The Annals of Statistics
2006, Vol. 34, No. 4, 1701–1732
DOI: 10.1214/009053606000000560

ASYMPTOTIC EQUIVALENCE OF NONPARAMETRIC AUTOREGRESSION AND NONPARAMETRIC REGRESSION

BY ION G. GRAMA AND MICHAEL H. NEUMANN

Université de Bretagne-Sud and Friedrich-Schiller-Universität Jena

It is proved that nonparametric autoregression is asymptotically equivalent in the sense of Le Cam's deficiency distance to nonparametric regression with random design as well as with regular nonrandom design.

1. Introduction. We assume that observations $X_0, \dots, X_n$ from a stationary autoregressive process $(X_i)_{i=0,\dots,n}$ are available which obey the model equation

$$X_i = f(X_{i-1}) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where $(\varepsilon_i)_{i=1,\dots,n}$ are i.i.d. random variables. The unknown autoregression function f is then the target of statistical inference and the development of efficient estimators is a natural task for theoretically oriented statisticians. On the one hand, it has been recognized for a long time that commonly used estimators in model (1) have the same asymptotic behavior as corresponding estimators in nonparametric regression. A result of Robinson [26] concerns the pointwise equivalence of nonparametric kernel estimators and Neumann and Kreiss [22] extended this equivalence to the global behavior of nonparametric estimators. On the other hand, despite these well-known similarities between estimators, there is still a certain discrepancy in the current state of available theory in both contexts. While there is a very well developed asymptotic theory for optimal estimation in nonparametric regression, even up to the level of exact asymptotics (see, e.g., [13] or [24], for an overview), there is considerably less theory available in the case of nonparametric autoregression.

The purpose of the present paper is to bridge this gap between the two settings of nonparametric regression and autoregression by showing asymptotic equivalence on an abstract level. The theory of asymptotic equivalence of statistical experiments has been developed in Le Cam's [19] work. In the framework of nonparametric statistics, Brown and Low [4] proved that the Gaussian white noise experiment and nonparametric regression with nonrandom design and Gaussian errors are asymptotically equivalent in the sense that Le Cam's deficiency distance between them tends to zero. In [12, 14, 15, 23] the scope of asymptotic

Received June 2004; revised September 2005.

AMS 2000 subject classifications. Primary 62B15; secondary 62G07, 62G20.

Key words and phrases. Asymptotic equivalence, deficiency distance, Gaussian approximation, nonparametric autoregression, nonparametric regression.

equivalence was extended to the nonparametric density estimation problem and to nonparametrically driven regression models. Moreover, asymptotic equivalence of nonparametric regression with random design and Gaussian white noise was shown in [2] while asymptotic equivalence of Poisson processes and Gaussian white noise was established in [3]. The issue of constructive asymptotic equivalence is considered in [25] and [5]. The asymptotic equivalence of a close relative of nonparametric autoregression, a diffusion experiment parametrized by the drift function, to Gaussian white noise experiments is proved in [8] and [7]. Milstein and Nussbaum [21] showed asymptotic equivalence of a nonparametric statistical model of small diffusion type and its discretization by a stochastic Euler difference scheme. These models deal with dependent observations in continuous time. However, asymptotic equivalence for models with dependent observations in discrete time where the noise is non-Gaussian seems to be a much more difficult issue.

In this paper we establish *local* equivalence of nonparametric autoregression (1) and nonparametric regression in the discrete-time setting. That is, the set of possible functions lies in a class $\Sigma_n(f_0)$ centered around some fixed function f_0 and shrinking in some appropriate norm as $n \to \infty$. Depending on additional prior smoothness assumptions on f, this class will nevertheless be rich enough for the transfer of minimax lower bounds from one to the other model. Under mild regularity assumptions stated below, the process $(X_i)_{i=0,\dots,n}$ corresponding to f_0 has a stationary density ψ_{f_0}, say. We show asymptotic equivalence of the experiment given by (1) to nonparametric regression with random design as well as with regular nonrandom design. The former experiment corresponds to i.i.d. observations $(Y_1, \xi_1), \dots, (Y_n, \xi_n)$ with

$$Y_i = f(\xi_i) + \eta_i, \qquad i = 1, \dots, n, \tag{2}$$

where $E(\eta_i|\xi_i) \equiv 0$. The basic assumption on the errors η_i is that their Fisher information is the same as that of the ε_i's. This includes the case of Gaussian errors as well as of errors having the same distribution as the ε_i. The ξ_i are distributed according to the stationary density ψ_{f_0} of the process corresponding to the central function f_0, regardless of the actual value of f.

We show also equivalence to nonparametric regression with regular nonrandom design, which corresponds to independent observations $Y_{n,1}, \dots, Y_{n,n}$ obeying the model

$$Y_{n,i} = f(t_{n,i}) + \eta_i, \qquad i = 1, \dots, n, \tag{3}$$

where $E\eta_i = 0$. Here we will assume that the design points are regularly spaced with density ψ_{f_0}, that is, $\int_{-\infty}^{t_{n,i}} \psi_{f_0}(x)\,dx = (i - 1/2)/n$. We assume again that the Fisher information of η_i is the same as the Fisher information of ε_i. Since Le Cam's equivalence relation is transitive, we also obtain as an immediate by-product asymptotic equivalence of nonparametric regression with random and regular nonrandom design. In the special case of Gaussian errors but under weaker smoothness assumptions on f, this equivalence also follows from the asymptotic equivalence of nonparametric regression with nonrandom design and Gaussian white noise [4] and the asymptotic equivalence of nonparametric regression with random design and Gaussian white noise [2].

At the end of Section 2 we discuss briefly how our results on asymptotic equivalence can be used to transfer well-known lower asymptotic bounds for the minimax risk in nonparametric regression to the case of nonparametric autoregression. Our local version of asymptotic equivalence does not allow an immediate transfer of upper asymptotic bounds; however, they could be independently proved by appeal to strong approximations of nonparametric estimators in both models (see [22] for details) or by direct computation of the risk of asymptotically optimal estimators.

2. Assumptions and main results.

We start by introducing an appropriate functional parameter set. Consider the set of functions

$$\mathcal{F} = \left\{ f : \mathbb{R} \to \mathbb{R} : \sup_{x \in \mathbb{R}} |f(x)| \le M \right\},$$

where $M < \infty$ is a constant. For any constants $\beta > 0$ and $L > 0$, let $\mathcal{H} = \mathcal{H}(\beta, L)$ be a Hölder ball, that is, the set of functions $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$|f| \le L, \qquad \left| f^{\lfloor \beta \rfloor}(x) - f^{\lfloor \beta \rfloor}(y) \right| \le L|x - y|^{\beta - \lfloor \beta \rfloor}, \qquad x, y \in \mathbb{R}.$$

Here $\lfloor \beta \rfloor$ denotes the largest integer strictly less than β. The set of functional parameters is defined as

$$\Sigma = \mathcal{F} \cap \mathcal{H}(\beta, L).$$

Let X_0 be a random variable on the probability space $(\Omega, \mathcal{A}, P)$. Assume that we observe a sequence $X_1, \dots, X_n$ which obeys

$$X_i = f(X_{i-1}) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{4}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. with a given density p that is continuous and positive on $\mathbb{R}$ and the function $f \in \Sigma$ is assumed to be unknown. It is easy to see that, for any $f \in \Sigma$, $P(X_{i+1} \in B | X_i = x) \ge \mu(B)$ holds for all $B \in \mathcal{B}$ and $x \in \mathbb{R}$, where μ is some measure not depending on f with $\mu(\mathbb{R}) = \mu_0 > 0$. From Theorem 2.4.1 in [10] it follows that the uniform mixing coefficients (see Section 6.2) decay geometrically and, therefore, there exists a stationary density which we shall denote ψ_f.

Throughout the paper we shall assume that the observations (4) satisfy the following assumption:

(A1) The random variable X_0 has the stationary density $\psi_f(\cdot)$, which implies that the sequence $(X_i)_{i=0,\dots,n}$ is in the stationary regime.

Note that the stationary density $\psi_f(\cdot)$ satisfies $\int_B \psi_f(x)\,dx \ge \mu(B)$, for all $B \in \mathscr{B}$.

Before we can state our main results on the approximation of the nonparametric autoregressive model by a nonparametric regression model, we have to introduce the basic concepts of *asymptotic equivalence*. Let $\mathscr{E}_l^n = (\Omega_l^n, \mathscr{A}_l^n, \{P_{l,f}^n, f \in \Sigma'\})$, $l = 1, 2$, be two sequences of statistical experiments indexed by f in a subset $\Sigma' \subset \Sigma$. *The deficiency* of $\mathscr{E}_1^n$ with respect to $\mathscr{E}_2^n$ is defined as

$$\delta(\mathscr{E}_1^n, \mathscr{E}_2^n) = \sup_L \inf_{\delta^{(1)}} \sup_{\delta^{(2)}} \sup_{f \in \Sigma'} \bigl|E_{1,f}^n L(f, \delta^{(1)}) - E_{2,f}^n L(f, \delta^{(2)})\bigr|,$$

where the first supremum is taken over all decision problems with loss function L with $0 \le L \le 1$, and the minimax value of the maximum difference in risks over $f \in \Sigma'$ is computed over all randomized statistical procedures $\delta^{(l)}$ for $\mathscr{E}_l^n$, $l = 1, 2$. According to Theorem 2 on page 15 in [20], the deficiency distance can alternatively be written as

$$\delta(\mathscr{E}_1^n, \mathscr{E}_2^n) = \inf_M \sup_{f \in \Sigma'} \tfrac{1}{2}\|M \cdot P_{1,f}^n - P_{2,f}^n\|_{\mathrm{Var}},$$

where $\|\cdot\|_{\mathrm{Var}}$ denotes the total variation distance and the infimum is taken over all Markov kernels M on $\Omega_1^n \times \mathscr{A}_2^n$. *Le Cam's pseudodistance* between $\mathscr{E}_1^n$ and $\mathscr{E}_2^n$ is

$$\Delta(\mathscr{E}_1^n, \mathscr{E}_2^n) = \max\{\delta(\mathscr{E}_1^n, \mathscr{E}_2^n), \delta(\mathscr{E}_2^n, \mathscr{E}_1^n)\}.$$

Following [4], we say that the sequences $\mathscr{E}_1^n$, $n = 1, 2, \ldots,$ and $\mathscr{E}_2^n$, $n = 1, 2, \ldots,$ are *asymptotically equivalent* if

$$\Delta(\mathscr{E}_1^n, \mathscr{E}_2^n) \to 0 \qquad \text{as } n \to \infty.$$

To formulate our results we also need to impose the following regularity assumptions on the density $p(\cdot)$ of the innovations:

(A2) (i) The density p is positive on $\mathbb{R}$.

(ii) The log-likelihood function $l_p(x) = \log p(x)$ has three derivatives and satisfies, for some $\epsilon > 0$,

$$\int_{\mathbb{R}} \sup_{|u| \le \epsilon} l_p''(x+u)^2 p(x)\,dx < \infty, \qquad \sup_{x \in \mathbb{R}} |l_p'''(x)| \le c_1 < \infty.$$

(iii) The score $l_p'(x) = p'(x)/p(x)$ satisfies, for some $\epsilon > 0$ and any $\lambda < \infty$,

$$\int_{\mathbb{R}} \sup_{|u| \le \epsilon} |l_p'(x+u)|^\lambda p(x)\,dx < \infty.$$

Assumption (A2) mainly requires the existence of three derivatives of $p(\cdot)$ and of the absolute moments of the corresponding scores. These types of assumptions can be related to the so-called Cramér conditions (see [20], page 102). Assumption (A2) is used here just for the sake of simplifying the proofs, but it is clear that they could be relaxed substantially. We refer to [15] for a relevant exposition of sufficient assumptions in the case of nonparametric models with independent observations.

In the sequel $q(\cdot)$ denotes a positive density which satisfies the following assumptions:

(A3) (i) The log-likelihood function $l_q(x) = \log q(x)$ has three derivatives and satisfies, for some $\epsilon > 0$,

$$\int_{\mathbb{R}} \sup_{|u|\le\epsilon} l_q''(x+u)^2 q(x)\,dx < \infty, \qquad \sup_{x\in\mathbb{R}} |l_q'''(x)| \le c_1 < \infty.$$

(ii) The score $l_q'(x) = q'(x)/q(x)$ satisfies, for some $\epsilon > 0$ and any $\lambda < \infty$,

$$\int_{\mathbb{R}} \sup_{|u|\le\epsilon} |l_q'(x+u)|^\lambda q(x)\,dx < \infty.$$

(iii) The Fisher information corresponding to the density $q(\cdot)$ is the same as that corresponding to $p(\cdot)$, that is,

$$I = \int_{\mathbb{R}} l_p'(x)^2 p(x)\,dx = \int_{\mathbb{R}} l_q'(x)^2 q(x)\,dx.$$

We state *local* versions of asymptotic equivalence, that is, we additionally assume that f lies in a shrinking (as $n \to \infty$) neighborhood of some central function f_0. To get a meaningful result, we have to choose this neighborhood large enough such that it can be reached with a probability tending to 1 by an appropriate preliminary estimator. We fix any $\beta > 5/2$ and define

$$\gamma_n = c\left(\frac{\log n}{n}\right)^{\beta/(2\beta+1)}, \qquad \gamma_n' = c'\left(\frac{\log n}{n}\right)^{(\beta-1)/(2\beta+1)}. \tag{5}$$

Here γ_n and γ_n' are the rates at which the function f and its derivative f' can be estimated in the model (4) and in the corresponding regression models. For any $f_0 \in \Sigma$, introduce the neighborhood

$$\Sigma_{f_0}^n = \{f \in \Sigma : f(x) = f_0(x), x \notin [A, B],$$
$$\|f - f_0\|_\infty \le \gamma_n, \|f' - f_0'\|_\infty \le \gamma_n'\},$$

where $A < B$ are two constants.

Our main results are the following two theorems which state the local asymptotic equivalence of our nonparametric autoregressive model to a nonparametric regression with random and nonrandom designs. We start with the case of random design.

THEOREM 2.1. *Let $\mathcal{E}_{f_0}^n = (\mathbb{R}^n, \mathcal{B}^n, \{P_f^n, f \in \Sigma_{f_0}^n\})$ be the local experiment based on observations X_i, $i = 0, \dots, n$, obeying* (A1) *with $f \in \Sigma_{f_0}^n$. Suppose that* *the density* $p(\cdot)$ *satisfies assumption* (A2). *Let* $\mathcal{G}^n_{f_0} = (\mathbb{R}^n, \mathcal{B}^n, \{Q^n_f, f \in \Sigma^n_{f_0}\})$ *be the nonparametric regression model in which we observe*

$$Y_i = f(\xi_i) + \eta_i, \qquad i = 1, \dots, n, \tag{6}$$

where $\eta_1, \dots, \eta_n$ *are i.i.d. with density* $q(\cdot)$ *obeying* (A3), $\xi_1, \dots, \xi_n$ *are i.i.d. with the common density* $\psi_{f_0}(\cdot)$, *independent of* $\eta_1, \dots, \eta_n$, *and* $f \in \Sigma^n_{f_0}$ *is unknown. Then, for all* $\beta > 5/2$, *the sequences of experiments* $\mathcal{E}^n_{f_0}$, $n = 1, 2, \dots$, *and* $\mathcal{G}^n_{f_0}$, $n = 1, 2, \dots$, *are asymptotically equivalent uniformly in* $f_0 \in \Sigma$:

$$\sup_{f_0 \in \Sigma} \Delta(\mathcal{E}^n_{f_0}, \mathcal{G}^n_{f_0}) \to 0 \qquad \text{as } n \to \infty.$$

Our second local result states asymptotic equivalence to the regression model with nonrandom design.

THEOREM 2.2. *Let* $\mathcal{E}^n_{f_0} = (\mathbb{R}^n, \mathcal{B}^n, \{P^n_f, f \in \Sigma^n_{f_0}\})$ *be the local experiment based on observations* X_i, $i = 0, \dots, n$, *obeying assumption* (A1) *with* $f \in \Sigma^n_{f_0}$. *Assume that the density* $p(\cdot)$ *satisfies assumption* (A2). *Let* $\mathcal{G}^n_{f_0} = (\mathbb{R}^n, \mathcal{B}^n, \{Q^n_f, f \in \Sigma^n_{f_0}\})$ *be the nonparametric regression model in which we observe*

$$Y_{n,i} = f(t_{n,i}) + \eta_i, \qquad i = 1, \dots, n, \tag{7}$$

where $\eta_1, \dots, \eta_n$ *are i.i.d. with density* $q(\cdot)$ *obeying assumption* (A3). *Furthermore,* $t_{n,1}, \dots, t_{n,n}$ *are nonrandom design points chosen according to the density* $\psi_{f_0}(\cdot)$, *that is,* $(i - 1/2)/n = \int_{-\infty}^{t_{n,i}} \psi_{f_0}(x)\, dx$, $i = 1, \dots, n$, *and* $f \in \Sigma^n_{f_0}$ *is unknown. Then, for all* $\beta > 5/2$, *the sequences of experiments* $\mathcal{E}^n_{f_0}$, $n = 1, 2, \dots$, *and* $\mathcal{G}^n_{f_0}$, $n = 1, 2, \dots$, *are asymptotically equivalent uniformly in* $f_0 \in \Sigma$:

$$\sup_{f_0 \in \Sigma} \Delta(\mathcal{E}^n_{f_0}, \mathcal{G}^n_{f_0}) \to 0 \qquad \text{as } n \to \infty.$$

REMARK 1. As a by-product of our main results, we obtain also asymptotic equivalence of nonparametric regression with random and regular nonrandom design. However, since we used a construction of the likelihood ratios based on a Skorokhod embedding rather than a KMT construction, the rate for the approximation error between the likelihood ratios of both models is presumably not the best possible one. We conjecture that the constraint $\beta > 5/2$ that was imposed for proving asymptotic equivalence of nonparametric autoregression and nonparametric regression can be further relaxed for the case of asymptotic equivalence of nonparametric regression with random and regular nonrandom design. It follows from the results in [4] and [2] that in the special case of Gaussian errors this equivalence holds even for $\beta > 1/2$.

REMARK 2. Our results on asymptotic equivalence in the Le Cam sense of nonparametric regression and autoregression can be used to transfer existing lower asymptotic efficiency bounds (when the loss is measured in the supremum norm) in nonparametric regression to the case of nonparametric autoregression. Indeed, it can be seen from the calculations in [9], Section 5, that a shrinking neighborhood of size $O((\log n/n)^{\beta/(2\beta+1)})$ around some central function f_0 is large enough for generating the desired risk bound. Hence, we can actually deduce these lower asymptotic efficiency bounds in the cases $\beta > 5/2$ which are covered by our results.

Owing to the local character of our results (asymptotic equivalence is proved for shrinking neighborhoods of f_0), we cannot directly use them for transferring upper asymptotic risk bounds. However, such bounds can be easily derived by straightforward calculations or by using asymptotic equivalence results between nonparametric estimators in both settings as given by strong approximations in [22].

The possibility of transferring asymptotic efficiency bounds on the basis of the asymptotic equivalence of experiments has been already known for a long time. This principle was applied by Korostelev and Nussbaum [18] for deducing asymptotic minimax bounds in nonparametric density estimation from known results in signal estimation in Gaussian white noise. On the basis of local equivalence results, Drees [11] transferred available lower asymptotic risk bounds from the Gaussian white noise model to the case of estimating an extreme value index.

3. Proofs of the main theorems.

In this section we shall prove Theorem 2.1. Theorem 2.2 can be derived in the same way.

Our method of estimating the Le Cam distance $\Delta(\mathscr{E}^n_{f_0}, \mathscr{G}^n_{f_0})$ runs as follows. Let $X_0, \dots, X_n$ be the observations obeying assumption (A1) with $f \in \Sigma^n_{f_0}$ and let $(Y_1, \xi_1), \dots, (Y_n, \xi_n)$ be the observations defined in Theorem 2.1. Denote by $L^{1,n}_{f,f_0}$ and $L^{2,n}_{f,f_0}$ the likelihood ratio processes of the experiments $\mathscr{E}^n_{f_0}$ and $\mathscr{G}^n_{f_0}$, respectively,

$$L^{1,n}_{f,f_0} = \frac{\psi_f(X_0)}{\psi_{f_0}(X_0)} \prod_{i=1}^n \frac{p(X_i - f(X_{i-1}))}{p(X_i - f_0(X_{i-1}))}$$

and

$$L^{2,n}_{f,f_0} = \prod_{i=1}^n \frac{q(Y_i - f(\xi_i))}{q(Y_i - f_0(\xi_i))}.$$

According to Proposition 2.2 in [23] (see also [20], page 16, for a similar assertion in the parametric context), the deficiency distance can be estimated as

$$\Delta(\mathscr{E}^n_{f_0}, \mathscr{G}^n_{f_0}) \le \sup_{f \in \Sigma^n_{f_0}} E_{\mathbf{P}} |\widetilde{L}^{1,n}_{f,f_0} - \widetilde{L}^{2,n}_{f,f_0}|, \tag{8}$$

where $\widetilde{L}^{1,n}_{f,f_0}$ and $\widetilde{L}^{2,n}_{f,f_0}$ are arbitrary versions of the likelihood ratios $L^{1,n}_{f,f_0}$ and $L^{2,n}_{f,f_0}$ constructed on a common probability space $(\Omega, \mathscr{F}, \mathbf{P})$ and distributed according to the central measure P_{f_0}. The versions $\widetilde{L}^{1,n}_{f,f_0}$ and $\widetilde{L}^{2,n}_{f,f_0}$ will be constructed in such a way that the right-hand side of (8) tends to zero as $n \to \infty$. Since this will hardly cause any confusion, we drop the tildes in the notation of $\widetilde{L}^{1,n}_{f,f_0}$ and $\widetilde{L}^{2,n}_{f,f_0}$. With this agreement inequality (8) can be written as

$$\Delta(\mathcal{E}^n_{f_0}, \mathcal{G}^n_{f_0}) \le \sup_{f \in \Sigma^n_{f_0}} E_{f_0} |L^{1,n}_{f,f_0} - L^{2,n}_{f,f_0}|. \tag{9}$$

The subscript f_0 at the expectation indicates that the measure $\mathbf{P}$ corresponds to the central measure P_{f_0}.

First, we give a bound for the L_1-distance on the right-hand side of (9) in terms of the Hellinger distance:

$$\tfrac{1}{2} E_{f_0} |L^{1,n}_{f,f_0} - L^{2,n}_{f,f_0}| \le H(P^n_f, Q^n_f) \equiv \sqrt{E_{f_0}\Bigl(\sqrt{L^{1,n}_{f,f_0}} - \sqrt{L^{2,n}_{f,f_0}}\Bigr)^2}, \tag{10}$$

where $L^{1,n}_{f,f_0}$ and $L^{2,n}_{f,f_0}$ mean the corresponding versions of the likelihood ratios. Here $H(P, Q)$ denotes the Hellinger distance between two probability measures P and Q. Following an idea originating from [23] in the context of density estimation and from [14] in the context of regression with independent observations, we shall use an analogue of the following property of the Hellinger distance for product measures (see Lemma 2.17 in [27]):

$$H^2\Biggl(\bigotimes_{l=1}^{K_n} P^{(l)}, \bigotimes_{l=1}^{K_n} Q^{(l)}\Biggr) \le \sum_{l=1}^{K_n} H^2(P^{(l)}, Q^{(l)}), \tag{11}$$

where $P^{(l)}$ and $Q^{(l)}$ are the measures corresponding to certain disjoint blocks of observations and K_n is a sequence satisfying $K_n \to \infty$ and $K_n/n \to 0$. The size of these blocks will be chosen small enough so that one can get reasonable estimates for $H^2(P^{(l)}, Q^{(l)})$. It is clear that the estimate (11) is essentially based on the product structure of the measures $\bigotimes_{l=1}^{K_n} P^{(l)}$ and $\bigotimes_{l=1}^{K_n} Q^{(l)}$, and in general does not directly apply to the case of dependent observations.

In the particular context of the dependent data under consideration, we proceed as follows. Set $K_n = [n^{1/6}]$. Split the set of indices $\{1, \dots, n\}$ into K_n blocks,

$$\mathcal{I}_l = \Biggl\{ i : (l-1)\frac{n}{K_n} < i \le l\frac{n}{K_n} \Biggr\}, \qquad l = 1, \dots, K_n.$$

Denote by m_l the number of elements in the block $\mathcal{I}_l$, that is, $m_l = \#\mathcal{I}_l = O(n^{5/6})$. Let i_l be the first element in the set $\mathcal{I}_l$. Furthermore, let $\mathcal{F}_0$ be the trivial σ-field and, for $1 \le l \le K_n$,

$$\mathcal{F}_l = \sigma\bigl(X_0, \dots, X_{i_l - 1}; (Y_1, \xi_1), \dots, (Y_{i_l-1}, \xi_{i_l-1})\bigr).$$

The likelihood ratio corresponding to the observations $X_0, \dots, X_n$ can be written as the product

$$L^{1,n}_{f,f_0} = \prod_{l=0}^{K_n} L^{1,(l)}_{f,f_0}, \qquad L^{1,(l)}_{f,f_0} = \prod_{i \in \mathcal{I}_l} \frac{p(X_i - f(X_{i-1}))}{p(X_i - f_0(X_{i-1}))}, \qquad 1 \le l \le K_n,$$

where $L_{f,f_0}^{1,(0)} = \psi_f(X_0)/\psi_{f_0}(X_0)$ and $L_{f,f_0}^{1,(l)}$ is the conditional (given $\mathcal{F}_l$) likelihood ratio generated by $(X_i : i \in \mathcal{I}_l)$. Analogously, in the case of a regression experiment with random design, we have that

$$L_{f,f_0}^{2,n} = \prod_{l=0}^{K_n} L_{f,f_0}^{2,(l)}, \qquad L_{f,f_0}^{2,(l)} = \prod_{i \in \mathcal{I}_l} \frac{q(Y_i - f(\xi_i))}{q(Y_i - f_0(\xi_i))}, \qquad 1 \le l \le K_n,$$

where $L_{f,f_0}^{2,(0)} = 1$. A generalization of (11) to our setting with dependent random variables is given by the following result.

LEMMA 3.1.

$$H^2(P_{f,f_0}^n, Q_{f,f_0}^n) \le \sum_{l=0}^{K_n} \operatorname{ess\,sup} E_{f_0}((\sqrt{L_{f,f_0}^{1,(l)}} - \sqrt{L_{f,f_0}^{2,(l)}})^2 | \mathcal{F}_l).$$

PROOF. The proof of this assertion is adapted from that of Lemma 2.17 in [27]. We rewrite the Hellinger distance as

$$\tfrac{1}{2} H^2(L_{f,f_0}^{1,n}, L_{f,f_0}^{2,n}) = \tfrac{1}{2} E_{f_0}(\sqrt{L_{f,f_0}^{1,n}} - \sqrt{L_{f,f_0}^{2,n}})^2 = 1 - E_{f_0} \prod_{l=0}^{K_n} \sqrt{L_{f,f_0}^{1,(l)} L_{f,f_0}^{2,(l)}}.$$

For the last term one easily deduces

$$\begin{aligned} E_{f_0} \prod_{l=0}^{K_n} \sqrt{L_{f,f_0}^{1,(l)} L_{f,f_0}^{2,(l)}} &= E_{f_0}\Biggl[\Biggl(\prod_{l=0}^{K_n-1} \sqrt{L_{f,f_0}^{1,(l)} L_{f,f_0}^{2,(l)}}\Biggr) E_{f_0}(\sqrt{L_{f,f_0}^{1,(K_n)} L_{f,f_0}^{2,(K_n)}} | \mathcal{F}_{K_n})\Biggr] \\ &\ge E_{f_0}\Biggl[\prod_{l=0}^{K_n-1} \sqrt{L_{f,f_0}^{1,(l)} L_{f,f_0}^{2,(l)}}\Biggr] \operatorname{ess\,inf} E_{f_0}(\sqrt{L_{f,f_0}^{1,(K_n)} L_{f,f_0}^{2,(K_n)}} | \mathcal{F}_{K_n}). \end{aligned}$$

Continuing in the same way we obtain

$$\begin{aligned} E_{f_0} \prod_{l=0}^{K_n} \sqrt{L_{f,f_0}^{1,(l)} L_{f,f_0}^{2,(l)}} &\ge E_{f_0}[\sqrt{L_{f,f_0}^{1,(0)} L_{f,f_0}^{2,(0)}}] \prod_{l=1}^{K_n} \operatorname{ess\,inf} E_{f_0}(\sqrt{L_{f,f_0}^{1,(l)} L_{f,f_0}^{2,(l)}} | \mathcal{F}_l) \\ &\ge \prod_{l=0}^{K_n} (1 - \operatorname{ess\,sup} \tfrac{1}{2} E_{f_0}((\sqrt{L_{f,f_0}^{1,(l)}} - \sqrt{L_{f,f_0}^{2,(l)}})^2 | \mathcal{F}_l)). \end{aligned}$$

Using the inequality $1 - \prod(1 - a_i) \le \sum a_i$, which is true for all $0 \le a_i \le 1$, we obtain the assertion of the lemma. □

Hence, we have an analogue of (11) for the case of dependent random variables. Separability, which is equivalent to independence of the factors in $\prod_{l=0}^{K_n} \sqrt{L_{f,f_0}^{1,(l)} L_{f,f_0}^{2,(l)}}$, is just achieved by transition to the "worst case" which is appropriately expressed by $\operatorname{ess\,sup} E_{f_0}((\sqrt{L_{f,f_0}^{1,(l)}} - \sqrt{L_{f,f_0}^{2,(l)}})^2 | \mathcal{F}_l)$.

Note that, since p is positive on $\mathbb{R}$ and $\sup_{f \in \Sigma_{f_0}^n} \|f\|_\infty \le M < \infty$, the condition $\rho < 1$ of Lemma 6.1 below is satisfied with some ρ depending on p and M. Then Lemma 6.1 and assumption (A2) imply, as $n \to \infty$,

$$\begin{aligned} &\sup_{f_0 \in \Sigma} \sup_{f \in \Sigma_{f_0}^n} E_{f_0}((\sqrt{L_{f,f_0}^{1,(0)}} - \sqrt{L_{f,f_0}^{2,(0)}})^2 | \mathcal{F}_0) \\ &\qquad = \sup_{f_0 \in \Sigma} \sup_{f \in \Sigma_{f_0}^n} E_{f_0}\big(\sqrt{\psi_f(X_0)/\psi_{f_0}(X_0)} - 1\big)^2 \to 0. \end{aligned} \tag{12}$$

Now Theorem 2.1 follows from Lemma 3.1, (12) and from the following assertion which provides us with bounds for the conditional Hellinger distance.

PROPOSITION 3.1. *Suppose that assumptions* (A1)–(A3) *are satisfied. Then there exists a construction of the sequences* $X_0, \dots, X_n$ *and* $(Y_1, \xi_1), \dots, (Y_n, \xi_n)$ *on a common probability space such that*

$$\max_{1 \le l \le K_n} \sup_{f_0 \in \Sigma} \sup_{f \in \Sigma_{f_0}^n} \operatorname{ess\,sup} E_{f_0}((\sqrt{L_{f,f_0}^{1,(l)}} - \sqrt{L_{f,f_0}^{2,(l)}})^2 | \mathcal{F}_l) = o(K_n^{-1}).$$

The proof of this proposition is postponed to Section 4.

In the case of comparing nonparametric autoregression and regression with regular nonrandom design, we proceed analogously. We use the same splitting of the set of indices $\{1, \dots, n\}$ into blocks $\mathcal{I}_1, \dots, \mathcal{I}_{K_n}$ as above. The pairs $(Y_{n,1}, t_{n,1}), \dots, (Y_{n,n}, t_{n,n})$ are rearranged in such a way that

$$\left| \int_{-\infty}^{t_{n,n_l+i-1}} \psi_{f_0}(x)\, dx - \frac{i - 1/2}{m_l} \right| = O(m_l^{-1}),$$

for all $i \in \{1, \dots, m_l\}$, $l \in \{1, \dots, K_n\}$. Then we write the likelihood ratio as

$$L_{f,f_0}^{3,n} = \prod_{l=0}^{K_n} L_{f,f_0}^{3,(l)}, \qquad L_{f,f_0}^{3,(l)} = \prod_{i \in \mathcal{I}_l} \frac{q(Y_{n,i} - f(t_{n,i}))}{q(Y_{n,i} - f_0(t_{n,i}))}, \qquad 1 \le l \le K_n,$$

where $L_{f,f_0}^{3,(0)} = 1$. Let $\mathcal{F}_0'$ be the trivial σ-field and, for $l = 1, \dots, K_n$,

$$\mathcal{F}_l' = \sigma\big(X_0, \dots, X_{i_l-1}; Y_{n,1}, \dots, Y_{n,i_l-1}\big).$$

Using the same arguments as in the proof of Proposition 3.1 we obtain the following assertion.

PROPOSITION 3.2. *Suppose that assumptions* (A1)–(A3) *are satisfied. Then there exists a construction of the sequences $X_0, \dots, X_n$ and $Y_{n,1}, \dots, Y_{n,n}$ on a common probability space such that, as $n \to \infty$,*

$$\max_{1 \le l \le K_n} \sup_{f_0 \in \Sigma} \sup_{f \in \Sigma_{f_0}^n} \operatorname{ess\,sup} E_{f_0}\Big(\Big(\sqrt{L_{f,f_0}^{1,(l)}} - \sqrt{L_{f,f_0}^{3,(l)}}\Big)^2 \big| \mathcal{F}_l'\Big) = o(K_n^{-1}).$$

Theorem 2.2 follows from Lemma 3.1, (12) and Proposition 3.2.

4. Proofs of Propositions 3.1 and 3.2.

PROOF OF PROPOSITION 3.1. Let $X_0, \dots, X_n$ and $(Y_1, \xi_1), \dots, (Y_n, \xi_n)$ be the observations generated according to (4) and (6). According to Theorem 5.1, there is a construction of the sequences $X_0, \dots, X_n$ and $(Y_1, \xi_1), \dots, (Y_n, \xi_n)$ on a common probability space which are coupled in such a way that the assertion of Theorem 5.1 holds true. Without loss of generality, we can assume that the sequences $X_0, \dots, X_n$ and $(Y_1, \xi_1), \dots, (Y_n, \xi_n)$ are already constructed on the probability space $(\Omega, \mathcal{F}^n, P_{f_0})$ endowed with the central measure P_{f_0}.

Recall that $m_l = \#\mathcal{I}_l = O(n^{5/6})$ is the number of indices in the set $\mathcal{I}_l$ and that $K_n = [n^{1/6}]$ is the number of blocks. Set, for brevity, $g(x) = f(x) - f_0(x)$. Since $f \in \Sigma_{f_0}^n$, we have $\|g\|_\infty \le \gamma_n$ and $\|g'\|_\infty \le \gamma_n'$. Since $\sup_x |l_p'''(x)| \le c_1$ [by assumption (A2)(i)] and $\gamma_n^3 m_l = o(K_n^{-1/2})$, we obtain by a Taylor series expansion that

$$\begin{aligned}\log L_{f,f_0}^{1,(l)} &= \sum_{i \in \mathcal{I}_l} g(X_{i-1}) l_p'(\varepsilon_i) + \tfrac{1}{2} \sum_{i \in \mathcal{I}_l} g(X_{i-1})^2 l_p''(\varepsilon_i) + o(K_n^{-1/2}) \\ &= T_1^{1,(l)} + T_2^{1,(l)} + o(K_n^{-1/2})\end{aligned} \tag{13}$$

and, in the same way,

$$\begin{aligned}\log L_{f,f_0}^{2,(l)} &= \sum_{i \in \mathcal{I}_l} g(\xi_i) l_q'(\eta_i) + \tfrac{1}{2} \sum_{i \in \mathcal{I}_l} g(\xi_i)^2 l_q''(\eta_i) + o(K_n^{-1/2}) \\ &= T_1^{2,(l)} + T_2^{2,(l)} + o(K_n^{-1/2}).\end{aligned} \tag{14}$$

We introduce the set $A_l = A_{l,1} \cap A_{l,2}$, where

$$\begin{aligned}A_{l,1} &= \{|T_1^{1,(l)} - T_1^{2,(l)}| \le c_1 (\gamma_n)^{1/4} (\gamma_n')^{3/4} m_l^{1/4} \log m_l\}, \\ A_{l,2} &= \{|T_2^{1,(l)} - T_2^{2,(l)}| \le v_n K_n^{-1/2}\},\end{aligned}$$

and $v_n \to 0$ sufficiently slowly. An appropriate choice of the sequence v_n is described in the course of the proof of Lemma 4.1 below. We bound the Hellinger distance between the partial likelihoods $L^{1,(l)}_{f,f_0}$ and $L^{2,(l)}_{f,f_0}$ as

$$E_{f_0}\big(\big(\sqrt{L^{1,(l)}_{f,f_0}} - \sqrt{L^{2,(l)}_{f,f_0}}\big)^2 | \mathcal{F}_l\big) \le R_1 + R_2, \tag{15}$$

say, where

$$R_1 = E_{f_0}\big(\big(\sqrt{L^{1,(l)}_{f,f_0}} - \sqrt{L^{2,(l)}_{f,f_0}}\big)^2 I_{A_l} | \mathcal{F}_l\big), \qquad R_2 = E_{f_0}\big(2\big(L^{1,(l)}_{f,f_0} + L^{2,(l)}_{f,f_0}\big) I_{\overline{A_l}} | \mathcal{F}_l\big).$$

First we bound R_1. On the set A_l, we get

$$|\log L^{1,(l)}_{f,f_0} - \log L^{2,(l)}_{f,f_0}| = O\big((\gamma_n)^{1/4}(\gamma_n')^{3/4} m_l^{1/4} \log m_l\big) + o(K_n^{-1/2}).$$

Since $\beta > 5/2$, we have $(\gamma_n)^{1/4}(\gamma_n')^{3/4} m_l^{1/4} \log m_l = o(K_n^{-1/2})$, which in turn implies that

$$\Big|\sqrt{L^{1,(l)}_{f,f_0}/L^{2,(l)}_{f,f_0}} - 1\Big| = \big|\exp\big(\tfrac{1}{2}\log L^{1,(l)}_{f,f_0} - \tfrac{1}{2}\log L^{2,(l)}_{f,f_0}\big) - 1\big| = o(K_n^{-1/2}).$$

Taking into account that $E_{f_0}(L^{2,(l)}_{f,f_0} | \mathcal{F}_l) = 1$, we get

$$\begin{aligned} R_1 &= E_{f_0}\Big(L^{2,(l)}_{f,f_0}\Big(\sqrt{L^{1,(l)}_{f,f_0}/L^{2,(l)}_{f,f_0}} - 1\Big)^2 I_{A_l} | \mathcal{F}_l\Big) \\ &= o\big(K_n^{-1} E_{f_0}\big(L^{2,(l)}_{f,f_0} | \mathcal{F}_l\big)\big) = o(K_n^{-1}). \end{aligned} \tag{16}$$

Now we shall bound R_2. Set $B_l = \{\log L^{1,(l)}_{f,f_0} \le 1\}$ and $C_l = \{\log L^{2,(l)}_{f,f_0} \le 1\}$. Then

$$\begin{aligned} R_2 &\le E_{f_0}\big(2\big(L^{1,(l)}_{f,f_0} I_{B_l} + L^{2,(l)}_{f,f_0} I_{C_l}\big) I_{\overline{A_l}} | \mathcal{F}_l\big) \\ &\quad + E_{f_0}\big(2\big(L^{1,(l)}_{f,f_0} I_{\overline{B_l}} + L^{2,(l)}_{f,f_0} I_{\overline{C_l}}\big) | \mathcal{F}_l\big) \\ &\le 4e P_{f_0}(\overline{A}_l | \mathcal{F}_l) + 2E_{f_0}\big(L^{1,(l)}_{f,f_0} I_{\overline{B}_l} | \mathcal{F}_l\big) + 2E_{f_0}\big(L^{2,(l)}_{f,f_0} I_{\overline{C}_l} | \mathcal{F}_l\big). \end{aligned} \tag{17}$$

We will prove that

$$P_{f_0}(\overline{A}_l | \mathcal{F}_l) = o(K_n^{-1}), \qquad P_{f_0}\text{-a.s.}, \tag{18}$$

and that

$$\begin{aligned} &E_{f_0}\big(L^{1,(l)}_{f,f_0} I_{\overline{B}_l} | \mathcal{F}_l\big) = o(K_n^{-1}), \\ &E_{f_0}\big(L^{2,(l)}_{f,f_0} I_{\overline{C}_l} | \mathcal{F}_l\big) = o(K_n^{-1}), \qquad P_{f_0}\text{-a.s.} \end{aligned} \tag{19}$$

Then, in conjunction with (15)–(17), we obtain the desired bound

$$E_{f_0}\big(\big(\sqrt{L^{1,(l)}_{f,f_0}} - \sqrt{L^{2,(l)}_{f,f_0}}\big)^2 | \mathcal{F}_l\big) = o(K_n^{-1}).$$

Hence, it remains to prove (18) and (19).

First we prove (18). By Theorem 5.1 (with some λ large enough) we have that

$$P_{f_0}(\overline{A}_{l,1}|\mathcal{F}_l) = O(m_l^{-\lambda}) = o(K_n^{-1}). \tag{20}$$

To complete the proof of (18) we shall prove the following bound.

LEMMA 4.1.

$$P_{f_0}(\overline{A}_{l,2}|\mathcal{F}_l) = o(K_n^{-1}). \tag{21}$$

PROOF. We shall use the fact that the Markov chain $X_0, \dots, X_n$ is ϕ-mixing. Decompose the set $\mathcal{I}_l$ as $\mathcal{I}_l = \mathcal{I}_l^{(1)} \cup \mathcal{I}_l^{(2)}$ so that $\mathcal{I}_l^{(1)}$ contains the first $c_0 \log m_l$ elements of the set $\mathcal{I}_l$ and $\mathcal{I}_l^{(2)}$ the remaining ones, where the positive constant c_0 will be chosen below. Let $l \in \{1, \dots, K_n\}$ and let $i_{l,2}$ be the first element of $\mathcal{I}_l^{(2)}$. According to Lemma 6.2 in Section 6.2, we can construct a version $\widetilde{X}_{i_{l,2}-1}$ of the r.v. $X_{i_{l,2}-1}$ on the same probability space, such that $\widetilde{X}_{i_{l,2}-1}$ is independent of $\mathcal{F}_l$ and

$$P_{f_0}\big(\widetilde{X}_{i_{l,2}-1} \neq X_{i_{l,2}-1}|\mathcal{F}_l\big) \le \phi\big(X_{i_{l,2}-1}, \mathcal{F}_l\big) \le c\rho^{c_0 \log m_l}, \tag{22}$$

for some large enough constant c_0 and for some $\rho < 1$. Having constructed $\widetilde{X}_{i-1}$ for some $i \in \mathcal{I}_l^{(2)}$, we define recursively a version $\widetilde{X}_i$ of the r.v. X_i on the same probability space, such that $\widetilde{X}_i$ is independent of $\mathcal{F}_l$ and of $\widetilde{X}_{i-c_0\log m_l}, \dots, \widetilde{X}_{i_l}, \varepsilon_{i-c_0\log m_l}, \dots, \varepsilon_{i_l}$ and

$$P_{f_0}(\widetilde{X}_i \neq X_i|\mathcal{F}_l) = \rho^{c_0 \log m_l}.$$

Choosing c_0 large enough, $\widetilde{X}_{i_l-1}, \dots, \widetilde{X}_{i_{l+1}-1}$ satisfy

$$\begin{aligned} P_{f_0}(\widetilde{X}_i \neq X_i, \forall i \in \{i_l - 1, \dots, i_{l+1} - 1\}|\mathcal{F}_l) &= m_l \rho^{c_0 \log m_l} \\ &= o(K_n^{-1}). \end{aligned} \tag{23}$$

Denote

$$\widetilde{T}_2^{1,(l)} = \tfrac{1}{2} \sum_{i \in \mathcal{I}_l} g(\widetilde{X}_{i-1})^2 l_p''(\varepsilon_i) = \tfrac{1}{2} \sum_{i \in \mathcal{I}_l} \zeta_i,$$

where $\zeta_i = g(\widetilde{X}_{i-1})^2 l_p''(\varepsilon_i)$. Since $\widetilde{T}_2^{1,(l)}$ is a sum of $c_0 \log m_l$-dependent r.v.'s, using Chebyshev's inequality, we obtain

$$\begin{aligned} &P_{f_0}\big(\big|\widetilde{T}_2^{1,(l)} - E_{f_0}\big(\widetilde{T}_2^{1,(l)}|\mathcal{F}_l\big)\big| > v_n K_n^{-1/2}|\mathcal{F}_l\big) \\ &\quad \le K_n/v_n E_{f_0}\big(\big(\widetilde{T}_2^{1,(l)} - E_{f_0}\big(\widetilde{T}_2^{1,(l)}|\mathcal{F}_l\big)\big)^2|\mathcal{F}_l\big) \\ &\quad = O\left(\frac{K_n}{v_n^2} \sum_{i \in \mathcal{I}_l} \sum_{j \in \mathcal{I}_l} E_{f_0}\big((\zeta_i - E_{f_0}(\zeta_i|\mathcal{F}_l))(\zeta_j - E_{f_0}(\zeta_j|\mathcal{F}_l))|\mathcal{F}_l\big)\right) \\ &\quad = O(v_n^{-2} K_n \gamma_n^4 m_l \log m_l) = o(v_n^{-2} n^{-1/3}). \end{aligned}$$

Choosing v_n such that $v_n \to 0$ and $o(v_n^{-2} n^{-1/3}) = o(K_n^{-1})$ we get

$$P_{f_0}\big(|\widetilde{T}_2^{1,(l)} - E_{f_0}(\widetilde{T}_2^{1,(l)}|\mathcal{F}_l)| > v_n K_n^{-1/2}|\mathcal{F}_l\big) = o(K_n^{-1}). \tag{24}$$

By similar arguments for sums of independent random variables, we can show that

$$P_{f_0}\big(|T_2^{2,(l)} - E_{f_0} T_2^{2,(l)}| > v_n K_n^{-1/2}|\mathcal{F}_l\big) = o(K_n^{-1}). \tag{25}$$

Taking into account that $E_{f_0} l_p''(\varepsilon_i) = E_{f_0} l_q''(\eta_i) = I$ [by Assumption (A3)(iii)] we obtain

$$\begin{aligned} &E_{f_0}(\widetilde{T}_2^{1,(l)}|\mathcal{F}_l) - E_{f_0}(T_2^{2,(l)}) \\ &\quad = O(\gamma_n^2 \log m_l) + \frac{I}{2} \sum_{i \in \mathcal{I}_l^{(2)}} E_{f_0}(g(\widetilde{X}_{i-1})^2|\mathcal{F}_l) - \frac{I}{2} \sum_{i \in \mathcal{I}_l^{(2)}} E_{f_0} g(\xi_i)^2. \end{aligned}$$

Since X_{i-1} and ξ_i have the same density ψ_f we get $E_{f_0}[g(X_{i-1})^2] = E_{f_0}[g(\xi_i)^2]$, and thus

$$E_{f_0}(\widetilde{T}_2^{1,(l)}|\mathcal{F}_l) - E_{f_0}(T_2^{2,(l)}) = O(\gamma_n^2 \log m_l) = o(K_n^{-1}). \tag{26}$$

By (23)–(26) we get

$$\begin{aligned} P_{f_0}&\big(|T_2^{1,(l)} - T_2^{2,(l)}| \geq v_n K_n^{-1/2}|\mathcal{F}_l\big) \\ &\leq P_{f_0}\big(|\widetilde{T}_2^{1,(l)} - T_2^{2,(l)}| \geq v_n K_n^{-1/2}|\mathcal{F}_l\big) \\ &\qquad + P_{f_0}\big(\widetilde{X}_i \neq X_i \ \forall i \in \{i_l - 1, \ldots, i_{l+1} - 1\}|\mathcal{F}_l\big) \\ &= o(K_n^{-1}), \end{aligned}$$

which proves (21). □

Now we prove (19). We give a proof for the first bound; the second one can be proved in the same way. Changing the probability measure we obtain that

$$E_{f_0}\big(L_{f,f_0}^{1,(l)} I_{\overline{B}_l}|\mathcal{F}_l\big) = P_f(\overline{B}_l|\mathcal{F}_l) = P_f\big(\log L_{f,f_0}^{1,(l)} > 1|\mathcal{F}_l\big).$$

We shall prove that

$$P_f\big(\log L_{f,f_0}^{1,(l)} > 1|\mathcal{F}_l\big) = o(K_n^{-1}). \tag{27}$$

Indeed, proceeding as in the proof of (24) and using the fact that $\varepsilon_i = X_i - f_0(X_{i-1}) = X_i - f(X_{i-1}) + o(\gamma_n)$ and assumption (A2)(ii), one gets

$$P_f\big(|T_2^{1,(l)} - E_f(T_2^{1,(l)}|\mathcal{F}_l)| \geq c K_n^{-1/2}|\mathcal{F}_l\big) = o(K_n^{-1}). \tag{28}$$

Since $E_f(T_2^{1,(l)}|\mathcal{F}_l) = o(1)$, we get from (13) and (28),

$$P_f\big(\log L_{f,f_0}^{1,(l)} > 1|\mathcal{F}_l\big) \leq o(K_n^{-1}) + P_f\big(T_1^{1,(l)} > \tfrac{1}{2}|\mathcal{F}_l\big). \tag{29}$$

If we prove that

$$P_f\big(T_1^{1,(l)} > \tfrac{1}{2}|\mathcal{F}_l\big) = o(K_n^{-1}), \tag{30}$$

then we get, in conjunction with (29), that (27) holds.

To prove (30) we use the exponential Chebyshev's inequality for martingales. Since $\beta > 5/2$, by (5), we have $\gamma_n = o(n^{-5/12-3\delta})$, for some $\delta > 0$ small enough. Recall that $m_l = O(n^{5/6})$ and $\|g\|_\infty \le \gamma_n$. Assume first that $n^{-\delta}|l'_p(\varepsilon_i)| \le const$. Using Lemma 6.3,

$$\begin{aligned}
&P_f\big(T_1^{1,(l)} > \tfrac{1}{2}|\mathcal{F}_l\big)\\
&\quad\le e^{-n^\delta} E_f\Bigg(\exp\Bigg(2\sum_{i\in\mathcal{I}_l} n^\delta g(X_{i-1})l'_p(\varepsilon_i)\Bigg)\Bigg|\mathcal{F}_l\Bigg)\\
&\quad= e^{-n^\delta} E_f\Bigg(\prod_{i\in\mathcal{I}_l} E_f\big(\exp(2n^\delta g(X_{i-1})l'_p(\varepsilon_i))|X_{i-1}\big)|\mathcal{F}_l\Bigg)\\
&\quad\le e^{-n^\delta}\prod_{i\in\mathcal{I}_l}\exp(cn^{2\delta}\gamma_n^2 E_f l'_p(\varepsilon_i)^2)\\
&\quad\le e^{-n^\delta}\exp(cn^{2\delta}\gamma_n^2 m_l E_f l'_p(\varepsilon_i)^2),
\end{aligned}$$

where c is a constant. The latter implies (30). If $n^{-\delta}|l'_p(\varepsilon_i)| \le const$ is not satisfied, we use the same arguments with truncated scores $\hat{l}_i = \bar{l}_i - E_f(\bar{l}_i|X_{i-1})$, $\bar{l}_i = l'_p(\varepsilon_i)1(|l'_p(\varepsilon_i)| \le n^\delta)$ instead of the true scores $l'_p(\varepsilon_i)$. The term with the difference $l'_p(\varepsilon_i) - \hat{l}_i$ is bounded easily as before, using Chebyshev's inequality, the fact that $\varepsilon_i = X_i - f_0(X_{i-1}) = X_i - f(X_{i-1}) + o(\gamma_n)$ and assumption (A2)(iii):

$$\begin{aligned}
&P_f\Bigg(\sum_{i\in\mathcal{I}_l}\big(g(X_{i-1})l'_p(\varepsilon_i) - \hat{l}_i\big) > \tfrac{1}{2}|\mathcal{F}_l\Bigg)\\
&\qquad = O\big(\gamma_n^2 m_l E_f l'_p(\varepsilon_i)^2\mathbb{1}\big(|l'_p(\varepsilon_i)| \ge n^\delta\big)\big) = O(K_n^{-1}).
\end{aligned}$$

Using the same types of arguments for sums of independent random variables we obtain

$$E_{f_0}\big(L_{f,f_0}^{2,(l)} I_{\overline{C}_l}|\mathcal{F}_l\big) = P_f(\overline{C}_l|\mathcal{F}_l) = P_f\big(\log L_{f,f_0}^{2,(l)} > 1|\mathcal{F}_l\big) = o(K_n^{-1}),$$

which completes the proof of the first bound in (19). □

PROOF OF PROPOSITION 3.2. This proof is analogous to that of Proposition 3.1 and requires only a few minor modifications. Analogously to (13) and (14), we use the Taylor expansion

$$\begin{aligned}
\log L_{f,f_0}^{3,(l)} &= \sum_{i\in\mathcal{I}_l} g(t_{n,i})l'_q(\eta_i) + \tfrac{1}{2}\sum_{i\in\mathcal{I}_l} g(t_{n,i})^2 l''_q(\eta_i) + o(K_n^{-1/2})\\
&= T_1^{3,(l)} + T_2^{3,(l)} + o(K_n^{-1/2}).
\end{aligned} \tag{31}$$

Similarly to the calculations in the proof of Proposition 3.1, the closeness of $T_1^{1,(l)}$ and $T_1^{3,(l)}$ follows from Theorem 5.2, while that of $T_2^{1,(l)}$ and $T_2^{3,(l)}$ follows in complete analogy to the derivation of (21). □

5. A functional strong approximation result. In the proof of the main results we use the following strong approximation theorem. It can be viewed as an analogue of the functional strong approximation result established in [16] for sums of independent random variables.

Let X_i, $i = 1, \dots, n$, and (Y_i, ξ_i), $i = 1, \dots, n$, be defined according to (4) and (6), respectively. Let $f_0 \in \Sigma$ and $f \in \Sigma_{f_0}^n$. We set

$$S_f^{1,(l)} = \sum_{i \in \mathcal{I}_l} (f - f_0)(X_{i-1}) l_p'(\varepsilon_i), \qquad S_f^{2,(l)} = \sum_{i \in \mathcal{I}_l} (f - f_0)(\xi_i) l_q'(\eta_i).$$

THEOREM 5.1. *Suppose that assumptions* (A1)–(A3) *are satisfied. Let* $\lambda > 1$ *be a constant. Then there are versions of the random variables* $X_0, \dots, X_n$ *and* $(Y_1, \xi_1), \dots, (Y_n, \xi_n)$ *on a common probability space such that, for* $1 \le l \le K_n$,

$$\sup_{f_0 \in \Sigma} \sup_{f \in \Sigma_{f_0}^n} \operatorname{ess\,sup} P_{f_0}\big(|S_f^{1,(l)} - S_f^{2,(l)}| > c(\lambda) r_n | \mathcal{F}_l\big) = O(m_l^{-\lambda}),$$

where $r_n = (\gamma_n)^{1/4} (\gamma_n')^{3/4} m_l^{1/4} \log m_l + m_l^{-\lambda}$ *and* $c(\lambda)$ *is a constant depending only on* λ.

The proof of this functional approximation result is based on a truncated Haar series expansion of $f - f_0$ and Lemma 5.1 below which provides a strong approximation result for partial sums with respect to a system of dyadic subintervals of $[A, B]$.

Define, for $j \ge 0$ and $k = 0, \dots, 2^j$, $s_{j,k} = A + k 2^{-j}(B - A)$, and

$$I_{j,k} = (s_{j,k-1}, s_{j,k}], \qquad k = 1, \dots, 2^j.$$

The Haar basis functions are defined via indicators as

$$h_0 = (B - A)^{-1/2} \mathbf{1}_{I_{0,1}},$$
$$h_{j,k} = (B - A)^{-1/2} 2^{j/2} (\mathbf{1}_{I_{j+1,2k-1}} - \mathbf{1}_{I_{j+1,2k}}) \qquad (j \ge 0;\ k = 1, \dots, 2^j).$$

With a choice of the finest scale of the expansion, $j^* = j^*(n)$, described at the end of the proof of Theorem 5.1, we obtain a truncated Haar series expansion of $g = f - f_0$ as

$$g(x) = c_0(g) h_0(x) + \sum_{j=0}^{j^*} \sum_{k=1}^{2^j} c_{j,k}(g) h_{j,k}(x) + r_{j^*}(x),$$

where $c_0(g) = \int_A^B g(t) h_0(t)\, dt$, $c_{j,k}(g) = \int_A^B g(t) h_{j,k}(t)\, dt$, and $r_{j^*}(x)$ is the residual term. This yields that

$$
\begin{aligned}
&|S_f^{1,(l)} - S_f^{2,(l)}| \\
&\quad \le |c_0(g)| \left| \sum_{i \in \mathcal{I}_l} h_0(X_{i-1}) l_p'(\varepsilon_i) - \sum_{i \in \mathcal{I}_l} h_0(\xi_{i-1}) l_q'(\eta_i) \right| \\
&\qquad + \sum_{j=0}^{j^*} \sum_{k=1}^{2^j} |c_{j,k}(g)| \left| \sum_{i \in \mathcal{I}_l} h_{j,k}(X_{i-1}) l_p'(\varepsilon_i) - \sum_{i \in \mathcal{I}_l} h_{j,k}(\xi_{i-1}) l_q'(\eta_i) \right| \\
&\qquad + \left| \sum_{i \in \mathcal{I}_l} r_{j^*}(X_{i-1}) l_p'(\varepsilon_i) - r_{j^*}(\xi_{i-1}) l_p'(\eta_i) \right| \\
&\quad \le (B-A)^{-1/2} |c_0(g)| |Z_{0,1}^{1,(l)} - Z_{0,1}^{2,(l)}| \\
&\qquad + \sum_{j=0}^{j^*} 2^{j/2} \sum_{k=1}^{2^j} |c_{j,k}(g)| \bigl(|Z_{j+1,2k-1}^{1,(l)} - Z_{j+1,2k-1}^{2,(l)}| + |Z_{j+1,2k}^{1,(l)} - Z_{j+1,2k}^{2,(l)}| \bigr) \\
&\qquad + \left| \sum_{i \in \mathcal{I}_l} r_{j^*}(X_{i-1}) l_p'(\varepsilon_i) - r_{j^*}(\xi_{i-1}) l_p'(\eta_i) \right|,
\end{aligned}
$$

where

$$
Z_{j,k}^{1,(l)} = \sum_{i \in \mathcal{I}_l} I(X_{i-1} \in I_{j,k}) l_p'(\varepsilon_i), \qquad Z_{j,k}^{2,(l)} = \sum_{i \in \mathcal{I}_l} I(\xi_{i-1} \in I_{j,k}) l_q'(\eta_i).
$$

While the approximation-theoretic calculations are rather straightforward, the strong approximation result will require a lengthy proof based on Skorokhod embedding techniques. Let $\mathcal{I}_n = \{(j,k) : 0 \le j \le j^*, k = 1, \dots, 2^j\}$.

LEMMA 5.1. *Suppose that assumptions* (A1)–(A3) *are satisfied. Then there exists a construction of the random variables* $X_0, \dots, X_n$ *and* $(Y_1, \xi_1), \dots, (Y_n, \xi_n)$ *on a common probability space such that, for* $1 \le l \le K_n$,

$$
\begin{aligned}
&\inf_{f_0 \in \Sigma} \operatorname{ess\,sup} P_{f_0} \bigl(|Z_{j,k}^{1,(l)} - Z_{j,k}^{2,(l)}| \le C_\lambda (m_l 2^{-j})^{1/4} \log m_l, \forall (j,k) \in \mathcal{I}_n | \mathcal{F}_l \bigr) \\
&\qquad = 1 - O(m_l^{-\lambda}).
\end{aligned}
$$

To formulate the next theorem, we define

$$
S_f^{3,(l)} = \sum_{i \in \mathcal{I}_l} (f - f_0)(t_{n,i}) l'(\eta_i).
$$

THEOREM 5.2. *Suppose that asumptions* (A1)–(A3) *are satisfied. Let* $\lambda > 1$ *be a constant. Then there are versions of the random variables* $X_0, \dots, X_n$ *and* $Y_1, \dots, Y_n$ *on a common probability space such that, for* $1 \le l \le K_n$,

$$\sup_{f_0 \in \Sigma} \sup_{f \in \Sigma_{f_0}^n} \operatorname{ess\,sup} P_{f_0}\big(|S_f^{1,(l)} - S_f^{3,(l)}| > c(\lambda) r_n | \mathcal{F}_l'\big) = O(m_l^{-\lambda}),$$

where $r_n = (\gamma_n)^{1/4}(\gamma_n')^{3/4} m_l^{1/4} \log m_l + m_l^{-\lambda}$ *and* $c(\lambda)$ *is a constant depending only on* λ.

The assertion of this theorem is a consequence of the following lemma. Set

$$Z_{j,k}^{3,(l)} = \sum_{i \in \mathcal{I}_l} I(t_{n,i} \in I_{j,k}) l'(\eta_i).$$

LEMMA 5.2. *Suppose that assumptions* (A1)–(A3) *are satisfied. Then there exists a construction of the random variables* $X_0, \dots, X_n$ *and* $Y_{n,1}, \dots, Y_{n,n}$ *on a common probability space such that, for* $1 \le l \le K_n$,

$$\inf_{f_0 \in \Sigma} \operatorname{ess\,sup} P_{f_0}\big(|Z_{j,k}^{1,(l)} - Z_{j,k}^{3,(l)}| \le C_\lambda (m_l 2^{-j})^{1/4} \log m_l, \forall (j,k) \in \mathcal{I}_n | \mathcal{F}_l'\big)$$
$$= 1 - O(m_l^{-\lambda}).$$

The proofs of Lemmas 5.1 and 5.2 make use of a multiscale version of the Skorokhod embedding and are similar to the construction in [22]. We postpone these proofs to Section 5.2. Now we shall give proofs of Theorems 5.1 and 5.2.

5.1. *Proofs of Theorems* 5.1 *and* 5.2. As already indicated, the proofs of the theorems split into an approximation-theoretic and a stochastic part. The following lemma contains the approximation-theoretic facts needed for the proofs of Theorems 5.1 and 5.2.

LEMMA 5.3. *Let* $c_0(g)$ *and* $c_{j,k}(g)$ *be the Haar coefficients of a function* g *defined above. Then*:

(i) $|c_0(g)| \le (B-A)^{1/2} \|g\|_\infty$,
(ii) $|c_{j,k}(g)| \le \min\{(B-A)^{1/2} 2^{-j/2} \|g\|_\infty (B-A)^{3/2} 2^{-3j/2-2} \|g'\|_\infty\}$,
(iii) $\|g - (c_0(g)h_0 + \sum_{j=0}^{j^*} \sum_{k=1}^{2^j} c_{j,k}(g) h_{j,k})\|_\infty \le (B-A) 2^{-j^*-2} \|g'\|_\infty$.

PROOF. Assertion (i) follows from

$$|c_0(g)| \le \|g\|_\infty \int |h_0(t)|\, dt \le (B-A)^{1/2} \|g\|_\infty.$$

Analogously, we obtain that

$$|c_{j,k}(g)| \le \|g\|_\infty \int |h_{j,k}(t)|\, dt \le (B-A)^{1/2} 2^{-j/2} \|g\|_\infty.$$

Furthermore, it follows that

$$\begin{aligned}|c_{j,k}(g)| &\le (B-A)^{-1/2}2^{j/2}\left|\int_A^B g(t)\big(\mathbf{1}_{I_{j+1,2k-1}}(t)-\mathbf{1}_{I_{j+1,2k}}(t)\big)\,dt\right|\\ &\le (B-A)^{-1/2}2^{j/2}\int_{I_{j+1,2k-1}}\big|g(t)-g\big(t+(B-A)2^{-j-1}\big)\big|\,dt\\ &\le (B-A)^{3/2}2^{-3j/2-2}\|g'\|_\infty,\end{aligned}$$

which yields (ii).

Finally, we obtain from (ii) that

$$\begin{aligned}&\left|g(x)-\left(c_0(g)h_0(x)+\sum_{j=0}^{j^*}\sum_{k=1}^{2^j}c_{j,k}(g)h_{j,k}(x)\right)\right|\\ &\qquad\le\sum_{j=j^*+1}^{\infty}\sum_{k=1}^{2^j}|c_{j,k}(g)h_{j,k}(x)|\\ &\qquad\le\sum_{j=j^*+1}^{\infty}(B-A)^{3/2}2^{-3j/2-2}\|g'\|_\infty(B-A)^{-1/2}2^{j/2}\\ &\qquad=(B-A)2^{-j^*-2}\|g'\|_\infty.\end{aligned}$$ □

Now we are in a position to prove Theorems 5.1 and 5.2.

PROOF OF THEOREM 5.1. Define

$$S_{f,j^*}^{1,(l)}=\sum_{i\in\mathcal{I}_l}\left[c_0(g)h_0(X_{i-1})+\sum_{j=0}^{j^*}\sum_{k=1}^{2^j}c_{j,k}(g)h_{j,k}(X_{i-1})\right]l_p'(\varepsilon_i),$$

$$S_{f,j^*}^{2,(l)}=\sum_{i\in\mathcal{I}_l}\left[c_0(g)h_0(\xi_i)+\sum_{j=0}^{j^*}\sum_{k=1}^{2^j}c_{j,k}(g)h_{j,k}(\xi_i)\right]l_q'(\eta_i)$$

and

$$R_{f,j^*}^{i,(l)}=S_f^{i,(l)}-S_{f,j^*}^{i,(l)},\qquad i=1,2.$$

Define the event

$$D_l:=\{|Z_{j,k}^{1,(l)}-Z_{j,k}^{2,(l)}|\le C_\lambda(m_l2^{-j})^{1/4}\log m_l,\forall(j,k)\in\mathcal{I}_n\},$$

where C_λ is a constant. By Lemma 5.1, $P_{f_0}(\overline{D_l}|\mathcal{F}_l)=O(m_l^{-\lambda})$ with some choice of C_λ. By (i) and (ii) of Lemma 5.3, on the set D_l it holds that

$$\begin{aligned}
&|S^{1,(l)}_{f,j^*} - S^{2,(l)}_{f,j^*}| \\
&\quad \le (B-A)^{-1/2}\Biggl(|c_0(g)||Z^{1,(l)}_{0,1} - Z^{2,(l)}_{0,1}| \\
&\qquad\qquad + \sum_{j=0}^{j^*} 2^{j/2} \sum_{k=1}^{2^j} |c_{j,k}(g)|\bigl(|Z^{1,(l)}_{j+1,2k-1} - Z^{2,(l)}_{j+1,2k-1}| \\
&\qquad\qquad\qquad + |Z^{1,(l)}_{j+1,2k} - Z^{2,(l)}_{j+1,2k}|\bigr)\Biggr) \\
&\quad \le C_\lambda \|g\|_\infty m_l^{1/4} \log m_l \\
&\qquad + C_\lambda \sum_{j=0}^{j^*} 2^{3j/2} \min\{2^{-j/2}\|g\|_\infty, (B-A)2^{-3j/2-2}\|g'\|_\infty\} \\
&\qquad\qquad \times (m_l 2^{-j})^{1/4} \log m_l \\
&\quad = O\bigl((\|g\|_\infty + \|g\|_\infty^{1/4}\|g'\|_\infty^{3/4}) m_l^{1/4} \log m_l\bigr).
\end{aligned}$$

The latter proves that, with some constant $c(\lambda)$ depending on λ,

$$P_{f_0}\bigl(|S^{1,(l)}_{f,j^*} - S^{2,(l)}_{f,j^*}| > c(\lambda) r'_n | \mathcal{F}_l\bigr) = O(m_l^{-\lambda}), \tag{32}$$

where $r'_n = (\gamma_n)^{1/4}(\gamma'_n)^{3/4} m_l^{1/4} \log m_l$. By (iii) of Lemma 5.3 it holds that

$$\begin{aligned}
P_{f_0}\bigl(|R^{1,(l)}_{f,j^*}| \ge m_l^{-\lambda}\bigr) &\le m_l^{2\lambda} E_{f_0}(R^{1,(l)}_{f,j^*})^2 \\
&\le m_l^{2\lambda}(B-A)2^{-j^*-2}\|g'\|_\infty \sum_{i\in\mathcal{I}_l} E_{f_0}(l'_p(\varepsilon_i))^2.
\end{aligned} \tag{33}$$

Choosing the finest level $j^*(n) = c^* \log m_l$, with some c^* large enough, we obtain that

$$P_{f_0}\bigl(|R^{1,(l)}_{f,j^*}| \ge m_l^{-\lambda}\bigr) = O(m_l^{-\lambda}). \tag{34}$$

Since the above bounds are uniform in $f_0 \in \Sigma$, from (32)–(34) and a similar bound for $R^{2,(l)}_{f,j^*}$ we conclude the assertion. □

Theorem 5.2 can be proved in a similar way.

5.2. *Proofs of Lemmas* 5.1 *and* 5.2. We prove Lemma 5.1 only for $l = 1$, since the proof for $l > 1$ is completely analogous. The proof of Lemma 5.2 then requires only some obvious modifications and therefore will not be described here. To simplify notation we drop the index l in the following, that is, we write $Z^1_{j,k}$, $Z^2_{j,k}$, m

instead of $Z_{j,k}^{1,(l)}$, $Z_{j,k}^{2,(l)}$, m_l, respectively.

PROOF OF LEMMA 5.1. Conditional on X_0 (which represents the information contained in $\mathcal{F}_0$), we construct a pairing of $X_1, \ldots, X_m$ and $(Y_1, \xi_1), \ldots, (Y_m, \xi_m)$ such that

$$\inf_{f_0 \in \Sigma} \operatorname{ess\,sup} P_{f_0}\big(|Z_{j,k}^1 - Z_{j,k}^2| \le C_\lambda (m2^{-j})^{1/4} \log m, (j,k) \in \mathcal{I}_n | X_0\big)$$
$$= 1 - O(m^{-\lambda})$$

is satisfied. In the following, all estimates are to be understood to hold uniformly in $f_0 \in \Sigma$.

The pairing of the random variables of both models is organized by a simultaneous Skorokhod embedding of $Z_{j,k}^1$ and $Z_{j,k}^2$ in a common set of Wiener processes $W_{j,k}$ assigned to the intervals $I_{j,k}$. We describe this embedding in detail for the autoregressive process. The embedding of $l_q(\eta_i)$ from the regression model is completely analogous and will be briefly mentioned only. Then we draw conclusions for the rate of approximation of $Z_{j,k}^1$ by $Z_{j,k}^2$, which will conclude the proof. An embedding scheme like this has already been developed in [22], in a different context. In view of some modifications and since we intend to provide a self-contained paper, we give a full proof of this lemma.

Let $W_{j,k}$, $(j,k) \in \mathcal{I}_n$, be independent Wiener processes. Apart from the coarsest resolution scale which corresponds to $j = 0$, we use each of these processes only on a finite time interval $[0, T_{j,k}]$, where the particular (nonrandom) values of the $T_{j,k}$ will be specified in part (iv) below. For the time being it is only important to know that $T_{0,k} = \infty$.

(i) *Embedding of $l_p(\varepsilon_1)$ and construction of X_1.*

First we define $l_p(\varepsilon_1)$ by a Skorokhod embedding in the Wiener processes mentioned above. Since $l_p(\varepsilon_1)$ does not necessarily define X_1 uniquely, we have to use perhaps an additional randomization to get X_1.

Let k_1 be that random number with $X_0 \in I_{j^*,k_1}$. Now we are going to represent $l_p(\varepsilon_1)$ by increments of the Wiener processes, preferably by those of W_{j^*,k_1}. However, since we want to use W_{j^*,k_1} up to some prespecified time T_{j^*,k_1} only, it might happen that this is not enough for representing $l_p(\varepsilon_1)$. In this case we additionally use a certain stretch of the process $W_{j^*-1,[k_1/2]}$, and so on. The Wiener processes which are potentially used for the representation of $l_p(\varepsilon_1)$ correspond to a containment relation of the dyadic intervals,

$$I_{j^*,k} \subseteq I_{j^*-1,[k/2]} \subseteq \cdots \subseteq I_{0,[k2^{-j^*}]},$$

where $[a]$ denotes the largest integer not greater than a. This means that we represent $l_p(\varepsilon_1)$ by the following Wiener process:

$$W^{(1)}(s) = \begin{cases} W_{j^*,k_1}(s), & \text{if} \quad 0 \le s \le T_{j^*,k_1}, \\ W_{j^*,k_1}(T_{j^*,k_1}) + \cdots + W_{j+1,[k_1 2^{j+1-j^*}]}(T_{j+1,[k_1 2^{j+1-j^*}]}) & \\ \qquad + W_{j,[k_1 2^{j-j^*}]}\left(s - \sum_{l=j+1}^{j^*} T_{l,[k_1 2^{l-j^*}]}\right), & \\ \qquad\qquad \text{if } T_{j^*,k_1} + \cdots + T_{j+1,[k_1 2^{j+1-j^*}]} & \\ \qquad\qquad\qquad < s \le T_{j^*,k_1} + \cdots + T_{j,[k_1 2^{j-j^*}]}. & \end{cases}$$

($W^{(1)}$ is indeed a Wiener process on $[0, \infty)$, since $T_{0,k} = \infty$.)

According to Lemma A.2 of [17], there exists a stopping time $\tau^{(1)}$ such that the distribution of $W^{(1)}(\tau^{(1)})$ is equal to the conditional distribution of $l_p(\varepsilon_1)$ given X_0. We define ε_1 in such a way that

$$l_p(\varepsilon_1) = W^{(1)}(\tau^{(1)}).$$

[This is achieved by first setting $l_p(\varepsilon_1)$ equal to $W^{(1)}(\tau^{(1)})$ and then defining ε_1 with the aid of an additional randomization according to its conditional distribution given $l_p(\varepsilon_1)$.] Finally, according to the model equation under f_0, we set $X_1 = f_0(X_0) + \varepsilon_1$.

To explain the following steps in a formally correct way, we introduce stopping times $\tau_{j,k}^{(i)}$, $i = 1, \dots, m$, assigned to the corresponding Wiener process $W_{j,k}$. Define

$$\tau_{j,k}^{(0)} = 0, \qquad (j,k) \in \mathcal{I}_n.$$

To get $\tau_{j,k}^{(1)}$, we redefine all those $\tau_{j,k}^{(0)}$ which are assigned to Wiener processes $W_{j,k}$ that were used for representing $l_p(\varepsilon_1)$. According to the above construction we set

$$\tau_{j^*,k_1}^{(1)} = \tau^{(1)} \wedge T_{j^*,k_1}.$$

We redefine further

$$\tau_{j,[k_1 2^{j-j^*}]}^{(1)} = \begin{cases} [\tau^{(1)} - T_{j^*,k_1} - \cdots - T_{j+1,[k_1 2^{j+1-j^*}]}] \wedge T_{j,[k_1 2^{j-j^*}]}, & \\ \qquad \text{if } T_{j^*,k_1} + \cdots + T_{j+1,[k_1 2^{j+1-j^*}]} < \tau^{(1)}, & \\ 0, \qquad \text{otherwise}. & \end{cases}$$

The remaining stopping times $\tau_{j,l}^{(1)}$ with $l \ne [k_1 2^{j-j^*}]$ keep their preceding values $\tau_{j,l}^{(0)} = 0$.

This procedure will be successively repeated for all other ε_i's with the modification that we use only those parts of the Wiener processes which are still untouched by the previous construction steps.

(ii) *Embedding of $l_p(\varepsilon_i)$ and construction of X_i.*

Assume that $X_0, \dots, X_{i-1}$ are already defined. Let k_i be that random number with $X_{i-1} \in I_{j^*,k_i}$. Now we represent $l_p(\varepsilon_i)$ by parts of W_{j^*,k_i}, $W_{j^*-1,[k_i/2]}, \dots,$ $W_{0,[2^{-j^*}]}$, which have not been used so far.

First note that, because of the strong Markov property, these remaining increments $W_{j,[k_i 2^{j-j^*}]}(s + \tau^{(i-1)}_{j,[k_i 2^{j-j^*}]}) - W_{j,[k_i 2^{j-j^*}]}(\tau^{(i-1)}_{j,[k_i 2^{j-j^*}]})$ form independent Wiener processes, also independent of $X_0, \dots, X_{i-1}$. Hence, gluing these parts together we obtain a Wiener process on $[0, \infty)$ which is independent of $X_0, \dots, X_{i-1}$. This process is given as

$$W^{(i)}(s) = \begin{cases} W_{j^*,k_i}\bigl(s + \tau^{(i-1)}_{j^*,k_i}\bigr) - W_{j^*,k_i}\bigl(\tau^{(i-1)}_{j^*,k_i}\bigr), \\ \qquad \text{if } 0 \le s \le T_{j^*,k_i} - \tau^{(i-1)}_{j^*,k_i}, \\ \bigl\{W_{j^*,k_i}\bigl(T_{j^*,k_i}\bigr) - W_{j^*,k_i}\bigl(\tau^{(i-1)}_{j^*,k_i}\bigr)\bigr\} + \cdots \\ \qquad + \bigl\{W_{j+1,[k_i 2^{j+1-j^*}]}\bigl(T_{j+1,[k_i 2^{j+1-j^*}]}\bigr) \\ \qquad\quad - W_{j+1,[k_i 2^{j+1-j^*}]}\bigl(\tau^{(i-1)}_{j+1,[k_i 2^{j+1-j^*}]}\bigr)\bigr\} \\ \qquad + \Biggl\{W_{j,[k_i 2^{j-j^*}]}\Biggl(s - \sum_{l=j+1}^{j^*} \bigl(T_{l,[k_i 2^{l-j^*}]} - \tau^{(i-1)}_{l,[k_i 2^{j-l}]}\bigr) \\ \qquad\qquad\qquad\qquad\qquad\qquad + \tau^{(i-1)}_{j,[k_i 2^{j-j^*}]}\Biggr) \\ \qquad\qquad\qquad\qquad - W_{j,[k_i 2^{j-j^*}]}\bigl(\tau^{(i-1)}_{j,[k_i 2^{j-j^*}]}\bigr)\Biggr\}, \\ \qquad \text{if } \displaystyle\sum_{l=j+1}^{j^*} \bigl(T_{l,[k_i 2^{j-l}]} - \tau^{(i-1)}_{l,[k_i 2^{j-l}]}\bigr) < s \\ \qquad\qquad \displaystyle\le \sum_{l=j}^{j*} \bigl(T_{l,[k_i 2^{j-l}]} - \tau^{(i-1)}_{l,[k_i 2^{j-l}]}\bigr). \end{cases}$$

There exists a stopping time $\tau^{(i)}$ such that $W^{(i)}(\tau^{(i)})$ has the same distribution as the conditional distribution of $l_p(\varepsilon_i)$. We define ε_i in such a way that

$$l_p(\varepsilon_i) = W^{(i)}(\tau^{(i)}),$$

and set $X_i = f_0(X_{i-1}) + \varepsilon_i$. [The definition of ε_i is again achieved in two steps by first setting $l_p(\varepsilon_i)$ equal to $W^{(i)}(\tau^{(i)})$ and then defining ε_i according to its conditional distribution.]

To complete this construction, it remains to define the stopping times $\tau^{(i)}_{j,k}$. These stopping times indicate up to which point the Wiener processes have been used in the first i steps. Accordingly we set

$$\tau^{(i)}_{j,[k_i 2^{j-j^*}]} = \begin{cases} \bigl[\tau^{(i-1)}_{j,[k_i 2^{j-j^*}]} + \bigl(\tau^{(i)} - \bigl(T_{j^*,k_i} - \tau^{(i-1)}_{j^*,k_i}\bigr) - \cdots \\ \qquad\qquad - \bigl(T_{j+1,[k_i 2^{j+1-j^*}]} - \tau^{(i-1)}_{j+1,[k_i 2^{j+1-j^*}]}\bigr)\bigr)\bigr] \\ \quad \wedge T_{j,[k_i 2^{j-j^*}]}, \\ \qquad\qquad \text{if } \bigl(T_{j^*,k_i} - \tau^{(i-1)}_{j^*,k_i}\bigr) + \cdots \\ \qquad\qquad\qquad + \bigl(T_{j+1,[k_i 2^{j+1-j^*}]} \\ \qquad\qquad\qquad\quad - \tau^{(i-1)}_{j+1,[k_i 2^{j+1-j^*}]}\bigr) < \tau^{(i)}, \\ \tau^{(i-1)}_{j,[k_i 2^{j-j^*}]}, \qquad \text{otherwise.} \end{cases}$$

For all (j,l) with $l \neq [k_i 2^{j-j^*}]$ we define

$$\tau^{(i)}_{j,l} = \tau^{(i-1)}_{j,l}.$$

After embedding $l_p(\varepsilon_1), \dots, l_p(\varepsilon_m)$ we arrive at stopping times $\tau^{(m)}_{j,k}$. The partial sums are connected to the Wiener processes by the relation

$$\begin{aligned} Z^1_{j,k} = {} & \sum_{(u,v)\colon I_{u,v} \subseteq I_{j,k}} W_{u,v}\bigl(\tau^{(m)}_{u,v}\bigr) \\ & + \sum_{i\colon 1 \le i \le m, X_{i-1} \in I_{j,k}} \sum_{(u,v)\colon I_{u,v} \supset I_{j,k}} W_{u,v}\bigl(\tau^{(i)}_{u,v}\bigr) - W_{u,v}\bigl(\tau^{(i-1)}_{u,v}\bigr). \end{aligned} \tag{35}$$

(iii) *Embedding of* $l_q(\eta_1), \dots, l_q(\eta_m)$ *and construction of* $(Y_1, \xi_1), \dots, (Y_m, \xi_m)$.

This will be done in complete analogy to the construction described above. We define again stopping times $\tilde{\tau}^{(i)}_{j,k}$ and obtain the following representation of the partial sums:

$$\begin{aligned} Z^2_{j,k} = {} & \sum_{(u,v)\colon I_{u,v} \subseteq I_{j,k}} W_{u,v}\bigl(\tilde{\tau}^{(m)}_{u,v}\bigr) \\ & + \sum_{i\colon 1 \le i \le m, Y_i \in I_{j,k}} \sum_{(u,v)\colon I_{u,v} \supset I_{j,k}} W_{u,v}\bigl(\tilde{\tau}^{(i)}_{u,v}\bigr) - W_{u,v}\bigl(\tilde{\tau}^{(i-1)}_{u,v}\bigr). \end{aligned} \tag{36}$$

(iv) *Choice of the values for* $T_{j,k}$.

To motivate our particular choice of the $T_{j,k}$ described below, we consider first two extreme cases. If $T_{j^*,k} = \infty$ for all k, then $Z^1_{j^*,k}$ and $Z^2_{j^*,k}$ are both completely represented by $W_{j^*,k}$. This leads indeed to a satisfactorily close connection of $Z^1_{j^*,k}$ and $Z^2_{j^*,k}$. On the other hand, this choice is unfavorable at scales $j \ll j^*$. Although we get immediately the upper estimate

$$|Z^1_{j,k} - Z^2_{j,k}| \le \sum_{l\colon I_{j^*,l} \subseteq I_{j,k}} |Z^1_{j^*,l} - Z^2_{j^*,l}|,$$

the difference between $Z_{j,k}^1$ and $Z_{j,k}^2$ will be unnecessarily large. This is because, for $j \ll j^*$, $Z_{j,k}^1$ and $Z_{j,k}^2$ are then represented by too many different stretches of the Wiener processes $W_{j^*,l}$ with $I_{j^*,l} \subseteq I_{j,k}$.

On the other hand, if the $T_{j^*,k}$ are rather small, then $Z_{j^*,k}^1$ and $Z_{j^*,k}^2$ will be represented in large parts by stretches of Wiener processes $W_{u,v}$ which correspond to intervals $I_{u,v} \supset I_{j^*,k}$ with $j < j^*$. Once we are on a coarser scale $j < j^*$, we cannot guarantee that $Z_{j^*,k}^1$ and $Z_{j^*,k}^2$ are (mostly) generated by identical parts of the Wiener processes. Consequently, we would also get a suboptimal connection, this time for $Z_{j^*,k}^1$ and $Z_{j^*,k}^2$.

To find a good compromise between these two conflicting aims, we choose the $T_{j,k}$ as large as possible, however, with the additional property that, for $j \neq 0$, the stretches $[0, T_{j,k}]$ are used up with a high probability in the representation of both $l_p(\varepsilon_1), \dots, l_p(\varepsilon_m)$ and $l_q(\eta_1), \dots, l_q(\eta_m)$. Strictly speaking, we choose the $T_{j,k}$ in such a way that

$$\begin{aligned}
&\operatorname{ess\,sup} P_{f_0}\Biggl(\sum_{i=1}^{m} \tau^{(i)} I(X_{i-1} \in I_{j,k}) < \sum_{(u,v)\colon I_{u,v} \subseteq I_{j,k}} T_{u,v}, \\
&\qquad\qquad \forall (j,k) \in \mathcal{I}_n \setminus \{(0,k)\} \,\Big|\, X_0\Biggr) = O(m^{-\lambda})
\end{aligned} \tag{37}$$

and

$$\begin{aligned}
&P_{f_0}\Biggl(\sum_{i=1}^{m} \widetilde{\tau}^{(i)} I(Y_i \in I_{j,k}) < \sum_{(u,v)\colon I_{u,v} \subseteq I_{j,k}} T_{u,v}, \forall (j,k) \in \mathcal{I}_n \setminus \{(0,k)\}\Biggr) \\
&\quad = O(m^{-\lambda}).
\end{aligned} \tag{38}$$

To this end, we study first the stochastic behavior of the above sums of stopping times assigned to the interval $I_{j,k}$.

Recall that the innovations ε_i are assumed to be independent. According to the construction of the Skorokhod embedding described in [17], Appendix A.1, the randomness of $\tau^{(i)}$ is driven by some $U_i \sim \mathit{Uniform}[0, 1]$ from a sequence of independent random variables and by $\{W^{(i)}(s), 0 \le s \le \tau^{(i)}\}$. The vectors (X_{i-1}, U_i) are of course also ϕ-mixing as the X_i. Since, for $i \neq i'$, $\{W^{(i)}(s), 0 \le s \le \tau^{(i)}\}$ and $\{W^{(i')}(s), 0 \le s \le \tau^{(i')}\}$ are composed of disjoint stretches of the Wiener processes $W_{j,k}$ separated by stopping times, the random variables $\tau^{(i)} I(X_{i-1} \in I_{j,k})$ inherit the ϕ-mixing property from the process $\{X_i\}$. Hence, we obtain by a Bernstein-type inequality for sums of ϕ-mixing random variables (see, e.g., [10]) that

$$\begin{aligned}
&\operatorname{ess\,sup} P_{f_0}\Biggl(\Biggl|\sum_{i=1}^{m} \{\tau^{(i)} I(X_{i-1} \in I_{j,k}) - E[\tau^{(i)} I(X_{i-1} \in I_{j,k})]\}\Biggr| \\
&\qquad\qquad > C_\lambda \sqrt{m 2^{-j}} \log m \,\Big|\, X_0\Biggr) = O(m^{-\lambda})
\end{aligned} \tag{39}$$

and, analogously,

$$P_{f_0}\left(\left|\sum_{i=1}^{m}\{\widetilde{\tau}^{(i)}I(Y_i \in I_{j,k}) - E[\widetilde{\tau}^{(i)}I(Y_i \in I_{j,k})]\}\right| > C_\lambda\sqrt{m2^{-j}}\log m\right) \tag{40}$$
$$= O(m^{-\lambda}).$$

Define

$$S_{j,k} = \sum_{i=1}^{m} E\tau^{(i)}I(X_{i-1} \in I_{j,k}) - C_\lambda\sqrt{m2^{-j}}\log m.$$

Furthermore, we define

$$T_{j,k} = S_{j,k} - \sum_{(u,v)\,:\, I_{u,v}\subset I_{j,k}} S_{u,v}.$$

Then $S_{j,k} = \sum_{(u,v)\,:\, I_{u,v}\subset I_{j,k}} T_{u,v}$. By (39) and (40) we obtain (37) and (38).

(v) *Conclusion for* $Z^1_{j,k} - Z^2_{j,k}$.

By (35)–(38) we obtain that

$$Z^1_{j,k} = \sum_{(u,v)\,:\, I_{u,v}\subseteq I_{j,k}} W_{u,v}(T_{u,v}) \tag{41}$$
$$+ \sum_{i\,:\, 1\le i\le m, X_{i-1}\in I_{j,k}}\ \sum_{(u,v)\,:\, I_{u,v}\supset I_{j,k}} W_{u,v}(\tau^{(i)}_{u,v}) - W_{u,v}(\tau^{(i-1)}_{u,v})$$

and

$$Z^2_{j,k} = \sum_{(u,v)\,:\, I_{u,v}\subseteq I_{j,k}} W_{u,v}(T_{u,v}) \tag{42}$$
$$+ \sum_{i\,:\, 1\le i\le m, Y_i\in I_{j,k}}\ \sum_{(u,v)\,:\, I_{u,v}\supset I_{j,k}} W_{u,v}(\widetilde{\tau}^{(i)}_{u,v}) - W_{u,v}(\widetilde{\tau}^{(i-1)}_{u,v})$$

are satisfied with a probability exceeding $1 - O(m^{-\lambda})$. At this point we see why our particular pairing of the random variables provides a close connection between $Z^1_{j,k}$ and $Z^2_{j,k}$: most of the randomness of $Z^1_{j,k}$ and $Z^2_{j,k}$ is contained in the first terms on the right-hand sides of (41) and (42), respectively. These terms are random, but identical to each other.

To analyze the difference between the right-hand sides of (41) and (42), we compose the pieces $\{W_{u,v}(s), \tau^{(i-1)}_{u,v} \le s \le \tau^{(i)}_{u,v}\}$ and $\{W_{u,v}(s), \widetilde{\tau}^{(i-1)}_{u,v} \le s \le \widetilde{\tau}^{(i)}_{u,v}\}$

corresponding to intervals $I_{u,v} \supset I_{j,k}$ to Wiener processes. For fixed i, we define

$$W_{j,k}^{\mathrm{res},i}(s) = \begin{cases} W_{j-1,[k/2]}\bigl(s + \tau_{j-1,[k/2]}^{(i-1)}\bigr) - W_{j-1,[k/2]}\bigl(\tau_{j-1,[k/2]}^{(i-1)}\bigr), \\ \qquad \text{if } 0 \le s \le \tau_{j-1,[k/2]}^{(i)} - \tau_{j-1,[k/2]}^{(i-1)}, \\ \bigl[W_{j-1,[k/2]}\bigl(\tau_{j-1,[k/2]}^{(i)}\bigr) - W_{j-1,[k/2]}\bigl(\tau_{j-1,[k/2]}^{(i-1)}\bigr)\bigr] + \cdots \\ \qquad + \bigl[W_{l+1,[k2^{l+1-j}]}\bigl(\tau_{l+1,[k2^{l+1-j}]}^{(i)}\bigr) \\ \qquad\qquad - W_{l+1,[k2^{l+1-j}]}\bigl(\tau_{l+1,[k2^{l+1-j}]}^{(i-1)}\bigr)\bigr] \\ \qquad + \bigl[W_{l,[k2^{l-j}]}(u) - W_{l,[k2^{l-j}]}\bigl(\tau_{l,[k2^{l-j}]}^{(i-1)}\bigr)\bigr], \\ \qquad \text{if } s = \bigl(\tau_{j-1,[k/2]}^{(i)} - \tau_{j-1,[k/2]}^{(i-1)}\bigr) + \cdots \\ \qquad\qquad\qquad + \bigl(\tau_{l+1,[k2^{l+1-j}]}^{(i)} - \tau_{l+1,[k2^{l+1-j}]}^{(i-1)}\bigr) + \bigl(u - \tau_{l,[k2^{l-j}]}^{(i-1)}\bigr) \\ \qquad \text{with } \tau_{l,[k2^{l-j}]}^{(i-1)} < u \le \tau_{l,[k2^{l-j}]}^{(i)}. \end{cases}$$

It is clear that $W_{j,k}^{\mathrm{res},i}$ is a Wiener process on the interval $[0, \tau_{j,k}^{\mathrm{res},i}]$, where $\tau_{j,k}^{\mathrm{res},i} = \sum_{(u,v)\,:\, I_{j,k} \subset I_{u,v}} (\tau_{u,v}^{(i)} - \tau_{u,v}^{(i-1)})$.

By the strong Markov property, the remaining parts of the Wiener processes $W_{j,k}$ again form independent Wiener processes, also independent of $\{W_{j,k}^{\mathrm{res},i}, 0 \le s \le \tau_{j,k}^{\mathrm{res},i}\}$. Therefore, we can compose all these latter parts to one Wiener process by setting

$$W_{j,k}^{\mathrm{res}}(s) = \begin{cases} W_{j,k}^{\mathrm{res},1}(s), & \text{if } 0 \le s \le \tau_{j,k}^{\mathrm{res},1}, \\ W_{j,k}^{\mathrm{res},1}(\tau_{j,k}^{\mathrm{res},1}) + \cdots + W_{j,k}^{\mathrm{res},u-1}(\tau_{j,k}^{\mathrm{res},u-1}) \\ \qquad + W_{j,k}^{\mathrm{res},u}(s - \tau_{j,k}^{\mathrm{res},1} - \cdots - \tau_{j,k}^{\mathrm{res},u-1}), \\ \qquad\qquad \text{if } \tau_{j,k}^{\mathrm{res},1} + \cdots + \tau_{j,k}^{\mathrm{res},u-1} < s \le \tau_{j,k}^{\mathrm{res},1} + \cdots + \tau_{j,k}^{\mathrm{res},u}. \end{cases}$$

An analogous construction can be made for the $\tilde{\tau}_{u,v}^{(i)}$, leading to Wiener processes $\widetilde{W}_{j,k}^{\mathrm{res}}$.

If both $\sum_{i=1}^m \tau^{(i)} I(X_{i-1} \in I_{j,k}) \ge S_{j,k}$ and $\sum_{i=1}^m \tilde{\tau}^{(i)} I(Y_i \in I_{j,k}) \ge S_{j,k}$ are satisfied, then

$$\begin{aligned} &\sum_{i\,:\,1 \le i \le m, X_{i-1} \in I_{j,k}} \; \sum_{(u,v)\,:\, I_{j,k} \subset I_{u,v}} W_{u,v}(\tau_{u,v}^{(i)}) - W_{u,v}(\tau_{u,v}^{(i-1)}) \\ &\qquad = W_{j,k}^{\mathrm{res}}\Biggl(\sum_{i\,:\,X_{i-1} \in I_{j,k}} \tau^{(i)} - S_{j,k}\Biggr) \end{aligned}$$

and

$$\begin{aligned} &\sum_{i\,:\,1 \le i \le m, Y_i \in I_{j,k}} \; \sum_{(u,v)\,:\, I_{j,k} \subset I_{u,v}} W_{u,v}(\tilde{\tau}_{u,v}^{(i)}) - W_{u,v}(\tilde{\tau}_{u,v}^{(i-1)}) \\ &\qquad = \widetilde{W}_{j,k}^{\mathrm{res}}\Biggl(\sum_{i\,:\,X_{i-1} \in I_{j,k}} \tilde{\tau}^{(i)} - S_{j,k}\Biggr). \end{aligned}$$

Hence, we obtain by (37), (38) and Lemma 1.2.1 in [6], page 29, that, for all $(j,k) \in \mathcal{I}_n$,

$$\begin{aligned}
&\operatorname{ess\,sup} P_{f_0}(|Z^1_{j,k} - Z^2_{j,k}| > r''_n | X_0) \\
&\quad \le \operatorname{ess\,sup} P_{f_0}\Biggl(\Biggl| W^{\mathrm{res}}_{j,k}\Biggl(\sum_{i\,:\,X_{i-1}\in I_{j,k}} \tau^{(i)} - S_{j,k}\Biggr)\Biggr| > r''_n/2 \Bigg| X_0\Biggr) \\
&\qquad + \operatorname{ess\,sup} P_{f_0}\Biggl(\Biggl| \widetilde{W}^{\mathrm{res}}_{j,k}\Biggl(\sum_{i\,:\,X_{i-1}\in I_{j,k}} \widetilde{\tau}^{(i)} - S_{j,k}\Biggr)\Biggr| > r''_n/2 \Bigg| X_0\Biggr) \\
&\qquad + O(m^{-\lambda}) = O(m^{-\lambda}),
\end{aligned}$$

where $r''_n = C_\lambda (m2^{-j})^{1/4} \log m$. This completes the proof. □

6. Some auxiliary results.

6.1. *Convergence of stationary distributions.*

LEMMA 6.1. *Suppose that $(X^f_i)_{i\ge0}$ and $(X^{f_0}_i)_{i\ge0}$ are stationary processes obeying (4) with autoregression functions f and f_0, respectively, where $|f|$, $|f_0| \le M$. Assume that the innovations $(\varepsilon_i)_{i\ge1}$ are i.i.d. with a density p such that*

$$\rho = \tfrac{1}{2} \sup_{-M\le x_1\le x_2\le M} \int_{-\infty}^{\infty} |p(x-x_1) - p(x-x_2)|\,dx < 1.$$

Then, for the stationary densities ψ_f and ψ_{f_0}, it holds that

$$\begin{aligned}
&\int_{-\infty}^{\infty} \bigl(\sqrt{\psi_f(x)} - \sqrt{\psi_{f_0}(x)}\bigr)^2\,dx \\
&\qquad \le \frac{1}{1-\rho} \sup_{u\in[0,\|f-f_0\|_\infty]} \int_{-\infty}^{\infty} |p(x) - p(x-u)|\,dx.
\end{aligned}$$

PROOF. We denote by $p^f(x|y) = p(x - f(y))$ and $p^{f_0}(x|y) = p(x - f_0(y))$ the transition densities of the processes $(X^f_i)_{i\ge0}$ and $(X^{f_0}_i)_{i\ge0}$, respectively. It holds that

$$\int_{-\infty}^{\infty} \bigl(\sqrt{\psi_f(x)} - \sqrt{\psi_{f_0}(x)}\bigr)^2\,dx \le \int_{-\infty}^{\infty} |\psi_f(x) - \psi_{f_0}(x)|\,dx.$$

Let, for brevity, $\Psi_{f,f_0}(x) = \psi_f(x) - \psi_{f_0}(x)$. From

$$\Psi_{f,f_0}(x) = \int [p^f(x|y) - p^{f_0}(x|y)]\psi_f(y)\,dy + \int p^{f_0}(x|y)\Psi_{f,f_0}(y)\,dy$$

we deduce that

$$\begin{aligned}
&\int_{-\infty}^{\infty} |\Psi_{f,f_0}(x)|\,dx \\
&\quad \le \int\left[\int |p^f(x|y) - p^{f_0}(x|y)|\psi_f(y)\,dy\right]dx \\
&\qquad + \int\left|\int p^{f_0}(x|y)[\Psi_{f,f_0}(y)]_+\,dy - \int p^{f_0}(x|y)[\Psi_{f,f_0}(y)]_-\,dy\right|dx \\
&\quad \le \int\left[\int |p^f(x|y) - p^{f_0}(x|y)|\,dx\right]\psi_f(y)\,dy \\
&\qquad + \sup_{y_1,y_2}\int |p^{f_0}(x|y_1) - p^{f_0}(x|y_2)|\,dx \int [\Psi_{f,f_0}(y)]_+\,dy \\
&\quad \le \sup_y \int |p^f(x|y) - p^{f_0}(x|y)|\,dx \\
&\qquad + \sup_{y_1,y_2}\int |p^{f_0}(x|y_1) - p^{f_0}(x|y_2)|\,dx \int [\Psi_{f,f_0}(y)]_+\,dy.
\end{aligned}$$

The latter implies

$$\begin{aligned}
\int_{-\infty}^{\infty} |\Psi_{f,f_0}(x)|\,dx \le{}& \sup_{0\le u\le \|f-f_0\|_\infty} \int |p(x) - p(x-u)|\,dx \\
&+ \tfrac{1}{2}\sup_{y_1,y_2}\int |p^{f_0}(x|y_1) - p^{f_0}(x|y_2)|\,dx \int |\Psi_{f,f_0}(x)|\,dy.
\end{aligned}$$

Rearranging the terms we obtain the assertion. □

6.2. *An analogue of Berbee's lemma.*

DEFINITION 6.1. The uniform ϕ-mixing coefficient between r.v.'s ξ and η is defined to be the number

$$\phi(\xi,\eta) = \sup\{|P(A) - P(A|B)| : A \in \sigma(\xi), B \in \sigma(\eta), P(B) \ne 0\}.$$

LEMMA 6.2. *Suppose that ξ and η are two random variables with values in $\mathbb{R}^1$ and $\mathbb{R}^d$, respectively, given on the probability space $(\Omega, \mathcal{F}, P)$. Furthermore, we assume that ξ and η possess a joint density and that the probability space is rich enough for the definition of a random variable $\Delta \sim$ Uniform$[0,1]$ which is independent of ξ and η. Then we can construct a random variable $\widetilde{\xi} = \widetilde{\xi}(\xi,\eta,\Delta)$ such that*:

(i) $\mathcal{L}(\widetilde{\xi}|\eta) = \mathcal{L}(\xi)$ *a.s., that is, $\widetilde{\xi}$ is independent of η and has the same distribution as ξ,*

(ii) $P(\widetilde{\xi} \ne \xi|\eta) \le \phi(\xi,\eta)$ *a.s.*

PROOF. The idea of the proof is of course closely related to that of the proof of Theorem 2 in [1]. However, since the formulation of our result differs slightly from theirs (they constructed $\widetilde{\xi}$ in such a way that it is close to ξ with a high probability, whereas we are interested in an exact coincidence of $\widetilde{\xi}$ and ξ) we decided not to omit this proof.

We denote by $p_\xi(\cdot)$ the marginal density of ξ and by $p_{\xi|\eta}(\cdot|y)$ the conditional density of ξ given $\eta = y$. Define

$$\phi_y = \tfrac{1}{2} \int \bigl|p_\xi(x) - p_{\xi|\eta}(x|y)\bigr|\, dx = 1 - \int p_\xi(x) \wedge p_{\xi|\eta}(x|y)\, dx.$$

Then $\phi_\eta \le \phi(\xi, \eta)$ a.s.

If $\phi_\eta = 0$, then $p_\xi(\cdot)$ and $p_{\xi|\eta}(\cdot \mid y)$ coincide and we set $\widetilde{\xi} \equiv \xi$. Otherwise we proceed as follows. With a random variable $\Delta \sim$ *Uniform*[0, 1] which is independent of ξ and η, we set

$$\widetilde{\xi} = \widetilde{\xi}(\xi, \eta, \Delta) = \begin{cases} \xi, & \text{if } p_\xi(\xi) \ge \Delta p_{\xi|\eta}(\xi|\eta), \\ \overline{\xi}, & \text{otherwise,} \end{cases}$$

where $\overline{\xi}$ is an appropriate random variable having the density $[p_\xi(\cdot) - p_\xi(\cdot) \wedge p_{\xi|\eta}(\cdot|\eta)]/\phi_\eta$. The random variable $\overline{\xi}$ is defined via a quantile transform as $G_\eta^{-1}(\frac{\Delta p_{\xi|\eta}(\xi|\eta) - p_\xi(\xi)}{p_{\xi|\eta}(\xi|\eta) - p_\xi(\xi)})$, where

$$G_\eta(y) = \frac{1}{\phi_\eta} \int_{-\infty}^{y} [p_\xi(x) - p_{\xi|\eta}(x|\eta)]_+ \, dx.$$

Now we have

$$\begin{aligned} P(\widetilde{\xi} = \xi|\eta) &= P\bigl(p_{\xi|\eta}(\xi|\eta) \wedge p_\xi(\xi) \ge \Delta p_{\xi|\eta}(\xi|\eta)|\eta\bigr) \\ &= E\biggl(\frac{p_{\xi|\eta}(\xi|\eta) \wedge p_\xi(\xi)}{p_{\xi|\eta}(\xi \mid \eta)} I\bigl(p_{\xi|\eta}(\xi|\eta) > 0\bigr)\Big|\eta\biggr) \\ &= \int p_{\xi|\eta}(x|\eta) \wedge p_\xi(x)\, dx = 1 - \phi_\eta, \end{aligned}$$

which implies (ii). Part (i) follows from the construction. □

6.3. *An exponential inequality.* We made use of the following inequality whose proof can be found in [15].

LEMMA 6.3. *Let ξ be a r.v. such that $E\xi = 0$ and $|\xi| \le a$, for some positive constant a. Then*

$$E \exp(\lambda\xi) \le \exp(c\lambda^2 E\xi^2), \qquad |\lambda| \le 1,$$

where $c = e^a/2$.

Acknowledgments. We thank the referees and an Associate Editor for helpful comments on the paper.

REFERENCES

[1] Berkes, I. and Philipp, W. (1979). Approximation theorems for independent and weakly dependent random vectors. *Ann. Probab.* **7** 29–54. MR0515811

[2] Brown, L. D., Cai, T., Low, M. G. and Zhang, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30** 688–707. MR1922538

[3] Brown, L. D., Carter, A. V., Low, M. G. and Zhang, C.-H. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.* **32** 2074–2097. MR2102503

[4] Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958

[5] Carter, A. V. (2002). Deficiency distance between multinomial and multivariate normal experiments. *Ann. Statist.* **30** 708–730. MR1922539

[6] Csörgő, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics*. Academic Press, New York. MR0666546

[7] Dalalyan, A. and Reiss, M. (2006). Asymptotic statistical equivalence for scalar ergodic diffusions. *Probab. Theory Related Fields* **134** 248–282. MR2222384

[8] Delattre, S. and Hoffmann, M. (2002). Asymptotic equivalence for a null recurrent diffusion. *Bernoulli* **8** 139–174. MR1895888

[9] Donoho, D. L. (1994). Asymptotic minimax risk for sup-norm loss: Solution via optimal recovery. *Probab. Theory Related Fields* **99** 145–170. MR1278880

[10] Doukhan, P. (1994). *Mixing*: *Properties and Examples*. *Lecture Notes in Statist.* **85**. Springer, New York. MR1312160

[11] Drees, H. (2001). Minimax risk bounds in extreme value theory. *Ann. Statist.* **29** 266–294. MR1833966

[12] Genon-Catalot, V., Laredo, C. and Nussbaum, M. (2002). Asymptotic equivalence of estimating a Poisson intensity and a positive diffusion drift. *Ann. Statist.* **30** 731–753. MR1922540

[13] Golubev, G. K. and Nussbaum, M. (1990). A risk bound in Sobolev class regression. *Ann. Statist.* **18** 758–778. MR1056335

[14] Grama, I. G. and Nussbaum, M. (1998). Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Related Fields* **111** 167–214. MR1633574

[15] Grama, I. G. and Nussbaum, M. (2002). Asymptotic equivalence for nonparametric regression. *Math. Methods Statist.* **11** 1–36. MR1900972

[16] Grama, I. G. and Nussbaum, M. (2002). A functional Hungarian construction for sums of independent random variables. *Ann. Inst. H. Poincaré Probab. Statist.* **38** 923–957. MR1955345

[17] Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York. MR0624435

[18] Korostelev, A. and Nussbaum, M. (1995). Density estimation in the uniform norm and white noise approximation. Preprint No. 154, Weierstrass Institute, Berlin.

[19] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York. MR0856411

[20] Le Cam, L. and Yang, G. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. Springer, New York. MR1784901

[21] MILSTEIN, G. and NUSSBAUM, M. (1998). Diffusion approximation for nonparametric autoregression. *Probab. Theory Related Fields* **112** 535–543. MR1664703

[22] NEUMANN, M. H. and KREISS, J.-P. (1998). Regression-type inference in nonparametric autoregression. *Ann. Statist.* **26** 1570–1613. MR1647701

[23] NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and white noise. *Ann. Statist.* **24** 2399–2430. MR1425959

[24] NUSSBAUM, M. (1999). Minimax risk: Pinsker bound. *Encyclopedia of Statistical Sciences* **Update 3** 451–460. Wiley, New York.

[25] NUSSBAUM, M. and KLEMELÄ, J. (1998). Constructive asymptotic equivalence of density estimation and Gaussian white noise. Discussion Paper No. 53, Sonderforschungsbereich 373, Humboldt Univ., Berlin.

[26] ROBINSON, P. M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.* **4** 185–207. MR0732897

[27] STRASSER, H. (1985). *Mathematical Theory of Statistics*. de Gruyter, Berlin. MR0812467

UNIVERSITÉ DE BRETAGNE-SUD
CENTRE DE RECHERCHE YVES COPPENS
CAMPUS DE TOHANNIC
56000 VANNES
FRANCE
E-MAIL: ion.grama@univ-ubs.fr

INSTITUT FÜR STOCHASTIK
FRIEDRICH-SCHILLER-UNIVERSITÄT JENA
ERNST-ABBE-PLATZ 2
07743 JENA
GERMANY
E-MAIL: mneumann@mathematik.uni-jena.de