n^2.$$

Note that

$$Ed_H^2(p_{z_{i+1}|Z^i;u}, p_{z_{i+1}|Z^i;\overline{u}_i^*;v_*^i}) = 2E\int h(x_{i+1}|X^i)\left(1 - e^{-\left(m_{i,u}-\widetilde{m}_i^*\right)^2/\left(8\left(\sigma_{i+1}^2 - \beta_i'\Omega_i^{-1}\beta_i\right)\right)}\right) d\mu.$$

Using the fact that for any $A > 0$, there exists a constant $c > 0$ such that $1 - e^{-x^2} \geq c\min\left(x^2, A\right)$, together with (11), we have

$$E \parallel m_{i,u} - \widetilde{m}_i^* \parallel_{L_2(\nu_i),A}^2 \leq \widetilde{c}E\int h(x_{i+1}|X^i)\left(1 - e^{-\left(m_{i,u}-\widetilde{m}_i^*\right)^2/\left(8\left(\sigma_{i+1}^2 - \beta_i'\Omega_i^{-1}\beta_i\right)\right)}\right) d\mu$$

for a constant $\widetilde{c}$ depending only on A and an upper bound on the quantity in (11). Averaging over $0 \leq i \leq n-1$, we have ϵ_n^2 as an upper rate on the average cumulative prediction risk. This completes the proof of Theorem 3.

5.1 Proofs of the technical lemmas

Let $P_{Z^n,u}$ denote the distribution of $Z^n = (X_i, Y_i)_{i=1}^n$ when the regression function is u. The density of $P_{Z^n,u}$ is

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

Let $\omega_{i,j}^{-1}$ denote the (i,j)-element of Ω_n^{-1}. Recall that the Kullback-Leibler divergence $D(P \parallel Q)$ between two distributions P and Q with densities p and q with respect to μ is defined as $D(P \parallel Q) = \int p \log(p/q)\,d\mu$.

Lemma 2: *The K-L divergence between $P_{Z^n,u}$ and $P_{Z^n,v}$ is*

$$D\left(P_{Z^n,u} \parallel P_{Z^n,v}\right) = (1/2)\,Tr\left(\Omega_n^{-1}\right) \parallel u - v \parallel_{L_2(h)}^2 + (1/2)\left(\sum_{i\neq j} \omega_{i,j}^{-1}\right)(Eu - Ev)^2.$$

Proof: We have

$$2\log\frac{p_u(z^n)}{p_v(z^n)} = 2\,(u^n - v^n)'\,\Omega_n^{-1} y^n - (u^n)'\,\Omega_n^{-1} u^n + (v^n)'\,\Omega_n^{-1} v^n.$$

Given X^n,

$$\begin{aligned} 2E_{Z^n|X^n;u}\log\frac{p_u(Z^n)}{p_v(Z^n)} &= 2\,(u^n - v^n)\,\Omega_n^{-1}\,(u^n)' - u^n\Omega_n^{-1}\,(u^n)' + v^n\Omega_n^{-1}\,(v^n)' \qquad (24)\\ &= (u^n - v^n)\,\Omega_n^{-1}\,(u^n - v^n)'. \end{aligned}$$

Then

$$\begin{aligned} 2E_{Z^n,u}\log\frac{p_u(Z^n)}{p_v(Z^n)} &= E\left(\sum_{i,j}\omega_{i,j}^{-1}\,(u(X_i) - v(X_i))\,(u(X_j) - v(X_j))\right)\\ &= \sum_{i=1}^n \omega_{i,i}^{-1} \parallel u - v \parallel_{L_2(h)}^2 + \sum_{i\neq j}\omega_{i,j}^{-1} E\,(u(X_i) - v(X_i))\,(u(X_j) - v(X_j)). \end{aligned}$$

Under the i.i.d. assumption on $X_1, ..., X_n$, we have

$$2E_{Z^n,u}\log\frac{p_u(Z^n)}{p_v(Z^n)} = \sum_{i=1}^n \omega_{i,i}^{-1} \parallel u - v \parallel_{L_2(h)}^2 + \left(\sum_{i\neq j}\omega_{i,j}^{-1}\right)(E\,(u(X) - v(X)))^2.$$

This completes the proof of the Lemma 2.

Lemma 3: *Assume* $\sup_x |g(x)| \leq A$ *for some constant* A *and* $\sigma^2 \leq \sigma_0^2$. *Let* $h(x)$ *be a probability density function. Then*

$$\min_{\theta\in R}\int h(x)\left(1 - e^{-(g(x)-\theta)^2/\sigma^2}\right)d\mu \geq c\int h(x)\left(g(x) - \int h(x)g(x)d\mu\right)^2 d\mu,$$

where the constant c *depends only on* A *and* σ_0^2.

Proof: It is easy to prove that for $|g| \leq A$, $1 - e^{-(g(x)-\theta)^2/\sigma^2} \geq \begin{cases} c(g-\theta)^2 & |\theta| \leq 2A \\ cg^2 & |\theta| > 2A \end{cases}$ for some constant c depends only on A and σ_0^2. It follows that $\int h(x)\left(1 - e^{-(g-\theta)^2/\sigma^2}\right) d\mu \geq \begin{cases} c\int h(x)\left(g(x)-\theta\right)^2 d\mu & |\theta| \leq 2A \\ c\int h(x) g(x)^2 d\mu & |\theta| > 2A \end{cases}$. Since $\int h(x)\left(g(x)-a\right)^2 d\mu$ is minimized when $a = \int h(x)g(x)d\mu$, the conclusion of the lemma follows.

Lemma 4: *Let $\hat{u}_1, ..., \hat{u}_k$ be k estimators of u. Then the estimator $\widehat{\hat{u}}_k = (1/k)\sum_{i=1}^k \hat{u}_i$ satisfies*

$$E \parallel u - \widehat{\hat{u}}_k \parallel_{L_2(h)}^2 \leq (1/k)\sum_{i=1}^k E \parallel u - \hat{u}_i \parallel_{L_2(h)}^2 .$$

Proof: The result follows from that $\parallel u - v \parallel_{L_2(h)}^2$ is convex in v.

Lemma 5: *Let Ω_n be the $n \times n$ finite section of the covariance matrix of a stationary process. Assume Ω_n is invertible for $n \geq 1$. Then $Tr\left(\Omega_n^{-1}\right)$ is at least of order n. More generally, if $\sup_{i\geq 1}\sigma_i^2 < \infty$, then $Tr\left(\Omega_n^{-1}\right) \succeq n$.*

Proof: Let $\Omega_n = \begin{pmatrix} \Omega_{n-1} & \beta_{n-1} \\ \beta_{n-1}' & \sigma_n^2 \end{pmatrix}$. Then

$$\Omega_n^{-1} = \begin{pmatrix} \Omega_{n-1}^{-1} + \left(\sigma_n^2 - \beta_{n-1}'\Omega_{n-1}^{-1}\beta_{n-1}\right)^{-1} \Omega_{n-1}^{-1}\beta_{n-1}\beta_{n-1}'\Omega_{n-1}^{-1} & -\left(\sigma_n^2 - \beta_{n-1}'\Omega_{n-1}^{-1}\beta_{n-1}\right)^{-1} \Omega_{n-1}^{-1}\beta_{n-1} \\ -\left(\sigma_n^2 - \beta_{n-1}\Omega_{n-1}^{-1}\beta_{n-1}'^{-1}\right)^{-1} \beta_{n-1}'\Omega_{n-1}^{-1} & \left(\sigma_n^2 - \beta_{n-1}'\Omega_{n-1}^{-1}\beta_{n-1}\right)^{-1} \end{pmatrix}.$$

It follows that

$$Tr\left(\Omega_n^{-1}\right) \geq Tr\left(\Omega_{n-1}^{-1}\right) + \left(\sigma_n^2 - \beta_{n-1}'\Omega_{n-1}^{-1}\beta_{n-1}\right)^{-1} \geq Tr\left(\Omega_{n-1}^{-1}\right) + \sigma_n^{-2}.$$

The conclusion follows from induction. This completes the proof of Lemma 5.

Let r_n be defined as in (2) and let $\widetilde{r}_n = \min_{\widehat{\eta}}\max_{\eta\in\Delta} E(\widehat{\eta} - \eta)^2$ be the minimax risk for estimating η based on $(Y_i)_{i=1}^n$ under model $Y_i = \eta + \varepsilon_i$, $1 \leq i \leq n$.

Lemma 6: *Under Assumption A3, we have $r_n \geq \widetilde{r}_n$.*

Proof: Under Assumption A3, r_n decreases when $u \in \mathcal{U}$ is instead restricted to the set of constant functions $\{\eta, \eta \in \Delta\}$. For the restricted model, it is easy to see by factorization theorem that $(Y_1, ..., Y_n)$ is a sufficient statistic for η. Then for any estimator $\widehat{\eta}$ based on $(X_i, Y_i)_{i=1}^n$, we may take $\widehat{\widehat{\eta}} = E\left(\widehat{\eta}|Y_1, ..., Y_n\right)$ to get an estimator based only on $Y_1, ..., Y_n$ with no bigger mean squared error. The conclusion follows.

The following two lemmas give sufficient conditions for $Tr(\Omega_n^{-1}) \asymp n$ as used in Section 3.

Lemma 7: *Assume $\sup_i \sigma_i^2 < \infty$ and that Ω_n can be expressed as the sum of two components $\Omega_n = \Omega_n^{(1)} + \Omega_n^{(2)}$, where $\Omega_n^{(1)} = diag\left(\omega_{1,n}, ..., \omega_{n,n}\right)$ with $\min_{1\leq i\leq n}\omega_{i,n} \geq c > 0$ for some constant $c > 0$ independent of n, and $\Omega_n^{(2)}$ is nonnegative definite. Then $Tr(\Omega_n^{-1}) \asymp n$.*

Proof: By Lemma 5, under the condition $\sup_i \sigma_i^2 < \infty$, we have $Tr(\Omega_n^{-1}) \succeq n$. Under the other condition, we have $\Omega_n \geq \Omega_n^{(1)}$ and hence $\Omega_n^{-1} \leq \left(\Omega_n^{(1)}\right)^{-1}$. So $Tr\left(\Omega_n^{-1}\right) \leq Tr\left(\Omega_n^{(1)}\right)^{-1} \preceq n$. This completes the proof of Lemma 7.

LEMMA 8: *For stationary serially correlated errors with spectral density bounded away from zero, one has* $Tr(\Omega_n^{-1}) \asymp n$.

PROOF: From Lemma 5, $Tr\,(\Omega_n^{-1})$ is at least of order n. From Grenander and Szegö (1958, p. 64), the minimum eigenvalue of Ω_n is uniformly bounded away from zero for $n \geq 1$. Since $Tr\,(\Omega_n^{-1})$ is the sum of the reciprocals of the eigenvalues of Ω_n, we have $Tr\,(\Omega_n^{-1}) \preceq n$. This completes the proof of Lemma 8.

ACKNOWLEDGMENTS. The author is very grateful to Andrew Barron for his comments, suggestions, and encouragement. He thanks Wayne Fuller, Cui-Hui Zhang, Zhiliang Ying, and Nicolas Hengartner for helpful discussions.

References

[1] R.K. Adenstedt (1974). "On large sample estimation for the mean of a stationary sequence," *Ann. Statist.* **2**, 1095-1107.

[2] A.R. Barron (1987). "Are Bayes rules consistent in information?" *Open Problems in Communication and Computation*, 85-91. T.M. Cover and B. Gopinath eds., Springer, NY.

[3] J. Beran (1986). "Estimation, testing and prediction for self-similar and related processes," Ph.D Thesis, ETH, Zürich.

[4] J. Beran (1994). *Statistics for Long-Memory Processes.* Chapman and Hall, New York.

[5] L. Birgé (1983). "Approximation dans les espaces metriques et theorie de l'estimation," *Z. Wahrscheinlichkeitstheor. Verw. Geb.* **65**, 181-237.

[6] L. Birgé (1986). "On estimating a density using Hellinger distance and some other strange facts," *Probability Theory and Related Fields* **71**, 271-291.

[7] J. Bretagnolle and C. Huber (1979). "Estimation des densites: risque minimax," *Z. Wahrscheinlichkeitstheor. Verw. Geb.,* **47**, 119-137.

[8] D.R. Cox (1984). "Long-range dependence: a review," in *Statistics: An Appraisal. Proceedings 50th Anniversary Conference.* H.A. David and H.T. David (eds.). The Iowa State University Press, 55-74.

[9] R. Dahlhaus (1989). "Efficient parameter estimation for self-similar processes," *Ann. Statist.* **17**, 1749-1766.

[10] R.A. DeVore and G.G. Lorentz (1993). *Constructive Approximation*, Springer-Verlag, New York.

[11] S. Efromovich (1999). "How to overcome curse of long-memory," *IEEE Trans. Inform. Theory* **45**, 1735-1741.

[12] R. Fox and M.S. Taqqu (1985). "Non-central limit theorems for quadratic forms in random variables having long-range dependence," *Ann. Probab.* **13**, 428-446.

[13] R. Gay and C.C. Heyde (1990) . "On a class of random field models which allow long range dependence," *Biometrika* **77**, 401-403.

[14] L. Giraitis and D. Surgailis (1990). "A central limit theorem for quadratic forms in strongly dependent linear variables and application to asymptotical normality of Whittle's estimate," *Probab. Th. Rel. Fields* **86**, 87-104.

[15] C.W.J. Granger and R. Joyeux (1980). "An introduction to long-range time series models and fractional differencing," *J. Time Ser. Anal.,* **1**, 15-30.

[16] U. Grenander and G. Szegö (1958). *Toeplitz Forms and Their Applications,* University of California Press.

[17] P. Hall and J.D. Hart (1990a). "Nonparametric regression with long-range dependence," *Stochastic Process. Appli.* **36**, 339-351.

[18] P. Hall and J.D. Hart (1990b). "Convergence rates in density estimation for data from infinite-order moving average processes," *Probab. Th. Rel. Fields* **87**, 253-274.

[19] J.R.M. Hosking (1981). "Fractional differencing," *Biometrika* **68**, 165-176.

[20] I.A. Ibragimov and R.Z. Hasminskii (1977). "On the estimation of an infinite-dimensional parameter in Gaussian white noise," *Soviet Math. Dokl.,* **18**, 1307-1309.

[21] I.M. Johnstone and B.W. Silverman (1997). "Wavelet threshold estimators for data with correlated noise," *Journal of the Royal Statistical Association, Series B*, **59**, 319–351.

[22] A.N. Kolmogorov and V.M. Tihomirov (1959). "ϵ-entropy and ϵ-capacity of sets in function spaces," *Uspehi Mat. Nauk* **14**, 3-86; English transl. (1961). *Amer. Math. Soc. Transl.* **17**, 277-364.

[23] H. Künsch, J. Beran, and F. Hampel (1993). "Contrasts under long-range correlations," *Ann. Statist.* **21**, 943-964.

[24] L.M. Le Cam (1975). "On local and global properties in the theory of asymptotic normality of experiments," *Stochastic Processes and Related Topics,* Vol. 1 (M. Puri, ed.), 13-54. Academic Press, New York.

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

[26] G.G. Lorentz, M.v. Golitschek, and Y. Makovoz (1996). *Constructive Approximation: Advanced Problems*, Springer-Verlag, New York.

[27] B.B. Mandelbrot and J.W. van Ness (1968). "Fractional Brownian motions, fractional noises and applications," *Siam Rev.* **10,** 422-437.

[28] E. Renshaw (1994). "The linear spatial-temporal interaction process and its relation to $1/\omega$-noise," *J. Roy. Statist. Soc., Ser. B* **56,** 75-91.

[29] P.M. Robinson (1995). "Gaussian semiparametric estimation of long-range dependence," *Ann. Statist.* **23**, 1630-1661.

[30] P.M. Robinson (1997). "Large-sample inference for nonparametric regression with dependent errors," *Ann. Statist.* **25**, 2054-2083.

[31] A. Samarov and M.S. Taqqu (1988). "On the efficiency of sample mean in long memory noise," *J. Time Ser. Anal.* **9,** 191-200.

[32] Ya.G. Sinai (1976). "Self-similar probability distributions," *Theory Probab. Appl.,* **21**, 64-80.

[33] H. Triebel (1975). "Interpolation properties of ϵ-entropy and diameters. Geometric characteristics of embedding for function spaces of Sobolev-Besov type," *Mat. Sbornik* **98**, 27-41; English Transl. in *Math. USSR Sb.,* **27**, 23-37, 1977.

[34] Y. Wang (1996). "Function estimation via wavelet shrinkage for long-memory data," *Ann. Statist.* **24**, 466-484.

[35] P. Whittle (1962). "Topographic correlation, power-law covariance functions, and diffusion," *Biometrika* **49**, 304-314.

[36] Y. Yang and A.R. Barron (1999). "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.,* **27**, 1564-1599.

[37] Y. Yajima (1991) "Asymptotic properties of LSE in a regression model with long-memory stationary errors," *Ann. Statist.,* **19**, 158-177.

[38] Y.G. Yatracos (1988) "A lower bound on the error in nonparametric regression type problems," *Ann. Statist.,* **16**, 1180-1187.

[39] B. Yu (1996). "Assouad, Fano, and Le Cam," in *Research Papers in Probability and Statistics: Festschrift in honor of Lucien Le Cam* (D. Pollard, E. Torgersen and G. Yang, eds.), Springer, New York.