# On robustness of model-based bootstrap schemes in nonparametric time series analysis

Michael H. Neumann
Humboldt-Universität
Sonderforschungsbereich 373
Spandauer Straße 1
D – 10178 Berlin
Germany

*1991 Mathematics Subject Classification.* Primary 62G09, 62M10; secondary 62G07.
*Keywords and Phrases.* Bootstrap, nonparametric autoregression, nonparametric regression, strong approximation, weak dependence, whitening by windowing.
*Short title.* Robustness of bootstrap.

I thank H. Herwartz for helpful comments on this paper. The research was carried out within the Sonderforschungsbereich 373 at Humboldt University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft.

ABSTRACT. Theory in time series analysis is often developed in the context of finite-dimensional models for the data generating process. Whereas corresponding estimators such as those of a conditional mean function are reasonable even if the true dependence mechanism is of a more complex structure, it is usually necessary to capture the whole dependence structure asymptotically for the bootstrap to be valid. However, certain model-based bootstrap methods remain valid for some interesting quantities arising in nonparametric statistics. We generalize the well-known "whitening by windowing" principle to *joint* distributions of nonparametric estimators of the autoregression function. As a consequence, we obtain that model-based nonparametric bootstrap schemes remain valid for supremum-type functionals as long as they mimic the corresponding finite-dimensional joint distributions consistently. As an example, we investigate a finite order Markov chain bootstrap in the context of a general stationary process.

## 1. INTRODUCTION

One of the major merits of the bootstrap is its universality: it is valid for a variety of different purposes (statistics) and under quite general assumptions on the distributions of the observations. For i.i.d. data, it is easy to implement and usually one does not need severe conditions for its validity.

Without the assumption of independence of the observations, the construction of valid resampling schemes becomes more difficult since one has to appropriately mimic the dependence mechanism. Also in this context, there exist nearly assumption-free methods. Hall (1985), Carlstein (1986) and Shi (1986) proposed resampling from nonoverlapping blocks of increasing length which was later refined by Künsch (1989). Other modifications are the circular block bootstrap proposed by Politis and Romano (1992) and Shao and Yu (1993), the stationary bootstrap of Politis and Romano (1994) and the matched-block bootstrap of Carlstein, Do, Hall, Hesterberg and Künsch (1996).

On the other hand, there exists an extensive literature on model-based bootstrap methods in the time series context. Under the assumption of i.i.d. innovations in a linear autoregressive model, Efron and Tibshirani (1986) proposed to generate bootstrap series by drawing bootstrap innovations independently with replacement from the set of mean-adjusted residuals. Kreiss and Franke (1992) generalized this to autoregressive moving average models. Furthermore, there exists a series of proposals for bootstrapping Markov chains; see the brief survey in Section 3. There also exist several semiparametric methods. For example, Kreiss (1988) approximated linear autoregressive processes by a bootstrap process of finite, but increasing order. Franke and Wendel (1992) and Kreutzberger (1993) generalized the method of Efron and Tibshirani (1986) to the case of nonlinear autoregressive processes.

Concerning universality, blockwise bootstrap schemes with a block length tending to infinity dominate model-based methods since they do not require structural assumptions on the data generating process to be fulfilled. They are nearly assumption-free regarding both the distributions of the observations as well as the dependence structure between them. These methods are shown to be asymptotically correct for a number of important statistics; see, for example, Künsch (1989), Bühlmann (1994) and Götze and Künsch (1996). In contrast, model-based methods reflect the dependence mechanism of a general process only partially, even if the sample size tends to infinity. They are usually more powerful than model-free methods – at least as long as the data generating process obeys indeed the assumed structure.

In view of possible gains of power, one might be tempted to prefer model-based methods whenever there is some evidence for the (exact or approximate) correctness of a certain simple form of the dependence mechanism. However, often such a structure is at best approximately true. In such a case, if one had actually applied such a model-based bootstrap, it seems that one had risked too much in order to benefit from a supposed gain by this method. This is indeed the case with (finite-dimensional) parametric problems where even weak dependence of the observations influences first-order asymptotics of corresponding parameter estimates. In sharp contrast, first-order asymptotics of nonparametric estimators is often not affected by weak dependence. Robinson (1983) established corresponding results for the pointwise behaviour of nonparametric estimators under mixing in conjunction with an additional condition on the boundedness of joint densities. Hart (1995) coined the term "whitening by windowing" for this effect. This suggests that model-based bootstrap methods can correctly imitate the pointwise properties of nonparametric estimators based on the $m$-dimensional joint distribution of the observations if only these $m$-dimensional distributions are correctly retained, no matter whether or not the dependence mechanism of the bootstrap process actually coincides with that of the original process.

On the other hand, many methods of statistical inference are based on the whole nonparametric estimator rather than on an estimate at a single point. To benefit from the whitening by windowing principle, it is necessary to generalize it beyond the pointwise case. Such a result was established by Neumann (1996) for a kernel estimator of the stationary density of a weakly dependent process. By embedding both the observations from the time series model and the observations from a corresponding i.i.d. model in a common Poisson process, Neumann obtained a pairing of these random variables such that the unordered sets of observations are nearly the same. This led to a strong approximation of a kernel estimator in the time series model by an analogous kernel estimator in the i.i.d. model, and allowed to apply bootstrap techniques that were originally developed under the assumption of independence.

In the present paper we intend to establish a version of the whitening by windowing principle that concerns the joint distribution of nonparametric estimators of the conditional mean function, $m(x_1, \ldots, x_d) = E(X_t \mid X_{t-l_1}, \ldots, X_{t-l_d})$ . The result is again formulated in terms of a strong approximation of a nonparametric estimator in the time series model by an analogous estimator in a regression model with independent errors. To this end, we establish first a strong approximation of partial sums

with respect to small hypercubes $I_{\underline{k}} = [(k_1-1)g_1, k_1 g_1) \times \cdots \times [(k_d-1)g_d, k_d g_d)$ ,

$$Z_{\underline{k}} = \sum_{t:\,(X_{t-l_1},\ldots,X_{t-l_d}) \in I_{\underline{k}}} [X_t - E(X_t \mid X_{t-l_1}, \ldots, X_{t-l_d})] ,$$

by corresponding partial sums in a regression experiment. The link is achieved by embedding the summands from both models in a common set of Wiener processes $W_{\underline{k}}$ assigned to the intervals $I_{\underline{k}}$. As described in Subsection 2.1, attempts to embed the summands $I((X_{t-l_1}, \ldots, X_{t-l_d}) \in I_{\underline{k}})[X_t - E(X_t \mid X_{t-l_1}, \ldots, X_{t-l_d})]$ in their natural order failed. Quite surprisingly, an embedding of these quantities in reverse time order turned out to be successful and led to an approximation with a sufficiently small error. The derivation of the strong approximation is mainly achieved by this construction, whereas the analytical part of the proofs is comparatively simple.
This theoretical result can be applied as the first and most important step in proving robustness of certain model-based bootstrap methods against deviations from the underlying structural assumptions of the data generating process. As a particular model-based method we investigate in Section 3 a local bootstrap which is motivated by a Markov chain approximation of the process. According to our main result, this bootstrap procedure remains valid even if the true data generating process is of a more complex form. The same can be assumed for other model-based methods as well, for example, the moving blocks bootstrap with a fixed length of the blocks. We apply our Markov chain bootstrap to the construction of supremum-type tests in the context of general (not necessarily Markovian) processes. The proofs are deferred to the Appendix.

## 2. Whitening by windowing for the joint distribution of nonparametric estimators

We make the following basic assumption for the process $\{X_t\}$ under consideration:

**(A1)** $\{X_t : t \geq 0\}$ is a (strictly) stationary process. Furthermore, we assume absolute regularity (i. e. $\beta$-mixing) for $\{X_t\}$ and that the $\beta$-mixing coefficients decay at an exponential rate.

Throughout the whole paper, we do not impose any kind of *structural* assumptions on the data generating process. Although those assumptions are often made in the time series literature, it is always a potential source of an inadequate analysis and erroneous conclusions because they are rarely exactly fulfilled in practical applications.
Even if the dependence mechanism is much more complex, it makes sense to look at conditional expectations in finitely many lags, for example for the purpose of displaying marginal effects. Let

$$m(x_1, \ldots, x_d) = E(X_t \mid X_{t-l_1} = x_1, \ldots, X_{t-l_d} = x_d) \tag{2.1}$$

be the conditional expectation of $X_t$ given $X_{t-l_1} = x_1, \ldots, X_{t-l_d} = x_d$ , where $1 \leq l_1 < l_2 < \ldots < l_d$ .
There are several different options to estimate $m$. One may do this in a fully nonparametric manner, for example, by a multivariate kernel estimator as proposed by Robinson (1983) or by a local polynomial estimator as considered by Härdle and Tsybakov (1995) and Masry (1996). Alternative methods include semiparametric

estimators, for example those based on additive models for $m$, or even parametric estimators. For example, Yang and Härdle (1996) investigated a nonparametric estimator in an additive model. Unfortunately, up to now theory in this field is often developed under the assumption that the data generating process obeys indeed the structure underlying the fitted model. Nevertheless, it is reasonable to fit such dimension-reduced schemes, although it is often hard to believe that the true process follows actually such a rather specific structure.

After defining any point estimate for $m$, the next step in a statistical data analysis consists of making assertions which characterize the accuracy of these estimates like, for example, confidence intervals or bands for $m$, or forecast intervals for future observations. Generally, one would always prefer estimators based on lower-dimensional models over fully nonparametric methods as long as the corresponding model is indeed adequate. Therefore, testing a lower-dimensional against a fully nonparametric hypothesis is an important step in data analysis. In principle, these problems can be tackled by methods based on asymptotic theory. However, sometimes such asymptotic theory is not easily available, and there also exist cases where first-order asymptotic theory is known to provide rather poor approximations. A familiar example are simultaneous confidence bands in nonparametric regression, where it is known that their actual coverage probability converges to the nominal one with the rather slow rate of $(\log T)^{-1}$; see Hall (1991) for details. This is a typical case where bootstrap methods are really important.

Unless we have extraordinarily large sample sizes, we always have to take care of the curse of dimensionality. Owing to the sparsity of data in high dimensions, the performance of nonparametric estimators deteriorates rather quickly as the dimension increases. If we intend to generate a bootstrap process without structural assumptions on the original process like linearity, we are essentially in the same situation as with nonparametric estimators. Hence, such methods necessarily suffer from the curse of dimensionality. Therefore, it is tempting to implement a fully nonparametric bootstrap with almost the same dimensionality as the fitted model. Even if the dimension of the bootstrap model is slightly larger than that of $m$, an almost adequate asymptotics is that for a finite-dimensional bootstrap model. In order to show that such simplified bootstrap methods which imitate only the dependence from a fixed number of lagged variables are asymptotically valid, we prove first an even more rigorous result: We show that, in our nonparametric context, the dependence between the observations can be completely neglected. This is formalized in terms of a strong approximation of statistics connected with nonparametric estimators by corresponding statistics in a regression model. More exactly, the random variables of both models are paired in such a way that the error of approximation is of smaller order than the stochastic fluctuations of the statistic of interest. This means in principle that the dependence of the data generating process can be completely neglected when one intends to devise valid bootstrap methods.

An appropriate candidate for a model that is asymptotically equivalent to the process $\{X_t\}$ concerning nonparametric inference on $m$ is the nonparametric regression model

$$Z_t = m(x_{t1}, \ldots, x_{td}) + \eta_t, \tag{2.2}$$

where $(x_{t1}, \ldots, x_{td}) = (X_{t-l_1}, \ldots, X_{t-l_d})$ corresponds to a fixed realization of $\{X_t\}$. The errors $\eta_t$ are assumed to be independent with $E\eta_t = 0$ and $E\eta_t^2 = E((X_t -$

$m(x_{t1},\dots,x_{td}))^2 \mid X_{t-l_1} = x_{t1},\dots,X_{t-l_d} = x_{td})$ . Model (2.2) is an analogue to bootstrap methods that do not mimic the randomness in the lagged variables as for example the wild bootstrap. This method was investigated in the context of nonparametric autoregressive processes by Franke, Kreiss and Mammen (1996) and Neumann and Kreiss (1997). Another natural counterpart to $\{X_t\}$ is a nonparametric regression model with random design,

$$Z_t = m(Y_{t1},\dots,Y_{td}) + \eta_t, \tag{2.3}$$

where $(Y_{t1},\dots,Y_{td})$ are i.i.d. with the same marginal distribution as $(X_{t-l_1},\dots,X_{t-l_d})$ and $\eta_t$ as above. For definiteness of our presentation, we will stick to model (2.3) in the following.

In the present paper we focus on the joint distribution of nonparametric estimators or similar statistics. A usual kernel estimator of Nadaraya-Watson type, with a product kernel, has the form

$$\widehat{m}_{h_1,\dots,h_d}(x_1,\dots,x_d) = \frac{\sum_t K((x_1 - X_{t-l_1})/h_1)\cdots K((x_d - X_{t-l_d})/h_d)X_t}{\sum_t K((x_1 - X_{t-l_1})/h_1)\cdots K((x_d - X_{t-l_d})/h_d)}. \tag{2.4}$$

The bandwidths $h_1,\dots,h_d$ may take different values, preferably in accordance with the smoothness properties of the function $m$ in the respective coordinates. One may indeed expect that different degrees of smoothness in different coordinates are present: Since the dependence on higher lags is decaying, one might suppose that $m$ shows less variability in coordinates corresponding to such high lags.

To reduce the burden of multiple indices, we use the following shorthands: $\underline{h} = (h_1,\dots,h_d)$ , $\underline{x} = (x_1,\dots,x_d)$ , $\underline{X}_t = (X_{t-l_1},\dots,X_{t-l_d})$ , $\underline{Y}_t = (Y_{t1},\dots,Y_{td})$ and $w(\underline{x},\underline{y}) = K((x_1 - y_1)/h_1)\cdots K((x_d - y_d)/h_d)$ . The deviation of $\widehat{m}_{\underline{h}}(\underline{x})$ from the conditional mean $m(\underline{x})$ can be decomposed into a stochastic term,

$$\left(\sum_t w(\underline{x},\underline{X}_t)\right)^{-1} \sum_t w(\underline{x},\underline{X}_t)\,[X_t - m(\underline{X}_t)]\,, \tag{2.5}$$

and a bias-type term,

$$\left(\sum_t w(\underline{x},\underline{X}_t)\right)^{-1} \sum_t w(\underline{x},\underline{X}_t) m(\underline{X}_t) \quad - \quad m(\underline{x}). \tag{2.6}$$

(We call the latter expression "bias-type term" rather than "bias-term" since it is only asymptotically nonstochastic.)

It can be seen that the bias-type term is rather close to the corresponding term for the regression model. The more difficult part in proving asymptotic equivalence concerns the stochastic term. Since the denominator in (2.5) converges to its expectation with a sufficiently fast rate, we focus on the numerator in the following. We will show that this term can be approximated by its analogue in the regression model (2.3),

$$\left(\sum_t w(\underline{x},\underline{Y}_t)\right)^{-1} \sum_t w(\underline{x},\underline{Y}_t)\eta_t. \tag{2.7}$$

To formalize such an approximation, we construct, on a sufficiently rich probability space, a pairing of the random vector $(X'_{1-l_d}, \ldots, X'_T)$ with another vector $(\underline{Y}'_1, \ldots, \underline{Y}'_T, \eta'_1, \ldots, \eta'_T)$ such that

1) $(X'_{1-l_d}, \ldots, X'_T) \stackrel{d}{=} (X_{1-l_d}, \ldots, X_T)$,

2) $(\underline{Y}'_1, \ldots, \underline{Y}'_T, \eta'_1, \ldots, \eta'_T) \stackrel{d}{=} (\underline{Y}_1, \ldots, \underline{Y}_T, \eta_1, \ldots, \eta_T)$

and

3) $\sup_{x \in \mathbb{R}^d} \{|\sum \{w(\underline{x}, \underline{X}'_t)[X'_t - m(\underline{X}'_t)] - w(\underline{x}, \underline{Y}'_t)\eta'_t\}|\}$ is small with high probability.

To facilitate notation, we do not distinguish between the original random variables $X_t, \underline{Y}_t, \eta_t$ and their artificial counterparts $X'_t, \underline{Y}'_t, \eta'_t$. A first step toward an approximation as in 3) is an approximation of partial sums $\sum_{t:\underline{X}_t \in I_{\underline{k}}}[X_t - m(\underline{X}_t)]$ on small hypercubes $I_{\underline{k}}$ by their respective counterparts $\sum_{t:\underline{Y}_t \in I_{\underline{k}}} \eta_t$ .

The construction of the desired pairing of $(X_{1-l_d}, \ldots, X_T)$ and $(\underline{Y}_1, \ldots, \underline{Y}_T, \eta_1, \ldots, \eta_T)$ is based on a Skorokhod embedding of the random variables $v_{t,\underline{k}} = I(\underline{X}_t \in I_{\underline{k}})[X_t - m(\underline{X}_t)]$ as well as $\tilde{v}_{t,\underline{k}} = I(\underline{Y}_t \in I_{\underline{k}})\eta_t$ in a common set of independent Wiener processes $W_{\underline{k}}$. A similar method was used in Neumann and Kreiss (1997) to prove asymptotic equivalence of nonparametric estimators of the autoregression function in a nonparametric autoregressive model and analogous estimators in a usual regression model. In this paper we develop an embedding scheme which deviates from approaches that people would most probably first try in this context. Before we describe this method in detail, we explain in the next subsection why the seemingly most natural attempt fails.

**2.1. Failure of a natural attempt.** The ultimate goal of our construction of an embedding of the $X_t$ in the Wiener processes is to obtain an (at least approximate) representation

$$Z_{\underline{k}} = \sum_{t:\underline{X}_t \in I_{\underline{k}}} [X_t - m(\underline{X}_t)] \approx W_{\underline{k}}(\tau_{\underline{k}}). \tag{2.8}$$

In a similar manner, the $\underline{Y}_t$ and $\eta_t$ from model (2.3) will be embedded in the same Wiener processes, namely

$$\tilde{Z}_{\underline{k}} = \sum_{t:\underline{Y}_t \in I_{\underline{k}}} \eta_t = W_{\underline{k}}(\tilde{\tau}_{\underline{k}}). \tag{2.9}$$

Provided we can show that $|\tau_{\underline{k}} - \tilde{\tau}_{\underline{k}}|$ is small compared to the magnitude of either one of these stopping times, then most of the randomness of both partial sums is driven by the same stretch of $W_{\underline{k}}$. Hence, the difference between $Z_{\underline{k}}$ and $\tilde{Z}_{\underline{k}}$ is small as compared to the standard deviation of either one of these quantities.

It is quite natural to try to construct a representation of the $X_t$'s leading to (2.8) by a successive embedding of $I(\underline{X}_t \in I_{\underline{k}})[X_t - m(\underline{X}_t)]$ in the Wiener processes. We explain in this subsection why an embedding of the $X_t$ in their natural order fails. A successful embedding in reverse time order will be described in the next subsection. To simplify notation, we restrict our considerations in the next two subsections to the case of one lagged variable with $l_1 = 1$ . Assume for a moment that $\{X_t\}$ is a Markov chain. Then the observations obey the model

$$X_t = m(X_{t-1}) + \varepsilon_t,$$

where $E(\varepsilon_t \mid \mathcal{F}_{t-1}) \equiv 0$ , $\mathcal{F}_s = \sigma(X_0, \ldots, X_s)$ . Define $I_k = [(k-1)g, kg)$ and let $W_k$ be independent Wiener processes. Now we can embed the $\varepsilon_t$'s successively in the Wiener processes. Given $X_0$ falls into $I_{k_1}$, we can represent $\varepsilon_1$ by $W_{k_1}$ with the aid of a stopping time $\tau_1$, that is $\varepsilon_1 = W_{k_1}(\tau_1)$ . $\tau_1$ has to be chosen such that $\mathcal{L}(W_{k_1}(\tau_1)) = \mathcal{L}(\varepsilon_1 \mid X_0)$ . Such a representation is called Skorokhod embedding; cf. Hall and Heyde (1980, Appendix I). Since $E(\varepsilon_1 \mid X_0) \equiv 0$ , the stopping time $\tau_1$ has a certain number of finite moments, in dependence on the number of finite moments of $\varepsilon_1$. The next steps can be defined recursively. Assume that $\varepsilon_1, \ldots, \varepsilon_{t-1}$ are already embedded. By the strong Markov property, the remaining parts of the Wiener processes, $\{W_k(s + \sum_{u:u\le t, X_{u-1}\in I_k} \tau_k) - W_k(\sum_{u:u\le t, X_{u-1}\in I_k} \tau_k), s \ge 0\}$ , are again Wiener processes. Given $X_{t-1}$ falls into $I_{k_t}$, then we embed $\varepsilon_t$ in the remaining part of $W_{k_t}$, and so on. The same is done with the $\eta_t$'s which are embedded in the same set of Wiener processes by means of stopping times $\widetilde{\tau}_k$. Finally, we obtain that

$$Z_k \;=\; \sum_{t:X_{t-1}\in I_k} \varepsilon_t \;=\; W_k(\tau^{(k)}), \quad \text{where } \tau^{(k)} = \sum_{t:X_{t-1}\in I_k} \tau_t,$$

and

$$\widetilde{Z}_k \;=\; \sum_{t:Y_{t-1}\in I_k} \eta_t \;=\; W_k(\widetilde{\tau}^{(k)}), \quad \text{where } \widetilde{\tau}^{(k)} = \sum_{t:Y_{t-1}\in I_k} \widetilde{\tau}_t.$$

As it was shown in Neumann and Kreiss (1997), $\tau^{(k)}$ and $\widetilde{\tau}^{(k)}$ are close to their respective expectations and, moreover, these expectations coincide. Hence, most of the randomness of $Z_k$ and $\widetilde{Z}_k$ is driven by the *same* stretch of $W_k$, which finally leads to the closeness of $Z_k$ and $\widetilde{Z}_k$.

Now it is tempting to generalize this idea to our case of a general, not necessarily Markovian process. Since $E(I(X_{t-1} \in I_k)[X_t - E(X_t \mid X_{t-1})] \mid \mathcal{F}_{t-1})$ is in general not equal to 0, one might decompose the vector $v_t = (I(X_{t-1} \in I_k)[X_t - E(X_t \mid X_{t-1})])_k$ into sums of martingale differences, that is

(2.10)
$$v_t \;=\; [E(v_t \mid \mathcal{F}_t) \;-\; E(v_t \mid \mathcal{F}_{t-1})] + [E(v_t \mid \mathcal{F}_{t-1}) \;-\; E(v_t \mid \mathcal{F}_{t-2})] + \ldots$$

This is a well-known standard approach to derive central limit theorems for dependent random variables. In the case of only *one* interval $I_1$ we could indeed perform such an embedding. However, in our multivariate context with more than one interval, we run into serious problems with joint distributions. According to (2.10), at the transition from $\mathcal{F}_{t-1}$ to $\mathcal{F}_t$, we have to embed the vector $E(w_t \mid \mathcal{F}_t) - E(w_t \mid \mathcal{F}_{t-1})$ , where $w_t = v_t + v_{t+1} + \ldots$ , in the Wiener processes. The obstacle is that the components of $w_t$ are dependent in a manner which is difficult to handle. For example, the value of the "active" component of $v_t$ (which has the index $k_t$ where $X_{t-1} = I_{k_t}$) determines which component of $v_{t+1}$ will be active. A subsequent embedding of these components, as proposed by Kiefer (1972) for vectors with a martingale structure, is not possible since $E(v_{t+1} \mid \mathcal{F}_{t-1}, v_t)$ is nonzero in general. Moreover, although the conditional expectation $E(v_{t+1} \mid \mathcal{F}_{t-1})$ is of negligible order, the conditional expectation $E(v_{t+1} \mid \mathcal{F}_{t-1}, v_t)$ is of a nonnegligible order of magnitude. In view of these difficulties, we did not found an appropriate way to embed the vectors $w_t$ in the Wiener processes $W_k$.

**2.2. Backward embedding.** In order to present the essential ideas in an clear as a possible manner, we restrict our considerations again to the one-dimensional case. Moreover, we consider only a finite number of intervals $I_k = [(k-1)g, kg)$ , $k = 1, \ldots, K_T$ . In order to obtain a useful result, we let $g$ tend to 0, which leads to $K_T \to \infty$ as $T \to \infty$ . The generalization to the general case requires only a few technical modifications and is described in the proof of Theorem 2.1.
In contrast to the unsuccessful attempt of the previous subsection, it will turn out that an embedding in reverse time order does lead to a useful approximation. Define $\mathcal{G}_t = \sigma(X_t, \ldots, X_T)$ . At the transition from $\mathcal{G}_t$ to $\mathcal{G}_{t-1}$, we have to represent the vector $v_t = (v_{t,k})_{k=1,\ldots,K_T}$ by the Wiener processes $W_k$. Again, $E(v_{t,k} \mid \mathcal{G}_t)$ is not guaranteed to be 0. However, at least for a single $k$, we can embed the mean-corrected quantity $v_{t,k} - E(v_{t,k} \mid \mathcal{G}_t)$ in $W_k$. Under natural conditions concerning the boundedness of conditional moments of $v_{t,k}$ under $\mathcal{G}_t$, it turns out that

$$\sum_t E(v_{t,k} \mid \mathcal{G}_t) \;=\; O_P(g\sqrt{T}), \tag{2.11}$$

which is negligible compared to the stochastic fluctuations of $Z_k = \sum_t v_{t,k}$ that are of order $O_P(\sqrt{Tg})$. The simultaneous embedding of the whole vector will be described below.
Here is just the point where the essential difference to our attempt of a forward embedding becomes visible: When we embed the vector $v_t$ at the transition from $\mathcal{G}_t$ to $\mathcal{G}_{t-1}$, we have correction terms of order $O_P(g)$ for all components. These correction terms have an *unconditional* mean 0 and are weakly dependent, which leads to a total sum of order $O_P(g\sqrt{T})$. In contrast, at the transition from $\mathcal{F}_{t-1}$ to $\mathcal{F}_t$, the vector $v_t$ has only one nonzero component, $v_{t,k_t}$, where $k_t$ is defined by $X_{t-1} \in I_{k_t}$ . Accordingly, we need only a single correction term, now of order $O_P(1)$, for the $k_t$-th component of $v_t$. Although these correction terms are again weakly dependent with unconditional mean 0, they sum up to $\sum_t E(v_{t,k} \mid \mathcal{F}_{t-1}) = O_P(\sqrt{Tg})$ , which is no longer negligible.
With our backward embedding we are in a situation similar to density estimation from weakly dependent data. Besides the additional random factor $[X_t - E(X_t \mid X_{t-1})]$ , we have to determine the position of $X_{t-1}$. The above discussion sheds also some new light on a result of Neumann (1996) for the case of density estimation, where also a uniform version of the whitening by windowing principle was derived. There the link to the independent case was established via an embedding of the observations both from the time series model and from the i.i.d. model in a common Poisson process, which led to the even stronger result that the unordered sets of observations from both sets were almost identical. The perhaps more elegant application of such a Poisson embedding is not possible here since we have to deal with the additional factors $[X_t - E(X_t \mid X_{t-1})]$ .
Now we describe how the whole vector $v_t$ can be embedded in the Wiener processes $W_k$, $k = 1, \ldots, K_T$ . So far there does not exist an appropriate generalization of the Skorokhod embedding to random vectors with an arbitrary joint distribution. However, Kiefer (1972) developed such an embedding in the special case that the components of the vector form a martingale. We will use this idea as a starting point for the construction of an approximation of $v_t$ by the Wiener processes $W_k$. Remember that we are finally interested in a close connection of $Z_k = \sum_t I(X_{t-1} \in$

$I_k)[X_t - m(X_{t-1})]$ and $\widetilde{Z}_k = \sum_t I(Y_t \in I_k)\eta_t$ . Conditioned on $\mathcal{G}_t$, the vector $v_{t,k}$ does not have a martingale structure. However, we can appropriately generalize our mean-correction and obtain finally a sufficiently close approximation by a vector that can indeed be embedded.

Suppose that we have already defined $X_t, \ldots, X_T$ . Now we have to determine $X_{t-1}$ which will be attained by an embedding of $v_t$ in the $W_k$'s. We begin with the first component $v_{t,1}$. As described above, we represent instead of $v_{t,1}$ the mean-corrected quantity $v_{t,1} - E(v_{t,1} \mid \mathcal{G}_t)$ by means of an appropriate stopping time $\tau_1^{(t)}$ as

$$v_{t,1} - E(v_{t,1} \mid \mathcal{G}_t) \;=\; W_1\left(\tau_1^{(t)} + \tau_1^{(t+1)} + \ldots + \tau_1^{(T)}\right) \;-\; W_1\left(\tau_1^{(t+1)} + \ldots + \tau_1^{(T)}\right).$$

Suppose now that $v_{t,1}, \ldots, v_{t,k-1}$ have been defined with the aid of $W_1, \ldots, W_{k-1}$ , respectively. If $X_{t-1} \notin I_1 \cup \ldots \cup I_{k-1}$ , then we proceed with our embedding and represent the mean-corrected quantity

$$v_{t,k} - E(v_{t,k} \mid \mathcal{G}_t, X_{t-1} \notin I_1 \cup \ldots \cup I_{k-1})$$

with the aid of a stopping time $\tau_k^{(t)}$ as

$$W_k\left(\tau_k^{(t)} + \tau_k^{(t+1)} + \ldots + \tau_k^{(T)}\right) \;-\; W_k\left(\tau_k^{(t+1)} + \ldots + \tau_k^{(T)}\right).$$

If $X_{t-1} \in I_1 \cup \ldots \cup I_{k-1}$ , then we set $v_{t,k}$ equal to zero, that is we set $\tau_k^{(t)} = 0$ . This can be done for all $k = 1, \ldots, K_T$ .

After performing this procedure for all $t$ down to 1, we hope to get finally a significant approximation of $Z_k = \sum_t v_{t,k}$ by $W_k(\tau_k^{(1)} + \ldots + \tau_k^{(T)})$. The mean-corrected terms satisfy

$$\begin{aligned} & E\left(v_{t,k} \mid \mathcal{G}_t, X_{t-1} \notin I_1 \cup \cdots \cup I_{k-1}\right) \\ &= \frac{E(v_{t,k} \mid \mathcal{G}_t)}{P(X_{t-1} \notin I_1 \cup \cdots \cup I_{k-1} \mid \mathcal{G}_t)} \\ &= O\left(g / P(X_{t-1} \notin I_1 \cup \cdots \cup I_{k-1} \mid \mathcal{G}_t)\right). \end{aligned} \tag{2.12}$$

Since $v_{t,k}$ has an unconditional mean 0 one might expect that the sum of these terms satisfies

$$\sum_t E\left(v_{t,k} \mid \mathcal{G}_t, X_{t-1} \notin I_1 \cup \cdots \cup I_{k-1}\right) \;=\; O_P(g\sqrt{T}). \tag{2.13}$$

This would be indeed enough, since the stochastic fluctuations of $\sum_t v_{t,k}$ are of order $O_P(\sqrt{Tg})$. However, the right-hand side of (2.12) deteriorates as $k \to K_T$, since then $P(X_{t-1} \notin I_1 \cup \cdots \cup I_{k-1} \mid \mathcal{G}_t)$ becomes small. In order to keep all these mean-correction terms uniformly small, we use a simple modification: We introduce an additional bin, $I_{K_T+1}$, and generate $X_{t-1}$ according to the law $P'$, where

$$P'(X_{t-1} \in A) \;=\; P(X_{t-1} \in A \mid \mathcal{G}_t)/2$$

and $P'(X_{t-1} \in I_{K_T+1}) = 1/2$. Now it happens with a probability of $1/2$ that $X_{t-1}$ does not fall into one of the intervals $I_1, \ldots, I_{K_T}$ . In this case we just repeat the whole procedure once more, and so on. In a similar manner to the description above, we compose $\tau_k^{(t)}$ perhaps from more than one stopping times, say $\tau_{k,1}^{(t)}, \ldots, \tau_{k,r_t}^{(t)}$ , where $r_t$ is the number of trials needed to hit $I_1 \cup \cdots \cup I_{K_T}$ . The number of these loops has a geometric distribution with parameter $1/2$. We have, as above,

$v_{t,k} = W_k(\tau_k^{(t)} + \ldots + \tau_k^{(T)}) - W_k(\tau_k^{(t+1)} + \ldots + \tau_k^{(T)}) + O_P(g)$ . With this modification, we are able to show in the proof of Theorem 2.1 that (2.13) is indeed true up some logarithmic factor.
The random variables from the regression model (2.3) can be connected with the Wiener processes $W_k$ in an analogous manner, which leads to the desired strong approximation of the partial sums. As a result of this construction, we obtain $Z_k = \sum v_{t,k} = W_k(\tau_k) + O_P(g\sqrt{T})$ and $\tilde{Z}_k = \sum \tilde{v}_{t,k} = W_k(\tilde{\tau}_k)$ , where $\tau_k = \sum \tau_k^{(t)}$ and $\tilde{\tau}_k = \sum \tilde{\tau}_k^{(t)}$ .
Now it is easy to find an appropriate generalization to the case of more than one lags $l_1, \ldots, l_d$ as well as to the case of an infinite number of intervals $I_{\underline{k}}$. The corresponding modifications are described in the proof of Theorem 2.1.
Before we formalize this result by a theorem, we introduce two more assumptions. Remember that (2.11) requires some condition on the conditional distribution of $v_{t,k}$ under $\mathcal{G}_t$. Let $p_{X_{t-1}|G}$ be the conditional density of $X_t$ given the event $G$. Moreover, remember that $\underline{X}_t = (X_{t-l_1}, \ldots, X_{t-l_d})$ . We will assume

**(A2)** (i) $\sup_t \sup_{G \in \mathcal{G}_t} \sup_v \{p_{X_{t-1}|G}(v)\} \leq C$ ,
(ii) $\forall\ M < \infty\ \exists\ C_M < \infty$ such that

$$E(X_t - m(\underline{X}_t))^M \leq C_M,$$

(iii) $\sup_{\underline{x}} \left\{ \left| E(X_t \mid X_{t-l_1} = x_1, \ldots, X_{t-l_{d-1}} = x_{d-1}) - E(X_t \mid \underline{X}_t = \underline{x}) \right| \right\} \leq C.$

**(A3)** $K$ is Lipschitz and compactly supported.

In order to derive rates for our approximation rather than only consistency, but also for deriving uniform results from pointwise approximations, we will frequently use the fact that some remainder terms are smaller than certain bounds with a high probability. For notational convenience, we introduce the following notation:

**Definition 2.1.** Let $\{Z_T\}$ be a sequence of random variables and let $\{\alpha_T\}$ and $\{\beta_T\}$ be sequences of positive reals. We write

$$Z_T = \tilde{O}(\alpha_T, \beta_T),$$

if

$$P(|Z_T| > C\alpha_T) \leq C\beta_T \tag{2.14}$$

holds for $T \geq 1$ and some $C < \infty$ .

This definition is obviously stronger than the usual $O_P$ and it is well suited for our particular purpose of constructing confidence bands and critical values for tests; see the applications in Subsection 3.2.
Whenever we claim that $\tilde{O}$ holds uniformly over a certain set, we mean that (2.14) is true for a unique constant $C$. Moreover, we use the letter $C$ to denote any constants whose exact value is not important, and which may attain different values at different places. Here and in the following we make the convention that $\delta$ denotes a positive but arbitrarily small, and $\lambda$ an arbitrarily large constant.

**Theorem 2.1.** *Suppose that (A1) to (A3) are fulfilled. Furthermore, we assume that* $(Th_1 \cdots h_d)^{-1} = O(T^{-\delta})$ . *On an appropriate probability space, there exists a pairing of the random variables from (2.1) with those from (2.3) such that*

$$\sup_{\underline{x} \in \mathbb{R}^d} \left\{ \left| \sum_t w(\underline{x}, \underline{X}_t)[X_t - m(\underline{X}_t)] \; - \; \sum_t w(\underline{x}, \underline{Y}_t)\eta_t \right| \right\} \; = \; \tilde{O}\left( \sqrt{Th_1 \cdots h_d} \left[ \sqrt{h_d} \log T \; + \; T^{-\delta} \right], T^{-\lambda} \right).$$

Under quite natural assumptions on the bandwidth $h_d$, Theorem 2.1 provides a significant approximation: If

$$h_d \; = \; o((\log T)^{-2}), \tag{2.15}$$

then the error of approximation is below the level of pointwise fluctuations of $\sum w(\underline{x}, \underline{X}_t)[X_t - m(\underline{X}_t)]$ . Moreover, if

$$h_d \; = \; o((\log T)^{-3}), \tag{2.16}$$

then the error of approximation is below the level of fluctuations of $\sup_{\underline{x} \in \mathbb{R}^d} \{ | \sum w(\underline{x}, \underline{X}_t)[X_t - m(\underline{X}_t)] | \}$ .

There are some interesting implications from this approximation. First, on a more abstract level, it formalizes in some sense that nonparametric inference from weakly dependent data is asymptotically equivalent to nonparametric inference from i.i.d. data. For example, with some additional considerations, we immediately obtain the equivalence of risks of nonparametric estimators in both models. Second, it means that we can neglect the dependence beyond those within the blocks of observations of length $m$ when we intend to devise bootstrap methods for nonparametric statistics that depend only on a $m$-dimensional joint distribution. In particular, this delivers a justification for model-based bootstrap schemes which usually capture only some part of the dependence mechanism. To prevent possible misunderstandings, we do not propagate to neglect uncritically the whole dependence structure. Sometimes it needs quite large sample sizes to make this effect really significant. Therefore, it is certainly important to spend some efforts to capture the dependence structure as good as possible.

## 3. Finite order Markov chain bootstrap for general stationary processes

It is quite a popular practice to use semiparametric models in time series analysis. Such models can provide useful approximations to perhaps more complex processes if the dependence between the observations is rapidly decaying. Especially for moderate sample sizes, the application of such finite-dimensional models is a reasonable compromise between the two requirements of imitating the true dependence structure as good as possible and of avoiding the curse of dimensionality by too complex models.

On the other hand, with semiparametric models such as nonparametric autoregressive models or Markov chains of fixed order, a rather strong structural assumption on the dependence mechanism is imposed, whereas the distribution of the innovations

or the transition probabilities are modeled nonparametrically. Since it is rather unlikely in practical applications that the true data generating process actually obeys a (finite-dimensional) semiparametric model exactly, it is of considerable interest what happens with the validity of corresponding bootstrap methods. According to the uniform version of the whitening by windowing principle derived in the previous section, there is some hope that certain model-based bootstrap methods which capture only some part of the whole dependence mechanism remain valid for certain purposes in nonparametric statistics. In what follows we analyze a finite-order Markov chain bootstrap in the context of a general stationary process.

There is already an extensive literature on bootstrap methods for Markov chains. Kulperger and Prakasa Rao (1989) and Basawa, Green, McCormick and Taylor (1990) devised methods for finite state Markov chains. Athreya and Fuh (1992a, 1992b) considered the countable case. Furthermore, Rajarshi (1990) proposed a valid bootstrap for the case of a general state space based on nonparametric kernel estimators of the transition probabilities while Lall and Sharma (1996) and Paparoditis and Politis (1997) discussed a Markov chain bootstrap without explicit nonparametric estimation of the transition probabilities. Within this list of methods, the "nearest neighbors" to our proposal below are the methods of Lall and Sharma (1996) and Paparoditis and Politis (1997). However, whereas all of the above Markov chain bootstrap methods are derived under the assumption that the data generating process has indeed a Markovian structure, we do not impose any kind of structural assumptions and show the validity of our Markov chain bootstrap in this more general context.

Now we describe our bootstrap proposal in detail. Denote by $\pi_{(s_1,\dots,s_m)}$ the stationary distribution of $(X_{t-s_1},\dots,X_{t-s_m})$. Notice that the approximation of the distribution of $\widehat{m}_{\underline{h}}(\underline{x})$ requires at least a consistent reproduction of $\pi_{(0,l_1,\dots,l_d)}$. Hence, we have to generate a Markov chain of order at least $l_d$ which is based on reasonable estimates of the transition probabilities with respect to the lags $l_1,\dots,l_d$. Moreover, as can be seen from the proof of Proposition 3.3, the consistency of the stationary distribution requires that the Markov chain is based on lags that are consecutive multiples of a certain natural number.

According to this discussion, we take lags $r_1,\dots,r_\Delta$ such that

$$r_i = ir_1, \quad i = 1,\dots,\Delta, \tag{3.1}$$

and

$$\{l_1,\dots,l_d\} \subseteq \{r_1,\dots,r_\Delta\}. \tag{3.2}$$

We denote the vectors of lagged variables $(X_{t-r_1}, X_{t-r_2},\dots,X_{t-r_\Delta})$ and $(X^*_{t-r_1}, X^*_{t-r_2},\dots,X^*_{t-r_\Delta})$ by $\mathbb{X}_t$ and $\mathbb{X}^*_t$, respectively. Moreover, we use the symbols $\widetilde{x} = (x_1,\dots,x_\Delta)$ and $\widetilde{y} = (y_1,\dots,y_\Delta)$. To initialize the recursive scheme, we draw $(X^*_{1-r_\Delta}, X^*_{1-r_{\Delta-1}},\dots,X^*_{1-r_1})$, $(X^*_{2-r_\Delta}, X^*_{2-r_{\Delta-1}},\dots,X^*_{2-r_1})$, $\dots$, $(X^*_{r_1-r_\Delta}, X^*_{r_1-r_{\Delta-1}},\dots,X^*_0)$ independently, according to their stationary distribution $\pi^*_{(r_\Delta,\dots,r_1)}$.

Let $D(.,.) : \mathbb{R}^\Delta \times \mathbb{R}^\Delta \longrightarrow [0,\infty)$ be any distance function. Further, let $N_T$ be chosen such that $N_T \to \infty$ and $N_T/T \to 0$. Given $X^*_{1-l_d},\dots,X^*_{t-1}$, we draw $X^*_t$ with respective probabilities of $1/N_T$ from the set

$$\widehat{\mathcal{U}}_T(\mathbb{X}^*_t, N_T/T) = \{X_s \,|D(\mathbb{X}^*_t, \mathbb{X}_s) \le c_T\},$$

where $c_T = c_T(\mathbb{X}_t^*)$ is chosen such that $\{\dots\}$ contains exactly $N_T$ elements. This set is the empirical counterpart to $\mathcal{U}(\widetilde{x}, N_T/T)$, where $\mathcal{U}(\widetilde{x}, N_T/T) = \{\widetilde{y} \mid D(\widetilde{x}, \widetilde{y}) \le c_T\}$ and $c_T$ is chosen such that $P(\underline{X}_t \in \mathcal{U}(\widetilde{x}, c_T)) = N_T/T$ . Although other choices of $D$ are possible as well, we restrict our considerations to the case of

$$D(\widetilde{x}, \widetilde{y}) \;=\; \max_{1 \le i \le \Delta} \{|x_i - y_i|/f_i\}\,,$$

where $f_1, \dots, f_\Delta$ are certain bandwidths and $N_T = [T f_1 \cdots f_\Delta]$ .
Such a nearest neighbor bootstrap has already been considered on a heuristical level by Lall and Sharma (1996). Paparoditis and Politis (1997) proposed a similar version of a Markov chain bootstrap where the transition probabilities are determined by kernel weights. The nearest neighbor approach is an alternative, which circumvents the risk that conditional distributions deteriorate to one-point measures in regions of sparse data. A related idea of a local bootstrap has been used by Shi (1991), Rutherford and Yakowitz (1991) and Falk and Reiss (1992) in the regression context, in order to deal with conditional heteroscedasticity. Moreover, Paparoditis and Politis (1996) implemented such an idea in the frequency domain, for bootstrapping the periodogram.

**3.1. Some important properties of the bootstrap process.** In the following we intend to show some important properties of the bootstrap process $\{X_t^*\}$. First we prove the consistency of the transition probabilities with respect to the lags $r_1, \dots, r_\Delta$ . Then we intend to derive an appropriate mixing property for the bootstrap process. Such properties are important for the wide applicability of particular bootstrap methods, and have been the subject of recent research; see, for example, Rajarshi (1990) and Paparoditis and Politis (1997) for Markov chain bootstrap, Bickel and Bühlmann (1995) for a sieve bootstrap for linear processes, and Franke, Kreiss, Mammen and Neumann (1997) for a nonparametric autoregressive bootstrap. Finally, we show the consistency of the stationary distribution $\pi^*_{(r_1, r_2, \dots, r_\Delta)}$ for $\pi_{(r_1, r_2, \dots, r_\Delta)}$, which implies the consistency of $\pi^*_{(l_1, l_2, \dots, l_d)}$ for $\pi_{(l_1, l_2, \dots, l_d)}$.
Before we turn to an assertion about the consistency of the transition probabilities, we first state a useful lemma about the empirical process indexed by hyperrectangles of an $m$-dimensional stationary process.

**Lemma 3.1.** *Suppose that the $m$-dimensional random vectors $(Z_t)_{t=1,\dots,T}$ form a stationary, exponentially $\beta$-mixing process. Denote by $\mathcal{C}_m$ the set of all hyperrectangles in $\mathbb{R}^m$. Then*

$$P\left(\sup_{C \in \mathcal{C}_m} \left\{ \frac{|\#\{t \mid Z_t \in C\} \;-\; TP(Z_1 \in C)|}{\sqrt{TP(Z_1 \in C)}\log T \;+\; (\log T)^2} \right\} \;>\; C_\lambda \right) \;=\; O(T^{-\lambda}).$$

The important fact is that the supremum is inside the probability, that is with a probability exceeding $1 - O(T^{-\lambda})$ the deviations of $\#\{t \mid Z_t \in C\}$ from $TP(Z_1 \in C)$ can be *simultaneously* bounded by the above bounds.
In what follows we also assume

**(A4)** $\sup_{\widetilde{x}} \sup_{c,d} \{|P(X_t \in [c,d] \mid \mathbb{X}_t = \widetilde{x})|\} \leq C P(X_t \in [c,d])$.

**Proposition 3.1.** *Suppose that (A1) and (A4) are fulfilled. Then*

$$\sup_{\widetilde{x}} \sup_{c<d} \left\{ \frac{|P^*(X_t^* \in [c,d] \mid \mathbb{X}_t^* = \widetilde{x}) - P(X_t \in [c,d] \mid \mathbb{X}_t \in \mathcal{U}(\widetilde{x}, N_T/T))|}{\sqrt{P(X_t \in [c,d])} \log T/\sqrt{N_T} + (\log T)^2/N_T} \right\} = \widetilde{O}\left(1, T^{-\lambda}\right).$$

To keep the technicalities as simple as possible, we impose for the original process the following conditions:

**(A5)** There exists some interval $[c,d]$ such that the joint density of $X_{t-r_1}, \ldots, X_{t-r_\Delta}$ fulfills

$$p_{X_{t-r_1},\ldots,X_{t-r_\Delta}}(\widetilde{x}) \geq C > 0 \quad \text{for all } \widetilde{x} \in [c,d]^\Delta.$$

Moreover, there exists a constant $\eta > 0$ such that for all $\widetilde{x}, \widetilde{y} \in \mathbb{R}^\Delta$

$$\int_c^d \left( p_{X_t|\mathbb{X}_t=\widetilde{x}}(x) \wedge p_{X_t|\mathbb{X}_t=\widetilde{y}}(x) \right) dx \geq \eta \tag{3.3}$$

holds.

In order to have a property similar to (3.3) for the bootstrap process, we impose the following condition on the bandwidths $f_i$:

**(A6)** $T f_1 \cdots f_\Delta \min_i\{f_i\}/(\log T)^2 \longrightarrow \infty$.

It will be shown in the proof of the next proposition that the bootstrap process satisfies

$$\sup_{A_1, A_2 \in \sigma(X_{t-1}^*, \ldots)} \sup_B \left\{ \left| P\left( (X_{t+r_\Delta}^*, \ldots, X_{t+2r_\Delta-1}^*) \in B \mid A_1 \right) - P\left( (X_{t+r_\Delta}^*, \ldots, X_{t+2r_\Delta-1}^*) \in B \mid A_2 \right) \right| \right\} \leq 1 - \eta',$$

where $\eta' > 0$, for an appropriate set of events $(X_{1-l_d}, \ldots, X_T) \in \Omega_T$ with $P(\Omega_T^c) = O(T^{-\lambda})$. This will imply uniform mixing ($\phi$-mixing) for the bootstrap process.
The $\phi$-mixing coefficients of a process $Z_1, Z_2, \ldots$ are defined as

$$\phi(n) = \sup_t \sup_{\substack{U \in \sigma(Z_1, \ldots, Z_t),\ P(U)>0 \\ V \in \sigma(Z_{t+n}, \ldots)}} \{|P(V) - P(V \mid U)|\}.$$

The next proposition states the announced mixing property of the bootstrap process, which in particular implies absolute regularity.

**Proposition 3.2.** *Suppose that (A1) and (A4) to (A6) are fulfilled. Then there exists a constant $\rho < 1$ such that the $\phi$-mixing coefficients of the bootstrap process satisfy*

$$\phi(n) \leq C\rho^n \quad \textit{for all } n,$$

*provided* $(X_{1-l_d}, \ldots, X_T) \in \Omega_T$ *for some appropriate set* $\Omega_T$ *with* $P(\Omega_T^c) = O(T^{-\lambda})$ .

Proposition 3.1 provides information about the size of the fluctuations of $P^*(X_t^* \in [c,d] \mid \mathbb{X}_t^* = \widetilde{x})$ about a smoothed version of the original transition probabilities. To get significant results for the smoothing bias, we have to impose certain conditions on the smoothness of the transition probabilities as functions in the lagged variables. The necessary strength of such conditions depends on the size of the neighborhoods $\mathcal{U}(\widetilde{x}, N_T/T)$. If, for example, the density of $\mathbb{X}_t$ is bounded away from 0 on a set $K$, then

$$\sup_{\widetilde{x} \in K} \{\mathrm{diam}(\mathcal{U}(\widetilde{x}, N_T/T))\} \; = \; O\left(\max_i \{f_i\}\right).$$

Hence, we obtain under Lipschitz continuity of the transition densities in the lagged variables that

(3.4)

$$\sup_{\widetilde{x} \in K} \{|P(X_t \in [c,d] \mid \mathbb{X}_t = \widetilde{x}) \; - \; P(X_t \in [c,d] \mid \mathbb{X}_t \in \mathcal{U}(\widetilde{x}, N_T/T))|\} \; = \; O\left((d-c)\max_i \{f_i\}\right).$$

Rajarshi (1990) states the consistency of the estimated transition probabilities just under the condition that the stationary density of $\mathbb{X}_t$ is bounded away from zero on a certain set $K$. Without this somewhat restrictive condition, one may develop an analogous asymptotics on growing sets $K_T$, where the stationary density is supposed to fulfill

$$\inf_{\widetilde{x} \in K_T} \left\{p_{X_{t-r_1}, \ldots, X_{t-r_\Delta}}(\widetilde{x})\right\} \; \geq \; T^{-\rho};$$

see Remark 2.1 in Rajarshi (1990) and Remark 2.4 in Paparoditis and Politis (1997). In order to avoid some nasty technicalities, we adapt the smoothness condition for the transition probabilities directly to the size of $\mathcal{U}(\widetilde{x}, N_T/T)$. For the sake of further simplification, we focus on the special case of $f_1 = \ldots = f_\Delta$ . We will assume

**(A7)**

$$\sup_{\widetilde{x}} \sup_{c,d} \{|P(X_t \in [c,d] \mid \mathbb{X}_t = \widetilde{x}) \; - \; P(X_t \in [c,d] \mid \mathbb{X}_t \in \mathcal{U}(\widetilde{x}, p))|\} \; \leq \; CP(X_t \in [c,d])p^{1/\Delta}.$$

Notice that the exponent $1/\Delta$ on the right-hand side reflects the fact that $\mathcal{U}(\underline{x}, p)$ is a hypercube rather than merely a hyperrectangle.

**Proposition 3.3.** *Suppose that (A1) and (A4) to (A7) are fulfilled. Then*

$$\begin{aligned}
&\left|\pi_{(r_1,\ldots,r_\Delta)}([a_1,b_1] \times \cdots \times [a_\Delta, b_\Delta]) \; - \; \pi^*_{(r_1,\ldots,r_\Delta)}([a_1,b_1] \times \cdots \times [a_\Delta, b_\Delta])\right| \\
&= \; O\Bigg([(N_T/T)^{1/\Delta} \; + \; (\log T/\sqrt{N_T})^{2/3}] \prod_{i=1}^{\Delta} [P(X_t \in [a_i,b_i]) \; + \; (\log T)^2/N_T] \\
&\qquad + \sum_{i=1}^{\Delta} \left[\sqrt{F(b_i) - F(a_i)}\frac{\log T}{\sqrt{N_T}} \; + \; \frac{(\log T)^2}{N_T}\right] \prod_{j \neq i} \left[P(X_t \in [a_j,b_j]) \; + \; \frac{(\log T)^2}{N_T}\right]\Bigg)
\end{aligned}$$

*holds uniformly in* $(X_{1-l_d},\dots,X_T)\in\Omega_T$ *for some appropriate set* $\Omega_T$ *with* $P(\Omega_T^c)=O(T^{-\lambda})$ .

The proof of this proposition runs as follows. First, we consider a Markov chain of order $r_\Delta$, $\{X'_t\}$, with transition probabilities

$$P\left(X'_t\in A\mid (X'_{t-r_1},\dots,X'_{t-r_\Delta})\in B\right) \;=\; \pi_{(0,r_1,\dots,r_\Delta)}(A\times B)/\pi_{(r_1,\dots,r_\Delta)}(B).$$

It is easy to see that the stationary distribution of $(X_{t-r_1},\dots,X_{t-r_\Delta})$, $\pi_{(r_1,\dots,r_\Delta)}$, is equal to the stationary distribution of $(X'_{t-r_1},\dots,X'_{t-r_\Delta})$. To prove the closeness of $\pi_{(r_1,\dots,r_\Delta)}$ to $\pi^*_{(r_1,\dots,r_\Delta)}$, we start both chains with the same random sequence according to the stationary distribution of $(X'_1,\dots,X'_{r_\Delta})$. Then we study the "decoupling" of the two Markov chains $\{X'_t\}$ and $\{X^*_t\}$: Since $(X^*_{t-r_1},\dots,X^*_{t-r_\Delta})$ reaches its own stationary distribution with an exponential rate, the distribution of $(X^*_{t_0-r_1},\dots,X^*_{t_0-r_\Delta})$ is sufficiently close to $\pi^*_{(r_1,\dots,r_\Delta)}$, for some $t_0\asymp\log T$ . On the other hand, since by Proposition 3.1 the transition probabilities of $\{X'_t\}$ and $\{X^*_t\}$ are quite similar, we can find a pairing of both chains such that $P((X'_{t_0-r_1},\dots,X'_{t_0-r_\Delta})\neq(X^*_{t_0-r_1},\dots,X^*_{t_0-r_\Delta}))$ is still small. This gives finally the desired upper bound for the difference between $\pi_{(r_1,\dots,r_\Delta)}$ and $\pi^*_{(r_1,\dots,r_\Delta)}$.

**3.2. Application to nonparametric supremum-type tests.** Theorem 2.1 and Propositions 3.1 to 3.3 imply that the Markov chain bootstrap consistently estimates the pointwise distribution as well as those of supremum-type functionals of nonparametric estimators of the conditional mean function. Whereas the pointwise case can be tackled in a straightforward manner, one may develop theory for simultaneous confidence bands and supremum-type tests analogously to Neumann and Kreiss (1997) and Neumann (1996).
We allow a composite hypothesis, that is

$$H_0: \quad m\in\mathcal{M},$$

where the only requirement is that the function class $\mathcal{M}$ allows a faster rate of convergence than the nonparametric model. We will assume that

**(A8)** There exists an estimator $\widehat{\widehat{m}}$ of $m$ such that

(3.5)
$$\sup_{\underline{x}\in\mathbb{R}^d}\left\{\left|\sum_{t=1}^T w(\underline{x},\underline{X}_t)\left[\widehat{\widehat{m}}(\underline{X}_t)-m(\underline{X}_t)\right]\right|\right\} \;=\; o_P\left(\sqrt{Th_1\cdots h_d}(\log T)^{-1/2}\right),$$

where, as in Section 2, $w(\underline{x},\underline{y})=K((x_1-y_1)/h_1)\cdots K((x_d-y_d)/h_d)$ .

A sufficient condition for (A8) is obviously that $\widehat{\widehat{m}}$ itself converges in the supremum norm to $m$ with a faster rate than $(Th_1\cdots h_d)^{-1/2}(\log T)^{-1/2}$, which can be expected to hold in certain parametric models, $\mathcal{M}=\{m_\theta\mid\theta\in\Theta\}$. For the particular purpose of testing, there is no reason to use an explicit nonparametric estimator of $m$. Rather,

one may choose the test statistic under the aspect of convenience, for example,

$$W_T = \sup_{\underline{x}\in\mathbb{R}^d} \left\{ \left| \sum_{t=1}^{T} w(\underline{x}, \underline{X}_t) \left[ X_t - \widehat{\widehat{m}}(\underline{X}_t) \right] \right| \right\}. \tag{3.6}$$

This roughly corresponds to a contrast function which weights the difference between $m$ and $\mathcal{M}$ with a factor proportional to the stationary density $\pi_{(l_1,\ldots,l_d)}$.
Let $t_\alpha$ be the $(1-\alpha)$-quantile of the (random) distribution of

$$W_T^* = \sup_{\underline{x}\in\mathbb{R}^d} \left\{ \left| \sum_{t=1}^{T} w(\underline{x}, \underline{X}_t^*) \left[ X_t^* - E(X_t^* \mid \underline{X}_t^*) \right] \right| \right\}.$$

Suppose now that (A1) to (A7) are fulfilled. It can be shown that, for arbitrary $m \in \mathcal{M}$,

$$P_m \left( W_T > t_\alpha \right) = 1 - \alpha + o(1).$$

The Markov chain bootstrap can also be used for the construction of simultaneous confidence bands. There are several options to deal with the usual bias problem. To get an asymptotically bias-free situation like under the null hypothesis in testing, one may establish a confidence band for a smoothed version of $m$, $\sum w(\underline{x}, \underline{X}_t) m(\underline{X}_t) / \sum w(\underline{x}, \underline{X}_t)$ . To get confidence bands directly for $m$, one may use an undersmoothed estimator for $m$ or apply a subsequent explicit bias correction. A more detailed discussion of these issues can be found, for example, in Neumann and Kreiss (1997) and Neumann (1996).

## Appendix

**Proof of Theorem 2.1.** First we describe the necessary modifications of the construction explained in Subsection 2.2. Then we turn to the analytical part of the proof and develop estimates for the error terms that occur in our construction.

(i) *Modifications of the construction*

In order to avoid problems with an infinite number of hyperrectangles $I_{\underline{k}}$, we focus our primary attention to points $\underline{x}$ from the set

$$\mathcal{X}_0 = \left\{ \underline{x} \,\middle|\, P(\underline{X}_t \in \operatorname{supp}(w(\underline{x}, .))) \geq T^{-1} \right\}. \tag{A.1}$$

It is easy to see that $\mathcal{X}_0$ can be covered by a finite number of hyperrectangles, $\{ I_{\underline{k}} \mid \underline{k} = \underline{k}_1, \ldots, \underline{k}_{c_T} \}$ with $c_T = O(T^\kappa)$ for some constant $\kappa$. The indices of the remaining intervals are combined to disjoint sets $\mathcal{K}_1, \ldots, \mathcal{K}_{d_T}$ such that

$$T^{-1} \leq P\left( \underline{X}_t \in \bigcup_{\underline{k}\in\mathcal{K}_i} I_{\underline{k}} \right) \leq 2T^{-1}.$$

We set

$$(I_1, \ldots, I_{K_T}) = (I_{\underline{k}_1}, \ldots, I_{\underline{k}_{c_T}}, \bigcup_{\underline{k}\in\mathcal{K}_1} I_{\underline{k}}, \ldots, \bigcup_{\underline{k}\in\mathcal{K}_{d_T}} I_{\underline{k}}). \tag{A.2}$$

Another modification concerns the time point at which an appropriate approximation to $v_t = (I(\underline{X}_t \in I_1)[X_t - m(\underline{X}_t)], \ldots, I(\underline{X}_t \in I_{K_T})[X_t - m(\underline{X}_t)])$ is embedded in $(W_k)_{k=1,\ldots,K_T}$. This is just done when the most lagged variable defining $v_t$, $X_{t-l_d}$, has to be determined, that is with the transition from $\mathcal{G}_{t-l_d+1}$ to $\mathcal{G}_{t-l_d}$. According to the description in Subsection 2.2, we embed $v_{t,k} - E(v_{t,k} \mid \mathcal{G}_{t-l_d+1}, X_{t-l_d} \notin I_1 \cup \ldots \cup I_{k-1})$ in the remaining part of $W_k$. This has to be done for the first time at the transition from $\mathcal{G}_{T-l_d+1}$ to $\mathcal{G}_{T-l_d}$. Before that we generate $(X_{T-l_d+1}, \ldots, X_T)$ according to the $l_d$-dimensional stationary distribution. Again as in Subsection 2.2, we introduce an additional bin $I_{K_T+1}$ and generate $X'_{T-l_d,1}, X'_{T-l_d,2}, \ldots$ with

$$P\left(X'_{T-l_d,i} \in A \mid \mathcal{G}_{T-l_d+1}\right) = P\left(X_{T-l_d} \in A \mid \mathcal{G}_{T-l_d+1}\right)/2,$$

and $P(X'_{T-l_d,i} \in I_{K_T+1}) = 1/2$. If $X'_{T-l_d,1}$ falls into $I_1 \cup \ldots \cup I_{K_T}$, then we set $X_{T-l_d} = X'_{T-l_d,1}$. Otherwise we repeat this procedure until $X'_{T-l_d,i} \in I_1 \cup \ldots \cup I_{K_T}$ is achieved. Then we apply the same method to embed successively $X_{T-l_d-1}, \ldots, X_{1-l_d}$.

Let $r_t$ be the number of trials needed to get the event $X_{t-l_d,i} \in I_1 \cup \ldots \cup I_{K_T}$. Then $X_{t-l_d} = X'_{t-l_d,r_t}$ and $\tau_k^{(t)} = \sum_{i=1}^{r_t} \tau_{k,i}^{(t)}$, where $\tau_{k,i}^{(t)}$ is the stopping time connected with $X'_{t-l_d,i}$.

(ii) *Embedding of the* $(\underline{Y}_t, \eta_t)$

In principle, this embedding could be performed in the same manner as above. However, since the vectors $(\underline{Y}_t, \eta_t)$ are independent, we can proceed in a much simpler way. Assume that $(\underline{Y}_1, \eta_1), \ldots, (\underline{Y}_{t-1}, \eta_{t-1})$ are already embedded. Given $\underline{Y}_t$ falls into $I_{k_t}$, then we represent $\eta_t$ by the remaining part of $W_{k_t}$, $\{W_{k_t}(s + \sum_{u:u\le t-1} \widetilde{\tau}_{k_t}^{(u)}) - W_{k_t}(\sum_{u:u\le t-1} \widetilde{\tau}_{k_t}^{(u)}), s \ge 0\}$, with the aid of a stopping time $\widetilde{\tau}_{k_t}^{(t)}$. If $\underline{Y}_t \notin I_k$, we set $\widetilde{\tau}_k^{(t)} = 0$. As a result, we get $\widetilde{Z}_k = W_k(\widetilde{\tau}_k)$, where $\widetilde{\tau}_k = \sum \widetilde{\tau}_k^{(t)}$.

(iii) *Difference of* $\tau_k$ *and* $\widetilde{\tau}_k$

Our estimate of the difference between $Z_k$ and $\widetilde{Z}_k$ will be based on upper estimates of the difference between $\tau_k$ and $\widetilde{\tau}_k$.

Remember that we have to generate successively independent copies of $X'_{t-l_d}$, say $X'_{t-l_d,i}$, until the first of these copies falls into the target set $I_1 \cup \ldots \cup I_{K_T}$. Let $\mathcal{G}_{t,1} = \mathcal{G}_t$ and $\mathcal{G}_{t,i} = \sigma\left(X_t, \ldots, X_T, X'_{t-1,1}, \ldots, X'_{t-1,i-1}\right)$. Then

$$v_{t,k} - \sum_{i=1}^{r_t} E\left(v_{t,k} \mid \mathcal{G}_{t-l_d+1,i}\right) = W_k\left(\tau_k^{(t)} + \tau_k^{(t+1)} + \ldots + \tau_k^{(T)}\right) - W_k\left(\tau_k^{(t+1)} + \ldots + \tau_k^{(T)}\right).$$

Since the event $\{\omega \mid r_t \geq i\}$ is $\mathcal{G}_{t-l_d+1,i}$-measurable, we obtain

$$\begin{aligned}
E\tau_k^{(t)} &= \sum_{i=1}^{\infty} E\left[E\left(I(r_t \geq i)\tau_{k,i}^{(t)} \mid \mathcal{G}_{t-l_d+1,i}\right)\right] \\
&= \sum_{i=1}^{\infty} E\left[I(r_t \geq i)E\left(\tau_{k,i}^{(t)} \mid \mathcal{G}_{t-l_d+1,i}\right)\right] \\
&= \sum_{i=1}^{\infty} 2^{-(i-1)} E\left(\tau_{k,1}^{(t)} \mid \mathcal{G}_{t-l_d+1}\right) \\
&= 2EE\left(\tau_{k,1}^{(t)} \mid \mathcal{G}_{t-l_d+1}\right) \\
&= 2EE\left(\left\{[X_t - m(\underline{X}_t)]I(X_{t-l_1} \in I_{k_1}) \cdots I(X_{t-l_{d-1}} \in I_{k_{d-1}})I(X'_{t-l_d,1} \in I_{k_d})\right\}^2 \middle| \mathcal{G}_{t-l_d+1}\right) \\
&\quad - 2E\left\{E\left([X_t - m(\underline{X}_t)]I(X_{t-l_1} \in I_{k_1}) \cdots I(X_{t-l_{d-1}} \in I_{k_{d-1}})I(X'_{t-l_d,1} \in I_{k_d}) \middle| \mathcal{G}_{t-l_d+1}\right)\right\}^2 \\
&= Ev_{t,k}^2 - 2R_{t,k}.
\end{aligned} \tag{A.3}$$

Furthermore, we have

$$\begin{aligned}
R_{t,k} &= E\left\{I(X_{t-l_1} \in I_{k_1}) \cdots I(X_{t-l_{d-1}} \in I_{k_{d-1}})E\left([X_t - m(\underline{X}_t)]I(X'_{t-l_d} \in I_{k_d}) \middle| \mathcal{G}_{t-l_d+1}\right)\right\}^2 \\
&= EI(X_{t-l_1} \in I_{k_1}) \cdots I(X_{t-l_{d-1}} \in I_{k_{d-1}})\left\{E\left([X_t - m(\underline{X}_t)]I(X'_{t-l_d} \in I_{k_d}) \middle| \mathcal{G}_{t-l_d+1}\right)\right\}^2 \\
&\leq 2E\left\{I(X_{t-l_1} \in I_{k_1}) \cdots I(X_{t-l_{d-1}} \in I_{k_{d-1}})[X_t - E(X_t \mid X_{t-l_1}, \ldots, X_{t-l_{d-1}})]^2 \times \right. \\
&\qquad \left. \times [P(X'_{t-l_d} \in I_{k_d} \mid \mathcal{G}_{t-l_d+1})]^2\right\} \\
&\quad + 2E\left\{I(X_{t-l_1} \in I_{k_1}) \cdots I(X_{t-l_{d-1}} \in I_{k_{d-1}}) \times \right. \\
&\qquad \left. \times \left[E\left([E(X_t \mid X_{t-l_1}, \ldots, X_{t-l_{d-1}}) - m(\underline{X}_t)]I(X'_{t-l_d} \in I_{k_d}) \mid \mathcal{G}_{t-l_d+1}\right)\right]^2\right\} \\
&= O\left(g_1 \cdots g_{d-1} g_d^2\right).
\end{aligned} \tag{A.4}$$

Since

$$E\widetilde{\tau}_k^{(t)} = Ev_{t,k}^2$$

is obviously fulfilled, we obtain from (A.3) and (A.4) that

$$E\tau_k - E\widetilde{\tau}_k = O\left(Tg_1 \cdots g_{d-1} g_d^2\right). \tag{A.5}$$

To derive an upper estimate for the deviation of $\tau_k$ from its mean, we intend to use Bernstein's inequality. The necessary reduction to sums of independent random variables is achieved by a well-known blocking technique. We consider *overlapping* blocks of indices,

$$\mathcal{J}_i = \{(i-1)\rho_T - l_d + 1, \ldots, i\rho_T\}, \quad i = 1, \ldots, [(T-1)/\rho_T + 1],$$

where $\rho_T = [C_\lambda \log T]$ . Now we split the sum over $t$ into sums of blocks with odd numbers and sums of blocks with even numbers.

Without loss of generality, we consider the blocks with odd numbers. By Proposition 2 in Doukhan, Massart and Rio (1995), we can successively replace the blocks $\{X_t, t \in \mathcal{J}_i\}$, $i$ odd, by independent blocks $\{X'_t, t \in \mathcal{J}_i\}$, $i$ odd, with the property

(A.6)
$$P\left((X'_t, t \in \mathcal{J}_i) \neq (X_t, t \in \mathcal{J}_i) \quad \text{for any odd } \ i \leq [(T-1)/\rho_T + 1]\right) \ = \ O(T^{-\lambda}),$$

where the value of $\lambda$ may be chosen arbitrarily large, in dependence on $C_\lambda$.
After this reduction to the independent case, we will obtain the assertion from Bernstein's inequality, which we quote for reader's convenience from Shorack and Wellner (1986, p. 855):
Let $U_1, \ldots, U_n$ be independent random variables with $EU_i = 0$ and $|U_i| \leq K_n$ almost surely. Then, for $U = \sum U_i$ ,

$$\begin{aligned} P(U > c) \ &\leq \ \exp\left(-\frac{c^2/2}{\mathrm{var}(U) + (K_n c)/3}\right) \\ &\leq \ \exp\left(-\frac{c^2}{4\,\mathrm{var}(U)}\right) \ + \ \exp\left(-\frac{3c}{4K_n}\right) \end{aligned}$$

holds for arbitrary $c > 0$ .
Setting

$$c_\lambda \ = \ \sqrt{\mathrm{var}(U)}\sqrt{4\lambda \log(n)} \ + \ (4/3)K_n \lambda \log(n)$$

we get

$$P\left(|U| > c_\lambda\right) \ \leq \ 4\exp(-\lambda \log(n)).$$

In other words, we have that

(A.7)
$$U \ = \ \widetilde{O}\left(\sqrt{\mathrm{var}(U)}\sqrt{\log(n)} \ + \ K_n \log(n), n^{-\lambda}\right).$$

Instead of $(\tau_k^{(t)} - E\tau_k^{(t)})$ we consider the truncated random variables

$$\mu_{k,t} \ = \ (\tau_k^{(t)} - E\tau_k^{(t)}) I(|\tau_k^{(t)} - E\tau_k^{(t)}| < T^\delta).$$

Since all moments of $(\tau_k^{(t)} - E\tau_k^{(t)})$ are bounded, it follows from Markov's inequality that

(A.8)
$$P\left(\mu_{t,k} \neq (\tau_k^{(t)} - E\tau_k^{(t)})\right) \ = \ O(T^{-\lambda}).$$

Moreover, we have

$$\begin{aligned} \mathrm{var}\left(\sum_{t \in \mathcal{J}_i} \mu_{t,k}\right) \ &\leq \ \rho_T \sum_{t \in \mathcal{J}_i} \mathrm{var}(\mu_{t,k}) \\ &\leq \ \rho_T \sum_{t \in \mathcal{J}_i} E(\tau_k^{(t)} - E\tau_k^{(t)})^2 \ = \ O(\rho_T^2 g_1 \cdots g_d). \qquad \text{(A.9)} \end{aligned}$$

We obtain from (A.7) to (A.9)

(A.10)
$$\sum_{i \ odd} \sum_{t \in \mathcal{J}_i} (\tau_k^{(t)} - E\tau_k^{(t)}) \ = \ O\left(\sqrt{T g_1 \cdots g_d} \log T \ + \ T^\delta, T^{-\lambda}\right),$$

and, analogously,

(A.11)
$$\sum_i \sum_{even\ t\in\mathcal{J}_i} (\tau_k^{(t)} - E\tau_k^{(t)}) \;=\; O\left(\sqrt{Tg_1\cdots g_d}\log T \;+\; T^{\delta}, T^{-\lambda}\right).$$

Again by (A.7) and (A.8), we get

(A.12)
$$\sum_t (\widetilde{\tau}_k^{(t)} - E\widetilde{\tau}_k^{(t)}) \;=\; O\left(\sqrt{Tg_1\cdots g_d}\sqrt{\log T} \;+\; T^{\delta}, T^{-\lambda}\right).$$

(iv) *Conclusion for the difference of the weighted sums*

Let $\underline{x} \in \mathcal{X}_0$ . Recall that, according to (A.2), $\mathrm{supp}(w(\underline{x},.)) \subseteq \bigcup_{k=1}^{c_T} I_k$ . Define $w_k = |I_k|^{-1}\int_{I_k} w(\underline{x},\underline{y})d\underline{y}$ . We consider the following decomposition of the approximation error:

(A.13)
$$\begin{aligned}
&\sum_t w(\underline{x},\underline{X}_t)[X_t - m(\underline{X}_t)] \;-\; \sum_t w(\underline{x},\underline{Y}_t)\eta_t \\
&= \sum_k \sum_{t:\underline{X}_t\in I_k} \{w(\underline{x},\underline{X}_t) - w_k(\underline{x})\}[X_t - m(\underline{X}_t)] \\
&\quad + \sum_k w_k(\underline{x}) \left\{ \sum_{t:\underline{X}_t\in I_k} [X_t - m(\underline{X}_t)] \right\} \;-\; W_k(\tau_k) \\
&\quad + \sum_k w_k(\underline{x}) \{W_k(\tau_k) - W_k(E\tau_k)\} \\
&\quad + \sum_k w_k(\underline{x}) \{W_k(E\tau_k) - W_k(E\widetilde{\tau}_k)\} \\
&\quad + \sum_k w_k(\underline{x}) \{W_k(E\widetilde{\tau}_k) - W_k(\widetilde{\tau}_k)\} \\
&\quad + \sum_k \sum_{t:\underline{Y}_t\in I_k} \{w_k(\underline{x}) - w(\underline{x},\underline{Y}_t)\}\eta_t \\
&= T_1(\underline{x}) \;+\; \ldots \;+\; T_6(\underline{x}).
\end{aligned}$$

Since the kernel is Lipschitz continuous, we have $\left|w(\underline{x},\underline{y}) - w(\underline{x},\underline{z})\right| \;\le\; C\sum |y_i - z_i|/h_i$ , which implies, for *fixed* $\underline{x} \in \mathcal{X}_0$ ,

(A.14)
$$T_1(\underline{x}) \;+\; T_6(\underline{x}) \;=\; \widetilde{O}\left(\max_i\{g_i/h_i\}\sqrt{Th_1\cdots h_d}\log T \;+\; T^{\delta}, T^{-\lambda}\right).$$

Again for fixed $\underline{x} \in \mathcal{X}_0$ , we obtain, by making use of the blocking technique

(A.15)
$$\begin{aligned}
T_2(\underline{x}) &= \sum_k w_k(\underline{x}) \sum_{i=1}^{r_t} E(v_{t,k} \mid \mathcal{G}_{t-l_d+1,i}) \\
&= \widetilde{O}\left(\sqrt{Th_1\cdots h_{d-1}}h_d \log T \;+\; T^{\delta}, T^{-\lambda}\right).
\end{aligned}$$

By Lemma 1.2.1 in Csörgő and Révész (1981, p. 29) we get, in conjunction with (A.10) to (A.12),

$$\left|W_{\underline{k}}(\tau_{\underline{k}}) - W_{\underline{k}}(E\tau_{\underline{k}})\right| \, + \, \left|W_{\underline{k}}(\widetilde{\tau}_{\underline{k}}) - W_{\underline{k}}(E\widetilde{\tau}_{\underline{k}})\right| \, = \, \widetilde{O}\left((Tg_1 \cdots g_d)^{1/4} \log T \, + \, T^{\delta}, T^{-\lambda}\right),$$

which yields that

(A.16)

$$\sup_{\underline{x}} \{|T_3(\underline{x})| \, + \, |T_5(\underline{x})|\} \, = \, \widetilde{O}\left(\frac{h_1 \cdots h_d}{g_1 \cdots g_d}\left[(Tg_1 \cdots g_d)^{1/4} \log T \, + \, T^{\delta}\right], T^{-\lambda}\right).$$

Finally, we obtain by (A.5), for fixed $\underline{x} \in \mathcal{X}_0$ ,

$$T_4(\underline{x}) \, = \, \widetilde{O}\left(\sqrt{Th_1 \cdots h_d}\sqrt{g_d}\sqrt{\log T} \, + \, T^{\delta}, T^{-\lambda}\right). \tag{A.17}$$

By proving (A.14), (A.15) and (A.17) on a sufficiently fine grid, we obtain that these results remain true uniformly over $\underline{x} \in \mathcal{X}_0$ . Moreover, we can choose the $g_i$ in such a way that

$$\begin{aligned} &\sup_{\underline{x} \in \mathcal{X}_0} \{|T_1(\underline{x})| \, + \, \ldots \, + \, |T_6(\underline{x})|\} \\ &= \, \widetilde{O}\left(\sqrt{Th_1 \cdots h_d}\left[\max\{g_i/h_i\} \, + \, \frac{h_1 \cdots h_d}{g_1 \cdots g_d}\frac{(Tg_1 \cdots g_d)^{1/4} \, + \, T^{\delta}}{(Th_1 \cdots h_d)^{1/2}} \, + \, \sqrt{h_d}\right] \log T, T^{-\lambda}\right) \\ &= \, \widetilde{O}\left(\sqrt{Th_1 \cdots h_d}\left[\sqrt{h_d} \log T + T^{-\delta}\right], T^{-\lambda}\right). \end{aligned} \tag{A.18}$$

Finally, since $\#\{t \mid \underline{X}_t \in \mathcal{K}_i\} + \#\{t \mid \underline{Y}_t \in \mathcal{K}_i\} = O(\log T, T^{-\lambda})$ , we get

$$\sup_{\underline{x} \in \mathbb{R}^d \setminus \mathcal{X}_0} \left\{\left|\sum w(\underline{x}, \underline{X}_t)[X_t - m(\underline{X}_t)] \, - \, \sum w(\underline{x}, \underline{Y}_t)\eta_t\right|\right\} \, = \, \widetilde{O}\left(T^{\delta}, T^{-\lambda}\right), \tag{A.19}$$

which completes the proof. □

**Proof of Lemma 3.1.** (i) *Reduction to the case of independent rv's*

To handle the dependence, we consider instead of the whole set of rv's $\{Z_1, \ldots, Z_T\}$ $\rho_T$ subsets, $\{Z_t, t \in \mathcal{J}_i\}$ , where $\mathcal{J}_i = \{i, \rho_T + i, 2\rho_T + i, \ldots\} \cap \{1, \ldots, T\}$ . According to Proposition 2 in Doukhan et al. (1995), there exist sequences of independent random vectors, $\{Z'_t, t \in \mathcal{J}_i\}$ , such that

$$\mathcal{L}(Z'_t) \, = \, \mathcal{L}(Z_t)$$

and

$$P\left(Z_t \neq Z'_t \quad \text{for any} \quad t \in \mathcal{J}_i\right) \, = \, O(T^{-\lambda})$$

if $\rho_T \asymp C_\lambda \log T$ and $C_\lambda$ is appropriately chosen. Hence, we have with a probability exceeding $1 - O(T^{-\lambda})$ that

$$\begin{aligned} |\#(\{Z_t\} \cap C) - TP(Z_1 \in C)| &\leq \sum_{i=1}^{\rho_T} |\#(\{Z_t, t \in \mathcal{J}_i\} \cap C) - \#\mathcal{J}_i P(Z_1 \in C)| \\ &= \sum_{i=1}^{\rho_T} |\#(\{Z'_t, t \in \mathcal{J}_i\} \cap C) - \#\mathcal{J}_i P(Z'_1 \in C)| \end{aligned} \tag{A.20}$$

is satisfied for all $C \in \mathcal{C}_m$ .

(ii) *An upper bound for the fluctuations of the empirical process*

Let $F_k$ be the cumulative distribution function of the $k$th component of $Z_1$. We consider the following hyperrectangles:

$$I_{\underline{i},\underline{j}} = [F_1^{-1}(i_1/T), F_1^{-1}(j_1/T)] \times \ldots \times [F_m^{-1}(i_m/T), F_m^{-1}(j_m/T)],$$

where $0 \leq i_k < j_k \leq T$ and $F_k^{-1}(0) = -\infty$ , $F_k^{-1}(1) = \infty$ . (W.l.o.g., we prove the assertion for the case that $F$ is continuous. The result in the general case follows by simple modifications of the arguments.)
Since the number of the above hyperrectangles is of algebraic order, we obtain from (A.7) that

$$P\left(\max_{0 \leq i_k < j_k \leq T} \left\{ \frac{|\sum_{t \in \mathcal{J}_i} I(Z'_t \in I_{\underline{i},\underline{j}}) - \#\mathcal{J}_i P(Z_1 \in I_{\underline{i},\underline{j}})|}{\sqrt{\#\mathcal{J}_i P(Z_1 \in I_{\underline{i},\underline{j}})}\sqrt{\log T} + \log T} \right\} \geq C'_\lambda \right) = O(T^{-\lambda}). \tag{A.21}$$

Let

$$I_i^{(k)} = (-\infty, \infty)^{k-1} \times [F_k^{-1}((i-1)/T), F_k^{-1}(i/T)] \times (-\infty, \infty)^{m-k}.$$

From (A.21) we obtain

$$P\left(\max_{1 \leq i \leq T, 1 \leq k \leq m} \left\{ \sum_{t \in \mathcal{J}_i} I(Z'_t \in I_i^{(k)}) \right\} \geq C''_\lambda \log T \right) = O(T^{-\lambda}). \tag{A.22}$$

Let now $C \in \mathcal{C}_m$ be arbitrary. Then there exist $i_1, \ldots, i_m, j_1, \ldots, j_m$ such that

$$I_{\underline{i},\underline{j}} \subseteq C \subseteq I_{\underline{i},\underline{j}} \cup \left( \bigcup_{k=1}^{m} I_{i_k}^{(k)} \cup I_{j_k+1}^{(k)} \right).$$

Hence, we obtain from (A.21) and (A.22) that

$$P\left(\sup_{C \in \mathcal{C}_m} \left\{ \frac{|\#\{t \in \mathcal{J}_i \mid Z_t \in C\} - \#\mathcal{J}_i P(Z_1 \in C)|}{\sqrt{\#\mathcal{J}_i P(Z_1 \in C)}\sqrt{\log T} + \log T} \right\} > C_\lambda \right) = O(T^{-\lambda}), \tag{A.23}$$

which implies, in conjunction with (A.20), the assertion. □

**Proof of Proposition 3.1.** We split up

$$
\begin{aligned}
&P^*\left(X_t^* \in [c,d] \mid \mathbb{X}_t^* = \widetilde{x}\right) \;-\; P\left(X_t \in [c,d] \mid \mathbb{X}_t \in \mathcal{U}(\widetilde{x}, N_T/T)\right) \\
&\quad = \; N_T^{-1}\left[\#\{X_t \in [c,d], \mathbb{X}_t \in \mathcal{U}(\widetilde{x}, N_T/T)\} \;-\; T\; P\left(X_t \in [c,d], \mathbb{X}_t \in \mathcal{U}(\widetilde{x}, N_T/T)\right)\right] \\
&\qquad + N_T^{-1}\left[\#\{X_t \in [c,d], \mathbb{X}_t \in \widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T)\} \;-\; \#\{X_t \in [c,d], \mathbb{X}_t \in \mathcal{U}(\widetilde{x}, N_T/T)\}\right] \\
&\quad = \; R_1(\widetilde{x}) \;+\; R_2(\widetilde{x}).
\end{aligned}
$$

(A.24)

We obtain from Lemma 3.1 and (A4) that

(A.25)

$$
\sup_{\widetilde{x}} \{|R_1(\widetilde{x})|\} \;=\; \widetilde{O}\left(\sqrt{P(X_t \in [c,d])}\log T/\sqrt{N_T} \;+\; (\log T)^2/N_T, T^{-\lambda}\right).
$$

To bound $R_2(\widetilde{x})$, we use the estimate

$$
|R_2(\widetilde{x})| \;\leq\; N_T^{-1} \#\left\{t : (\mathbb{X}_t, X_t) \in \Delta_T\right\}, \tag{A.26}
$$

where $\Delta_T = (\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T) \Delta \mathcal{U}(\widetilde{x}, N_T/T)) \times [c,d]$ . Notice that $\Delta_T$ can be decomposed into a bounded number of hyperrectangles, which will allow for the application of Lemma 3.1.

According to Lemma 3.1, we obtain

$$
\begin{aligned}
&\#\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T) \;-\; T\; P(\mathbb{X}_t \in S)|_{S=\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T)} \\
&\quad = \; \widetilde{O}\left(\sqrt{T\; P(\mathbb{X}_t \in S)|_{S=\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T)}}\log T \;+\; (\log T)^2, T^{-\lambda}\right),
\end{aligned}
$$

which implies $|\sqrt{\#\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T)} - \sqrt{T\; P(\mathbb{X}_t \in S)|_{S=\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T)}}| = \widetilde{O}(\log T, T^{-\lambda})$ , and therefore

(A.27)

$$
T\; P(\mathbb{X}_t \in S)|_{S=\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T)} \;=\; N_T \;+\; \widetilde{O}\left(\sqrt{N_T}\log T \;+\; (\log T)^2, T^{-\lambda}\right).
$$

Since either $\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T) \subseteq \mathcal{U}(\widetilde{x}, N_T/T)$ or $\mathcal{U}(\widetilde{x}, N_T/T) \subseteq \widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T)$ , we get

$$
T\; P(\mathbb{X}_t \in S\Delta\mathcal{U}(\widetilde{x}, N_T/T))|_{S=\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T)} \;=\; \widetilde{O}\left(\sqrt{N_T}\log T \;+\; (\log T)^2, T^{-\lambda}\right)
$$

and, by (A4),

$$
\begin{aligned}
&T\; P\left((\mathbb{X}_t, X_t) \in (S\Delta\mathcal{U}(\widetilde{x}, N_T/T)) \times [c,d]\right)|_{S=\widehat{\mathcal{U}}_T(\widetilde{x}, N_T/T)} \\
&\quad = \; \widetilde{O}\left(P(X_t \in [c,d])\sqrt{N_T}\log T \;+\; (\log T)^2, T^{-\lambda}\right).
\end{aligned}
$$

Using again Lemma 3.1, we obtain

(A.28)

$$
\#\left\{t : (\mathbb{X}_t, X_t) \in \Delta_T\right\} \;=\; \widetilde{O}\left(P(X_t \in [c,d])\sqrt{N_T}\log T \;+\; (\log T)^2, T^{-\lambda}\right).
$$

The assertion follows now from (A.24), (A.26) and (A.28). □

**Proof of Proposition 3.2.** Let $\{X'_t\}$ and $\{X''_t\}$ be two Markov chains of order $r = r_\Delta$ with the same transition probabilities as $\{X^*_t\}$. Furthermore, let $x^{[r]} = (x_1, \dots, x_r)$ and $y^{[r]} = (y_1, \dots, y_r)$ be arbitrary. We show that there exists a pairing of $\{X'_t\}$ with $\{X''_t\}$, and a constant $\delta' > 0$ such that

$$P\left((X'_{t+r}, \dots, X'_{t+2r-1}) = (X''_{t+r}, \dots, X''_{t+2r-1}) \,\middle|\, (X'_{t-r}, \dots, X'_{t-1}) = x^{[r]}, (X''_{t-r}, \dots, X''_{t-1}) = y^{[r]}\right) \geq \delta'. \tag{A.29}$$

According to Theorem 2.1.35 of Iosifescu and Theodorescu (1969, p. 93), this will immediately imply the assertion of the proposition.
The pairing of the two versions of $\{X^*_t\}$, $\{X'_t\}$ and $\{X''_t\}$, leading to (A.29) will be constructed in two steps. First, we exploit (A5) and (A6) to get, with a probability bounded away from 0, an *approximate* pairing of $(X'_t, \dots, X'_{t+r-1})$ and $(X''_t, \dots, X''_{t+r-1})$. Given $(X'_t, \dots, X'_{t+r-1})$ and $(X''_t, \dots, X''_{t+r-1})$ are sufficiently close to each other, we obtain that $\widehat{\mathcal{U}}_T((X'_{t+r-r_1}, \dots, X'_{t+r-r_\Delta}), N_T/T) \cap \widehat{\mathcal{U}}_T((X''_{t+r-r_1}, \dots, X''_{t+r-r_\Delta}), N_T/T)$ contains at least $N_T/2$ elements. This is the basis for getting an *exact* pairing of $(X'_{t+r}, \dots, X'_{t+2r-1})$ with $(X''_{t+r}, \dots, X''_{t+2r-1})$.

(i) *Approximate pairing*

The first step proceeds as follows. Let $f = K \min_i\{f_i\}$, where $K$ will be specified later. We divide the interval $[c, d]$ into subintervals $I_l = [c + (l-1)f, c + lf)$, $l = 1, 2, \dots$. Now we construct the paring of $(X'_t, \dots, X'_{t+r-1})$ with $(X''_t, \dots, X''_{t+r-1})$ recursively.
According to Proposition 3.1, there exists a set of events $\Omega_T$ with $P(\Omega_T^c) = O(T^{-\lambda})$ such that, for $(X_{1-l_d}, \dots, X_T) \in \Omega_T$,

$$\begin{aligned}
&\sum_l \min\left\{P^*\left(X^*_t \in I_l \mid (X^*_{t-r}, \dots, X^*_{t-1}) = x^{[r]}\right),\right. \\
&\qquad \left. P^*\left(X^*_t \in I_l \mid (X^*_{t-r}, \dots, X^*_{t-1}) = y^{[r]}\right)\right\} \\
&\geq \sum_l \min\left\{P\left(X_t \in I_l \mid \underline{X}_t \in \mathcal{U}((x_{r_1}, \dots, x_{r_\Delta}), N_T/T)\right),\right. \\
&\qquad \left. P\left(X_t \in I_l \mid \underline{X}_t \in \mathcal{U}((y_{r_1}, \dots, y_{r_\Delta}), N_T/T)\right)\right\} \\
&\qquad - C\frac{1}{f}\left[\sqrt{f}\log T/\sqrt{N_T} + (\log T)^2/N_T\right].
\end{aligned} \tag{A.30}$$

By (A5) and (A6), this can be further estimated by

$$\sum_l \min\left\{P^*\left(X^*_t \in I_l \mid (X^*_{t-r}, \dots, X^*_{t-1}) = x^{[r]}\right), P^*\left(X^*_t \in I_l \mid (X^*_{t-r}, \dots, X^*_{t-1}) = y^{[r]}\right)\right\} \geq \eta'/2$$

if $T$ is sufficiently large. Hence, provided $(X_{1-l_d}, \dots, X_T) \in \Omega_T$, there exists a pairing of $X'_t$ with $X''_t$ such that

(A.31)
$$P\left(X'_t, X''_t \in [c, d] \quad \text{and} \quad |X'_t - X''_t| \leq f \,\middle|\, (X'_{t-r}, \dots, X'_{t-1}) = x^{[r]}, (X''_{t-r}, \dots, X''_{t-1}) = y^{[r]}\right) \geq \eta'/2.$$

Now we can pair $(X'_{t+1}, X''_{t+1}), \ldots, (X'_{t+r-1}, X''_{t+r-1})$ in the same manner such that

$$P\Big(X'_{t+i}, X''_{t+i} \in [c,d] \quad \text{and} \quad |X'_{t+i} - X''_{t+i}| \le f \quad \text{for all } i = 0, \ldots, r-1 \Big|$$

$$(X'_{t-r}, \ldots, X'_{t-1}) = x^{[r]}, (X''_{t-r}, \ldots, X''_{t-1}) = y^{[r]}\Big) \ \ge \ (\eta'/2)^r. \tag{A.32}$$

(ii) *Exact pairing*

Provided we chose $K$ small enough, then there exists a set of events $\Omega'_T$ with $P(\Omega'_T) \ge 1 - O(T^{-\lambda})$ such that

$$\widehat{\mathcal{U}}_T(\underline{x}, N_T/T) \cap \widehat{\mathcal{U}}_T(\underline{y}, N_T/T) \ \ge \ N_T/2 \tag{A.33}$$

for all $\underline{x}, \underline{y} \in [c,d]^\Delta$ with $|x_i - y_i| \le f$ . We suppose for the rest of this proof that $(X_{1-l_d}, \ldots, X_T) \in \Omega_T \cap \Omega'_T$ . According to (A.33), there exists a pairing of $(X'_{t+r}, \ldots, X'_{t+2r-1})$ with $(X''_{t+r}, \ldots, X''_{t+2r-1})$ such that

$$P\Big((X'_{t+r}, \ldots, X'_{t+2r-1}) = (X''_{t+r}, \ldots, X''_{t+2r-1}) \Big| X'_{t+i}, X''_{t+i} \in [c,d]$$

$$\text{and} \quad |X'_{t+i} - X''_{t+i}| \le f \quad \text{for all } i = 0, \ldots, r-1\Big) \ \ge \ \left(\frac{1}{2}\right)^r. \tag{A.34}$$

(A.32) and (A.34) imply the desired relation (A.29), which completes the proof. □

**Proof of Proposition 3.3.** To prove the closeness of $\pi_{(r_1,\ldots,r_\Delta)}$ and $\pi^*_{(r_1,\ldots,r_\Delta)}$, we set $\{X^*_t\}$ in relation to another Markov chain of order $r_\Delta$, $\{X'_t\}$ with transition probabilities

$$P\Big(X'_t \in A \mid (X'_{t-r_1}, \ldots, X'_{t-r_\Delta}) \in B\Big) \ = \ \pi_{(0,r_1,\ldots,r_\Delta)}(A \times B)/\pi_{(r_1,\ldots,r_\Delta)}(B). \tag{A.35}$$

We start both chains with the same random sequence, $(X'_{1-l_d}, \ldots, X'_{r_\Delta - l_d}) = (X^*_{1-l_d}, \ldots, X^*_{r_\Delta - l_d})$ , being distributed according to the $r_\Delta$-dimensional stationary distribution of $\{X_t\}$. It is easy to see that $(X'_{t-r_1}, \ldots, X'_{t-r_\Delta})$ has the same stationary distribution as $(X_{t-r_1}, \ldots, X_{t-r_\Delta})$ . Whereas $(X'_{t-r_1}, \ldots, X'_{t-r_\Delta}) \sim \pi_{(r_1,\ldots,r_\Delta)}$ for all $t$, the Markov chain $\{X^*_t\}$ reaches its own stationary distribution $\pi^*_{(r_1,\ldots,r_\Delta)}$ to a sufficiently good approximation after $O(\log T)$ steps. Hence, the proof is reduced to a comparison of $\pi_{(r_1,\ldots,r_\Delta)}$ with $P((X^*_{t_0-r_1}, \ldots, X^*_{t_0-r_\Delta}) \in .)$, for some $t_0 \asymp \log T$ . This comparison is made by observing the "decoupling" of the Markov chains $\{X'_t\}$ and $\{X^*_t\}$. Since their transition probabilities are quite similar, we can pair both chains in such a way that $P(X'_t \ne X^*_t$ for any $t \le t_0 - r_1)$ is small, which proves the assertion. In what follows we describe this approach in detail.
As indicated above, we have

$$\begin{aligned}
&\Big|\pi_{(r_1,\ldots,r_\Delta)}([\widetilde{a},\widetilde{b}]) \ - \ \pi^*_{(r_1,\ldots,r_\Delta)}([\widetilde{a},\widetilde{b}])\Big| \\
&\quad \le \ \Big|P\Big((X'_{t_0-r_1}, \ldots, X'_{t_0-r_\Delta}) \in [\widetilde{a},\widetilde{b}]\Big) \ - \ P\Big((X^*_{t_0-r_1}, \ldots, X^*_{t_0-r_\Delta}) \in [\widetilde{a},\widetilde{b}]\Big)\Big| \\
&\qquad + \Big|P\Big((X^*_{t_0-r_1}, \ldots, X^*_{t_0-r_\Delta}) \in [\widetilde{a},\widetilde{b}]\Big) \ - \ \pi^*_{(r_1,\ldots,r_\Delta)}([\widetilde{a},\widetilde{b}])\Big| \\
&\quad = \ T_1 \ + \ T_2.
\end{aligned} \tag{A.36}$$

According to Theorem 2.1.35 of Iosifescu and Theodorescu (1969), we obtain from Proposition 3.2 that

$$T_2 \le \rho^{t_0} = O(T^{-\lambda}), \tag{A.37}$$

if $t_0 = [C_\lambda \log T]$.
Now we turn to the decoupling approach leading to an estimate for $T_1$. Let $\Delta_T = (\log T/\sqrt{N_T})^{2/3}$ and let $I_k = [F^{-1}((k-1)\Delta_T), F^{-1}(k\Delta_T \wedge 1)]$. (It will turn out below that this choice of $\Delta_T$ is optimal; see (A.41).)

(i) *Pairing of $X'_t$ and $X^*_t$ for $t = r_\Delta - l_d + 1, \dots, r_\Delta - l_d + r_1$*

Since $X'_{t-r_i} = X^*_{t-r_i}$ $(i = 1, \dots, \Delta)$, we obtain by Proposition 3.1 and (A7) that

(A.38)
$$\left|P\left(X'_{r_\Delta - l_d + i} \in I_k\right) - P\left(X^*_{r_\Delta - l_d + i} \in I_k\right)\right| = \tilde{O}\left(\sqrt{\Delta_T}\log T/\sqrt{N_T}, T^{-\lambda}\right) + O\left(\Delta_T (N_T/T)^{1/\Delta}\right).$$

Accordingly, we pair both Markov chains in such a way that

$$\begin{aligned} &P\left(|F(X'_{r_\Delta - l_d + i}) - F(X^*_{r_\Delta - l_d + i})| > \Delta_T \quad \text{for any } i = 1, \dots, r_1\right) \\ &= \tilde{O}\left(\frac{\log T}{\sqrt{\Delta_T}\sqrt{N_T}}, T^{-\lambda}\right) + O\left((N_T/T)^{1/\Delta}\right). \end{aligned} \tag{A.39}$$

(ii) *Pairing of $X'_t$ and $X^*_t$ for $t > r_\Delta - l_d + r_1$*

For $t > r_\Delta - l_d + r_1$, the situation is slightly different to the previous case since we can only guarantee that $|F(X'_{t-r_i}) - F(X^*_{t-r_i})| \le \Delta_T$ holds with a high probability. Assume that $|F(X'_{t-r_i}) - F(X^*_{t-r_i})| \le \Delta_T$ is actually satisfied for $i = 1, \dots, \Delta$. Then we obtain by Proposition 3.1 and (A7) that

$$\begin{aligned} &\left|P\left(X'_t \in I_k \mid X'_{t-r_1}, \dots, X'_{t-r_\Delta}\right) - P\left(X^*_t \in I_k \mid X^*_{t-r_1}, \dots, X^*_{t-r_\Delta}\right)\right| \\ &= \tilde{O}\left(\sqrt{\Delta_T}\log T/\sqrt{N_T}, T^{-\lambda}\right) + O\left(\Delta_T^2\right) + O\left(\Delta_T (N_T/T)^{1/\Delta}\right). \end{aligned} \tag{A.40}$$

Hence, we can pair $X'_t$ and $X^*_t$ in such a way that

$$\begin{aligned} &P\left(|F(X'_t) - F(X^*_t)| > \Delta_T \quad \text{for any } i = 1, \dots, r_1\right) \\ &= \tilde{O}\left(\frac{\log T}{\sqrt{\Delta_T}\sqrt{N_T}}, T^{-\lambda}\right) + O\left(\Delta_T\right) + O\left((N_T/T)^{1/\Delta}\right). \end{aligned} \tag{A.41}$$

This construction will be successively applied for $r_\Delta - l_d + r_1 < t < t_0 - r_1$ with $t \ne t_0 - r_i$. Note that the sum of the first two terms on the right-hand side of (A.41) is minimized by the above choice of $\Delta_T$.

(iii) *Pairing of $X'_{t_0-r_i}$ and $X^*_{t_0-r_i}$ $(i = 1, \dots, \Delta)$*

Whereas we were so far concerned with such a pairing of $X'_t$ and $X^*_t$ that $|F(X'_t) - F(X^*_t)| \leq \Delta_T$ with an as large as possible probability, we focus now on the events $\{\omega \mid X'_{t_0-r_i} \in [a_i, b_i]\}$ and $\{\omega \mid X^*_{t_0-r_i} \in [a_i, b_i]\}$. Provided that $|F(X'_{t_0-r_i-r_j}) - F(X^*_{t_0-r_i-r_j})| \leq \Delta_T$ $(j = 1, \ldots, \Delta)$, we can find a pairing of $X'_{t_0-r_i}$ and $X^*_{t_0-r_i}$ such that

$$P\left(X'_{t_0-r_i} \in [a_i, b_i], X^*_{t_0-r_i} \notin [a_i, b_i] \quad \text{or} \quad X'_{t_0-r_i} \notin [a_i, b_i], X^*_{t_0-r_i} \in [a_i, b_i] \middle| X'_{t_0-r_i-r_1}, \ldots, X'_{t_0-r_i-r_\Delta}; X^*_{t_0-r_i-r_1}, \ldots, X^*_{t_0-r_i-r_\Delta}\right)$$

$$= \tilde{O}\left(\sqrt{F(b_i) - F(a_i)}\frac{\log T}{\sqrt{N_T}} + \frac{(\log T)^2}{N_T}, T^{-\lambda}\right) + O\left([F(b_i) - F(a_i)]\Delta_T\right) \tag{A.42}$$

and

$$P\left(|F(X'_{t_0-r_i}) - F(X^*_{t_0-r_i})| > \Delta_T \quad \text{for any } i = 1, \ldots, r_1\right)$$

$$= \tilde{O}\left(\frac{\log T}{\sqrt{\Delta_T}\sqrt{N_T}}, T^{-\lambda}\right) + O\left(\Delta_T\right) \tag{A.43}$$

are simultaneously fulfilled.
Moreover, we have by (A4)

(A.44)

$$P\left(X'_{t_0-r_i} \in [a_i, b_i] \mid X'_{t_0-r_i-r_1}, \ldots, X'_{t_0-r_i-r_\Delta}\right) = O\left(F(b_i) - F(a_i)\right)$$

and

(A.45)

$$P\left(X^*_{t_0-r_i} \in [a_i, b_i] \mid X^*_{t_0-r_i-r_1}, \ldots, X^*_{t_0-r_i-r_\Delta}\right) = O\left([F(b_i) - F(a_i)] + \frac{(\log T)^2}{N_T}\right).$$

To complete the proof, we define the following events:

$$\begin{aligned}\Omega_{01} &= \{\omega \mid |X'_t - X^*_t| > \Delta_T \quad \text{for any } t < t_0 - r_1\}, \\ &\quad \cap \left\{\omega \mid (X'_{t_0-r_1}, \ldots, X'_{t_0-r_\Delta}) \in [\underline{a}, \underline{b}]\right\},\end{aligned}$$

$$\begin{aligned}\Omega_{02} &= \{\omega \mid |X'_t - X^*_t| > \Delta_T \quad \text{for any } t < t_0 - r_1\}, \\ &\quad \cap \left\{\omega \mid (X^*_{t_0-r_1}, \ldots, X^*_{t_0-r_\Delta}) \in [\underline{a}, \underline{b}]\right\},\end{aligned}$$

$$\begin{aligned}\Omega_{i1} &= \{\omega \mid |X'_t - X^*_t| \leq \Delta_T \quad \text{for all } t < t_0 - r_1\}, \\ &\quad \cap \left\{\omega \mid (X'_{t_0-r_1}, \ldots, X'_{t_0-r_\Delta}) \in [\underline{a}, \underline{b}]\right\} \\ &\quad \cap \left\{\omega \mid X^*_{t_0-r_i} \notin [a_i, b_i]\right\},\end{aligned}$$

and

$$\begin{aligned}\Omega_{i2} &= \{\omega \mid |X'_t - X^*_t| \leq \Delta_T \quad \text{for all } t < t_0 - r_1\}, \\ &\quad \cap \left\{\omega \mid (X^*_{t_0-r_1}, \ldots, X^*_{t_0-r_\Delta}) \in [\underline{a}, \underline{b}]\right\} \\ &\quad \cap \left\{\omega \mid X'_{t_0-r_i} \notin [a_i, b_i]\right\}.\end{aligned}$$

Then, by (A.39), (A.41) and (A.43) to (A.45),

$$\begin{aligned} T_1 &\leq P(\Omega_{01}) + P(\Omega_{02}) + \sum_{i=1}^{\Delta} P(\Omega_{i1}) + P(\Omega_{i2}) \\ &= O\Bigg( [(N_T/T)^{1/\Delta} + (\log T/\sqrt{N_T})^{2/3}] \prod_{i=1}^{\Delta} [P(X_t \in [a_i, b_i]) + (\log T)^2/N_T] \\ &\qquad + \sum_{i=1}^{\Delta} \left[ \sqrt{F(b_i) - F(a_i)} \frac{\log T}{\sqrt{N_T}} + \frac{(\log T)^2}{N_T} \right] \prod_{j \neq i} \left[ P(X_t \in [a_j, b_j]) + \frac{(\log T)^2}{N_T} \right] \Bigg) . \end{aligned}$$

□

## References


Athreya, K. B. and Fuh, C. D. (1992a). Bootstrapping Markov chains: Countable case. *J. Statist. Plan. Inference* **33**, 311–331.

Athreya, K. B. and Fuh, C. D. (1992b). Bootstrapping Markov chains. *In: Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 49–64, Wiley, New York.

Basawa, I. V., Green, T. A., McCormick, W. P. and Taylor, R. L. (1990). Asymptotic bootstrap validity for finite Markov chains. *Comm. Statist.-Theory Meth.* **19**, 1493–1510.

Bickel, P. J. and Bühlmann, P. (1995). Mixing property and functional central limit theorems for a sieve bootstrap in time series. Technical Report, University of California, Berkeley.

Bühlmann, P. (1994). Blockwise bootstrapped empirical process for stationary sequences. *Ann. Statist.* **22**, 995–1012.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14**, 1171–1194.

Carlstein, E., Do, K.-A., Hall, P., Hesterberg, T. and Künsch, H.-R. (1996). Matched-block bootstrap for dependent data. Research Report No. 74, ETH Zürich.

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

Doukhan, P., Massart, P. and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. Henri Poincaré* **31**, 393–427.

Efron, B. and Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Science* **1**, 54–77.

Falk, M. and Reiss, R.-D. (1992). Bootstrapping conditional curves. *In: Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 173–180, Wiley, New York.

Franke, J., Kreiss, J.-P. and Mammen, E. (1996). Bootstrap of kernel smoothing in nonlinear time series, Preprint.

Franke, J., Kreiss, J.-P., Mammen, E. and Neumann, M. H. (1997). Properties of the nonparametric autoregressive bootstrap. Manuscript.

Franke, J. and Wendel, M. (1992). A bootstrap approach for nonlinear autoregressions - some preliminary results. In: Jöckel, K. H., Rothe, G. and Sendler, W. eds.: Bootstrapping and Related Techniques. Lecture Notes in Economics and Mathematical Systems 376, Springer, Berlin Heidelberg.

Götze, F. and Künsch, H. R. (1996). Second-order correctness of the blockwise bootstrap for stationary observations. *Ann. Statist.* **24**, 1914–1933.

Hall, P. (1985). Resampling a coverage pattern. *Stoch. Proc. Appl.* **20**, 231–246.

Hall, P. (1991). On the distribution of suprema. *Probab. Theory Rel. Fields* **89**, 447–455.

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

Härdle, W. and Tsybakov, A. B. (1995). Local polynomial estimators of the volatility function in nonparametric autoregression, Discussion Paper 42, SFB 373, Berlin.

Hart, J. D. (1995). Some automated methods of smoothing time-dependent data. *J. Nonpar. Statist.* **6**, 115-142.

Iosifescu, M. and Theodorescu, R. (1969). *Random Processes and Learning*. Springer, New York.

Kiefer, J. (1972). Skorohod embedding of multivariate rv's, and the sample df. *Z. Wahrscheinlichkeitstheorie verw. Geb.* **24**, 1–35.

Kreiss, J.-P. (1988). Asymptotic statistical inference for a class of stochastic processes. Habilitationsschrift.

Kreiss, J.-P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving average models. *J. Time Ser. Anal.* **13**, 297–317.

Kreutzberger, E. (1993). Bootstrap für nichtlineare AR(1)-Prozesse. Thesis, University of Kaiserslautern, Germany.

Kulperger, P. J. and Prakasa Rao, B. L. S. P. (1989). Bootstrapping a finite state Markov chain. *Sankhyā A* **51**, 178–191.

Künsch, H. R. (1889). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217–1241.

Lall, U. and Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research* **32**, 679–693.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571–599.

Neumann, M. H. (1996). Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations, Preprint No. 295, Weierstrass Institute, Berlin.

Neumann, M. H. and Kreiss, J.-P. (1997). Regression-type inference in nonparametric autoregression, Manuscript.

Neumann, M. H. and Polzehl, J. (1995). Simultaneous bootstrap confidence bands in nonparametric regression, *J. Nonpar. Statist.*, to appear.

Paparoditis, E. and Politis, D. N. (1996). The local bootstrap for periodogram statistics. Manuscript.

Paparoditis, E. and Politis, D. N. (1997). The local bootstrap for Markov processes. Technical Report 11/97, University of Cyprus, Nicosia.

Politis, D. N. and Romano, J. P. (1992). A circular block-resampling procedure for stationary data. *In: Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 263–270, Wiley, New York.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *J. Amer. Statist. Assoc.* **89**, 1303–1313.

Rajarshi, M. B. (1990). Bootstrap in Markov-sequences based on estimates of transition density. *Ann. Inst. Statist. Math.* **42**, 253–268.

Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.* **4**, 185–207.

Rutherford, B. and Yakowitz, S. (1991). Error inference for nonparametric regression. *Ann. Inst. Statist. Math.* **43**, 115–129.

Shao, Q. and Yu, H. (1993). Bootstrapping the sample means for stationary mixing sequences. *Stoch. Proc. Appl.* **48**, 175–190.

Shi, S. G. (1991). Local bootstrap. *Ann. Inst. Statist. Math.* **43**, 667-676.

Shi, X. (1986). Bootstrap estimate for $m$-dependent sample means. *Kexue Tongbao (Chineses Bulletin of Science* **31**, 404–407.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

Skorokhod, A. V. (1965). *Studies in the Theory of Random Processes*. Addison-Wesley, Reading, Massachussetts.

Yang, L. and Härdle, W. (1996). Nonparametric autoregression with multiplicative volatility and additive mean. Discussion Paper No. 62/96, SFB 373, Humboldt University, Berlin.