Nearest Neighbor Conditional Estimation for Harris Recurrent Markov Chains

Alessio Sancetta*

Faculty of Economics, University of Cambridge, UK

July 12, 2007

Abstract

This paper is concerned with consistent nearest neighbor time series estimation for data generated by a Harris recurrent Markov chain. The goal is to validate nearest neighbor estimation in this general time series context, using simple and weak conditions. The framework considered covers, in a unified manner, a wide variety of statistical quantities, e.g. autoregression function, conditional quantiles, conditional tail estimators and, more generally, extremum estimators. The focus is theoretical, but examples are given to highlight applications.

Key Words: Nonparametric Estimation, Quantile Estimation, Semiparametric Estimation, Sequential Forecasting, Tail Estimation, Time Series.

1 Introduction

This paper is concerned with conditional nonparametric and semiparametric estimation from data generated by a stochastic process that can be represented as a Harris Recurrent Markov Chain (HRMC). The class of HRMC is quite general and includes

*Address for correspondence: Alessio Sancetta, Faculty of Economics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DE, UK. E-mail: alessio.sancetta@econ.cam.ac.uk.

processes that may not be stationary (e.g. univariate random walks). The basic interest of the paper is to consider a process $X = (X_i)_{i \in \mathbb{N}}$ with values in some set $E \subseteq \mathbb{R}^K$ ($K \geq 1$) and some measurable function f on E and to estimate $\mathbb{E}_{i-1} f(X_i)$ ($\mathbb{E}_{i-1}$ is expectation conditional on the sigma algebra generated by $(X_s)_{s<i}$) or some related quantity like $\inf_f \mathbb{E}_{i-1} f(X_i)$ over some class of functions from which we can derive conditional extremum estimators. Most common examples include conditional distribution function estimation ($f(x) = I\{x \leq y\}$, $y \in E$), regression estimation ($f(x) = x$) and, as just mentioned, conditional extremum estimators. The goal is not to derive new estimators, but to validate, in a unified manner, the application of nearest neighbor estimation to time series problems. Nevertheless, some of the applications consider estimators that might be new. The advantage of nearest neighbor estimators over kernel estimators is that they are usually more stable, as they automatically adapt to regions where there is sparsity of data.

Assuming the HRMC condition, the goal is to state simple general conditions that would imply consistency for nonparametric and/or semiparametric estimation, avoiding mixing conditions. When dealing with real data, it is often difficult to check mixing and/or dependence conditions. When the hypotesized data generating process (DGP) is available, computation of mixing conditions is difficult (Doukhan, 1994) and for this reason some new weak dependence coefficients are used (e.g. Doukhan and Louhichi, 1999, and Ango-Nze and Doukhan, 2004 for applications to econometrics). However, there are many weak dependence coefficients, and the choice of one condition among many may require ad hoc arguments for each different model. On the other hand, we may suppose that the data come from a given class of stochastic processes, but no other information is available. We may not even know if the process is stationary. The natural question to ask is the following: is it possible to identify a broad class of stochastic processes in which many econometric and statistical models can be embedded and such that nonparametric estimation is still consistent? This question has been positively answered by Yakowitz (1993), where a slightly more general class of stochastic processes than HRMC has been considered, but attention is limited to autoregression function estimation. Karlsen and Tjostheim (2001) slightly restricted the class of stochastic

processes, but considered more general nonparametric estimation problems. Karlsen and Tjostheim (2001) studied nonparametric kernel estimation, while Yakowitz (1993) used a nearest radii approach, also used here and to be described in due course. The nearest radii approach considerably simplifies the argument. Markov chains (MC) and in particular HRMC have also been considered as an important case of DGP around which to develop empirical methods for inference (e.g. Horowitz, 2003, Bertail and Clémençon, 2006).

Unlike Karlsen and Tjostheim (2001), the present paper is only concerned with consistency and weak conditions required to assure it. Inferential arguments in conditional nonparametric estimation have been carefully handled by Karlsen and Tjostheim (2001). Restricting our interest on consistency only, the conditions used here are particularly simple. Unlike Yakowitz (1993), this paper is not restricted to autoregression function estimation, but more general nonparametric and semiparametric procedures are studied. The main idea is to be able to consistently estimate the conditional distribution function. This allows us to derive consistency for a large number of nonparametric and semiparametric problems imposing mild smoothness conditions on the transition distribution only. Several examples will be given to highlight the number of possible applications of interest to the empirical researcher. In this respect, the class of problems considered includes extremum estimators, hence, it is more general than some of the problems considered by Karlsen and Tjostheim (2001).

The present paper complements the previous ones in an effort to provide nonparametric estimators for time series without the need to impose dependence conditions beyond the null recurrence hypothesis, hence allowing for nonstationary time series. It is remarkable that the proofs remain simple and do not require complicated technical conditions.

Note that mixing conditions are commonly used in the nonparametric literature (e.g. Robinson, 1983, for an early reference, see also the monograph of Pagan and Ullah, 1999), though recently, more general weak dependence conditions have also been employed (Ango-Nze et al., 2002). The present discussion does not only differ from previous work because mixing and weak dependence conditions are not used (and

stationarity is not always necessary), but, as already mentioned, because the class of problems is more general than the regression problem usually studied in the literature.

Section 2 discusses the nearest neighbor procedure and states minimal conditions under which the nonparametric estimator of the conditional distribution function is consistent. This result is then used to show consistency in a variety of cases with an illustrative example of conditional tail estimation and one of optimal sequential forecasting of conditional quantiles. Section 3 informally overviews issues of applied nature like neighbors' selection and dimensionality reduction. Its purpose is to provide suggestions for future research in nonparametric time series. Proofs of results can be found in the Appendix. Next we just mention a few models that can be embedded in HRMC.

1.1 Many Important Econometric and Statistical Models are HRMC

Recall that an MC is a discrete time process such that, conditioning on the present, the future and the past are independent. Then, an HRMC, say X, with state space E is an irreducible MC such that

$$\Pr\left(X_n \in B \ i.o. | X_0 = x\right) = 1, \ x \in E \ (i.o. \text{ stands for infinitely often})$$

for any set B of positive ψ measure, where ψ is some suitable sigma finite measure (e.g. Meyen and Tweedie, 1993, for details).

By suitable definition of the state space E, it is possible to embed many econometric and statistical models in the class of HRMC, under suitable restrictions (e.g. non-explosive coefficients). Linear autoregressive, SETAR, multilinear, and ARCH models, all fall within the class of HRMC. Many examples can be obtained by considering the class of models that can be embedded in the following multivariate stochastic difference equation

$$X_n = A_n X_{n-1} + B_n, \tag{1}$$

where $(A_n)_{n\in\mathbb{N}}$ and $(B_n)_{n\in\mathbb{N}}$ are iid matrix and vector valued random variables (Babillot et al., 1997, for details on recurrence and references)

Example 1 *Consider the ARCH(2) model*

$$\begin{aligned} Y_n &= Z_n\sigma_n, \\ \sigma_n^2 &= \alpha_0 + \alpha_1 Y_{n-1}^2 + \alpha_2 Y_{n-2}^2, \end{aligned}$$

where $(Z_n)_{n\in\mathbb{N}}$ *is a sequence of independent identically distributed (iid) random variables. Then,* Y_n^2 *admits the representation*

$$\begin{pmatrix} Y_n^2 \\ Y_{n-1}^2 \end{pmatrix} = \begin{pmatrix} Z_n^2\alpha_1 & Z_n^2\alpha_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} Y_{n-1}^2 \\ Y_{n-2}^2 \end{pmatrix} + \begin{pmatrix} \alpha_0 Z_n^2 \\ 0 \end{pmatrix}.$$

If Z_n *is symmetric around zero, there is a one to one mapping between* Y_n *and* Y_n^2*. This is in the form of (1).*

The above example extends to any finite order. As will become obvious, we need the HRMC to be observable, i.e. we must observe X_n at time n.

Example 2 *Consider the GARCH(1,1) model*

$$\begin{aligned} Y_n &= Z_n\sigma_n, \\ \sigma_n^2 &= \alpha_0 + \alpha_1 Y_{n-1}^2 + \beta\sigma_{n-1}^2, \end{aligned}$$

where $(Z_n)_{n\in\mathbb{N}}$ *is a sequence of iid random variables. Then,* Y_n^2 *admits the representation*

$$\begin{pmatrix} Y_n^2 \\ \sigma_n^2 \end{pmatrix} = \begin{pmatrix} Z_n^2\alpha_1 & Z_n^2\beta \\ \alpha_1 & \beta \end{pmatrix} \begin{pmatrix} Y_{n-1}^2 \\ \sigma_{n-1}^2 \end{pmatrix} + \begin{pmatrix} \alpha_0 Z_n^2 \\ \alpha_0 \end{pmatrix}.$$

In this case, $X_n = (Y_n^2, \sigma_n^2)'$*, but* σ_n^2 *is not observable.*

In the case of GARCH, the Markov chain contains the unobservable component σ_n^2 that can be approximated by a finite MA process in $(Y_n^2)_{n\in\mathbb{N}}$ if we assume the invertibility condition $\beta < 1$. Then, we could deal with this problem by method of sieves allowing the state space to increase with time. This could be done imposing suitable stationarity conditions. However, for general HRMC this is not possible, as this class includes null recurrent chains that are not stationary. Hence, in this general context, some important estimation procedures will not be discussed unless they can

be dealt within the unifying framework of the paper. This is restrictive, nevertheless, estimation for many models is accounted for. As just mentioned, the class of HRMC also includes models that do not possess a stationary distribution (e.g. the univariate random walk model commonly used in econometrics). Further details on examples can be found in Meyn and Tweedie (1993, ch.2).

As noted by Karlsen and Tjostheim (2001), the availability of large data-sets (e.g. high frequency financial data) makes the use of nonparametric methods for non-stationary time series a possibility. Hence, while not always efficient, we can even hope for successful nonparametric estimation for nonstationary time series. In these cases, the nonparametric approach could be used for preliminary data analysis and data exploration or as a preliminary stage for adaptive estimation. Going back to the GARCH example, there has been considerable interest in realized volatility estimation (e.g. Barndorff-Nielsen and Shephard, 2002). Nonparametric methods could be used to forecast volatility once an estimate of realized volatility is available. This would be a fully nonparametric alternative to parametric GARCH.

2 Conditional Estimation using Nearest Neighbors

Let $X = (X_n)_{n\in\mathbb{N}}$ be an aperiodic HRMC on a state space $\left(E \subseteq \mathbb{R}^K, \mathcal{E}\right)$ with transition probability $P(x, A)$ and invariant measure π. The Markovian probability with initial value x is denoted by P_x. We shall use linear functional notation, as commonly done in the MC literature, e.g. for some suitable function f, $Pf(x) := \int_E f(y) P(x, dy)$ and for some set $B \subset E$, $Pf(B) := \int_B \int_E f(y) P(x, dy) [\pi(dx)/\pi(B)]$ (and the use of this notation will not require further explanation). Note that if $\pi(E) < \infty$ the HRMC is said to be positive recurrent, while null recurrent if $\pi(E) = \infty$. Null recurrent MC do not possess stationary distribution. At first, we shall be concerned with estimation of

$$P(x, \{y \in E : y \leq s\}) = \Pr(X_n \leq s | X_{n-1} = x),$$

where for $K > 1$ the inequality is meant elementwise and the meaning of this notation will be assumed throughout without reminder. By relatively standard results, consis-

tent estimation of the transition distribution allow us to derive in a unified manner a wide variety of estimators which are discussed in the sequel.

For simplicity, but with abuse of notation, we shall write $P(s|x)$ as a short cut for $P(x, \{x : x \leq s\})$, the conditional distribution function.

2.1 The Estimator

We shall follow Yakowitz (1993). Denote by $m \to \infty$ the number of neighbors. The estimator is derived in terms of the recurrence times of X to some conditioning set $B(x, r_m) \to \{x\}$ as $r_m \to 0$, which is a ball of d-radius r_m (d usually being topologically equivalent to the Euclidean distance, and irrelevant to the development of the paper). To ease notation, we shall use B_m, $B_m(x)$ and $B(x, r_m)$ interchangeably, whichever is felt more appropriate. For any set $B \subseteq E$, define $T_B := \inf\{n > 0 : X_n \in B\}$ and $T_B(i) := \inf\{n > T_B(i-1) : X_n \in B\}$, $T_B(1) := T_B$, i.e. $T_B(i)$ is the time of the i^{th} visit to B. Hence,

$$\hat{P}_m(s|B_m) := \hat{P}_m(B_m, \{y \in E : y \leq s\}) = \frac{1}{m}\sum_{i=1}^{m} I\{X(T_{B_m}(i)+1) \leq s\} \qquad (2)$$

is an m nearest neighbor estimator for the one step ahead conditional distribution ($X(i) = X_i$ for typographical reasons). The same linear functional notation used for P will also be used for $\hat{P}_m$, e.g. $\hat{P}_m g(B_m) = \int_E g(y)\hat{P}_m(B_m(x), dy)$.

Note that by the Harris recurrence assumption, $T_B(i) < +\infty$ a.s. for each i. This means that as $n \to \infty$ we shall be able to allow $m \to \infty$ so that the estimation error goes to zero. However, for consistency, we shall also require $B(x, r_m) \to \{x\}$ so that the bias is vanishing (i.e. the conditioning set needs to shrink as the sample size increases). To this end, we shall first fix a sequence $r_m \to 0$ as $m \to \infty$. This means that fixed a radius r_m, we shall wait for the m visit to $B_m(x)$ in order to construct $\hat{P}_m$, which is an m nearest neighbor estimator. By Harris recurrence, this will happen a.s. in finite time for any m.

Let $L(n)$ be a slowly varying function of n at infinity (e.g. Bingham et al., 1987). If we assume X to be β-recurrent (using the terminology in Karlsen and Tjostheim, 2001), then, by Theorem 2.1 in Chen (1999), $\sum_{i=1}^{n} f(X_i) \asymp n^{\beta}L(n)$ in probability,

$\beta \in [0,1]$, for any non-negative π integrable f such that $\pi f > 0$. (Note that Chen, 1999, calls this MC regular and expresses the condition in terms of recurrent times of D-sets: using results about atoms and small functions, the two definitions are equivalent, e.g. Chen, 1999.) Clearly, $\beta = 1$ is the positive recurrent case. It is well known (e.g. Chen, 1999, Karlsen and Tjostheim, 2001) that a random walk is recurrent of index $\beta = 1/2$. Hence, if we knew β, we would know that $n^{\beta}/m_n \to \infty$ is necessary. (When $\beta = 1$, we recover the familiar necessary condition for consistency on the m neighbors.) Mutatis mutandis, this is the approach of Karlsen and Tjostheim (2001), where in practice a lower bound for β has to be estimated, though the formal approach requires the use of Nummelin splitting technique (e.g. Meyn and Tweedie, 1993) and considerable technicalities. Note that in Karlsen and Tjostheim (2001) the badnwidth is a function of β. Here, no assumption of regularity is made so that the estimator can be constructed only using the predetermined sequence of sets $B(x, r_m)$. Noting that $\pi(B(x, r_m)) < \infty$ because π is sigma finite, under the assumption of β recurrence in Karlsen and Tjostheim (2001), we could use Theorem 2.1 in Chen (1999) and impose conditions directly on the neighbors, without worrying about the choice of the radius r_m.

2.2 Consistency of the Conditional Empirical Distribution Function

The conditions used for consistency of the conditional empirical distribution are formally listed below. Further conditions might be required in the applications and these will be stated when needed.

Condition 3 *$X := (X_n)_{n\in\mathbb{N}}$ is an aperiodic Harris recurrent Markov chain on a state space $\left(E \subseteq \mathbb{R}^K, \mathcal{E}\right)$ with transition probability $P(x, A)$ and invariant measure π. The sigma algebra $\mathcal{E}$ is countably generated.*

Remark 4 *It is possible to allow for a more general state space and some details are given in the Appendix. If $E = \mathbb{R}^K$ equipped with its usual metric, $\mathcal{E}$ is countably generated.*

Condition 5 $\Pr(X_1 \leq s | X_0 = x)$ *is a.s. continuous in* $x \in E$ *for any* $s \in E$.

Remark 6 *If continuity does not hold, the results are still true for* π*-almost all* x.

Condition 7 $r_m \to 0$ *and* $m \to \infty$.

Remark 8 *By Condition 3, Condition 7 is always feasible.*

Theorem 9 *Under Conditions 3, 5 and 7,*

$$\sup_{s \in E} \left| \hat{P}_m(s | B_m(x)) - P(s|x) \right| \xrightarrow{a.s.} 0.$$

We now use this result to consider more interesting problems.

2.3 Estimation of Conditional Minimum Estimators

The following set up is a bit abstract. The reader mainly interested in examples might skim through the remaining of this section and look at Section 3 to get a feeling of the possible applications.

Consider the following problem

$$\inf_{f \in \mathfrak{F}} Pf(x)$$

where $\mathfrak{F}$ is some set of functions (and recall that $Pf(x)$ is the expectation of $f(X_n)$ conditioning on $X_{n-1} = x$). Suppose $f(y) = f_\theta(y)$ is convex in $\theta \in \Theta$ for some suitable set Θ. Then, the above problem can be seen as an abstract version of the more common problem of minimizing the risk $Pf_\theta(x)$ with respect to θ. Solution of this problem allows us to define population values for many statistical estimators.

Example 10 *Suppose* $f_\theta(x) = |x - \theta|^2$ *and* $x \in E \subseteq \mathbb{R}$. *Then,*

$$\arg \inf_{\theta \in \Theta} Pf_\theta(x) = \mathbb{E}(X_n | X_{n-1} = x),$$

i.e. the expectation of X_n *conditioning on* $X_{n-1} = x$.

Example 11 *Suppose* $f_\theta(x) = |x - \theta|$ *and* $x \in E \subseteq \mathbb{R}$. *Then,*

$$\arg\inf_{\theta\in\Theta} Pf_\theta(x) = M(X_n|x),$$

which denotes the median of X_n *conditioning on* $X_{n-1} = x$.

Example 12 *Suppose* $f_\theta(x) = u|x-\theta|^+ + (1-u)|x-\theta|^-$ *and* $x \in E \subseteq \mathbb{R}$, $u \in (0,1)$. *Then,*

$$\arg\inf_{\theta\in\Theta} Pf_\theta(x) = Q_u(X_n|x),$$

which denotes the u *quantile of* X_n *conditioning on* $X_{n-1} = x$.

For a general treatment of the problem, it is simpler to define minimization with respect to $f \in \mathfrak{F}$ rather than $\theta \in \Theta$. Examples will be given in due course.

We shall use standard concepts like integrability under the measure induced by the kernel P fixed at x.

Definition 13 *The measure induced by the transition kernel* P *at fixed* $x \in E$ *will be denoted by* π_x*:*

$$\pi_x(A) := P(x, A).$$

Note that π_x should not be confused with P_x, e.g. $\pi_x(A) = \Pr(X_n \in A|X_{n-1} = x)$, while $P_x(X_n \in A) = \Pr(X_n \in A|X_0 = x)$.

We also introduce the following definition.

We need to restrict the class of functions $\mathfrak{F}$ to be considered.

Condition 14 *For any* $x \in C \subseteq E$, *the following holds:*
i. $\mathfrak{F}$ *has envelope function* $F(x) := \sup_{f\in\mathfrak{F}}|f(x)|$ *such that* $\limsup_m PF^p(B_m(x)) < \infty$ *for some* $p > 1$;
ii. $\mathfrak{F}$ *is any family of* π_x*-a.s. equicontinuous functions on* E.

Remark 15 *We may have* $C = E$. *However, in some applications we may just want to consider* $C = \{x\}$, *i.e. a singleton or some other subset of* E.

Remark 16 *A family of equicontinous functions contains functions that are not necessarily Lipschitz for a given metric, e.g. any finite set of continuous functions. Moreover, we can allow for more general families of functions, possibly discontinuous. To limit the notational burden in the text, we do not pursue this generalization here, but detail can be found in the appendix.*

Corollary 17 *Under Conditions 3, 5, 7 and 14,*

$$\sup_{f\in\mathfrak{F}}\left|\hat{P}f\left(B_m\left(x\right)\right)-Pf\left(x\right)\right|\xrightarrow{a.s.}0,$$

for any $x\in C$.

Remark 18 *This result is a generalization of Theorem 2 in Yakowitz (1993), where, mutatis mutandis,* $p>2$ *is required. Moment conditions higher than 2 are also used for consistency in Theorem 5.2 of Karlsen and Tjostheim (2001), though their results are not directly comparable because they use a different nonparametric estimator. Note that these authors do not consider the uniform in* $\mathfrak{F}$ *case.*

The above result can be used to derive conditional extremum estimators. Define

$$\hat{f}_m\left(x\right):=\arg\inf_{f\in\mathfrak{F}}\hat{P}_mf\left(B_m\left(x\right)\right)\text{ and }f_0\left(x\right):=\arg\inf_{f\in\mathfrak{F}}Pf\left(x\right),$$

so that f_0 is the unfeasible optimal choice of $f\in\mathfrak{F}$ (i.e. unknown), while $\hat{f}$ is the feasible estimator. Then, under an additional identifiability condition, we have that $\hat{f}$ and f are close to each other for each fixed x. To formalize this we need the following additional condition, which is minimal.

Condition 19 *For any* $x\in C\subseteq E$, *let* $G=G_x$ *be any arbitrary open set that contains* $f_0\left(x\right)$ *and let* G^c *be its complement. Then,*

$$\inf_{f\in G^c}Pf\left(x\right)>Pf_0\left(x\right).$$

Corollary 20 *Suppose* $(\mathfrak{F},\rho)$ *is a metric space. Under Conditions 3, 5, 7, 14, and 19,*

$$\rho\left(\hat{f}_m\left(x\right),f_0\left(x\right)\right)\xrightarrow{p}0,$$

for any $x\in C$.

2.4 Sequential Forecasting

We now consider sequential forecasting. Define

$$\hat{f}_{m,n} := \hat{f}_m\left(X_{n-1}\right) \text{ and } f_n := f\left(X_{n-1}\right),$$

so that f_n is the unfeasible $\mathcal{F}_{n-1}$ measurable optimal choice of $f \in \mathfrak{F}$, while $\hat{f}_{m,n}$ is the feasible estimator. The goal is to strengthen Corollary 20 for the more general problem of sequential forecasting. A detailed example will be given in the next subsection in order to explain the abstract setup. We introduce a strengthening of Condition 14.

Condition 21 *Condition 14 holds with C such that for any $n \geq 1$ and any $\epsilon > 0$, $P^n(x, C) > 1-\epsilon$ (P^n is the n transition probability, e.g. $P^n(x, C) = \Pr(X_n \in C | X_0 = x)$).*

Remark 22 *Note that $P^n(x, E) \leq 1$ if $P(x, E) \leq 1$, which is the case by definition. Condition 21 might be helpful if Condition 14 does not hold for $C = E$ but still holds for some set of arbitrary smaller measure. Note that C is not required to be compact.*

Theorem 23 *Suppose $(\rho, \mathfrak{F})$ is a metric space and $\rho\left(\hat{f}_n, f_n\right)$ is P_x-uniformly integrable for any n. Under Conditions 3, 5, 7, 19, and 21,*

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_x \rho\left(\hat{f}_{m,n}, f_n\right) \to 0,$$

where $\mathbb{E}_x(X_n) = \mathbb{E}(X_n | X_0 = x)$, i.e. expectation w.r.t. P_x.

Theorem 23 says that the average loss incurred using the estimated forecast $\hat{f}_{m,n}$ is equivalent to the one incurred using the optimal unfeasible sequential forecast f_n.

To provide some understanding of the condition "$\rho\left(\hat{f}_{m,n}, f_n\right)$ is P_x-uniformly integrable for any n", suppose: $f_n := \mathbb{E}_{n-1} X_n$, X is a random walk with values in $\mathbb{R}$ and $\rho(x, y) = |x - y|$. Then, $\hat{f}_{m,n} := \sum_{i=1}^{m} X\left(T_B(i) + 1\right)/m$ where $B = B_m\left(X_{n-1}\right)$ and $\mathbb{E}_x\left|\hat{f}_{m,n} - f_n\right|^p < \infty$ under a p moment condition on the innovations of the random walk. Hence, $\rho\left(\hat{f}_{m,n}, f_n\right)$ is P_x-uniformly integrable for any n.

3 Applications

Nearest neighbor estimation is a standard statistical technique. We use this space to show some of its applications to problems covered by the previous results. These applications might not be that standard and are presented for illustrative purposes only. In particular, two applications are considered: conditional likelihood estimation and sequential forecasting of conditional quantiles.

3.1 Conditional Likelihood Estimation

Suppose that the transition kernel admits the following representation

$$P(x, A) = \int_A p(y; \theta(x)) \mu(dy),$$

where μ is a sigma finite measure and $\theta(x)$ is a function of x taking values in Θ . Then, $(p(y;\theta))_{\theta\in\Theta}$ is a model where $\theta = \theta(x)$ is unknown and we ignore a parametric form for $\theta(x)$. Hence the model $p(y;\theta(x))$ depends on the infinite dimensional parameter $\theta(x)$.

Example 24 *Suppose $X_n = \theta(X_{n-1}) Z_n$, where $(Z_n)_{n\in\mathbb{N}}$ is iid standard Gaussian noise and $\theta(X_{n-1})$ is a function of X_{n-1}. Then, $p(y;\theta(x)) = \phi(y/\theta(x))/\theta(x)$ denoting the standard Gaussian density by ϕ. This is a simple Markovian model for heteroskedastic data. If we are unable or unwilling to make a parametric assumption for $\theta(x)$, then, we could use nonparametric methods to estimate it. The conditionally Gaussian ARCH process of finite order is a special fully parametrized case of this model.*

In some models (notably the ones belonging to the exponential family), we also have that there is a function g such that

$$\theta(x) = \int_A g(y) p(y; \theta(x)) \mu(dy).$$

Example 25 *Suppose $p(y;\theta) = \exp\{\langle a(\theta), g(y)\rangle + b(\theta)\} c(y)$, for some positive functions a,b and c, where $\theta = \int g(y) p(y;\theta) d\mu(y)$. Clearly, a and g could be vector valued functions. This density is said to belong to the exponential family model, with natural parameter θ, canonical parameter $a(\theta)$ and canonical statistic $g(x)$. Properties of*

these models in relation to econometrics can be found in van Garderen (1997). The Gaussian, the Poisson and the Binomial distributions all belong to this family.

When $p(y;\theta(x))$ is the density kernel, it is natural to ask if nonparametric estimation can be used to consistently estimate $p(y;\theta(x))$ or $\theta(x)$. Clearly, the case

$$\theta(x) = Pg(x) = \int_A g(y)\, p(y;\theta(x))\, \mu(dy)$$

is dealt by Corollary 17. A general alternative to this method is to choose $\theta(x)$ to maximize

$$\mathbb{E}\left[\ln p(X_n;\theta)\,|X_{n-1}=x\right] \tag{3}$$

with respect to θ. Denoting the true unknown function to estimate by $\theta_0(x)$, the justification of (3) is the usual one via the scoring rule: under regularity conditions,

$$\begin{aligned}(\partial/\partial\theta)\,\mathbb{E}\left[\ln p(X_n;\theta)\,|X_{n-1}=x\right] &= \int_E \frac{(\partial p(y;\theta)/\partial\theta)}{p(y;\theta)} p(y;\theta_0(x))\,\mu(dy) \\ &= \int_E \left(\frac{\partial p(y;\theta_0(x))}{\partial\theta_0(x)}\right)\mu(dy) = 0\end{aligned}$$

if $\theta = \theta_0(x)$. Corollary 17 shows that, under regularity conditions,

$$\sup_{\theta\in\Theta}\left|\int_E \ln p(y;\theta)\, P_m(dy|B_m(x)) - \mathbb{E}\left[\ln p(X_n;\theta)\,|X_{n-1}=x\right]\right| \overset{a.s.}{\to} 0,$$

so that the semiparametric likelihood approach is consistent: this is just an application of Corollary 20. If Θ is compact or can be approximated a.s. by a compact set and $\ln p(y;\theta)$ is continuous in θ and in $L_p(\pi_x)$ for some $p > 1$ and for any x' in the neighborhood of x, and has a unique maximum, then Conditions 14 and 19 are satisfied and no further work is required for Corollary 17 and 20 to hold. For the sake of concreteness we give an example of semiparametric likelihood estimation in the case of tail estimation for extreme events.

3.1.1 Example: Estimating Tail Events for HRMC

Suppose that $E \subseteq \mathbb{R}$ and denote by $y_P := \sup\{y \in E : \Pr(X_n \le y|X_{n-1}=x) < 1\}$, i.e. the largest element in the support of P given x. Our goal is to find an estimator for

$$\Pr(X_n > z|X_{n-1}=x),$$

when z is very large. When z is quite large, the estimated survival function $1 - \hat{P}_m(z|B_m(x))$ can be a poor estimator of tail probabilities and clearly infeasible for events beyond the sample range. For this reason, we may use a semiparametric approach. We assume that the MC has peak over threshold function satisfying

$$\lim_{y \uparrow y_P} \sup_{0<s<y_P-y} \left| \Pr(X_n > y + s | X_n > y, X_{n-1} = x) - \left(1 + \alpha \frac{s}{\beta(y)}\right)^{-1/\alpha} \right| = 0 \quad (4)$$

for some positive function $\beta(s)$ and real α (Embrechts et al., 1997, ch.3, for details). In the iid case, this is the standard assumption that X_n is in the maximal domain of attraction of the generalized extreme value distribution and conditions can be used to assure that this is the case also in the dependent case (e.g. Leadbetter and Rootzen, 1988, Section 2). The difference here is that we are considering high levels conditioning on $X_{n-1} = x$ so that α and $\beta(y)$ may depend on $X_{n-1} = x$. When $y \uparrow y_P$ it is often the case that X_n and X_{n-1} are independent unless the kernel P exhibits tail dependence (e.g. Joe, 1997). The relevance of estimation of conditional tail events for time series as opposed to unconditional tail estimation is an empirical question that cannot be addressed here.

Using (4) we approximately have

$$\Pr(X_n > y + s | X_{n-1} = x) \simeq \Pr(X_n > y | X_{n-1} = x) \left(1 + \alpha \frac{s}{\beta(y)}\right)^{-1/\alpha} \quad (5)$$

for large fixed y. Using the fact that

$$\sup_{y \in E} \left| \hat{P}_m(y|B_m(x)) - \Pr(X_n \leq y | X_{n-1} = x) \right| \overset{a.s.}{\to} 0$$

by Theorem 9, and by the discussion about semiparametric conditional likelihood, we find α and β maximizing

$$-\ln \beta(y) - \int_{\{s>y\} \cap E} \left(1 + \frac{1}{\alpha}\right) \ln \left(1 + \alpha \frac{s-y}{\beta(y)}\right)^{-1/\alpha - 1} \hat{P}_m(ds|B_m(x)) \quad (6)$$

in place of

$$-\mathbb{E}\left[\ln \beta(y) + \left(1 + \frac{1}{\alpha}\right) \ln \left(1 + \alpha \frac{X_n - y}{\beta(y)}\right)^{-1/\alpha - 1} \middle| X_n > y, X_{n-1} = x \right],$$

as $\Pr(X_n > y | X_{n-1} = x)$ in (5) does not depend on α and $\beta(y)$. In the unconditional case, this procedure is standard (e.g. Embrechts et al., 1997, Ch.6) and for y large but such that $\hat{P}_m(y|B_m(x))$ can be reasonably estimated, we have

$$\Pr(X_n > y + s | X_{n-1} = x) \simeq \left[1 - \hat{P}_m(y|B_m(x))\right] \left(1 + \hat{\alpha}\frac{s}{\hat{\beta}(y)}\right)^{-1/\hat{\alpha}},$$

where $\hat{\alpha}$ and $\hat{\beta}(y)$ are the estimators from (6). The threshold level y is a crucial parameter to estimate and this problem is no different from the unconditional case: y should be large to minimize the bias in (4), but also small so that the estimation error is not too large (see Embrechts et al., 1997, for suggestions).

3.2 Sequential Forecasting of Conditional Quantiles

As an application of Theorem 23, we consider sequential forecasting of conditional quantiles of X. In order to avoid issues related to non-uniqueness of quantiles in high dimension, we assume that $E \subseteq \mathbb{R}$. This is just done for notational simplicity. If we required a larger state space to embed a higher order MC, the quantile of X_n would refer to the first entry in X_n, as all the other entries are past values for the original model. Hence, the conditional u quantile of X_n is given by

$$Q(u|x) := \inf_{s \in \mathbb{R}} \{\Pr(X_n \leq s | X_{n-1} = x) > u\}.$$

To apply the results of the previous subsections, we need to consider a loss function that once minimized gives the conditional population quantile. Hence, define $g(x) = (1-u)|x|^- + u|x|^+$ and $f_\theta(x) = g(x - \theta)$, so that f_θ is convex in θ. By Example 12,

$$\begin{aligned} Q(u|x) &= \inf_{\theta \in \mathbb{R}} \mathbb{E}\left[f_\theta(X_n) | X_{n-1} = x\right] \\ &= \inf_{\theta \in \mathbb{R}} P f_\theta(x), \end{aligned}$$

using compact notation. Therefore, the conditional quantile estimator is given by

$$\begin{aligned} \hat{Q}(u|B_m) &:= \inf_s \left\{\hat{P}(s|B_m) \geq u\right\} \\ &= \inf_{\theta \in \mathbb{R}} \hat{P}_m f_\theta(B_m(x)). \end{aligned} \tag{7}$$

(In practice, (7) is directly obtained from the order statistics, $X\left(T_B\left(i_1\right)+1\right) \leq .... \leq X\left(T_B\left(i_m\right)+1\right)$.) By convexity of g, an application of Jensen's inequality gives the following bound on quantile sequential forecasting

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{n-1}g\left(X_n-\hat{Q}\left(u|B_m\left(X_{n-1}\right)\right)\right) \leq \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{n-1}g\left(X_n-Q\left(u|X_{n-1}\right)\right)+error, \tag{8}$$

$$error=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{n-1}g\left(Q\left(u|X_{n-1}\right)-\hat{Q}\left(u|B_m\left(X_{n-1}\right)\right)\right), \tag{9}$$

where $\mathbb{E}_{n-1}$ is expectation conditional on $\mathcal{F}_{n-1}$, the sigma algebra generated by $\left(X_s\right)_{s<n}$. Our goal is to apply Theorem 23 to show that $error=o_p\left(1\right)$. To this end, we state the relevant conditions.

Condition 26 *$Q\left(u|x\right)$ is the unique solution $\theta_0\left(x\right)$ of*

$$\Pr\left(X_n<\theta_0\left(x\right)|X_{n-1}=x\right) \leq u \leq \Pr\left(X_n \leq \theta_0\left(x\right)|X_{n-1}=x\right),\ x\in E.$$

Condition 27 *For any n, some $\alpha>0$, and $x\in E$, $P_x\left(|X_n| \geq z\right)=O\left(z^{-(1+\alpha)}\right)$.*

In words the above condition requires the MC not to drift away from its central values (tightness), and it is stronger than the more general condition of not being evanescent (see Meyen and Tweedie, 1993). For example a random walk in $\mathbb{R}$ is not evanescent, but does not satisfy Condition 27, as it is not bounded in probability (see Nicolau, 2002, for a discussion of this and related models that are bounded in probability and embedded to the class of HRMC, under suitable restrictions).

We have the following.

Corollary 28 *Under Conditions 3, 5, 7, 26 and 27, for any $u\in[a,b]\subset(0,1)$,*

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{n-1}g\left(X_n-\hat{Q}\left(u|B_m\left(X_{n-1}\right)\right)\right) \leq \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{n-1}g\left(X_n-Q\left(u|X_{n-1}\right)\right)+o_p\left(1\right).$$

By Corollary 28 we can expect to forecast non extreme quantiles almost as well as if we used the true conditional quantiles.

4 Discussion

The goal of this paper is to identify general weak conditions that allow us to solve a broad class of nonparametric and semiparametric time series problems by nearest neighbor estimation. However, some issues of practical nature, whose detailed account is beyond the scope of this paper, deserve some mention. One is the choice of neighbors and the second is related to the curse of dimensionality and dimensionality reduction techniques. We briefly consider these two problems mainly relating to some existing results in the literature.

4.1 Choosing the Number of Neighbors by Prequential Validation

One fundamental issue in smoothing methods is the choice of smoothing parameter. In the present context, we considered two parameters: r_m and m_n, i.e. the radius of the d-ball and the number of required neighbors in this ball. The need to choose these two parameters makes practical implementation complex. Further restricting the class of HRMC, we may assume β recurrence so that the HRMC makes $O_p\left(n^\beta\right)$ visits to the set $B\left(x, r_m\right)$. Under this condition, we only need to discuss choice of m_n. The reader will see that the argument and the notation can then be simplified, with little loss of generality.

Suppose that we have preselected J sequences $\left\{m_n^{(j)}, j = 1, ..., J\right\}$, where $m_n^{(j)}$ is an increasing (sub-linear) function of n. Our goal is to identify the j^{th} sequence that gives the best relative performance for some given criterion. If we selected a large enough number of sequences, we could be confident that one of them would satisfy Condition 7.

To be specific, select a measurable loss function $\mathcal{R}$ for the estimator $\hat{P}_m$. Note that all the estimators we consider are functions of $\hat{P}_m$, where $m = m_n$ and for ease of notation the subscript is often omitted. The loss function at time n is a function of $(X_i)_{i \leq n}$and we shall generically write $\mathcal{R}_{n+1}^{(j)} := \mathcal{R}\left(\hat{P}_{m^{(j)}}\left(X_{n+1} | B_{m^{(j)}}\left(X_n\right)\right)\right)$ for the loss incurred at time $n+1$ when we use $m^{(j)} = m_n^{(j)}$ neighbors and only observations up to

time n to construct $\hat{P}_{m^{(j)}}$ so that when X_{n+1} is revealed we incur the loss $\mathcal{R}_{n+1}^{(j)}$.

Example 29 *Let* $\hat{X}_{n+1}^{(j)} := \int_E s\hat{P}_{m^{(j)}}\left(ds|B_{m^{(j)}}\left(X_n\right)\right)$ *be the* $m_n^{(j)}$ *neighbor estimator for the mean of* X_{n+1} *conditional on* X_n*. Then,* $\mathcal{R}_{n+1}^{(j)} := \left|X_{n+1} - \hat{X}_{n+1}^{(j)}\right|^2$.

For a sample of N observations, we shall choose

$$\hat{j} := \arg \min_{j \in \{1,...,J\}} \sum_{n=1}^{N} \mathcal{R}_{n+1}^{(j)} \tag{10}$$

to be the optimal choice of j.

Example 30 *Let* $m_n^{(j)} := \alpha_j n^{\beta(j)}$*, where* α_j *is a small positive constant and* $\beta_j \in (0,1)$ *for* $j = 0, ..., J$*. Then,* $\hat{j}$ *identifies the sequence* $m_n^{(j)}$ *which gives smallest total loss* $\sum_{n=1}^{N} \mathcal{R}_{n+1}^{(j)}$.

It follows that the number of neighbors is the same irrespective of the conditioning value X_n, so that $\sum_{n=1}^{N} \mathcal{R}_{n+1}^{(j)}$ is a global criterion for the loss based on $m^{(j)}$ neighbors. Note that neighbors automatically adjust the level of smoothing depending on the sparsity of data in different regions.

The approach just described is based on the prequential (predictive sequential) principle of Dawid (e.g. Dawid, 1986, Dawid and Vovk, 1999, and Seillier-Moiseiwitsch and Dawid, 1993). Since $\hat{j}$ in (10) is a random variable, the above rule might not be satisfactory. If j_1 and j_2 identify two sequences in $\left\{m_n^{(j)}, j = 1, ..., J\right\}$ that lead to equivalent conditional losses,

$$M_N^{(j_1,j_2)} := \sum_{n=1}^{N} \left(\mathcal{R}_{n+1}^{(j_1)} - \mathcal{R}_{n+1}^{(j_2)}\right)$$

is a martingale and standard inference can be conducted (Seillier-Moiseiwitsch and Dawid, 1993).

Proposition 31 *Suppose* $\mathbb{E}_n\left(\mathcal{R}_{n+1}^{(j_1)} - \mathcal{R}_{n+1}^{(j_2)}\right) = 0$ *and* $\left(\mathcal{R}_{n+1}^{(j_1)} - \mathcal{R}_{n+1}^{(j_2)}\right)^2$ *is uniformly integrable, and* $N^{-1}\sum_{n=1}^{N} \mathbb{E}_n \left|\mathcal{R}_{n+1}^{(j_1)} - \mathcal{R}_{n+1}^{(j_2)}\right|^2 \to \sigma^2 < \infty$*. Then,*

$$N^{-1/2} M_N^{(j_1,j_2)} \xrightarrow{w} N\left(0, \sigma^2\right) \ (\xrightarrow{w} \text{ is weak convergence}).$$

Given an apriori confidence level, the above result allows us to develop an automatic choice of sequence among $\left\{m_n^{(j)}, j = 1, ..., J\right\}$. Suppose that $[0, c_\alpha]$ is a $(1-\alpha)\,100\%$ one sided confidence interval for the standard normal distribution. For each j and given confidence level, choose j such that $N^{-1/2} M_N^{(j,\hat{j})} \in [0, \sigma c_\alpha]$ and $m_N^{(j)}$ is largest. By the assumptions of Proposition 31, σ^2 can be replaced by a consistent estimator. This approach allows us to impose maximum smoothing without significant increase in the approximation error. More refined approches based on sequential testing are possible (e.g. Belomestny and Spokoiny, 2005), but their description is beyond the scope of this paper.

A successful alternative to selection of j is averaging among different estimators based on different smoothing levels. This approach is widely used in different contexts, (e.g. Breiman, 1996, for linear model selection, Hoeting et al., 1999, for Bayesian model averaging, Polyak and Juditsky, 1992, for stochastic approximation estimators, Resnick and Starica, 1999, for tail index estimation) and can be seen as a special case of forecast combination (e.g. Timmermann, 2006, for a survey).

Example 32 *Suppose $\hat{X}_{n+1}^{(j)}$ is as in Example 29. Then, for $(w_1, ..., w_J)$ in the J dimensional unit simplex,*

$$\hat{X}_{n+1} := \sum_{j=1}^{J} w_j \hat{X}_{n+1}^{(j)}$$

is a combined estimator.

Estimation of the weights can be carried out by different methods.

Example 33 *Using the notation of Example 29, define*

$$\sum_{n=1}^{N} \mathcal{R}_{n+1}(w_1, ..., w_J) := \sum_{n=1}^{N} \left| X_{n+1} - \sum_{j=1}^{J} w_j \hat{X}_{n+1}^{(j)} \right|^2$$

and choose $(w_1, ..., w_J)$ such that the above is minimized.

Often it is not clear what is a best choice of weights. For this reason it is common to use equal weights as is done in several of the mentioned references, perhaps over the j's that have reasonable performance so not to increase the bias too much (e.g. Granger and Jeon, 2004)

Example 34 *Using the notation of Example 29, select all the j's such that $N^{-1/2}M_N^{(j,\hat{\jmath})} \in [0, \sigma c_\alpha]$ and to ease notation denote the selected sequences by $\left\{m_n^{(j)}, j = 1, ..., J'\right\}$. Then define the estimator by*

$$\sum_{j=1}^{J'} \frac{\hat{X}_{n+1}^{(j)}}{J'}.$$

The forecast combination literature is quite rich in examples of combined forecasts and other alternatives exist, but are not discussed here (e.g. Capistrán and Timmermann, 2006, for equally weighted forecasts). The relative merit of these approaches is both a theoretical and empirical question beyond the scope of this paper. We now turn to the problem of dimensionality reduction.

4.2 Imposing Restrictions on High Order MC

Some processes admit an MC representation only when embedded into a large state space. The effect of dimension on the neighbor's estimation is quite detrimental and this problem is common to all local methods. A way to mitigate this problem is to incorporate extra knowledge or assumptions in the metric d used to construct the neighbors. One simple way to do so is to consider different metrics that are topologically equivalent, but have different implications for the estimation. Recall that two metrics d_1 and d_2 on a set E are topologically equivalent if $y, x \in E$, $d_1(y, x) = 0$ if and only if $d_2(y, x) = 0$.

Example 35 *Suppose d is the Euclidean distance on $\mathbb{R}^K$ and d_λ is such that for x, $y \in \mathbb{R}^K$ and for positive λ bounded away from zero,*

$$d_\lambda(x, y) := \left(\sum_{k=1}^{K} \lambda^{k-1} |x_k - y_k|^2\right)^{1/2}.$$

Then, d_λ and d are topologically equivalent and in particular $d = d_1$.

Example 35 with $\lambda < 1$ can be used if there is higher order dependence, but with decreasing importance on the past, so that k represent the k^{th} lag. Then, the m_n neighbor using d_λ may vary considerably for different choices of λ. This approach

leads to an implicit dimensionality reduction. Clearly, we could directly restrict d to act on some manifold in E.

Example 36 *Suppose that for some function* $R : E \to \mathbb{R}$

$$\Pr\left(X_i \leq s | X_{i-1} = x\right) = \Pr\left(X_i \leq s | R\left(X_{i-1}\right) = R\left(x\right)\right),$$

then we can substitute the E *valued conditioning value* x *with the* $\mathbb{R}$ *valued* $R\left(x\right)$*. There is a clear advantage if* $E = \mathbb{R}^K$ *and* $K > 1$*. Hall and Yao (2005) have studied this problem when* $R\left(x\right)$ *is a linear function and need to be estimated.*

Another approach is to estimate the model with unrestricted d and combine it with a low dimensional parametric model via shrinkage.

Example 37 *Using the notation of Example 29, let* $\hat{X}_{n+1}$ *be the selected nearest neighbor estimator, and let* $\tilde{X}_{n+1} = \hat{a}_n + \hat{b}_n X_n$ *be the linear least square predictor of* $\mathbb{E}_n X_{n+1}$ *based on the sample* $(X_i)_{i \leq n}$*. Then, consider the estimator*

$$\left[w\hat{X}_{n+1} + (1-w)\,\tilde{X}_{n+1}\right]$$

where $w \in [0,1]$ *is chosen such that*

$$\sum_{n=1}^{N} \left| X_{n+1} - \left[w\hat{X}_{n+1} + (1-w)\,\tilde{X}_{n+1}\right]\right|^2$$

is minimized recursively. Shrunk estimators are commonly used in high dimensional problems (e.g. Ledoit and Wolf, 2004).

4.3 Final Remarks

The study of optimal selection of neighbors' size and dimensionality reduction are fundamental in practical situations. The above suggestions are mainly based on the author's preferences and experience in applied work. Many existing results in the literature should also be applicable to the general context of HRMC. However, in this more general context, formal justification is required for any of the existing approaches (including the ones mentioned here) and this will be subject of future research. The

main focus of this paper is on conditions that allow us to derive consistent estimators for HRMC without mixing conditions. It is hoped that the generality of these conditions and the general class of problems considered might be appealing to the time series analyst and forecaster.

A More General State Space

Many of the results of this paper hold for a state space more general than $E \subseteq \mathbb{R}^K$. We can consider a general state space E with a countably generated sigma algebra $\mathcal{E}$. A nice example is $E \subseteq \mathbb{R}^\infty$ equipped with the metric $d_\infty(x, y) = \sum_{i=1}^\infty 2^{-i} f(d(x_i, y_i))$ where $x_i, y_i \in \mathbb{R}$, $f(t) = t/(1+t)$ and d is any metric topologically equivalent to the Euclidean norm. Then, $\mathbb{R}^\infty$ is metrizable by d_∞ (Dudley, 2002, Proposition 2.4.4). Since a metrizable space is countably generated, mutatis mutandis, the results of the paper can be derived in this more general framework where the conditioning sets are balls of d_∞-radius r_m. Clearly, difficulties arise, e.g. in general Theorem 9 will not hold uniformly because the set $\{y \in E \subseteq \mathbb{R}^\infty : y \leq s\}$ does not have finite bracketing number. However, let $E = E_1 \times E_2$ where $E_1 \subseteq \mathbb{R}^K$ and $E_2 \subseteq \mathbb{R}^\infty$ (K finite). If we restrict attention to $\{y_1 \in E_1 : y_1 \leq s\}$ ($s \in E_1$) then Theorem 9 still holds when we want to estimate the conditional distribution

$$\Pr(X_n \in \{y_1 \in E_1 : y_1 \leq s\} \cap E_2 | X_{n-1} \in E)$$

using the nearest neighbor estimator based on balls of d_∞-radius r_m. One may proceed along these lines to partially rederive the other results of the paper.

B Remarks on Continuity in Condition 14

Condition 5 is used to show that the bias vanishes. This together with Condition 14 avoids assuming that $Pf(x)$ is smooth in x and allows us to disregard conditions on the bracketing numbers of $\mathfrak{F}$.

The approach of the paper is to use Theorem 9 to show that

$$\sup_{f} \left| \hat{P}_m f\left(B_m\left(x\right)\right) - Pf\left(x\right) \right| \overset{a.s.}{\to} 0$$

for any family of π_x-a.s. equicontinuous bounded functions. However, Theorem 9 implies more, as its statement holds for functions that are not continuous, i.e. $I\left\{x : x \leq s\right\}$ is discontinuous. Hence, there is some gain in deriving convergence as a corollary of Theorem 9 because, as mentioned in Remark 16, we could consider larger classes of functions (though in the statement of the results we refrained to do so to avoid extra notation). We recall the following definition.

Definition 38 *A bounded function f on E is of bounded variation in the sense of Vitali also called uniform bounded variation (UBV) (e.g. Clarkson and Adams, 1933, Lenze, 2003) if for any compact subset of E,*

$$f\left(x\right) = \mu_1\left(\left\{s : s \leq x\right\}\right) - \mu_2\left(\left\{s : s \leq x\right\}\right)$$

where μ_1 and μ_2 are finite measures on the compact sets of E.

Remark 39 *In one dimension this is the usual definition of bounded variation. In higher dimensions, there is no unique way to define bounded variation, though the usual modern definition is different and weaker (e.g. Ziemer, 1989).*

Then, we note the following.

Lemma 40 *Suppose UBV_b is the class of uniformly bounded functions in UBV. Under the Conditions of Theorem 9,*

$$\sup_{f \in UBV_b} \left| \hat{P} f\left(B_m\left(x\right)\right) - Pf\left(x\right) \right| \overset{a.s.}{\to} 0,$$

for any $x \in E$.

Proof. The uniform convergence of Theorem 9 is a.s. convergence under the Kolmogorov metric (e.g. Rachev, 1991). Let $\mathfrak{M}_b$ be the class of bounded monotone

increasing functions in each argument with domain E. It is sufficient to prove uniform convergence in $\mathfrak{M}_b$. Hence, by Lemma 10 in Sancetta (2007) deduce that

$$\sup_{f\in\mathfrak{M}_b}\left|\hat{P}f\left(B_m\left(x\right)\right)-Pf\left(x\right)\right|\overset{a.s.}{\to}0 \text{ if and only if } \sup_{s\in E}\left|\hat{P}\left(s|B_m\left(x\right)\right)-P\left(s|x\right)\right|\overset{a.s.}{\to}0$$

and the result is proved. ■

For definiteness let $\mathfrak{E}_b$ be an arbitrary, but fixed, family of uniformly bounded equicontinuous functions. Note that by equicontinuity, each element in $\mathfrak{E}_b$ can be turned into a Lipschitz function under the metric

$$d\left(x,y\right):=\sup_{f\in\mathfrak{E}}\left|f\left(x\right)-f\left(y\right)\right|$$

for each $x,y\in E$ (see the proof of Corollary 11.3.4 in Dudley, 2002). This shows that $\mathfrak{E}_b$ may contain many functions of interest on top of Lipschitz functions. However, by Lemma 40 we may further increase the set of functions allowed by Condition 14 *ii.* to $\mathfrak{F}\subseteq\mathfrak{E}_b\cup UBV_b$. A tail condition as in Condition 14 *i.* allows us to truncate so that we can avoid the uniform boundedness condition. Note that while the intersection of $\mathfrak{E}_b$ and UBV_b is not empty, it is not possible to establish an inclusion of one family into another. In fact there are uniformly continuous functions that are not of bounded variation (e.g. $f\left(x\right)=x\sin\left(1/x\right)$ for $x\in(0,2\pi]$, 0 elsewhere is not in UBV_b). Clearly, $f\left(x\right)=\left\{s\in E:s\leq x\right\}$ is in UBV_b but not in $\mathfrak{E}_b$. Hence $\mathfrak{E}_b\cup UBV_b$ is fairly rich.

Example 41 *Suppose* $f\left(x\right)=\sum_{i=1}^{I}f_i\left(x\right)I\left\{x\in A_i\right\}$, *where* $\left(A_i\right)_{i\in\{1,...,I\}}$ *are non overlapping hyper-rectangular sets, i.e.* $A_i:=[s_i,t_i]$, $s_i\leq t_i\in\mathbb{R}^K$ *and* $f_1,...,f_I$ *are* π_x*-a.s. uniformly bounded and absolutely continuous functions (e.g. Dudley, 2002, for definitions). Then,*

$$\left|\hat{P}f\left(B_m\left(x\right)\right)-Pf\left(x\right)\right|\overset{a.s.}{\to}0.$$

As already mentioned, we could truncate to allow for unbounded functions (see Lemma 46 below).

C Proofs

We recall the definition of bracketing numbers (e.g. van der Vaart and Wellner, 2000, for more details) to be used in the present context.

Definition 42 *For measurable functions l and u, the bracket $[l,u]$ is the set of all functions f such that $l \leq f \leq u$ and an $L_p(\pi_x)$ ϵ-bracket is a bracket such that $[P|u-l|^p(x)]^{1/p} \leq \epsilon$. The minimal number of $L_p(\pi_x)$ ϵ-brackets needed to cover a set $\mathfrak{F}$ is called the bracketing number and it will be denoted by $N_{\mathfrak{F}}(\epsilon, L_p(\pi_x))$.*

We can now turn to the proof of the results.

C.1 Proof of Theorem 9

The proof of Theorem 9 depends on some intermediary results. We split the proof in control over the estimation error (e.g. variance) and over the approximation error (e.g. bias). The estimation error is first.

Lemma 43 *Under Condition 3, for any $B \subset E$, such that $P(s|B)\pi(B) < \infty$,*

$$\sup_{s \in E} \left| \hat{P}_m(s|B) - P(s|B) \right| \xrightarrow{a.s.} 0,$$

as $m \to \infty$.

Proof. Note that

$$\begin{aligned} \hat{P}_m(s|B) &= \frac{1}{m} \sum_{i=1}^{m} I\{X(T_{B_m}(i)+1) \leq s\} \\ &= \frac{\sum_{i=1}^{n} I\{X_i \in B\}}{m} \frac{\sum_{i=1}^{n} I\{X_i \in B, X_{i+1} \leq s\}}{\sum_{i=1}^{n} I\{X_i \in B\}} \end{aligned}$$

where n is such that

$$m = \sum_{i=1}^{n} I\{X_i \in B\}.$$

Clearly, given m, n is random, and given n, m is random, but in any case one goes to infinity a.s. if the other does. Hence,

$$\frac{\sum_{i=1}^{n} I\{X_i \in B, X_{i+1} \leq s\}}{\sum_{i=1}^{n} I\{X_i \in B\}} \xrightarrow{a.s.} P(s|B)$$

following Duflo (1997),asymptotic property (c) of a recurrent chain, p.277. To obtain uniform convergence, note that we can find a finite number S of bracketing functions

$(I\{X_n \leq y_s\}, s = 1, ..., S)$ for the indicator function of sets of the form $\left\{y \in E \subseteq \mathbb{R}^K, y \leq s\right\}$ (K bounded away from ∞) such that

$$\mathbb{E}\left(\left|I\{X_n \leq y_{s+1}\} - I\{X_n \leq y_s\}\right| | X_{n-1} = x\right) \leq \epsilon,$$

where $y_{s+1} > y_s$. Hence, the convergence is also uniform (e.g. Theorem 2.4.1 in van der Vaart and Wellner, 2000, for further details). ■

We now consider the approximation error.

Lemma 44 *Set $B_m := B(x, r_m)$. By Conditions 5 and 7*

$$\sup_{s \in E} |P(s|B_m) - P(s|x)| \to 0.$$

Proof. Recall that $Pf(B) := \int_B \int_E f(y) P(x, dy) [\pi(dx)/\pi(B)]$. Then,

$$P(s|B_m) - P(s|x) = \int_{B_m} [P(s|y) - P(s|x)] \frac{\pi(ds)}{\pi(B_m)} \to 0$$

by strong differentiability of the integral and by Condition 5, (e.g. Theorem 1.3.8 in Ziemer, 1989). Note that, by the same cited theorem, the result holds true for π-almost all x if Condition 5 fails. Then, using a finite number of bracketing functions for the indicator function of sets $\{y : y \leq s\}$, as in Lemma 43,

$$\sup_{s \in E} |P(s|B_m) - P(s|x)| = \sup_{s \in E} |\Pr(X_n \leq s | X_{n-1} \in B(x, r_m)) - \Pr(X_n \leq s | X_{n-1} = x)| \overset{a.s.}{\to} 0.$$

■

Proof of Theorem 9. By the triangle inequality,

$$\begin{aligned} & \sup_{s \in E} \left|\hat{P}_m(s|B(x, r_m)) - P(s|x)\right| \\ \leq & \sup_{s \in E} \left|\hat{P}_m(s|B_m) - P(s|B_m)\right| + \sup_{s \in E} |P(s|B_m) - P(s|x)| \end{aligned}$$

and the terms on the r.h.s go to zero by Lemmata 43 and 44 respectively. ■

C.2 Proof of Corollaries

To prove Corollary 17 we need two lemmata.

Lemma 45 *Let $\mathfrak{E}_b$ be a family of π_x-a.s. uniformly bounded and equicontinuous functions. Under the conditions of Theorem 9,*

$$\sup_{f\in\mathfrak{E}_b}\left|\hat{P}_m f\left(B_m\left(x\right)\right)-Pf\left(x\right)\right|\overset{a.s.}{\to}0.$$

Proof. By Theorem 9, $\hat{P}_m\left(s|B_m\left(x\right)\right)$ converges weakly a.s. to $P\left(s|x\right)$. Then, uniform convergence in $\mathfrak{E}_b$ follows by Corollary 11.3.4 in Dudley (2002). ■

Lemma 46 *Suppose $\mathfrak{F}$ satisfies i. in Condition 14. Then, for any $\epsilon>0$, there is a large enough b such that*

$$\sup_{f\in\mathfrak{F}}\left|\hat{P}_m\left|fI_{\{|f|>b\}}\right|\left(B_m\left(x\right)\right)+P\left|fI_{\{|f|>b\}}\right|\left(B_m\left(x\right)\right)\right|\overset{a.s.}{\leq}\epsilon.$$

Proof. Set $F^b:=F\left\{F>b\right\}$, where F is the envelope of $\mathfrak{F}$. Mutatis mutandis, as in Lemma 43, for some ball B centred at x, define

$$M_m=\sum_{i=1}^{m}\left(1-\mathbb{E}_i\right)F^b\left(X\left(T_B\left(i\right)+1\right)\right),$$

where $\mathbb{E}_i$ is expectation conditional on the sigma algebra generated by $X\left(T_B\left(i\right)\right)$, $X\left(T_B\left(i\right)-1\right)$, ..., $X\left(0\right)$. Hence,

$$M_m/m=\hat{P}_mF^b\left(B\left(x\right)\right)-PF^b\left(B\left(x\right)\right),$$

where M_m is a martingale and

$$\begin{aligned}
&\left|\hat{P}_mF^b\left(B_m\left(x\right)\right)+PF^b\left(x\right)\right|\\
=\;&\left|M_m/m+2PF^b\left(B\left(x\right)\right)\right|\\
&\text{[by definition of } M_m/m\text{]}\\
\leq\;&\left|M_m/m\right|+2PF^b\left(B\left(x\right)\right)\\
=\;&:\mathrm{I}+\mathrm{II}.
\end{aligned}$$

By $i.$ in Condition 14

$$\sum_{i=0}^{\infty}\frac{\mathbb{E}_i\left|\left(1-\mathbb{E}_i\right)F^b\left(X\left(T_B\left(i\right)+1\right)\right)\right|^p}{\left(i+1\right)^p}\overset{a.s.}{<}\infty$$

as the numerator is π_B integrable for some $p > 1$ and B small enough. Therefore, by the strong law of large numbers for martingales, I$= M_m/m \overset{a.s.}{\to} 0$ (e.g. Chow and Teicher, 1988). Since $B_m(x) \to \{x\}$, $PF^p(x) \leq \limsup_m PF^p(B_m(x)) < \infty$ implies II$= PF^b(x) \leq \epsilon$, for any $\epsilon > 0$, by suitable choice of b. Noting that

$$\sup_{f \in \mathfrak{F}} \hat{P}_m \left| f I_{\{|f|>b\}} \right| (B_m(x)) \leq \hat{P}_m F^b(B_m(x)),$$

and similarly for P, the result follows. ■

Proof of Corollary 17. Set $f^b := f\{|f| > b\}$ and $f_b := f\{|f| \leq b\}$. Then,

$$\begin{aligned} & \sup_{f \in \mathfrak{F}} \left| \hat{P}_m f(B_m(x)) - Pf(x) \right| \\ \leq & \sup_{f \in \mathfrak{F}} \left| \hat{P}_m f_b(B_m(x)) - Pf_b(x) \right| + \sup_{f \in \mathfrak{F}} \left| \hat{P}_m \left| f^b \right| (B_m(x)) + P \left| f^b \right| (x) \right| \\ = & \text{ I} + \text{II}. \end{aligned}$$

Since $f_b \in \mathfrak{E}_b$ by *ii.* in Condition 14, Lemma 45 applies and I$\overset{a.s.}{\to} 0$. Since the envelop of $\mathfrak{F}$ satisfies suitable moment conditions, Lemma 46 applies as well and II$\overset{a.s.}{\leq} \epsilon$ where ϵ is arbitrary for b large enough. ■

Proof of Corollary 20. The proof can be deduced from the proof of Lemma 48 (below). ■

C.3 Proof of Theorem 23

Lemma 47 *Suppose $(Z_n)_{n \in \mathbb{N}}$ is a sequence of uniformly integrable positive random elements such that $Z_n \xrightarrow{p} 0$. Then,*

$$\frac{1}{N} \sum_{n=1}^{N} Z_n \to 0 \text{ in } L_1.$$

Proof. For any $N' < N$,

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E} Z_n &= \frac{1}{N} \sum_{n=1}^{N'} \mathbb{E} Z_n + \frac{1}{N} \sum_{n=N'}^{N} \mathbb{E} Z_n \\ &\leq \max_{1 \leq n \leq N'} \frac{N'}{N} \mathbb{E} Z_n + \max_{N' \leq n \leq N} \mathbb{E} Z_n \\ &= \text{ I} + \text{II}. \end{aligned}$$

Let $N' = o(N)$, so that by uniform integrability I$\to 0$. Recall that convergence in probability plus uniform integrability is equivalent to convergence in L_1 (e.g. Rogers and Williams, 2000, Theorem 21.2), so that $\mathbb{E}Z_n \to 0$. Letting $N' \to \infty$ we then have II$\to 0$. ■

Lemma 48 *Suppose $(\rho, \mathfrak{F})$ is a metric space. Under Conditions 3, 5, 7, 19, and 21, conditioning on $X_0 = x$,*

$$\rho\left(\hat{f}_{m,n}, f_n\right) \xrightarrow{p} 0.$$

Proof. Note that $f_n := f_n(X_{n-1})$ and $\hat{f}_{m,n} := \hat{f}_m(X_{n-1})$ are random, as they depend on X_{n-1}. Let $G^{(n)} = G^{(n)}(X_{n-1})$ be an arbitrary open set that contains f_n and let $\left[G^{(n)}\right]^c$ be its complement. It is enough to show that

$$\mathrm{I} := \Pr\left(f_n \in G^{(n)}, \hat{f}_n \in \left[G^{(n)}\right]^c\right) = o(1),$$

as $G^{(n)}$ is arbitrary. To this end note that

$$\mathrm{I} = \Pr\left(\inf_{f \in \left[G^{(n)}\right]^c} \hat{P}_m f(B_m(X_{n-1})) \leq \inf_{f \in G^{(n)}} \hat{P}_m f(B_m(X_{n-1})), f_n \in G^{(n)}\right)$$

because the infimum of $\hat{P}_m f(B_m(X_{n-1}))$ is attained in $\left[G^{(n)}\right]^c$. Moreover, note that for any set $A \subseteq \mathfrak{F}$

$$\begin{aligned} & \inf_{f \in A} Pf(X_{n-1}) - \sup_{f \in A} \left|\hat{P}_m f(B_m(X_{n-1})) - Pf(X_{n-1})\right| \\ \leq & \inf_{f \in A} \hat{P}_m f(B_m(X_{n-1})) \leq \inf_{f \in A} Pf(X_{n-1}) + \sup_{f \in A} \left|\hat{P}_m f(B_m(X_{n-1})) - Pf(X_{n-1})\right|. \end{aligned}$$

Define

$$R_n := \sup_{f \in G^{(n)}} \left|\hat{P}_m f(B_m(X_{n-1})) - Pf(X_{n-1})\right|,$$

and

$$R'_n := \sup_{f \in \left[G^{(n)}\right]^c} \left|\hat{P}_m f(B_m(X_{n-1})) - Pf(X_{n-1})\right|.$$

Then,

$$\begin{aligned}
\mathrm{I} &\leq \Pr\left(\inf_{f\in\left[G^{(n)}\right]^c} Pf\left(X_{n-1}\right) \leq \inf_{f\in G^{(n)}} Pf\left(X_{n-1}\right) + R_n + R_n', f_n \in G^{(n)}\right) \\
&\leq \Pr\left(\inf_{f\in\left[G^{(n)}\right]^c} Pf\left(X_{n-1}\right) \leq \inf_{f\in G^{(n)}} Pf\left(X_{n-1}\right) + 2\epsilon, f_n \in G^{(n)}\right) + \\
&\quad + \int_E \Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, dx_{n-1}\right) \\
&\quad + \int_E \Pr\left(R_n' \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, dx_{n-1}\right) \\
&= \mathrm{II} + \mathrm{III} + \mathrm{IV}.
\end{aligned}$$

Since ϵ is arbitrary, by Condition 19, II$= 0$ because either $f_n \in \left[G^{(n)}\right]^c$ or $f_n \in G^{(n)}$. Denoting by C^c the complement of C, where $C \cup C^c = E$, consider the following inequalities,

$$\begin{aligned}
\mathrm{III} &= \int_C \Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, dx_{n-1}\right) \\
&\quad + \int_{C^c} \Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, dx_{n-1}\right) \\
&\leq \int_C \Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, dx_{n-1}\right) + P^{n-1}\left(x, C^c\right) \\
&\leq \int_C \Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, dx_{n-1}\right) + \epsilon \\
&= \mathrm{V} + \epsilon,
\end{aligned}$$

using Condition 21. By Corollary 17, $\Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) \to 0$ for any $x_{n-1} \in C$. Moreover,

$$\int_C \Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, dx_{n-1}\right) \leq \int_C 1 P^{n-1}\left(x, dx_{n-1}\right) \leq P^{n-1}\left(x, E\right) = 1.$$

Hence V$\to 0$ by the Dominated Convergence Theorem, so that III$\to 0$ because ϵ is arbitrary. An identical argument shows that IV$\to 0$ as well. ■

Proof of Theorem 23. By Lemma 48, $\rho\left(\hat{f}_n, f_n\right) \xrightarrow{p} 0$ conditioning on $X_0 = x$. Then, apply Lemma 47. ■

C.4 Proof of Corollary 28

Proof of Corollary 28. Using (8) and (9) we note that the relevant quantity to bound is

$$\begin{aligned}
& \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{n-1}g\left(Q\left(u|X_{n-1}\right)-\hat{Q}\left(u|B_m\left(X_{n-1}\right)\right)\right) \\
= & \frac{1}{N}\sum_{n=1}^{N}g\left(Q\left(u|X_{n-1}\right)-\hat{Q}\left(u|B_m\left(X_{n-1}\right)\right)\right) \\
& \text{[because } Q\left(u|X_{n-1}\right) \text{ and } \hat{Q}\left(u|B_m\left(X_{n-1}\right)\right) \text{ are } \mathcal{F}_{n-1} \text{ measurable]} \\
< & 2\frac{1}{N}\sum_{n=1}^{N}\left|Q\left(u|X_{n-1}\right)-\hat{Q}\left(u|B_m\left(X_{n-1}\right)\right)\right| =: \mathrm{I}
\end{aligned}$$

by definition of the loss function g. By an application of Theorem 23, we shall show that I$= o_p(1)$. To this end, we check that Condition 21 is satisfied and that I is uniformly integrable. Let Θ be some compact set and define

$$Q'\left(u|B_m\left(x\right)\right) := \arg\inf_{\theta\in\Theta} Pf_\theta\left(x\right) \text{ and } \hat{Q}'\left(u|B_m\left(x\right)\right) := \arg\inf_{\theta\in\Theta} \hat{P}_m f_\theta\left(B_m\left(x\right)\right).$$

Then,

$$\begin{aligned}
\mathrm{I} \leq & 2\frac{1}{N}\sum_{n=1}^{N}\left|Q'\left(u|X_{n-1}\right)-\hat{Q}'\left(u|B_m\left(X_{n-1}\right)\right)\right| + 2\frac{1}{N}\sum_{n=1}^{N}\left|Q\left(u|X_{n-1}\right)-Q'\left(u|X_{n-1}\right)\right| \\
& +2\frac{1}{N}\sum_{n=1}^{N}\left|\hat{Q}\left(u|B_m\left(X_{n-1}\right)\right)-\hat{Q}'\left(u|B_m\left(X_{n-1}\right)\right)\right| \\
= & \mathrm{II}+\mathrm{III}+\mathrm{IV},
\end{aligned}$$

and we shall control each term separately.

Control over II.

The loss function is Lipschitz continuous so that *ii.* in Condition 14 is satisfied. By Condition 27, using the fact that Θ is compact, $\mathbb{E}\left[\sup_{\theta\in\Theta}\left|X_n-\theta\right|^{1+\alpha}|X_0=x\right]<\infty$, so that also *i.* in Condition 14 is satisfied. Since x was arbitrary, it follows that Condition 21 is also satisfied. To show Condition 19 use Condition 26. To show P_x-uniform integrability of I, note that $Q'\left(u|X_{n-1}\right)$, and $\hat{Q}'\left(u|B_m\left(X_{n-1}\right)\right)$ are in Θ, hence they are bounded. Therefore, by Theorem 23, II$\xrightarrow{p}0$.

Control over III.

Since u is bounded away from 0 and 1, by Condition 27, there is a compact set C such that $Q(u|X_{n-1}) \in C$. Therefore, for any $\Theta \supseteq C$, $Q(u|X_{n-1}) = Q'(u|X_{n-1})$ and III$= 0$.

Control over IV.

Since $\hat{P}_m(s|B_m(x))$ is an unbiased estimator of $P(s|B_m(x))$, by Theorem 1 in Rychlik (1994), for $u \in [a,b] \subset (0,1)$,

$$\frac{[P(\theta|B_m(x)) - u] + 1/m}{(1-u)+1/m} \leq \Pr\left(\hat{Q}(u|B_m(x_{n-1})) \leq \theta | X_{n-1} \in B_m(x_{n-1})\right) \leq \frac{P(\theta|B_m(x))}{u},$$

so that using the definition of quantile and the above bounds, it is not difficult to see that, by Condition 27, the law of $\hat{Q}(u|B_m(X_{n-1}))$ conditioning on $X_0 = x$ has tails proportional to the tails of the law of X_n conditioning on $X_0 = x$. Hence, by Condition 27,

$$\Pr\left(\hat{Q}(u|B_m(X_{n-1})) > \theta | X_0 = x\right) = o\left(\theta^{-(1+\alpha)}\right)$$

for θ large enough. This implies that $\hat{Q}(u|B_m(X_{n-1}))$ is uniformly integrable and there is a compact Θ such that $\Pr\left(\hat{Q}(u|B_m(X_{n-1})) \in \Theta | X_0 = x\right) > 1 - \epsilon$ for any $\epsilon > 0$. Since

$$\begin{aligned} & \Pr\left(\left|\hat{Q}(u|B_m(X_{n-1})) - \hat{Q}'(u|B_m(X_{n-1}))\right| > 0 | X_0 = x\right) \\ = \; & \Pr\left(\hat{Q}(u|B_m(X_{n-1})) \notin \Theta | X_0 = x\right) \\ \leq \; & \epsilon, \end{aligned}$$

the conditions of Lemma 47 are satisfied and IV$\xrightarrow{p} 0$. Putting everything together, it follows that I$\xrightarrow{p} 0$. ■

References

[1] Ango Nze, P., P. Bühlmann and P. Doukhan (2002) Weak Dependence Beyond Mixing and Asymptotics for Nonparametric Regression. Annals of Statistics 30, 397-430.

[2] Ango Nze, P. and P. Doukhan (2004) Weak Dependence: Models and Applications to Econometrics. Econometric Theory 20, 995-1045.

[3] Babillot, M., P. Bougerol and L. Elie (1997) The Random Difference Equation $X_n = A_n X_{n-1} + B_n$ in the Critical Case. Annals of Probability 25, 478-493.

[4] Barndorff-Nielsen, O.E. and N. Shephard (2002) Econometric Analysis of Realized Volatility and its Use in Estimating Stochastic Volatility Models. Journal of the Royal Statistical Society, Series B 64, 253-280.

[5] Belomestny, D. and V. Spokoiny (2005) Local Likelihood Modelling via Stagewise Aggregation. Preprint No. 1000, Weierstrass Institute for Applied Analysis and Stochastics.

[6] Bertail, P. and S. Clémençon (2006) Regenerative Block Bootstrap for Markov Chains. Bernoulli 12, 689-712.

[7] Bingham, N.H., C.M. Goldie and J.L. Teugels (1987) Regular Variation. Cambridge: Cambridge University Press.

[8] Breiman, L. (1996) Heuristics of Instability and Stabilization in Model Selection. Annals of Statistics 24, 2350-2383.

[9] Capistrán, C. and A. Timmermann (2006) Forecast Combination with Entry and Exit of Experts. Working Paper Rady School of Management.

[10] Chen, X. (1999) How Often Does a Harris Recurrent Markov Chain Recur? Annals of Probability 27, 1324-1346.

[11] Chow, Y.S. and H. Teicher (1997) Probability Theory: Independence, Interchangeability, Martingales. New York: Springer.

[12] Clarkson, J.A. and C.R. Adams (1933) On Definitions of Bounded Variation for Functions of Two Variables. Transactions of the American Mathematical Society 35, 824-854.

[13] Dawid, A.P. (1986) Probability Forecasting. In S. Kotz, N.L. Johnson and C.B. Read (eds.), Encyclopedia of Statistical Sciences Vol. 7, 210-218. Wiley.

[14] Dawid, A.P. and V. Vovk (1999) Prequential Probability: Principles and Properties. Bernoulli 5, 125-162.

[15] Doukhan, P. (1994) Mixing: Properties and Examples. Lecture Notes in Statistics 85. New York: Springer.

[16] Doukhan, P. and S. Louhichi (1999) A New Weak Dependence Condition and Applications to Moment Inequalities. Stochastic Processes and Applications 84, 313-342.

[17] Dudley, R.M. (2002) Real Analysis and Probability. Cambridge: Cambridge University Press.

[18] Embrechts, P., C. Klüppelberg and T. Mikosch (1997) Modelling Extremal Events. Berlin: Springer.

[19] Granger C.W.J. and Y. Jeon (2004) Thick Modeling. Economic Modelling 21, 323-343.

[20] Hall, P. and Q. Yao (2005) Approximating Conditional Distribution Functions Using Dimension Reduction. Annals of Statistics 33, 1404-1421.

[21] Horowitz, J.L. (2003) Bootstrap Methods for Markov Processes. Econometrica 71, 1049-1082.

[22] Joe, H. (1997) Multivariate Models and Dependence Concepts. London: Chapman and Hall Ltd.

[23] Karlsen, H.A. and D. Tjøstheim (2001) Nonparametric Estimation in Null Recurrent Time Series. Annals of Statistics 29, 372-416.

[24] Leadbetter, M.R. and H. Rootzén (1988) Extremal Theory for Stochastic Processes. Annals of Probability 16, 431-478.

[25] Ledoit, O. and M. Wolf (2004) A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. Journal Multivariate Analysis 88, 365-411.

Hoeting, J.A., D. Madigan, A.E. Raftery and C.T. Volinsky (1999) Bayesian Model Averaging: A Tutorial. Statistical Science 14, 382-417.

[26] Lenze, B (2003) On the Points of Regularity of Multivariate Functions of Bounded Variation. Real Analysis Exchange 29, 646-656.

[27] Meyn, S.P. and R.L. Tweedie (1993) Markov Chains and Stochastic Stability. London: Springer.

[28] Nicolau, J. (2002) Stationary Processes that Look Like Random Walks- The Bounded Random Walk Process in Descrete and Continuous Time. Econometric Theory 18, 99-118.

[29] Pagan, A. and A. Ullah (1999) Nonparametric Econometrics. Cambridge: Cambridge University Press.

[30] Polyak, B.T. and A.B. Juditsky (1992) Acceleration of Stochastic Approximation by Averaging. SIAM Journal of Control and Optimization 30, 838-855.

[31] Rachev, S.T. (1991) Probability Metrics and the Stability of Stochastic Models. Chichester: Wiley.

[32] Resnick, S. and C. Starica (1999) Smoothing the Moment Estimator of the Extreme Value Parameter. Extremes 1, 263-293.

[33] Robinson, P. M. (1983) Nonparametric Estimators for Time Series. Journal of Time Series Analysis 4, 185-207.

[34] Rogers, L.C.G. and D. Williams (2000) Diffusions, Markov Processes and Martingales. Cambridge: Cambridge University Press.

[35] Rychlik, T. (1994) Distributions and Expectations of Order Statistics for Possibly Dependent Random Variables. Journal of Multivariate Analysis 48, 31-42.

[36] Sancetta (2007) Weak Convergence of Laws on $\mathbb{R}^K$ with Common Marginals. Forthcoming in Journal of Theoretical Probability. Downloadable: http://arxiv.org/abs/math.PR/0606462

[37] Seillier-Moiseiwitsch, F. and A.P. Dawid (1993) On Testing the Validity of Sequential Probability Forecasts. Journal of American Statistical Association 88, 355-359.

[38] Timmermann, A. (2006) Forecast Combinations. Forthcoming in G. Elliott, C.W.J Granger and A. Timmermann (eds.) Handbook of Economic Forecasting. North Holland.

[39] Van der Vaart, A. and J.A. Wellner (2000) Weak Convergence of Empirical Processes. Springer Series in Statistics. New York: Springer.

[40] Van Garderen K-J. (1997) Curved Exponential Models in Statistics. Econometric Theory 13, 771-790.

[41] Ziemer, W. (1989) Weakly Differentiable Functions. New York: Springer.

[42] Yakowitz, Sid (1993) Nearest Neighbor Regression Estimation for Null-Recurrent Markov Time Series. Stochastic Processes and their Applications 48, 311-318.