The Annals of Statistics
1999, Vol. 27, No. 2, 480–513

VARIABLE LENGTH MARKOV CHAINS

BY PETER BÜHLMANN[1] AND ABRAHAM J. WYNER[2]

ETH Zürich and University of Pennsylvania

We study estimation in the class of stationary variable length Markov chains (VLMC) on a finite space. The processes in this class are still Markovian of high order, but with memory of variable length yielding a much bigger and structurally richer class of models than ordinary high-order Markov chains. From an algorithmic view, the VLMC model class has attracted interest in information theory and machine learning, but statistical properties have not yet been explored. Provided that good estimation is available, the additional structural richness of the model class enhances predictive power by finding a better trade-off between model bias and variance and allowing better structural description which can be of specific interest. The latter is exemplified with some DNA data.

A version of the tree-structured context algorithm, proposed by Rissanen in an information theoretical set-up is shown to have new good asymptotic properties for estimation in the class of VLMCs. This remains true even when the underlying model increases in dimensionality. Furthermore, consistent estimation of minimal state spaces and mixing properties of fitted models are given.

We also propose a new bootstrap scheme based on fitted VLMCs. We show its validity for quite general stationary categorical time series and for a broad range of statistical procedures.

1. Introduction. One of the most general models for a stationary process $(X_t)_{t\in\mathbb{Z}}$ assuming no particular underlying mechanistic system is a full Markov chain of high, but finite, order. The only implicit assumption aside from stationarity is the finite memory of the process. We consider here exclusively the case where X_t takes values in a finite categorical space $\mathscr{X}$. We always refer to a stationary full Markov chain of order k whenever the transition mechanism carries no specific structure; that is, the state space is the entire $\mathscr{X}^k$. Probabilistically a nice model, such full Markov chains are not very appropriate from the estimation point of view. Let us illustrate two main problems. To be more specific, we momentarily take for illustrative purposes cardinality $|\mathscr{X}| = 4$, for example, $\mathscr{X} = \{A, C, G, T\}$ being the letters of a DNA string (but all the discussed problems below apply to any finite space $\mathscr{X}$).

Received February 1997; revised February 1999.

[1] Supported in part by the Swiss National Science Foundation. Part of the work was done while visiting the University of Heidelberg, Germany.

[2] Supported by NSF Grant DMS-95-08933.

AMS 1991 *subject classifications*. Primary 62M05; secondary 60J10, 62G09, 62M10, 94A15.

Key words and phrases. Bootstrap, categorical time series, central limit theorem, context algorithm, data compression, finite-memory sources, FSMX model, Kullback–Leibler distance, model selection, tree model.

PROBLEM 1. The class of all finite-order $\mathscr{X}$-valued full Markov chains is not structurally rich, implying that there are not many members in the class. This structural poverty particularly implies that any kind of parsimonious representation of the state space is not possible. The table below additionally demonstrates such structural poverty in terms of the dimension of full Markov chain models (the number of free parameters) as a function of their orders k, that is, Dim. = $(|\mathscr{X}| - 1)|\mathscr{X}|^k$ with cardinality $|\mathscr{X}| = 4$.

k	0	1	2	3	4	5	10
Dim	3	12	48	192	768	3072	$\approx 3.1 \cdot 10^6$

There are no models "in between," for example, it is impossible to fit a model with, say, 72 parameters. Such a very "discontinuous" increase in dimensionality of the model does not allow a good trade-off between bias (being low with many parameters) and variance (being low with a few parameters) of a predictor.

PROBLEM 2. As seen from the table, the curse of dimensionality is particularly damaging when fitting high-order models, since the dimensionality increases exponentially with the order k. This then leads to highly variable estimates.

A practical example which illustrates the table and Problems 1 and 2 is the modeling of DNA sequences with full Markov chains [cf. Prum, Rodolphe and de Turckheim (1995) and Braun and Müller (1998)]. The class of models and the estimator which we study in this paper will lead to an alternative, and for many purposes a better, statistical description of DNA sequences. We give in Section 3.3 a real-data example from this field of applications. Other examples of applications where our modeling is potentially attractive are precipitation analysis [(Guttorp (1995)], flood analysis [Brillinger (1995)] or analysis of discrete directional data and repeated patterns of behavioral events [Raftery and Tavaré (1994)].

Problems 1 and 2 can be addressed with a very simple idea: the memory of a stationary Markov chain is allowed to be of variable length, a function of the values from the past. More precisely, the time-homogeneous transition probabilities $\mathbb{P}[X_t = x_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \dots]$ are functions depending only on a variable number ℓ of lagged values $\mathbb{P}[X_t = x_t \mid X_{t-1} = x_{t-1}, \dots, X_{t-\ell} = x_{t-\ell}]$, where $\ell = \ell(x_{t-1}, x_{t-2}, \dots)$ is itself a function of the past. If $\ell(x_{t-1}, x_{t-2}, \dots) \equiv k$ for all $x_{t-1}, x_{t-2}, \dots$, we obtain the full Markov chain model of order k. For variable $\ell(\cdot)$ with $\sup\{\ell(x_{t-1}, x_{t-2}, \dots); x_{t-1}, x_{t-2}, \dots\} = k$, we have an embedding full Markov chain of order k, but with an additional well-interpretable *structure of a variable length memory*: it implies that some transition probabilities of the embedding Markov chain are lumped together. We call such a process variable length Markov chain (VLMC). It is closely related to models in information theory like "tree models," "FSMX models" or "finite-memory sources" [cf. Rissanen (1986), Weinberger, Lempel and Ziv (1992), Weinberger, Rissanen and Feder (1995), Feder, Merhav and Gutman (1992)]. If one is able to choose in a data-driven way an appropriate member in the class of VLMCs, there is nothing to lose but only to gain in comparison with the class of ordinary full Markov chains of high order. We give in this paper new results for VLMCs, in particular also addressing the problem of how to select in a data-driven way an asymptotically correct member in the extremely large class of all VLMCs.

The merits of our results are in different areas. On a theoretical level we offer a better understanding of Problems 1 and 2. With this goal in mind, the study of VLMCs and their estimation is per se an interesting task. The practical merits of our results, which offer new insights into VLMCs, apply generally to problems involving categorical or binary time series (see the examples mentioned above). We also offer advances in statistical methodology; specifically through a new bootstrap scheme, based on VLMCs, for categorical time series. This will prove to be a very attractive alternative to the more general blockwise bootstrap [Künsch (1989)]. Moreover, fitted VLMCs can be used as an excellent exploratory tool for the dynamics of a categorical time series. This is accomplished by representing structural dependencies graphically and compactly (we demonstrate this for some DNA data in Section 3.3). Such explorative information could also be used to build a more specific parametric model in a second stage. Finally, we offer new insights into information theory, where our findings sharpen and extend existing results on VLMCs and compression rates; see Rissanen (1983) and Weinberger, Rissanen and Feder (1995).

The notion of a variable length memory in a Markov chain is particularly attractive when there is long memory in certain "directions." In such cases, the *minimal* state space is drastically smaller than the embedding state space of a full Markov chain (having many equivalent states which are lumped together in the VLMC); the VLMC yields a *parsimonious* parameterization of the state space. The difficulty with this attractive notion is then the estimation of that minimal state space. This can be seen as a model selection problem. However, due to the extremely large number of VLMC submodels of a high-order full Markov chain, global model selection techniques like AIC, BIC, or MDL cannot be used. However, estimation of the minimal state space and the probability distribution of a VLMC can be done with a tree structured scheme, called the context algorithm [Rissanen (1983)], which acts hierarchically on *local* pairwise decisions. Weinberger, Rissannen and Feder (1995) proved consistency and optimal compression rates for the context algorithm under the assumption that the true underlying process is a finite-dimensional VLMC.

We give an entirely new consistency result where the true underlying model is allowed to grow in dimensionality as sample size n increases. The growth rate is in probabilistic terms and can be as large as $O(n^{1/2}/\log(n)^s)$ for some $s > 1/2$; or it can be even larger. This describes much better the performance of the context algorithm, because with increasing dimensionality, consistency is much less obvious. As an important consequence, our result implies a nontrivial balance between over- and underestimation of the true model. It can be loosely translated to the fact that a bias-variance tradeoff in a possibly very high-dimensional problem is handled by the context algorithm in an appropriate way. Also, by allowing for asymptotically infinite-dimensional models, the new results contribute to explore the approximation of a general, sufficiently "nice" stationary process by an estimated VLMC. Consistency of the context algorithm for estimating stationary processes or minimal state spaces does not require a prespecified model structure. Thus, estimation with the context algorithm is robust against model misspecification.

We then make use of the general consistency result described above to propose a novel resampling scheme, the VLMC bootstrap. We prove asymptotic validity of the VLMC bootstrap for a whole class of estimators and argue, by proving a mixing property for estimated VLMCs that such a scheme works under very general conditions. The VLMC bootstrap is tailored for categorical time series; it has a nice probabilistic interpretation and it enjoys the advantage of being applicable as a simple plug-in rule, as Efron's (1979) original proposal for the independent case, which is more user-friendly than the blockwise bootstrap [Künsch (1989)]. Based on the results in theory and from a small simulation study, we conclude that the VLMC bootstrap is a new universal resampling tool for categorical time series which is often expected to be better than the blockwise bootstrap.

The paper is organized as follows. In Section 2 we give the definition of VLMCs, Section 3 describes the process of fitting such models and gives new asymptotic properties thereof. In Section 4 we discuss the VLMC bootstrap, its asymptotic validity and present results from a simulation study, including a comparison with the blockwise bootstrap. All the proofs are given in Section 5.

2. Variable length Markov chains. As a starting point, consider a stationary full Markov chain $(X_t)_{t \in \mathbb{Z}}$ of finite-order k with values in a finite categorical space $\mathscr{X}$. In the sequel, we denote by $x_i^j = x_j, x_{j-1}, \dots, x_i$ ($i < j$, $i, j \in \mathbb{Z} \cup \{-\infty, \infty\}$) a string whose components are written in reverse order, $wu = (w_{|w|}, \dots, w_2, w_1, u_{|u|}, \dots, u_2, u_1)$ is the concatenation of the strings w and u. We usually denote by capital letters X random variables and by small letters x fixed deterministic values. Thus,

$$(2.1)\quad \mathbb{P}\left[X_1 = x_1 \mid X_{-\infty}^0 = x_{-\infty}^0\right] = \mathbb{P}\left[X_1 = x_1 \mid X_{-k+1}^0 = x_{-k+1}^0\right] \quad \text{for all } x_{-\infty}^1.$$

Note that stationarity implies time-homogeneous transition probabilities so that the time indices $-\infty, \dots, 0, 1$ can be replaced by other indices $-\infty, \dots, t-1, t$ for any $t \in \mathbb{Z}$. We now introduce the idea of a variable length memory which can also be seen as lumping together irrelevant states in the history x_{-k+1}^0 in (2.1). Only some values from the infinite history $x_{-\infty}^0$ of the variable X_1 are relevant: these can be thought of as a *context* for X_1. To achieve a flexible model class, ranging from some type of sparse to full Markov chains, we let the length of a context depend on (the first few of) the actual values $x_{-\infty}^0$. We formalize this by defining below a VLMC. Related models have been introduced in information theory as tree models, FSMX models or finite-memory sources [cf. Rissanen (1986), Weinberger, Lempel and Ziv (1992), Weinberger, Rissanen and Feder (1995)].

DEFINITION 2.1. Let $(X_t)_{t \in \mathbb{Z}}$ be a stationary process with values $X_t \in \mathscr{X}$, $|\mathscr{X}| < \infty$. Denote by $c: \mathscr{X}^\infty \to \mathscr{X}^\infty$ a (variable projection) function which maps c: $x_{-\infty}^0 \mapsto x_{-\ell+1}^0$, where ℓ is defined by $\ell = \ell(x_{-\infty}^0) = \min\{k; \mathbb{P}[X_1 = x_1 \mid X_{-\infty}^0 = x_{-\infty}^0] = \mathbb{P}[X_1 = x_1 \mid X_{-k+1}^0 = x_{-k+1}^0]$ for all $x_1 \in \mathscr{X}\}$, where $\ell \equiv 0$ corresponds to independence.

Then $c(\cdot)$ is called a context function and for any $t \in \mathbb{Z}$, $c(x_{-\infty}^{t-1})$ is called the context for the variable x_t.

The name *context* refers to the portion of the past that influences the next outcome. The definition of ℓ implicitly reflects the fact that the context length of a variable X_t is $\ell = \ell(x_{-\infty}^{t-1}) = |c(x_{-\infty}^{t-1})|$, depending on the history $X_{-\infty}^{t-1} = x_{-\infty}^{t-1}$. By the projection structure of the context function $c(\cdot)$, the context-length $\ell(\cdot) = |c(\cdot)|$ determines $c(\cdot)$ and vice versa.

DEFINITION 2.2. Let $(X_t)_{t \in \mathbb{Z}}$ be a stationary process with values $X_t \in \mathscr{X}$, $|\mathscr{X}| < \infty$ and corresponding context function $c(\cdot)$ as given in Definition 2.1. Let $0 \le k \le \infty$ be the smallest integer such that

$$\left|c\left(x_{-\infty}^0\right)\right| = \ell\left(x_{-\infty}^0\right) \le k \quad \text{for all } x_{-\infty}^0 \in \mathscr{X}^\infty.$$

Then $c(\cdot)$ is called a context function of order k, and if $k < \infty$, $(X_t)_{t \in \mathbb{Z}}$ is called a stationary variable length Markov chain (VLMC) of order k.

We sometimes identify a VLMC $(X_t)_{t \in \mathbb{Z}}$ with its probability distribution P_c on $\mathscr{X}^{\mathbb{Z}}$. In the sequel, we often write for a probability measure P on $\mathscr{X}^{\mathbb{Z}}$, $P(x) = \mathbb{P}_P[X_1^m = x]$ $(x \in \mathscr{X}^m,\ m \ge 1)$ and $P(x \mid w) = P(xw)/P(w)$ $(x, w \in \bigcup_{m=1}^\infty \mathscr{X}^m)$. The transition probabilities for P_c are denoted by $p(x_1 \mid c(x_{-\infty}^0))$ and coincide with the conditional probabilities $P_c(x_1 \mid c(x_{-\infty}^0))$. Clearly, a VLMC of order k is a Markov chain of order k, now having a *memory of variable length* ℓ. If the context function $c(\cdot)$ of order k is the full projection $x_{-\infty}^0 \mapsto x_{-k+1}^0$ for all $x_{-\infty}^0$, the VLMC is a full Markov chain of order k. Often the range space of the context function $c(\cdot)$ is not the full space $\mathscr{X}^k$, but also not the empty space. The class of context functions of length k is structurally rich enough to obtain a broad class of Markov chains, including special sparse types. In particular, some context functions $c(\cdot)$ yield a substantial reduction in the number of states compared to a full Markov chain of the same order as the context function. Both of the latter phrases relate to solutions to Problems 1 and 2 mentioned in Section 1.

2.1. *Tree representation of minimal state space.* By requiring stationarity, a VLMC P_c is completely specified by its transition probabilities,

$$\mathbb{P}_{P_c}\left[X_1 = x_1 \mid X_{-\infty}^0 = x_{-\infty}^0\right] = p\left(x_1 \mid c\left(x_{-\infty}^0\right)\right), \qquad x_{-\infty}^1 \in \mathscr{X}^\infty.$$

The states determining these transition probabilities are thus given by the values of the context function $c(\cdot)$. It is most convenient to represent these states, that is, the minimal state space of the VLMC P_c, as a tree.

We consider trees with a root node on top, from which the branches are growing downwards, so that every internal node has at most $|\mathscr{X}|$ offspring. Then, each value of a context function $c(\cdot)$: $\mathscr{X}^{\infty} \to \mathscr{X}^k$ can be represented as a branch (or terminal node) of such a tree. The context $w = c(x_{-\infty}^0)$ is represented by a branch, whose subbranch on the top is determined by x_0, the next subbranch by x_{-1} and so on, and the terminal subbranch by $x_{-\ell(x_{-\infty}^0)+1}$. As we will exemplify, context trees do not have to be complete, that is, every internal node does not need to have exactly $|\mathscr{X}|$ offspring.

Example 2.1. $\mathscr{X} = \{0, 1\}$, $k = 3$. The function

$$c(x_{-\infty}^0) = \begin{cases} 0, & \text{if } x_0 = 0,\ x_{-\infty}^{1} \text{ arbitrary},\\ 1,0,0, & \text{if } x_0 = 1,\ x_{-1} = 0,\ x_{-2} = 0,\ x_{-\infty}^{-3} \text{ arbitrary},\\ 1,0,1, & \text{if } x_0 = 1,\ x_{-1} = 0,\ x_{-2} = 1,\ x_{-\infty}^{-3} \text{ arbitrary}\\ 1,1, & \text{if } x_0 = 1,\ x_{-1} = 1,\ x_{-\infty}^{-2} \text{ arbitrary}, \end{cases}$$

can be represented by the tree τ_c on the left-hand side in Figure 1. A "growing to the left" subbranch represents the symbol 0 and vice versa for the symbol 1.

Definition 2.3. Let $c(\cdot)$ be a context function of a stationary VLMC. The ($|\mathscr{X}|$-ary) context tree τ and terminal node context tree τ^T are defined as

$$\tau = \tau_c = \{w; w = c(x_{-\infty}^0),\ x_{-\infty}^0 \in \mathscr{X}^{\infty}\},$$
$$\tau^T = \tau_c^T = \{w; w \in \tau_c \text{ and } wu \notin \tau_c \text{ for all } u \in \mathscr{X}\}.$$

The notion of a terminal context tree is convenient when formulating an estimation procedure for a context tree (minimal state space); see Section 3.1. Definition 2.3 says that only terminal nodes in the tree representation τ are considered as elements of the terminal node context tree τ^T and the states $w \in \tau_c$ do not need to be terminal nodes in τ_c. However, we can reconstruct the context function $c(\cdot)$ from either τ_c or τ_c^T. The context tree τ_c is nothing else than the minimal state space of the VLMC P_c (we sometimes refer to the elements of τ_c as branches and sometimes as nodes in a tree). An internal node with $b < N = |\mathscr{X}|$ offspring implicitly adds one complementary offspring, lumping the $N - b$ nonpresent offspring together to a single new terminal node w_{new} which represents a single state in τ_c.

EXAMPLE 2.2. $\mathscr{X} = \{0, 1, 2, 3\}$, $k = 2$. The function,

$$c(x_{-\infty}^{0}) = \begin{cases} 0, & \text{if } x_0 = 0,\ x_{-\infty}^{1} \text{ arbitrary,} \\ 1, & \text{if } x_0 = 1,\ x_{-\infty} \text{ arbitrary,} \\ 2, & \text{if } x_0 = 2,\ x_{-\infty} \text{ arbitrary,} \\ 3, & \text{if } x_0 = 3,\ x_{-1} \in \{0, 1, 2\},\ x_{-\infty}^{2} \text{ arbitrary,} \\ 3, 3, & \text{if } x_0 = 3,\ x_{-1} = 3,\ x_{-\infty}^{-2} \text{ arbitrary,} \end{cases}$$

can be represented by the tree τ_c on the right-hand side in Figure 1. The round-edged rectangle, which we usually do not draw, symbolizes the absent nodes 0, 1 and 2 in depth 2, which can be thought of as a completion of the tree with nodes lumped together; in terms of transition probabilities, it means that $p(x \mid 3y)$ $(x \in \mathscr{X})$ is the same for all $y \in \{0, 1, 2\}$. The terminal node context tree is $\tau_c^T = \{0, 1, 2, 33\}$, whereas the context tree is $\tau_c = \{0, 1, 2, 3, 33\}$. The state 3 is represented by an internal node in the tree and hence is only an element of τ_c and not of τ_c^T. An alternative representation of the state 3 is given by the final complementary node, indicated by the rectangle, lumping the three nonpresent nodes together to a new terminal node.

2.2. *Semiparametric VLMC model and sequences of VLMCs.* Rather than one finite order VLMC, we consider a semiparametric model in the spirit of Ritov and Bickel (1990). The semiparametric VLMC model is

$$\mathscr{P} = \bigcup_{k=0}^{\infty} \mathscr{P}_k, \tag{2.2}$$

where $\mathscr{P}_k$ is the set of stationary VLMCs of order k,

$$\mathscr{P}_k = \{P_c;\ P_c \text{ a stationary VLMC of order } k\}.$$

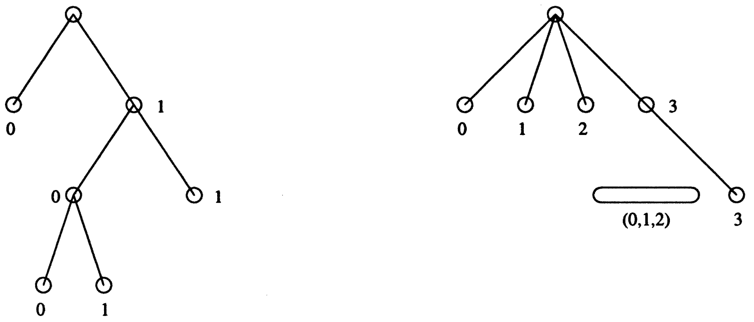

FIG. 1. *Tree representations of the context functions in Examples* 2.1 *and* 2.2.

Thus, every member of $\mathscr{P}$ belongs to a nice (parametric) VLMC model whose order can be arbitrarily large. We will study in Section 3 a consistent estimator for the semiparametric VLMC model $\mathscr{P}$ without using additional structural information such as the underlying context function $c(\cdot)$. Since we do not specify any particular model structure in $\mathscr{P}$, the estimator is *robust* against model misspecification in the set of all VLMCs (which is dense in the set of all stationary processes with respect to finite-dimensional weak convergence).

The asymptotic analysis of such a robust estimator in $\mathscr{P}$ is given for a framework where the underlying process changes with sample size, a so-called moving truth model; see (2.3). The reasons are twofold. First, it is much more interesting to see whether an estimation technique is still consistent in such a situation; the "nonmoving" case, considered by Weinberger, Rissanen and Feder (1995), is from an asymptotic point of view not so interesting, because the problem is finite (although high) dimensional. Second, the "moving truth" model has limiting elements on the boundary of the semiparametric model $\mathscr{P}$ which are infinite-dimensional non-VLMC models. In this sense, the "moving truth" model yields an interesting approximation for some general stationary $\mathscr{X}$-valued processes. The "moving truth" is a sequence $(P_n)_{n \in \mathbb{N}}$ of VLMCs in $\mathscr{P}$ from (2.2) from which the data are finite realizations in a triangular scheme,

$$\begin{aligned} &X_{1,n}, \dots, X_{n,n} \text{ a finite realization of } P_n, \\ &P_n \in \mathscr{P} \text{ from } (2.2) \text{ with context function } c_n(\cdot), \qquad n \in \mathbb{N}. \end{aligned} \tag{2.3}$$

The transition probabilities corresponding to P_n are denoted by $p_n(\cdot \mid \cdot)$. In the sequel, when writing data just as $X_1, \dots, X_n$, we usually think of a generating model as in (2.3).

3. Context algorithm and its consistency.

Given data $X_{1,n}, \dots, X_{n,n}$ as in (2.3), the aim is to find the underlying context function $c_n(\cdot)$ (the minimal state space) and an estimate of P_n. A version of the context algorithm [Rissanen (1983)] will be used to solve the problem. Besides obvious uses of the numerical estimate of the probability distribution including a resampling scheme as given in Section 4, the estimated context tree is an excellent exploratory tool for the dynamic structure of the underlying process; see Section 3.3.

3.1. *Context algorithm.*

We describe now the context algorithm for the aim mentioned above. The main strategy is as follows. First, a large context tree is grown, which represents an overfitted VLMC model. Since the value space $\mathscr{X}$ is finite, there aren't any sophisticated problems with finding accurate splits of the predictor space and the construction of such a large tree turns out to be simple and computationally fast. Second, the algorithm employs a backward tree-pruning procedure by considering a local decision criterion. Thus, on this very basic level, the context algorithm has an architecture similar to many other tree-fitting methods.

In the sequel we always make the convention that quantities involving time indices $\notin \{1, \dots, n\}$ equal zero (or are irrelevant). Let

$$N(w) = \sum_{t=1}^{n} 1_{[X_t^{t+|w|-1} = w]}, \qquad w \in \bigcup_{m=1}^{\infty} \mathscr{X}^m, \tag{3.1}$$

denote the number of occurrences of the string w in the sequence X_1^n. Moreover, let

$$\hat{P}(w) = \frac{N(w)}{n}, \qquad \hat{P}(x \mid w) = \frac{N(xw)}{N(w)}, \qquad x, w \in \bigcup_{m=1}^{\infty} \mathscr{X}^m. \tag{3.2}$$

The algorithm below constructs the estimated context tree $\hat{\tau}$ as the biggest context tree (with respect to the order "$\preccurlyeq$" defined in Step 1 below) such that

$$\Delta_{wu} = \sum_{x \in \mathscr{X}} \hat{P}(x \mid wu) \log\left(\frac{\hat{P}(x \mid wu)}{\hat{P}(x \mid w)} \right) N(wu) \geq K \tag{3.3}$$

$$\text{for all } wu \in \hat{\tau}^T \ (u \in \mathscr{X}),$$

with $K = K_n \sim C \log(n)$, $C > 2|\mathscr{X}| + 4$ a cutoff to be chosen by the user.

STEP 1. Given $\mathscr{X}$-valued data $X_1, \dots, X_n$, fit a maximal context tree, that is, search for the context function $c_{\max}(\cdot)$ with terminal node context tree representation $\tau_{\max}^T$ (see Definition 2.3), where $\tau_{\max}^T$ is the biggest tree such that every element (terminal node) in $\tau_{\max}^T$ has been observed at least twice in the data. This can be formalized as follows: $\tau_{\max}^T$ is such that $w \in \tau_{\max}^T$ implies $N(w) \geq 2$, and such that for every τ^T, where $w \in \tau^T$ implies $N(w) \geq 2$, it holds that $\tau^T \preccurlyeq \tau_{\max}^T$.

Here, $\tau_1 \preccurlyeq \tau_2$ means $w \in \tau_1 \Rightarrow wu \in \tau_2$ for some $u \in \bigcup_{m=0}^{\infty} \mathscr{X}^m$ ($\mathscr{X}^0 = \emptyset$). Set $\tau_{(0)}^T = \tau_{\max}^T$.

STEP 2. Examine every element (terminal node) of $\tau_{(0)}^T$ as follows (the order of examining is irrelevant; see Remark 3.3). Let $c(\cdot)$ be the corresponding context function of $\tau_{(0)}^T$ and let

$$wu = x_{-\ell+1}^0 = c\left(x_{-\infty}^0\right), \qquad u = x_{-\ell+1}, \qquad w = x_{-\ell+2}^0,$$

where wu is an element (terminal node) of $\tau_{(0)}^T$, which we compare with its pruned version $w = x_{-\ell+2}^0$ (if $\ell = 1$, the pruned version is the empty branch $\emptyset$, that is, the root node).

Prune $wu = x_{-\ell+1}^0$ to $w = x_{-\ell+2}^0$ if

$$\Delta_{wu} = \sum_{x \in \mathscr{X}} \hat{P}(x \mid wu) \log\left(\frac{\hat{P}(x \mid wu)}{\hat{P}(x \mid w)} \right) N(wu) < K,$$

with $K = K_n \sim C \log(n)$, $C > 2|\mathscr{X}| + 4$ and $\hat{P}(\cdot \mid \cdot)$ as defined in (3.2). Decision about pruning for every terminal node in $\tau_{(0)}^T$ yields a (possibly) smaller tree $\tau_{(1)} \preccurlyeq \tau_{(0)}^T$. Construct the terminal node context tree $\tau_{(1)}^T$.

STEP 3. Repeat Step 2 with $\tau_{(i)}, \tau_{(i)}^T$ instead of $\tau_{(i-1)}^T$ ($i = 1, 2, \dots$) until no more pruning is possible. Denote this maximal pruned context tree (not necessarily of terminal node type) by $\hat{\tau} = \tau_{\hat{c}}$ and its corresponding context function by $\hat{c}(\cdot)$.

STEP 4. If interested in probability distributions, estimate the transition probabilities $p(x_1 \mid c(x_{-\infty}^0))$ by $\hat{P}(x_1 \mid \hat{c}(x_{-\infty}^0))$, where $\hat{P}(\cdot \mid \cdot)$ is defined as in (3.2).

REMARK 3.1. The pruning decision in Step 2 can be related to the Kullback–Leibler distance and to the likelihood ratio test. By definition,

$$\begin{aligned}\Delta_{wu} &= \sum_{x \in \mathscr{X}} \hat{P}(x \mid wu) \log\left(\frac{\hat{P}(x \mid wu)}{\hat{P}(x \mid w)}\right) N(wu) \\ &= D\big(\hat{P}(\cdot \mid wu) \| \hat{P}(\cdot \mid w)\big) N(wu),\end{aligned} \tag{3.4}$$

where $N(wu)$ is defined in (3.1) and $D(P\|Q) = \sum_{x \in \mathscr{X}} P(x)\log(P(x)/Q(x))$ is the Kullback–Leibler distance between two probability measures P and Q on $\mathscr{X}$.

Denote the estimated likelihood function (conditioned on the first state), based on context function $c(\cdot)$ by

$$\hat{P}_c(X_1^n) = \prod_{t=k+1}^{n} \hat{P}\big(X_t \mid c(X_{-\infty}^{t-1})\big), \tag{3.5}$$

where k is the order of $c(\cdot)$ and $\hat{P}(X_t \mid c(X_{-\infty}^{t-1}))$ is defined in (3.2).

Denote by $c(\cdot)$ the context function of a nonpruned context tree and by $c'(\cdot)$ the context function of the subtree, pruned at one terminal node $wu = x_{-\ell+1}^0$ to its parent node $w = x_{-\ell+2}^0$. By the multiplicative structure in (3.5), many terms cancel in the likelihood ratio statistic and the only remaining term is at the node considered for pruning. One gets

$$\Delta_{wu} = \log\left(\frac{\hat{P}_c(X_1^n)}{\hat{P}_{c'}(X_1^n)}\right). \tag{3.6}$$

[If $c'(\cdot)$ is of lower order $k - 1$ than $c(\cdot)$, some minor edge effects due to conditioning on the first variables arise]. Formula (3.6) says that our pruning criterion is nothing else than a likelihood ratio test, but now with a large acceptance region $[0, C\log(n)]$ for the pruned (sub)tree. Our algorithm can be viewed as doing very many likelihood ratio tests.

REMARK 3.2. The cutoff value $K_n \sim C\log(n)$ in Step 2 for the pruning decision is chosen by an asymptotic consideration. Clearly, by the interpretation as likelihood ratio tests, small cutoff values will result in large context trees and overfitting occurs. Estimation of the cutoff value (the constant C in the present formulation) is given in Bühlmann (1999): it aims for optimality with respect to some loss function, whose specification allows tailoring the procedure for particular aims, for example, 0–1 prediction error loss. The cutoff value can be interpreted as a stepwise $1 - \alpha$ quantile for a multiple testing problem with $\alpha = \alpha_n \to 0$ $(n \to \infty)$. The necessity for α_n to converge to zero is, for example, explained in Rissanen (1989).

REMARK 3.3. For every tree $\tau_{(i)}$, the order of testing the terminal nodes wu in Step 2 (or Step 3) is irrelevant.

REMARK 3.4. The pruning criterion Δ_{wu} does not need to be based on the Kullback–Leibler distance. The quantity $D(\hat{P}(\cdot \mid wu)\|\hat{P}(\cdot \mid w))$ in (3.4) could be replaced by the squared L_1-distance $\|\hat{P}(\cdot \mid wu) - \hat{P}(\cdot \mid w)\|_1^2$ (for a definition of $\|\cdot\|_1$ see assumption (A2)). In this case, the cutoff in Step 2 of the context algorithm needs to satisfy $K_n \sim C\log(n)$, $C > 4|\mathscr{X}| + 8$. See the proof of Theorem 5.1 (which is mainly in terms of $\|\cdot\|_1^2$) and Theorem 5.3.

REMARK 3.5. The maximal tree $\tau_{\max}^T$ in Step 1 is constructed on the basis of at least two occurrences of every terminal node in the sequence. The number two is a low enough value in practice which guarantees a sufficiently large initial tree and at least two observations to estimate transition probabilities associated with terminal nodes (states) in $\tau_{\max}^T$. It is easy to show under the assumptions in Section 3.2 that $\mathbb{P}[N(w) \geq 2] \to 1$ $(n \to \infty)$ for $w \in \tau_n = \tau_{c_n}$. Asymptotic properties of the algorithm remain unchanged when replacing the number two by any finite number.

REMARK 3.6. The algorithm makes no a priori length restriction for long contexts (i.e., deep nodes in the tree) such as $\log(n)/\log(|\mathscr{X}|)$ employed by Weinberger, Rissanen and Feder (1995), which can be a severe restriction in practical applications.

Generally, the pruning in the context algorithm can be viewed as hierarchical backward selection. Dependence on some values further back in the history is weaker by considering deep nodes in the tree in a hierarchical way as less relevant. This hierarchic structure is a clear distinction to the CART algorithm [Breiman, Friedman, Olshen and Stone (1984)], where the tree architecture is binary and has no built-in time structure.

3.2. *Consistency under increasing dimensionality.* We give two results, both dealing with consistency when the dimension of the underlying model is allowed to increase. The first one shows consistency for finding the minimal state spaces and the second one describes properties of the estimated probability distributions.

We consider a sequence of VLMCs $(P_n)_{n \in \mathbb{N}}$ with $P_n \in \mathscr{P}$, as described in (2.3) with context tree $\tau_n = \tau_{c_n}$, induced by the context function $c_n(\cdot)$. We make the following assumptions.

(A1) For some $r \in \mathbb{N}$, $(P_n)_{n \in \mathbb{N}}$ satisfies

$$\sup_{n \in \mathbb{N}} \sup_{A \subseteq \mathscr{X}^{k_n}; w, w' \in \mathscr{X}^{k_n}} \left| P_n^{(r)}(A, w) - P_n^{(r)}(A, w') \right| < 1 - 2\kappa$$

$$\text{for some } \kappa > 0,$$

where k_n is the order of the VLMC P_n and $P_n^{(r)}(A, w) = \mathbb{P}_{P_n}[X_{r-k_n+1,n}^{r,n} \in A \mid X_{-k_n+1,n}^{0,n} = w]$ denotes the r-step transition kernel of the embedding Markov chain $(X_{t-k_n+1,n}^{t,n})_{t \in \mathbb{Z}}$.

(A2) Let $b_n = \min_{w \in \tau_n^T} P_n(w)$ and $\varepsilon_n = \min_{wu \in \tau_n^T, u \in \mathscr{X}} \| P_n(\cdot \mid wu) - P_n(\cdot \mid w) \|_1$ [with L_1-distance $\|f\|_1 = \sum_{x \in \mathscr{X}} |f(x)|$ for some $f: \mathscr{X} \to \mathbb{R}$]. Then

$$b_n^{-1} = O\left(\log(n)^{-(1/2+\beta)} n^{1/2}\right) \quad \text{for some } 0 < \beta < \infty \ (n \to \infty),$$

$$\varepsilon_n^{-1} = O\left(\left(\log(n)^{-(1+\delta)} n b_n\right)^{1/2}\right) \quad \text{for some } 0 < \delta < \infty \ (n \to \infty).$$

(A3) The minimal transition probabilities satisfy

$$\frac{1}{\min_{x \in \mathscr{X}, w \in \tau_n} p_n(x \mid w)} = O(n), \qquad n \to \infty.$$

REMARK 3.7. The assumption about transition kernels in (A1) is related to the ergodicity coefficient for stationary Markov processes; compare Iosifescu and Theodorescu (1969), Rajarashi (1990) or Doukhan (1994). It implies that the state processes $(Z_{t,n})_{t \in \mathbb{Z}}$ with $Z_{t,n} = c_n(X_{-\infty,n}^{t,n})$, and also the VLMCs $(X_{t,n})_{t \in \mathbb{Z}}$, are geometrically ϕ-mixing with mixing coefficients bounded by

$$\sup_{n \in \mathbb{N}} \phi_n(j) \le (1 - 2\kappa)^j \quad \text{for all } j \in \mathbb{N}.$$

REMARK 3.8. Assumption (A2) about the minimum stationary probability bounds the size of the terminal node context tree as $|\tau_n^T| \le b_n^{-1} = O(n^{1/2}/\log(n)^{1/2+\beta})$. Note that the number of transition probability parameters in the process P_n is $O(|\tau_n|) \ge O(|\tau_n^T|)$. The above bound, in probabilistic terms, is a weak condition for the number of parameters and there is no explicit restriction on the order $c_n(\cdot)$ (the depth of the context tree τ_n).

REMARK 3.9. For distinguishing a context wu from its parent node w in the terminal node context tree, assumption (A2) guarantees a minimal L_1 distance between the relevant conditional distributions.

In the special case with only one fixed VLMC $P = P_c$ with context tree τ_c, it is sufficient to only assume for the transition probabilities,

$$\min_{x \in \mathscr{X}, w \in \tau_c} p(x \mid w) > 0,$$

which implies assumptions (A1), (A2) with $b_n \equiv b > 0$, $\varepsilon_n \equiv \varepsilon > 0$ and (A3) with $O(1)$.

THEOREM 3.1. *Consider data* $X_{1,n}, \dots, X_{n,n}$ *as in* (2.3), *where* $c_n(\cdot)$ *denotes the context function and* τ_n *the context tree of the process* P_n, *satisfying* (A1)–(A3). *Let* $\hat{P}(\cdot \mid \cdot)$ *be defined as in* (3.2) *and* $\hat{c}(\cdot)$ *be the estimate in Step* 3 *of the context algorithm*. *Then*:

(i) $\lim_{n\to\infty} \mathbb{P}[\hat{c}(\cdot) = c_n(\cdot)] = 1$, *or equivalently* $\lim_{n\to\infty} \mathbb{P}[\hat{\tau} = \tau_n] = 1$;
(ii) $\sup_{x_{-\infty}^1 \in \mathscr{X}^\infty} |\hat{P}(x_1|\hat{c}(x_{-\infty}^0)) - p_n(x_1|c_n(x_{-\infty}^0))| = o_P(1)$ $(n \to \infty)$.

A proof of Theorem 3.1 is given in Section 5. There, more explicit bounds for the events of choosing too large or too small minimal state spaces are given. Theorem 3.1 explains why the context algorithm is a very powerful tool. Even if the dimensionality increases, the estimator $\hat{c}(\cdot)$ (or $\hat{\tau}$) neither chooses a too large nor a too small model asymptotically and is thus robust against model misspecification with respect to sequences in the broad semiparametric class $\mathscr{P}$. The increase in dimensionality of the underlying model is restricted in probabilistic terms but allows a growth as fast as $O(n^{1/2}/\log(n)^s)$ for some $s > 1/2$; or it can be even faster. The problem is thus highly nontrivial: there is possible failure with simple estimation rules which consider models within a fixed increase in dimensionality, independent of the underlying process, but otherwise quite general of the order $O(n^r)$ for some $0 < r < 1/2$. This relates to a good bias-variance trade-off, even for some very high dimensional VLMCs and for general stationary processes which are on the boundary of $\mathscr{P}$. Theorem 3.1 describes the solution of a model selection problem which is impossible to handle with a global selection criterion, due to the extremely large number of possible models. For example, the number of all VLMC submodels of a full $\mathscr{X}$-valued MC of order 4 with $|\mathscr{X}| = 4$ is $\approx 2 \cdot 10^{20}$. The selection criterion is based here on a hierarchical local criterion (Step 2 in the context algorithm) and, interestingly, it works also in the case where the model dimension increases. In theory but never practically feasible, a minimum description length estimator might yield consistent state estimation as well; compare Weinberger and Feder (1994) in the related class of finite-state models.

The next result describes the construction and properties of the estimator $\hat{P}_n$ for the underlying probability measure P_n. Define a metric for probability measures P, Q on $\mathscr{X}^\infty$,

$$\begin{aligned} d(P,Q) &= \sum_{m=1}^{\infty} 2^{-m} d_m\big(P \circ \pi_{1,\dots,m}^{-1}, Q \circ \pi_{1,\dots,m}^{-1}\big), \\ d_m\big(P \circ \pi_{1,\dots,m}^{-1}, Q \circ \pi_{1,\dots,m}^{-1}\big) &= \sup_{x_1^m \in \mathscr{X}^m} \big|P(x_1^m) - Q(x_1^m)\big|. \end{aligned} \tag{3.7}$$

where $\pi_{1,\dots,m}$: $x \mapsto x_1, \dots, x_m$, $(x \in \mathscr{X}^\infty)$ is the coordinate function.

THEOREM 3.2. *Consider data* $X_{1,n}, \dots, X_{n,n}$ *as in* (2.3) *with* P_n *satisfying* (A1)–(A3). *Then*:

(i) *for* $\hat{P}(\cdot \mid \cdot)$ *as in* (3.2) *and* $\hat{c}(\cdot)$ *the estimate in Step* 3 *of the context algorithm*,

$$\lim_{n\to\infty} \mathbb{P}\Big[\textit{the set } \big\{\hat{P}\big(\cdot \mid \hat{c}\big(x_{-\infty}^{0}\big)\big);\, x_{-\infty}^{0} \in \mathscr{X}^{\infty}\big\} \\ \textit{generates a unique stationary probability measure } \hat{P}_n \in \mathscr{P}\Big] = 1.$$

(ii) *For* $\hat{P}_n$ *in* (i) *and* $d(\cdot,\cdot)$ *as in* (3.7), $d(\hat{P}_n, P_n) = o_P(1)$, $n \to \infty$.

(iii) *The process* $\hat{P}_n$ *in* (i) *satisfies*

$$\mathbb{P}\Big[\hat{P}_n \textit{ is } \phi\textit{-mixing with mixing coefficients satisfying} \\ \phi_{\hat{P}_n}(j) \le (1-\kappa)^j \textit{ for all } j \in \mathbb{N}\Big] \to 1\ (n \to \infty).$$

In particular, *the bound for the mixing coefficients* $\phi_{\hat{P}_n}(j)$ *is nonrandom and the same for all* $n \in \mathbb{N}$.

A proof of Theorem 3.2 is given in Section 5. Statement (i) of Theorem 3.2 tells in a constructive way how to simulate the estimated underlying process, the fidi-convergence in (ii) is a minimal requirement for a reasonable estimator, whereas (iii) is important for simulation tasks like bootstrapping complicated statistics.

3.3. *DNA example*. We now present an interesting and instructive application of VLMC estimation. In particular, we demonstrate the usefulness of an estimated context tree as an excellent graphical tool for detecting structure in the time series. Our data consists of three distinct sequences of DNA from the Drosophila genome. We point out that genetic data is a natural candidate for modeling by a VLMC since it has a finite alphabet $\mathscr{X}$, there is a "time" index and its memory vanishes with increasing lag time. Furthermore, while it is known that the data is far from independent, high-order full Markov models are not easily fitted because of the explosion in the number of parameters; compare Braun and Müller (1998) and compare also with Problems 1 and 2 from Section 1. This is compounded by an observed degree of nonstationarity which prohibits estimation over very long sequences. In this environment, it is of paramount importance that the model parameters are used with great economy in order to capture any significant dependent structure.

Our data began as a single 100-thousand base contiguous stretch of DNA from the Drosophila genome (Genbank number DS02740). Each base is one of four possible DNA residues: Adenine, Cytosine, Guanine and Thymine, with abbreviations A, C, G and T, respectively. Using a variety of tools, biologists at Gerry Rubin's lab at the University of California at Berkeley segmented the sequence into genes (which code for amino acid sequences) and nongenes (so-called junk DNA) which are ignored by the cell chemistry. Physically, the genes are spaced apart and separated by junk DNA which we term "intergenes." Moreover, the genes are further segmented into coding regions called exons and noncoding regions called introns. The cell's engine for transcribing DNA first copies the gene (both intron and exon); it then splices out the intron sections. Each gene is in turn subdivided into alternating stretches of exon and intron. We form a single sequence of exons by concatenating all the exons (in the given order). Similarly, we form sequences of introns and intergenes.

Our goal is the application of the VLMC estimation algorithm to learn the dependence structure and to present the estimated minimal state space graphically as a tree, whose branches are the contexts. Application of the algorithm to each of the datasets suggests that complicated structures exist within the exons and the introns. On the other hand, the intergenes showed no complex structure (a first-order Markov model is a good fit). That exons exhibit such structure is not surprising due to constraints imposed by its coding function. The introns do not have a well-understood function, but evidence of structure suggests that the intron is constrained in some way and is thus unable to mutate freely.

We also consider the sequences under a reduction of the quaternary alphabet down to three possible binary alphabets, identifying (1) G with C; (2) G with A; (3) G with T. Equivalences (1) and (2) have genetic meaning, the third has none (reducing the data to random bits). As expected, this final equivalence (3) produces sequences with no dependence structure. The most dramatic finding was produced by the exon sequence reduced to a binary alphabet by identifying the base G with its bonding pair C (A is thus identified with T). The resulting context tree has branches of lengths 0, 3 and 6 only. Interestingly, we thus can represent it in terms of triplets, as shown in Figure 2. Because amino acids are known to be coded by triplets of DNA letters, the structure in Figure 2 has a beautiful biological interpretation. Our finding suggests that the triplet coding structure is strongly present

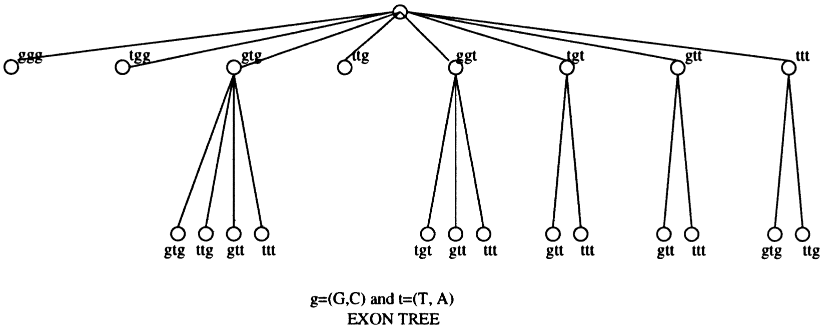

FIG. 2. *Triplet tree representation of the estimated minimal state space for exon sequence. The triplets are denoted in reverse order, for example, the terminal node with concatenation* (ggt)(gtt) *describes the context* $x_0 = g,\ x_{-1} = g,\ x_{-2} = t,\ x_{-3} = g,\ x_{-4} = t,\ x_{-5} = t$ *for the variable* x_1.

despite the dramatic processing of the data (the actual code is not recoverable from this binary reduction). We point out, for emphasis, that the VLMC estimation algorithm learned the triplet structure on its own, a discovery made only for the coding sequences reduced to binary alphabet.

4. The VLMC bootstrap.

Theorem 3.2 indicates that the estimate $\hat{P}_n$ of P_n can be used for resampling. Our proposal will be a bootstrap for stationary categorical time series. Since the semiparametric model (2.2) is dense [with respect to the metric in (3.7)] in the set of stationary $\mathscr{X}$-valued processes, this bootstrap is very general. It offers an attractive and often more accurate alternative to the model free blockwise bootstrap, which has been proposed by Künsch (1989).

STEP 1. Given $\mathscr{X}$-valued data $X_1, \dots, X_n$, fit a VLMC as described in Section 3.1, yielding a stationary probability measure $\hat{P}_n$ on $\mathscr{X}^{\mathbb{Z}}$; see Theorem 3.2.

STEP 2. Draw a finite realization

$$X_1^*, \dots, X_n^* \sim \hat{P}_n \circ \pi_{1,\dots,n}^{-1}.$$

The variables $X_1^*, \dots, X_n^*$ are called the VLMC bootstrap sample; they are nothing else than one random sample from the fitted VLMC. In practice, one would choose some starting values, generate a longer random sample via the estimated transition probabilities $\hat{P}(x_1 \mid \hat{c}(x_{-\infty}^0))$ and then use the last n elements as our bootstrap sample. Such a device tries to avoid nonstationarity effects due to starting values in a simulated Markov chain. Of course, one could also draw bootstrap samples of size $m \neq n$; compare Bickel, Götze and van Zwet (1997).

Given an estimator $T_n = h_n(X_1, \dots, X_n)$, which is a measurable function of $X_1, \dots, X_n$, the bootstrapped estimator is defined by the plug-in rule $T_n^* = h_n(X_1^*, \dots, X_n^*)$. This plug-in rule, which is also the basis to Efron's (1979) bootstrap for the independent case, is very convenient in practice. Once the bootstrap sample is constructed, bootstrapping can be done with exactly the same computing tools or programs as for the original estimator T_n. This is not the case with the blockwise bootstrap [Künsch (1989)] if the estimator T_n is nonsymmetric in the observations $X_1, \dots, X_n$, for example, the estimators in (S1) and (S2) from Section 4.2. Quantities induced by the resampling in Step 2 are denoted by an asterisk *.

4.1. *Consistency of the VLMC bootstrap under increasing dimensionality.*

We present here an asymptotic result justifying the use of the above-defined VLMC bootstrap for estimators T_n which are smooth functions of means. We will also discuss informally why the VLMC bootstrap should work in the more general framework of empirical processes, without giving the exact arguments.

We assume to have observations $X_{1,n},\dots,X_{n,n}\in\mathscr{X}$ from a sequence of VLMCs as given in (2.3). Consider the class of estimators which are smooth functions of means,

$$T_n = g\left\{(n-m+1)^{-1}\sum_{t=1}^{n-m+1} f\left(X_{t,n}^{t+m-1,n}\right)\right\},\qquad 1\le m<\infty, \tag{4.1}$$
$$f=(f_1,\dots,f_v)':\mathscr{X}^m\to\mathbb{R}^v,\ g=(g_1,\dots,g_w)':\mathbb{R}^v\to\mathbb{R}^w \text{ smooth.}$$

The function f is bounded, since $|\mathscr{X}|<\infty$. Examples include estimators of transition probabilities in full Markov chains of order $m-1$ or other functions of frequencies of tuples up to size m, such as the Z scores used in genetics [cf. Prum, Rudolphe and de Turckheim (1995)]. We usually make the following assumption.

(B1) The estimator T_n is given by (4.1) with g having continuous partial derivatives in a neighborhood of $\theta_n=\mathbb{E}[f(X_{1,n},\dots,X_{m,n})]$. Also, there exists an $n_0\in\mathbb{N}$, such that for every $n\ge n_0$,

$$\left[\sum_{k=-n+1}^{n-1}\operatorname{Cov}\left(f_i\left(X_{0,n}^{m-1,n}\right),f_j\left(X_{k,n}^{k+m-1,n}\right)\right)\right]_{i,j=1}^{v}\quad\text{is positive definite.}$$

REMARK 4.1. The assumption about positive definiteness of covariance matrices simplifies when assuming a limiting model P, where $\lim_{n\to\infty}d(P_n,P)=0$ for the metric $d(\cdot,\cdot)$ defined in (3.7). Generally, P is not a VLMC anymore. It is then sufficient to assume

$$\sum_{k=-\infty}^{\infty}\left|\operatorname{Cov}\left(f_i\left(X_0^{m-1}\right),f_j\left(X_k^{k+m-1}\right)\right)\right|<\infty,\qquad i,j\in\{1,\dots,v\},$$
$$\left[\sum_{k=-\infty}^{\infty}\operatorname{Cov}\left(f_i\left(X_0^{m-1}\right),f_j\left(X_k^{k+m-1}\right)\right)\right]_{i,j=1}^{v}\quad\text{is positive definite,}$$

where $(X_t)_{t\in\mathbb{Z}}\sim P$.

The following theorem justifies the VLMC bootstrap for smooth functions of means.

THEOREM 4.1. *Let* $X_{1,n},\dots,X_{n,n}$ *be as in* (2.3) *with* P_n *satisfying* (A1)–(A3). *Assume also that* (B1) *holds. Let the VLMC bootstrap be defined as in Section* 4 *and denote by* $\theta_n^*=\mathbb{E}^*[f((X^*)_1^m)]$. *Then*

$$\sup_{x\in\mathbb{R}^w}\left|\mathbb{P}^*\left[n^{1/2}(T_n^*-g(\theta_n^*))\le x\right]-\mathbb{P}\left[n^{1/2}(T_n-g(\theta_n))\le x\right]\right|=o_P(1),$$
$$n\to\infty.$$

The proof of Theorem 4.1 is given in Section 5. Our results can be generalized to consistency of the VLMC bootstrap for general empirical processes, because the VLMC bootstrap for categorical time series satisfies a ϕ-mixing property with exponentially decaying mixing coefficients; see Theorem 3.2(iii). These extensions are useful for studying the bootstrap consistency of estimators

$$T_n = T(\nu_n), \tag{4.2}$$

which are given as a smooth functional of a general empirical measure ν_n. The class of estimators in (4.2) is considerably larger than the class in (4.1). It includes as examples the maximum likelihood estimators in generalized linear models of autoregressive type with quite general link functions; compare Fahrmeir and Tutz (1994).

4.2. *Simulations.* We study here the VLMC bootstrap for variance estimation in various cases by simulation. We represent VLMC models by context trees and equip terminal nodes with tuples, describing the transition probabilities. A tuple $(i_0, \dots, i_{|\mathscr{X}|-1})$ corresponds to $p(j \mid w) = i_j / \sum_{j=0}^{|\mathscr{X}|-1} i_j$, $j \in \{0, \dots, |\mathscr{X}| - 1\}$ (without loss of generality we let $\mathscr{X} = \{0, \dots, |\mathscr{X}| - 1\}$). We consider the following models: (M1): full binary Markov chain of order 3; (M2) full quaternary Markov chain of order 2; (M3): semisparse binary VLMC of order 5; (M4): semisparse quaternary VLMC of order 3; (M5): sparse binary VLMC of order 8; (M6): sparse quaternary VLMC of order 4. The precise specifications are given by the trees and numbers shown in Figure 3.

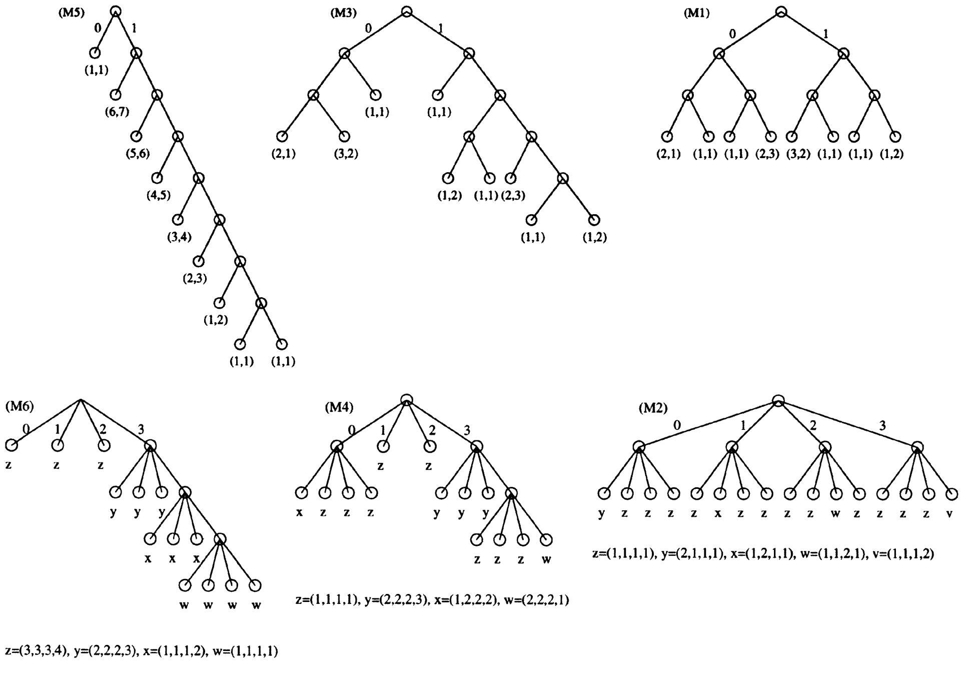

FIG. 3. *Tree representations of the VLMC models* (M1)–(M6). *Transition probabilities are specified by tuples at terminal nodes.*

We also consider (M7): $X_t = 1_{[Y_t > 0]}$ where $(Y_t)_{t \in \mathbb{Z}}$ is a stationary nonlinear process

$$Y_t = \big(0.5 + 0.9 \exp\big(-2.354 Y_{t-1}^2\big)\big) Y_{t-1}$$
$$-\big(0.8 - 1.8 \exp\big(-2.354 Y_{t-1}^2\big)\big) Y_{t-2} + Z_t$$

with $(Z_t)_{t \in \mathbb{Z}}$ an i.i.d. sequence, $Z_t \sim \mathscr{N}(0,1)$ and Z_t independent from $\{Y_s; s < t\}$. The process $(Y_t)_{t \in \mathbb{Z}}$ is also known as exponential AR(2). The quantized binary process $(X_t)_{t \in \mathbb{Z}}$ in (M7) is non-Markovian, although the $\mathbb{R}$-valued $(Y_t)_{t \in \mathbb{Z}}$ is Markov of order 2. It is interesting to see whether the VLMC bootstrap provides a good finite-sample approximation to this model which is not a VLMC. This is also an interesting test case to make a fair comparison between the VLMC and blockwise bootstrap [Künsch (1989)].

As sample sizes, we choose $n = 1000$ and $n = 2000$. We consider the following statistics.

(S1) $T_n = \hat{P}(1 \mid 0) = N_n(10)/N_n(0)$ for binary models (M1), (M3) and (M5).

(S2) $T_n = N_n(133)$, the frequency of the word $(x_3, x_2, x_1) = (1, 3, 3)$ as defined in (3.1), for quaternary models (M2), (M4) and (M6).

(S3) $T_n = \overline{X}_n = n^{-1} \sum_{t=1}^n X_t$, the relative frequency of symbol 1, for the binary model (M7).

The variance estimates are

$$\hat{\sigma}_n^2 = n \operatorname{Var}^*(T_n^*) \text{ for } \sigma_n^2 = n \operatorname{Var}(T_n) \text{ in (S1) and (S3)},$$
$$\hat{\sigma}_n^2 = \operatorname{Var}^*(T_n^*) \text{ for } \sigma_n^2 = \operatorname{Var}(T_n) \text{ in (S2)},$$

based on the VLMC bootstrap with 500 resamples (note the different standardizations).

The results are given in Tables 1 and 2. Moments of the bootstrap variances $\hat{\sigma}_n^2$ are estimated with 200 simulations over the different models; the true values of σ_n^2 are estimated with 1000 simulations. The relative mean square error is given by $\text{relMSE}(\hat{\sigma}_n^2) = \mathbb{E}|\hat{\sigma}_n^2 - \sigma_n^2|^2/\sigma_n^4$. Estimated standard errors for the bias and relMSE are given in parentheses. We tried different cutoff values K which act as tuning parameter; see also Remark 3.2. We report them as $\chi^2_{|\mathscr{X}|-1;\alpha}/2$ quantiles with different levels α, corresponding to the asymptotic distribution of *one* log-likelihood ratio statistic in (3.6). This is sometimes more intuitive than the numerical value of C in the cutoff $K_n \sim C \log(n)$ from theory. On the other hand, we point out the danger that direct interpretation of the cutoff as a $\chi^2_{|\mathscr{X}|-1;\alpha}/2$ quantile with α fixed and not depending on the sample size contradicts the very essence of the context algorithm in (3.3). For the binary models (M1), (M3), (M5), we used the cutoffs $\chi^2_{1;\alpha}/2$, for the quaternary models (M2), (M4), (M6) the cutoffs $\chi^2_{3;\alpha}/2$, both denoted in short by $\alpha 100\%$.

The results are promising in that the relative mean square error is most often smaller than 5%. Generally, the performance is better for sparse models. This indicates that the algorithm adapts to sparseness; it is exactly in these cases, where other methods are more likely to fail.

TABLE 1
VLMC bootstrap variance estimates, sample size $n = 1000$

	cutoff (in %)	σ_n^2	$E[\hat{\sigma}_n^2] - \sigma_n^2$	$\mathrm{Var}(\hat{\sigma}_n^2)$	$\mathrm{relMSE}(\hat{\sigma}_n^2)$
(M1, S1)	95	0.81	−0.04 (0.010)	0.02	0.033 (0.0035)
(M1, S1)	98	0.81	−0.10 (0.011)	0.02	0.053 (0.0040)
(M1, S1)	99.9	0.81	−0.26 (0.009)	0.02	0.127 (0.0038)
(M3, S1)	95	0.67	−0.02 (0.009)	0.01	0.031 (0.0029)
(M3, S1)	98	0.67	−0.05 (0.009)	0.01	0.036 (0.0030)
(M3, S1)	99.9	0.67	−0.17 (0.005)	0.01	0.078 (0.0027)
(M5, S1)	95	0.528	0.007 (0.0054)	0.006	0.021 (0.0027)
(M5, S1)	98	0.528	−0.005 (0.0031)	0.002	0.007 (0.0006)
(M5, S1)	99.9	0.528	0.003 (0.0027)	0.002	0.005 (0.0005)
(M2, S2)	95	14.5	−0.5 (0.21)	9.1	0.045 (0.0051)
(M2, S2)	98	14.5	0.1 (0.17)	5.8	0.028 (0.0027)
(M2, S2)	99.9	14.5	0.0 (0.14)	3.8	0.018 (0.0019)
(M4, S2)	95	14.1	−0.3 (0.18)	6.4	0.032 (0.0044)
(M4, S2)	98	14.1	−0.4 (0.17)	5.5	0.029 (0.0042)
(M4, S2)	99.9	14.1	−0.5 (0.12)	2.8	0.015 (0.0014)
(M6, S2)	95	11.2	0.0 (0.15)	4.8	0.038 (0.0043)
(M6, S2)	98	11.2	−0.1 (0.13)	3.1	0.025 (0.0029)
(M6, S2)	99.9	11.2	−0.3 (0.10)	2.0	0.017 (0.0026)
(M7, S3)	95	0.80	−0.03 (0.009)	0.02	0.029 (0.0029)
(M7, S3)	98	0.80	−0.11 (0.009)	0.02	0.043 (0.0030)
(M7, S3)	99.9	0.80	−0.24 (0.005)	0.01	0.100 (0.0032)

For comparison, we also tried the blockwise bootstrap [Künsch (1989)] in the cases (M5, S1) and (M7, S3) for sample size $n = 1000$ with different blocklengths l; see Table 3. A graphical representation is given in Figure 4. The comparison is at least fair in (M7, S3) where the model is not a VLMC [and by the structure of $(Y_t)_{t \in \mathbb{Z}}$, the quantized series $(X_t)_{t \in \mathbb{Z}}$ doesn't allow sparse approximation]. In this case, both bootstraps have similar performances; the best tuned VLMC bootstrap is about 15% better (in terms of relative mean square error) than the best-tuned blockwise bootstrap. In case (M5, S1) the blockwise bootstrap exhibits a serious bias and a large variability. The VLMC bootstrap is far better for this sparse VLMC (M5). We conclude that the VLMC bootstrap is at least as good as the blockwise bootstrap and enjoys the important practical advantage of being defined as a plug-in rule; see Section 4.

The role of the cutoff as tuning parameter of the VLMC bootstrap is found as follows: the absolute value of the bias of the bootstrap variance estimator increases and the variance decreases with increasing cutoff parameter. This is expected since a larger cutoff parameter leads to a lower dimensional fitted VLMC model, by design of the context algorithm. Note that in some simula-

TABLE 2
VLMC bootstrap variance estimates, sample size $n = 2000$

	cutoff (in %)	σ_n^2	$E[\hat{\sigma}_n^2] - \sigma_n^2$	$\mathrm{Var}(\hat{\sigma}_n^2)$	$\mathrm{relMSE}(\hat{\sigma}_n^2)$
(M1, S1)	95	0.82	−0.01 (0.009)	0.01	0.022 (0.0021)
(M1, S1)	98	0.82	−0.02 (0.007)	0.01	0.016 (0.0018)
(M1, S1)	99.9	0.82	−0.14 (0.011)	0.02	0.065 (0.0048)
(M3, S1)	95	0.67	0.00 (0.006)	0.01	0.014 (0.0013)
(M3, S1)	98	0.67	−0.03 (0.006)	0.01	0.017 (0.0016)
(M3, S1)	99.9	0.67	−0.09 (0.007)	0.01	0.042 (0.0033)
(M5, S1)	95	0.518	0.007 (0.0038)	0.003	0.011 (0.0012)
(M5, S1)	98	0.518	0.002 (0.0031)	0.002	0.007 (0.0008)
(M5, S1)	99.9	0.518	0.009 (0.0025)	0.001	0.005 (0.0004)
(M2, S2)	95	12.9	1.2 (0.16)	4.9	0.038 (0.0038)
(M2, S2)	98	12.9	1.4 (0.15)	4.4	0.039 (0.0042)
(M2, S2)	99.9	12.9	1.9 (0.13)	3.2	0.040 (0.0025)
(M4, S2)	95	14.7	−0.7 (0.15)	4.3	0.022 (0.0022)
(M4, S2)	98	14.7	−0.6 (0.11)	2.6	0.014 (0.0013)
(M4, S2)	99.9	14.7	−1.0 (0.09)	1.7	0.012 (0.0011)
(M6, S2)	95	11.5	−0.1 (0.14)	4.0	0.030 (0.0036)
(M6, S2)	98	11.5	−0.3 (0.10)	1.9	0.015 (0.0015)
(M6, S2)	99.9	11.5	−0.5 (0.08)	1.4	0.012 (0.0016)
(M7, S3)	95	0.81	−0.03 (0.008)	0.01	0.018 (0.0017)
(M7, S3)	98	0.81	−0.06 (0.007)	0.01	0.022 (0.0018)
(M7, S3)	99.9	0.81	−0.23 (0.006)	0.01	0.089 (0.0036)

TABLE 3
Blockwise bootstrap variance estimates; sample size $n = 1000$

	blocklength	σ_n^2	$E[\hat{\sigma}_n^2] - \sigma_n^2$	$\mathrm{Var}(\hat{\sigma}_n^2)$	$\mathrm{relMSE}(\hat{\sigma}_n^2)$
(M5, S1)	$\ell = 10$	0.528	0.106 (0.0057)	0.007	0.065 (0.0053)
(M5, S1)	$\ell = 20$	0.528	0.058 (0.0071)	0.010	0.049 (0.0066)
(M5, S1)	$\ell = 30$	0.528	0.030 (0.0084)	0.014	0.054 (0.0065)
(M7, S3)	$\ell = 10$	0.80	−0.17 (0.005)	0.01	0.051 (0.0027)
(M7, S3)	$\ell = 20$	0.80	−0.09 (0.008)	0.01	0.035 (0.0027)
(M7, S3)	$\ell = 30$	0.80	−0.07 (0.011)	0.02	0.045 (0.0038)

tion examples, this behavior is not significantly visible. For the blocklength in the blockwise bootstrap in turn, the general asymptotic behavior is observed, that is, the bias decreases, whereas the variance increases with growing l.

5. Proofs.

We first recall some notation. We usually denote by $w, u, v \in \bigcup_{m=0}^{\infty} \mathscr{X}^m$ sequences (written in reverse "time") $w = (w_{|w|}, \dots, w_2, w_1)$; the concatenation is $wu = (w_{|w|}, \dots, w_2, w_1, u_{|u|}, \dots, u_1) \in \bigcup_{m=0}^{\infty} \mathscr{X}^m$. Transition

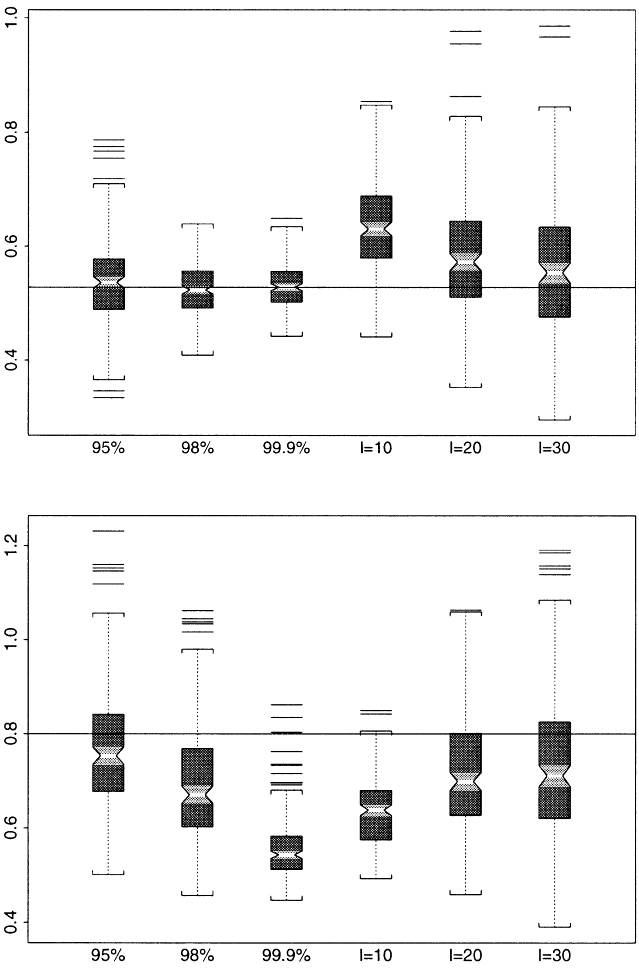

FIG. 4. *Boxplots of bootstrap variance estimates with* $n = 1000$, *for case* $(M5, S1)$ *on top and for case* $(M7, S3)$ *at the bottom. VLMC bootstrap estimates are denoted with* $\alpha 100\%$, *corresponding to their* $\chi^2_{1;\,\alpha}/2$-*quantiles as cutoff values. Blockwise bootstrap estimates are denoted with their blocklengths* ℓ. *The line denotes the true variance.*

probabilities in a context tree τ are indexed by $w \in \tau$ and abbreviated by $p_w(\cdot) = p(\cdot \mid w)$, estimated transition probabilities are denoted by $\hat{P}_w(x) = N(xw)/N(w)$ $(x \in \mathcal{X})$, $N(\cdot)$ as defined in (3.1). We also abbreviate by $P_w(x) = P(xw)/P(w)$ for general $w \in \bigcup_{m=1}^{\infty} \mathcal{X}^m$, $x \in \mathcal{X}$ (w is not necessarily a context in τ) and P a stationary probability measure on $\mathcal{X}^{\mathbb{Z}}$ with $P(x) = \mathbb{P}_P[X_1^m = x]$ $(x \in \mathcal{X}^m)$. We recall that for any $w' = wu$ $(u \in \mathcal{X})$, we have defined $\Delta_{w'} = D(\hat{P}_{wu} \| \hat{P}_w) N(w')$. When looking at a sequence $(P_n)_{n \in \mathbb{N}}$ of VLMCs, we sometimes drop the index n.

PROOF OF THEOREM 3.1. We define first the events of under- and overestimation for sample size n,

$$U_n = \left\{ \text{there exists } w \in \hat{\tau} \text{ with } wu \in \tau_n,\, wu \notin \hat{\tau} \text{ for some } u \in \bigcup_{m=1}^{\infty} \mathscr{X}^m \right\},$$

$$O_n = \left\{ \text{there exists } w \in \tau_n \text{ with } wu \in \hat{\tau},\, wu \notin \tau_n \text{ for some } u \in \bigcup_{m=1}^{\infty} \mathscr{X}^m \right\}.$$

Note that by (3.3) we can also characterize U_n and O_n in terms of the pruning criterion $\Delta_{wu} < K_n = C\log(n)$. The error event is

$$E_n = \{\hat{\tau} \neq \tau_n\} = U_n \cup O_n.$$

THEOREM 5.1. *Assume that* (A1) *and* (A2) *with* $0 < \beta,\ \delta < \infty$ *hold. Then*

$$\mathbb{P}[U_n] = O\Big(\max\{n^{-\log(n)^{2\beta} D_1}, n^{-\log(n)^{\delta} D_2}\}\Big), \qquad n \to \infty$$

for some constants $0 < D_1,\ D_2 < \infty$.

PROOF. We partition the underestimation event U_n using the event

$$D_n = \{N(w) \geq \rho_n \text{ for every } w \in \tau_n^T\},$$

where ρ_n is a constant to be chosen later. Thus $\mathbb{P}[U_n] \leq \mathbb{P}[U_n \cap D_n] + \mathbb{P}[D_n^c]$. We will pursue a bound on $\mathbb{P}[U_n]$ by bounding both $\mathbb{P}[U_n \cap D_n]$ and $\mathbb{P}[D_n^c]$. First, by restricting without loss of generality to underestimation of τ_n^T,

$$\begin{aligned}
\mathbb{P}[U_n \cap D_n] &\leq \sum_{wu \in \tau_n^T,\, u \in \mathscr{X}} \mathbb{P}[\Delta_{wu} < C\log(n),\, N(wu) \geq \rho_n] \\
&= \sum_{wu \in \tau_n^T,\, u \in \mathscr{X}} \sum_{k=\rho_n}^{n} \sum_{j=k}^{n} \mathbb{P}\Bigg[D\big(\hat{P}_{wu} \| \hat{P}_w\big) < \frac{C\log(n)}{k}, \\
&\qquad\qquad N(wu) = k,\, N(w) = j \Bigg].
\end{aligned} \tag{5.1}$$

It is well known [cf. Cover and Thomas (1991)], that the divergence can be lower bounded by the L_1 distance, $D(\hat{P}_{wu}\|\hat{P}_w) \geq \frac{1}{2}\|\hat{P}_{wu} - \hat{P}_w\|_1^2$ and that $\|\hat{P}_{wu} - \hat{P}_w\|_1^2 = 2(\hat{P}_{wu}(A) - \hat{P}_w(A))^2$, where $A = \{x \in \mathscr{X};\ \hat{P}_{wu}(x) > \hat{P}_w(x)\}$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left[D\big(\hat{P}_{wu} \| \hat{P}_w\big) < \frac{C\log(n)}{k},\, N(wu) = k,\, N(w) = j \right] \\
&\quad \leq \mathbb{P}\left[\big(\hat{P}_{wu}(A) - \hat{P}_w(A)\big)^2 < \frac{C\log(n)}{k},\, N(wu) = k,\, N(w) = j \right].
\end{aligned} \tag{5.2}$$

Now because of assumption (A2), it must be that either $\hat{P}_{wu}(A)$ or $\hat{P}_w(A)$ is far from $P_{wu}(A)$ or $P_w(A)$, respectively. We formalize this by letting $\gamma_n^2(k) = C\log(n)/k$ and $\hat{P}_{wu}(x) = a$, $\hat{P}_w(x) = b$, $p_{wu}(x) = r$ and $p_w(x) = s$, where

$x \in \mathscr{X}$. Our goal is to establish that if $|a - b|$ is small, then either $|r - a|$ is large or $|s - b|$ is large. First assume, without loss of generality, that $r > s$. We have by (A2) that $r - s > \varepsilon_n$. Now if $b < s$, then $|a - b| < \gamma_n(k)$ implies that $|a - r| > \varepsilon_n - \gamma_n(k)$. Furthermore, if $b > r$, then it must be that $|s - b| > \varepsilon_n$. Now if $s \le b \le r$ then either $s \le b < s + (r - s)/2$, in which case $|r - a| > \varepsilon_n/2 - \gamma_n(k)$ or $r - (r - s)/2 \le b \le r$, in which case $|s - b| > \varepsilon_n/2$. Taken together we have proved that if $|\hat{P}_{wu}(x) - \hat{P}_w(x)| < \gamma_n(k)$, then either $|\hat{P}_{wu}(x) - p_{wu}(x)| > (\varepsilon_n/2) - \gamma_n(k)$ or $|\hat{P}_w(x) - P_w(x)| > (\varepsilon_n/2) - \gamma_n(k)$. Thus, when applied to (5.2), we have proved that for

$$a_n(k) = \left(\frac{\varepsilon_n}{2} - \gamma_n(k)\right)^2, \tag{5.3}$$

it must be that

$$\begin{aligned}
&\mathbb{P}\left[D\left(\hat{P}_{wu} \| \hat{P}_w\right) < \frac{C\log(n)}{k}, N(wu) = k, N(w) = j\right] \\
&\quad \le \mathbb{P}\left[\sum_{x \in A} \left|\hat{P}_{wu}(x) - p_{wu}(x)\right| > a_n(k)^{1/2}, N(wu) = k\right] \\
&\qquad + \mathbb{P}\left[\sum_{x \in A} \left|\hat{P}_w(x) - P_w(x)\right| > a_n(k)^{1/2}, N(w) = j\right] \\
&\quad \le |\mathscr{X}| \max_{x \in \mathscr{X}} \mathbb{P}\left[\left|\hat{P}_{wu}(x) - p_{wu}(x)\right| > a_n(k)^{1/2}, N(wu) = k\right] \\
&\qquad + |\mathscr{X}| \max_{x \in \mathscr{X}} \mathbb{P}\left[\left|\hat{P}_w(x) - P_w(x)\right| > a_n(k)^{1/2}, N(w) = j\right].
\end{aligned} \tag{5.4}$$

We will now choose $\rho_n = b_n n/2 \ge \text{const. } n^{1/2} \log(n)^{1/2+\beta}$. Then, $\max_{k \ge \rho_n} \gamma_n^2(k) \le \text{const.}\, C \log(n)/(nb_n)$. Thus, it follows by assumption on ε_n,

$$\min_{k \ge \rho_n} a_n(k) = \min_{k \ge \rho_n}\left(\frac{\varepsilon_n}{2} - \gamma_n(k)\right)^2 \ge \text{const.} \frac{\log(n)^{1+\delta}}{nb_n},$$

and hence

$$\min_{k \ge \rho_n} k a_n(k) \ge \text{const.} \log(n)^{1+\delta}.$$

We treat the two cases on the right-hand side of (5.4) simultaneously by denoting $v = wu$ or $v = w$, respectively. Let $p = P_v(x)$ and let $\hat{p} = \hat{P}_v(x)$. We would like to find an upper bound for the probability of the event $\{|p - \hat{p}|^2 > a_n(k), N(v) = k\}$. Since there are a random number of terms in the denominator of $\hat{p}$ we cannot apply any large deviations bound directly. Instead we consider the extension of X_1^n to the infinite sequence $(X_t)_{t \in \mathbb{N}}$. Define

$$I_i = \left\{\text{the time of the } i\text{th occurrence of } v \text{ in } (X_t)_{t \in \mathbb{N}}\right\}, \qquad i \in \mathbb{N}.$$

Then let

$$W_i = X_{I_i + 1}, \text{ the symbol that occurs after the } i\text{th occurrence of } v.$$

The sequence $(W_i)_{i \in \mathbb{N}}$ is stationary ϕ-mixing with mixing coefficients bounded by the same bound as the original sequence $(X_t)_{t \in \mathbb{Z}}$. The marginal probability distribution of W_1 on $\mathscr{X}$ is equal to P_v. Let $Y_i = 1_{[W_i = x]}$. Now observe that

$$\left\{ \left| \sum_{i=1}^{N(v)} \frac{Y_i}{N(v)} - p \right|^2 > a_n(k),\, N(v) = k \right\} \subseteq \left\{ \left| \sum_{i=1}^{k} \frac{Y_i}{k} - p \right|^2 > a_n(k) \right\}.$$

Thus, we have established the upper bound,

$$\mathbb{P}\Big[|\hat{p} - p|^2 > a_n(k), N(v) = k \Big] \leq \mathbb{P}\left[\left| \sum_{i=1}^{k} \frac{Y_i}{k} - p \right|^2 > a_n(k) \right]. \tag{5.5}$$

At this point we are readily able to apply an exponential inequality.

Lemma 5.1. *Let $(Y_i)_{i \in \mathbb{N}}$ with $E[Y_1] = p$ be defined as above and $a_n(k)$ as in* (5.3). *Assume the conditions* (A1) *and* (A2) *with $0 < \beta$, $\delta < \infty$. Then, for $k \geq \rho_n = b_n n/2$,*

$$\sup_{0<p<1} \mathbb{P}\left[\left| \sum_{i=1}^{k} \frac{Y_i}{k} - p \right|^2 > a_n(k) \right] = O(n^{-\log(n)^{\delta} F})$$

for some constant $0 < F < \infty$.

Proof. By assumption (A1), the process $(X_t)_{t \in \mathbb{Z}}$ has mixing coefficients $\phi(j) \leq (1 - 2\kappa)^j$, and the same bound applies also for the mixing coefficients of the process $(Y_i)_{i \in \mathbb{N}}$. Now apply Theorem 4 from Doukhan [(1994), Chapter 1.4.2] with $q = \log(n)^{1+\delta}$ and note that $k \geq \rho_n = b_n n/2 \geq \text{const.} n^{1/2} \log(n)^{1/2+\beta}$. Using that $k a_n(k) \geq \text{const.} \log(n)^{1+\delta}$ for all $k \geq \rho_n$ completes the proof. □

Denote by $M_n = \exp(-F \log(n)^{1+\delta})$. A straightforward application of Lemma 5.1 to (5.5) proves that for $k, j \geq \rho_n$,

$$\max_{x \in \mathscr{X}} \mathbb{P}\Big[\big(\hat{P}_{wu}(x) - p_{wu}(x)\big)^2 > a_n(k), N(wu) = k \Big] = O(M_n),$$

$$\max_{x \in \mathscr{X}} \mathbb{P}\Big[\big(\hat{P}_{w}(x) - P_{w}(x)\big)^2 > a_n(k), N(w) = j \} = O(M_n).$$

Thus, together with (5.1), (5.2) and (5.4),

$$\mathbb{P}[U_n \cap D_n] \leq |\mathscr{X}| \sum_{wu \in \tau_n^T,\, u \in \mathscr{X}} \sum_{k=\rho_n}^{n} \sum_{j=k}^{n} O(M_n). \tag{5.6}$$

By Remark 3.8, (5.6) and assumption (A2),

$$\begin{aligned} \mathbb{P}[U_n \cap D_n] &= O\big(|\tau_n^T| n^2 M_n\big) \\ &= O\Big(b_n^{-1} n^2 \exp\big(-\text{const.} \log(n)^{1+\delta}\big)\Big) = O\big(n^{-\log(n)^{\delta} \text{const.}}\big). \end{aligned} \tag{5.7}$$

To complete the proof of Theorem 5.1, we need to bound $\mathbb{P}[D_n^c]$. Using the union bound and assumption (A2) we get

$$\begin{aligned}\mathbb{P}[D_n^c] &\le \sum_{w \in \tau_n^T} \mathbb{P}[N(w) < \rho_n] = \sum_{w \in \tau_n^T} \mathbb{P}[N(w) - E[N(w)] \\ &\qquad\qquad\qquad < \rho_n - (n - |w| + 1) P_n(w)] \\ &\le \sum_{w \in \tau_n^T} \mathbb{P}[N(w) - E[N(w)] < -b_n n/3] \\ &\le \sum_{w \in \tau_n^T} \mathbb{P}[|N(w) - E[N(w)]| > b_n n/3].\end{aligned}$$

We bound this quantity by the following exponential inequality.

LEMMA 5.2. *Assume that* (A1) *and* (A2) *with* $0 < \beta$, $\delta < \infty$ *hold. Then*

$$\max_{w \in \tau_n^T} \mathbb{P}[|N(w) - \mathbb{E}[N(w)]| \ge b_n n/3] = O(n^{-\log(n)^{2\beta} G})$$

for some constant $0 < G < \infty$.

PROOF. Since $w \in \tau_n$, we can write

$$N(w) = \sum_{t=1}^{n} 1_{[Z_{t,n} = w]}, \qquad Z_{t,n} = c(X_{-\infty,n}^{t,n}).$$

By assumption (A1), $(Z_{t,n})_{t \in \mathbb{Z}}$ is ϕ-mixing with mixing coefficients bounded by $\sup_{n \in \mathbb{N}} \phi_n(j) \le (1 - 2\kappa)^j$; see Remark 3.7. Now apply Theorem 4 in Doukhan [(1994), Chapter 1.4.2] with $q = \log(n)^{1+2\beta}$. Using that $nb_n \ge$ const. $n^{1/2} \log(n)^{1/2+\beta}$ completes the proof. □

By Lemma 5.2,

$$\mathbb{P}[D_n^c] = O\bigl(b_n^{-1} n^{-\log(n)^{2\beta}\,\text{const.}}\bigr) = O(n^{-\log(n)^{2\beta}\,\text{const.}}),$$

where the last estimate follows from (A2). Together with (5.7) we complete the proof of Theorem 5.1. □

We now consider the overestimation event $O_n = \{$there exists $w' \in \tau_n$ with $w = w'u \in \hat{\tau}$, $w = w'u \notin \tau_n$ for some $u \in \bigcup_{m=1}^{\infty} \mathscr{X}^m\}$. For a sequence w to be an element of $\hat{\tau}^T \subseteq \hat{\tau}$, it is necessary that $N(w) > 1$ and $\Delta_w \ge C \log(n)$. Now Weinberger, Rissanen and Feder (1995) establish for any $w = w'u$ ($w' \in \tau_n$, $u \in \bigcup_{m=1}^{\infty} \mathscr{X}^m$) with $w = w'u \notin \tau_n$,

$$\mathbb{P}[\Delta_w \ge C \log(n+1)] \le (n+1)^{2|\mathscr{X}|}(n+1)^{-C}.$$

In their algorithm, an overestimation event can only occur at any string w if $|w| \le \log(n)/\log(|\mathscr{X}|)$. Thus they establish that

$$\mathbb{P}[O_n] \le \sum_{|w| \le \log(n)/\log(|\mathscr{X}|)} (n+1)^{-C+2|\mathscr{X}|} \le n^{-C+2|\mathscr{X}|+1}.$$

The last inequality follows since, for any m there are no more than $|\mathscr{X}|^m$ distinct sequences w with length $|w| = m$.

It is possible to prove a stronger result, eliminating the need for a length restriction on $|w|$. We give only a detailed outline of such a proof.

LEMMA 5.3. *Let swv be any possible string with $s \in \tau_n$, $w \in \bigcup_{m=0}^{\infty} \mathscr{X}^m$, $v \in \mathscr{X}$ and $swv \notin \tau_n$. Let $O_n(swv) = \{\Delta_{swv} \geq C\log(n), N(swv) > 1\}$. Denote by $p_{\min}(n) = \min_{x \in \mathscr{X}, w \in \tau_n} p_w(x)$ and by $\hat{\tau}_{\max}$ the maximal context tree in Step* 1 *of the context algorithm. Then, under the assumptions* (A1)–(A3),

$$\mathbb{P}[O_n(swv)] \leq \frac{1}{p_{\min}(n)} \mathbb{P}[sw \in \hat{\tau}_{\max}] n^{-C+2|\mathscr{X}|}.$$

A proof is given below.

THEOREM 5.2. *Under the assumptions* (A1)–(A3),

$$\sum_{n=1}^{\infty} \mathbb{P}[O_n]\log(n) < \infty.$$

PROOF. We apply Lemma 5.3 for swv,

$$\mathbb{P}[O_n] \leq \sum_{swv} \mathbb{P}[O_n(swv)] = O(n^{-C+2|\mathscr{X}|+1}) \sum_{swv} \mathbb{P}[sw \in \hat{\tau}_{\max}],$$

where the last estimate follows from (A3).

Let L be the number of sequences which occur at least twice in the data X_1^n. Then

$$\sum_{swv} \mathbb{P}[sw \in \hat{\tau}_{\max}] \leq |\mathscr{X}| \mathbb{E}\left[\sum_{sw} 1_{[sw \text{ occurs at least twice in } X_1^n]}\right] \leq |\mathscr{X}| \mathbb{E}[L] \leq |\mathscr{X}| n^2.$$

Therefore, since $C > 2|\mathscr{X}| + 4$, we complete the proof. □

When defining the pruning criterion in Step 2 of the context algorithm in terms of the L_1 distance, we can sharpen Theorem 5.2. Let $\tilde{\Delta}_{wu} = \|\hat{P}_w(\cdot) - \hat{P}_{wu}(\cdot)\|_1^2$ and define $\tilde{O}_n = \{$there exists $w = w'u$ ($w' \in \tau_n$, $u \in \bigcup_{m=1}^{\infty} \mathscr{X}^m$), such that $\tilde{\Delta}_w \geq C\log(n)$, $N(w) > 1$ and $w \notin \tau_n\}$.

THEOREM 5.3. *Under the assumptions* (A1)–(A3) *but with cutoff in Step* 2 *of the context algorithm satisfying $K_n \sim C\log(n)$ for $C > 4|\mathscr{X}| + 8$,*

$$\sum_{n=1}^{\infty} \mathbb{P}[\tilde{O}_n]\log(n) < \infty.$$

PROOF. As used already in the proof of Theorem 5.1, $D(P\|Q) \geq \frac{1}{2}\|P - Q\|_1^2$. Thus, $\tilde{\Delta}_w \leq 2\Delta_w$. □

PROOF OF LEMMA 5.3. Let $s \in \tau_n$ be a context and $su = swv$ with $w \in \bigcup_{m=0}^{\infty} \mathscr{X}^m$, $v \in \mathscr{X}$ and $swv \notin \tau_n$. Our aim is to bound the probability of overestimation at su. We begin by recalling several inequalities and definitions from Weinberger, Rissanen and Feder (1995). First, we fix a sequence x_1^n, a realization from P_n. We can determine a probability law given by $Q_{su}(y_1^n \mid x_1^n)$ (on the set of sequences of length n), defined as follows:

$$\log(Q_{su}(y_1^n \mid x_1^n)) = R_{sw}(y_1^n \mid S_s) + \sum_{x \in \mathscr{X}} \sum_{b \neq v} N_{y_1^n}(x \mid swb) \log\Big(\hat{P}_{x_1^n}(x \mid sw)\Big)$$

$$+ \sum_{x \in \mathscr{X}} N_{y_1^n}(x \mid su) \log\Big(\hat{P}_{x_1^n}(x \mid su)\Big),$$

where $R_{sw}(y_1^n \mid S_s)$, defined formally in Weinberger, Rissanen and Feder (1995), is the sum of the log probability of all the symbols that occur in any context other than sw. An important observation is that for any sequence y_1^n with $N_{y_1^n}(sw) = 0$, the Q_{su} probability of y_1^n is the same as the P_n probability.

Now, for each x_1^n define $\sigma_{x_1^n}$ to be the set of all sequences y_1^n with $N_{y_1^n}(xsw) = N_{x_1^n}(xsw)$ and $N_{y_1^n}(xswv) = N_{x_1^n}(xswv)$ for all $x \in \mathscr{X}$. If $\Delta_{x_1^n}(swv) > C \log(n)$, it follows from (A9) in Weinberger, Rissanen and Feder (1995) that

$$P_n(\sigma_{x_1^n}) \leq Q_{su}\big(\sigma_{x_1^n} \mid x_1^n\big) n^{-C}. \tag{5.8}$$

At this point we need to introduce a new probability distribution given by Q' on the set of sequences of length n, closely related to Q_{su}. To that end, for every sequence y_1^t let x_0 be the symbol that occurs after the first occurrence of sw. Let b_0 be the symbol immediately preceding the first occurrence of sw. Thus x_0 occurs in the (extended) context swb_0. If $b_0 \neq v$, we define

$$\log(Q'(y_1^n \mid x_1^n)) = \log\big(Q_{su}(y_1^t \mid x_1^n)\big) + \log(P_n(x_0 \mid sw)) - \log\Big(\hat{P}_{x_1^n}(x_0 \mid sw)\Big).$$

If $b_0 = v$, then we define

$$\log(Q'(y_1^n \mid x_1^n)) = \log\big(Q_{su}(y_1^t \mid x_1^n)\big) + \log(P_n(x_0 \mid sw)) - \log\Big(\hat{P}_{x_1^n}(x_0 \mid swv)\Big).$$

Thus, if $N_{y_1^n}(sw) < 2$ it must be that $P_n(y_1^n) = Q'(y_1^n \mid x_1^n)$. It also follows from the definition of Q' that

$$Q_{su}(y_1^n \mid x_1^n) \leq \frac{1}{p_{\min}(n)} Q'(y_1^n \mid x_1^n).$$

Therefore, together with (5.8) we have the bound,

$$P_n(\sigma_{x_1^n}) \leq Q'(\sigma_{x_1^n}) \frac{1}{p_{\min}(n)} n^{-C}.$$

The construction of $\sigma_{x_1^n}$ and the fact that $N_{x_1^n}(sw) > 1$ implies that

$$Q'\big(\sigma_{x_1^n} \mid x_1^n\big) \leq Q'\big(y_1^n; N_{y_1^n}(sw) > 1 \mid x_1^n\big) = P_n\big(y_1^n; N_{x_1^n} > 1\big) = P_n(sw \in \hat{\tau}_{\max}).$$

Furthermore, since there are at most $n^{2|\mathscr{X}|}$ distinct classes $\sigma_{x_1^n}$, it follows that

$$\mathbb{P}[O_n(swv)] = P_n\big(y_1^n; \Delta_{y_1^n}(swv) > C \log(n)\big) \frac{1}{p_{\min}(n)}$$

$$\leq P_n(sw \in \hat{\tau}_{\max}) n^{-C + 2|\mathscr{X}|}. \qquad \square$$

Theorems 5.1 and 5.2 imply the assertion in Theorem 3.1(i). The assertion in Theorem 3.1(ii) follows from Theorem 3.1(i) and along the lines of the proof of Theorem 5.1; partition with the set D_n and use Lemmas 5.1 and 5.2. □

PROOF OF THEOREM 3.2. Statements (i) and (iii) follow from the general formula (5.10), statement (ii) is an immediate consequence of Theorem 3.1.

We give here the analogon of assumption (A1) for the estimated process $\hat{P}_n$. The r-step transition kernel $P_n^{(r)}(v, w) = \mathbb{P}_{P_n}[X_{r-k_n+i,n}^{r,n} = v \mid X_{-k_n+1,n}^{0,n} = w]$ (for some $r \in \mathbb{N}$) for the embedding Markov chain $(X_{t-k_n+1,n}^{t,n})_{t\in\mathbb{Z}}$ of the VLMC P_n can be characterized by the transition probabilities $p_n(\cdot \mid \cdot)$ and the context function $c_n(\cdot)$, that is,

$$
\begin{aligned}
&T(v \mid w; r, p_n(\cdot \mid \cdot), c_n(\cdot)) = P_n^{(r)}(v, w) \\
(5.9)\qquad &\qquad = \sum_{x_1^r \in \mathscr{X}^r, (x_1^r w)_{a-k_n+1}^a = v} \prod_{i=0}^{r-1} p_n\big(x_{r-i} \mid c_n\big(x_1^{r-i-1} w\big)\big),
\end{aligned}
$$

where $a = r + |w|$. For every $n \in \mathbb{N}$, the process $(Y_{t,n})_{t\in\mathbb{Z}} \sim \hat{P}_n$ is a VLMC. We consider its r-step transition kernel $\hat{P}_n^{(r)}(v, w) = \mathbb{P}_{\hat{P}_n}[Y_{r-\hat{k}_n+1,n}^{r,n} = v \mid Y_{-\hat{k}_n+1,n}{}^{0,n} = w]$ (for some $r \in \mathbb{N}$) for the embedding Markov chain $(Y_{t-\hat{k}_n+1,n}^{t,n})_{t\in\mathbb{Z}}$ of the VLMC $\hat{P}_n$ of order $\hat{k}_n$. This transition is characterized by

$$
T\big(v \mid w; r, \hat{P}(\cdot \mid \cdot), \hat{c}_n(\cdot)\big) = \hat{P}_n^{(r)}(v, w), \qquad r \geq 1.
$$

We now obtain an analogon of (A1) for $\hat{P}_n$. We consider sets

$$
A_n = \{\omega; \hat{c}(\cdot; \omega) = c_n(\cdot)\}.
$$

Thus by Theorem 3.1(i), $\hat{k}_n = k_n$ on A_n and for $w, w' \in \mathscr{X}^{\hat{k}_n}$,

$$
\begin{aligned}
&\big|T\big(v \mid w; r, \hat{P}(\cdot \mid \cdot), \hat{c}(\cdot)\big) - T\big(v \mid w'; r, \hat{P}(\cdot \mid \cdot), \hat{c}(\cdot)\big)\big| \\
&\quad \leq \big|T(v \mid w; r, p_n(\cdot \mid \cdot), c_n(\cdot)) - T(v \mid w'; r, p_n(\cdot \mid \cdot), c_n(\cdot))\big| \\
&\qquad + \big|T\big(v \mid w; r, \hat{P}(\cdot \mid \cdot), c_n(\cdot)\big) - T(v \mid w; r, p_n(\cdot \mid \cdot), c_n(\cdot))\big| \\
&\qquad + \big|T\big(v \mid w'; r, \hat{P}(\cdot \mid \cdot), c_n(\cdot)\big) - T(v \mid w'; r, p_n(\cdot \mid \cdot), c_n(\cdot))\big| \quad \text{on } A_n.
\end{aligned}
$$

We now invoke (A1) for $T(\cdot \mid \cdot; r, p_n(\cdot \mid \cdot), c_n(\cdot))$ about the true underlying process. For the other terms we use the finiteness of r and $\mathscr{X}$, together with (5.9) and Theorem 3.1(ii). We then obtain

$$
\begin{aligned}
&\sup_{v, w, w' \in \mathscr{X}^{\hat{k}_n}} \big|\hat{P}_n^{(r)}(v, w) - \hat{P}_n^{(r)}(v, w')\big| \\
(5.10)\qquad &\quad = \sup_{v, w, w' \in \mathscr{X}^{\hat{k}_n}} \big|T\big(v \mid w; r, \hat{P}(\cdot \mid \cdot), \hat{c}(\cdot)\big) \\
&\qquad\qquad - T\big(v \mid w'; r, \hat{P}(\cdot \mid \cdot), \hat{c}(\cdot)\big)\big| \\
&\quad \leq 1 - 2\kappa + o(1) \text{ on } A_n.
\end{aligned}
$$

Thus on A_n, $\hat{P}_n$ as constructed in Theorem 3.2(i) is uniquely determined, stationary and ϕ-mixing, with mixing coefficients bounded by

$$\phi_{\hat{P}_n}(j) \le (1-\kappa)^j \quad \text{for all } j \in \mathbb{N} \text{ on } A_n$$

[cf. Rajarshi (1990), Lemma 2.1, or Doukhan (1994)].

However, by Theorem 3.1(i), $\mathbb{P}[A_n] \to 1$ as $n \to \infty$, which completes the proof of Theorem 3.2(i) and (iii). □

Proof of Theorem 4.1. We usually suppress the index n when writing X_t instead of $X_{t,n}$. Consider

$$U_n = (n-m+1)^{-1} \sum_{t=1}^{n-m+1} f\big(X_t^{t+m-1}\big),$$

and denote by $\Sigma = \operatorname{Cov}[U_n]$ the covariance matrix of U_n.

Lemma 5.4. *Assume* (B1) *with* $(X_{t,n})_{t\in\mathbb{Z}} \sim P_n$ *satisfying* (A1). *Then*:

(i) *There exists* $n_0 \in \mathbb{N}$ *such that* $n\Sigma_n$ *is positive definite for all* $n \ge n_0$;
(ii) *For* $Z \sim \mathscr{N}_v(0, I)$,

$$\sup_{x\in\mathbb{R}^v} \left|\mathbb{P}\left[\Sigma_n^{-1/2}(U_n-\theta_n) \le x\right] - \mathbb{P}[Z \le x]\right| = o(1), \qquad n \to \infty.$$

Proof. For every $n \in \mathbb{N}$, the process $(X_{t,n})_{t\in\mathbb{Z}}$ is ϕ_n-mixing whose mixing coefficients are bounded by

$$\sup_{n\in\mathbb{N}} \phi_n(k) \le (1-2\kappa)^k \quad \text{for all } k \in \mathbb{N}; \tag{5.11}$$

see Remark 3.7.

Bounding covariances in terms of mixing coefficients [cf. Doukhan (1994)] and using the bound in (5.11) implies for $i, j \in \{1, \dots, v\}$.

(5.12)
$$(n-m+1)(\Sigma_n)_{i,j} = \sum_{k=-n+1}^{n-1} \operatorname{Cov}\big(f_i\big(X_0^{m-1}\big), f_j\big(X_k^{k+m-1}\big)\big) + O(n^{-1}).$$

Hence, assertion (i) follows from the assumption in (B1).

Assertion (i), assumption (B1) and (5.12) allow us to write

$$\Sigma_n^{-1/2} = n^{1/2}\Gamma_n, \ \sup_{n\in\mathbb{N}} \max_{1\le i,j\le v} \left|(\Gamma_n)_{i,j}\right| < \infty. \tag{5.13}$$

Now write

$$\Sigma_n^{-1/2}(U_n - \theta_n) = n^{-1/2} \sum_{t=1}^{n-m+1} \tilde{f}_n\big(X_t^{t+m-1}\big)(1+o(1)),$$

where $\tilde{f}_n(X_t^{t+m-1}) = \Gamma_n(f(X_t^{t+m-1}) - \theta_n)$.

By construction and (5.13),

$$
\begin{aligned}
&\mathbb{E}\big[\tilde{f}_n(X_1^m)\big] = 0,\\
(5.14)\qquad &\operatorname{Cov}\left(n^{-1/2}\sum_{t=1}^{n-m+1}\tilde{f}_n\big(X_t^{t+m-1}\big)\right) \to I(n\to\infty),\\
&\sup_{n\in\mathbb{N}}\mathbb{E}\big\|\tilde{f}_n(X_1^m)\big\|^2 < \infty.
\end{aligned}
$$

We can then apply Theorem 2.1 in Withers (1981) to $n^{-1/2}\sum_{t=1}^{n-m+1}\tilde{f}_n(X_t^{t+m-1})$. The conditions [version (A) or (B); note also the corrigendum (1983)] are easily verified by invoking the mixing bound in (5.11) and (5.14). Thus,

$$
n^{-1/2}\sum_{t=1}^{n-m+1}\tilde{f}_n\big(X_t^{t+m-1}\big) \to_p \mathscr{N}_v(0,I),
$$

and assertion (ii) follows by Pólya's theorem. □

By the smoothness assumption about g we use a first-order Taylor expansion,

$$
(5.15)\qquad n^{1/2}(T_n - g(\theta_n)) = n^{1/2}Dg\big(\tilde{\theta}_n\big)(U_n - \theta_n),
$$

where $Dg(\theta) = [\partial g_i(u)/\partial u_j]_{i,j}$, $(1 \le i \le w,\ 1 \le j \le v)$ and $\|\tilde{\theta}_n - \theta_n\| \le \|U_n - \theta_n\|$.

By (5.13) and Lemma 5.4(ii), $U_n - \theta_n = o_P(1)$, so that

$$
\big[Dg\big(\tilde{\theta}_n\big) - Dg(\theta_n)\big]_{i,j} = o_P(1), \qquad 1 \le i \le w, 1 \le j \le v.
$$

This, together with (5.15), the boundedness of $n^{1/2}\Sigma_n^{1/2}$ [use (5.12)] and Lemma 5.4(ii) implies

$$
\begin{aligned}
(5.16)\qquad &\sup_{x\in\mathbb{R}^w}\big|\mathbb{P}\big[n^{1/2}(T_n - g(\theta_n)) \le x\big]\\
&\qquad -\mathbb{P}\big[n^{1/2}\Sigma_n^{1/2}Dg(\theta_n)Z \le x\big]\big| = o(1), \qquad n\to\infty,
\end{aligned}
$$

where $Z \sim \mathscr{N}_v(0,I)$.

We are now going to show the bootstrap analog of (5.16). By Theorem 3.2(i) and (iii), the bootstrap process $(X_t^*)_{t\in\mathbb{Z}}$ is with high probability stationary and geometrically ϕ-mixing with mixing coefficients denoted by $\phi_n^*(k) = \phi_{\hat{P}_n}(k)$ from Theorem 3.2(iii). Note that the distribution of $(X_t^*)_{t\in\mathbb{Z}}$ depends again on the sample size n.

Denote by $U_n^* = (n-m+1)^{-1}\sum_{t=1}^{n-m+1} f((X^*)_t^{t+m-1})$ and let $\Sigma_n^* = \operatorname{Cov}^*[U_n^*]$ be the covariance matrix of U_n^* with respect to the bootstrap distribution.

LEMMA 5.5. *Assume the conditions of Theorem* 4.1. *Then*:

(i) $n(\Sigma_n^* - \Sigma_n)_{i,j} = o_P(1)$ $(n \to \infty)$, $i, j = 1, \dots, v$.
(ii) $\lim_{n\to\infty} \mathbb{P}[n\Sigma_n^*$ *is positive definite*$] = 1$.
(iii) *For* $Z \sim \mathscr{N}_v(0, I)$,

$$\sup_{x\in\mathbb{R}^v} \left|\mathbb{P}^*\left[(\Sigma_n^*)^{-1/2}(U_n^* - \theta_n^*) \le x\right] - \mathbb{P}[Z \le x]\right| = o_P(1), \qquad n \to \infty.$$

PROOF. For any $i, j \in \{1, \dots, v\}$,

$$\begin{aligned} n(\Sigma_n^*)_{i,j} &= \sum_{k=-n+m}^{n-m} \mathrm{Cov}^*\Big(f_i\big((X^*)_0^{m-1}\big), f_j\big((X^*)_k^{k+m-1}\big)\Big)\left(1 - \frac{|k|}{n-m+1}\right) \\ &= \sum_{k=-M}^{M} \mathrm{Cov}^*\Big(f_i\big((X^*)_0^{m-1}\big), f_j\big((X^*)_k^{k+m-1}\big)\Big)\left(1 - \frac{|k|}{n-m+1}\right) \\ &\quad + \Delta_{n,M}, \end{aligned} \tag{5.17}$$

where M is a finite constant.

By well-known bounds of covariances in terms of mixing coefficients [cf. Doukhan (1994)],

$$|\Delta_{n,M}| \le 2\,\mathrm{const.} \sum_{k=M+1}^{\infty} \phi_n^*(k).$$

Therefore by Theorem 3.2(iii),

$$\mathbb{P}\left[\lim_{M\to\infty} |\Delta_{n,M}| = 0\right] \to 1, \qquad n \to \infty. \tag{5.18}$$

By Theorem 3.2(ii),

$$\max_{x_1^d \in \mathscr{X}^d} \left|\mathbb{P}^*\left[(X^*)_1^d = x_1^d\right] - \mathbb{P}\left[X_1^d = x_1^d\right]\right| = o_P(1), \qquad d \in \mathbb{N}. \tag{5.19}$$

This, the boundedness of f and the finiteness of M imply

$$\begin{aligned} &\left| \sum_{k=-M}^{M} \mathrm{Cov}^*\Big(f_i\big((X^*)_0^{m-1}\big), f_j\big((X^*)_k^{k+m-1}\big)\Big) \right. \\ &\qquad \left. - \sum_{k=-M}^{M} \mathrm{Cov}\Big(f_i\big(X_0^{m-1}\big), f_j\big(X_k^{k+m-1}\big)\Big) \right| \\ &= o_P(1)\ (n \to \infty). \end{aligned} \tag{5.20}$$

By the geometric ϕ-mixing property of $(X_t)_{t\in\mathbb{Z}}$ [see (5.11)] and the boundedness of f,

$$\begin{aligned} &\left| \sum_{k=-M}^{M} \mathrm{Cov}\Big(f_i\big(X_0^{M-1}\big), f_j\big(X_k^{k+m-1}\big)\Big) \right. \\ &\qquad \left. - \sum_{k=-\infty}^{\infty} \mathrm{Cov}\Big(f_i\big(X_0^{m-1}\big), f_j\big(X_k^{k+m-1}\big)\Big) \right| \\ &= o(1)\ (M \to \infty). \end{aligned} \tag{5.21}$$

Thus, by (5.17)–(5.21), we have shown assertion (i). Assertion (ii) follows by (i) and Lemma 5.4(i). Assertion (iii) can be proved as was Lemma 5.4(ii); we now invoke the mixing bound in Theorem 3.2(iii) and use (i). □

By (5.19) and the finiteness of $|\mathscr{X}|$ we have,

$$\theta_n^* - \theta_n = \sum_{x_1^m \in \mathscr{X}^m} f(x_1^m)\Big(\mathbb{P}^*\big[(X^*)_1^m = x_1^m\big] - \mathbb{P}\big[X_1^m = x_1^m\big]\Big) = o_P(1), \tag{5.22}$$

and hence by the continuous differentiability of g,

$$\begin{aligned}\mathbb{P}^*\Big\{\big|\big[Dg(\tilde{\theta}_n^*) - Dg(\theta_n)\big]_{i,j}\big| > \eta\Big\} = o_P(1) \text{ for any } \eta > 0 \text{ and } \|\tilde{\theta}_n^* - \theta_n^*\| \\ \le \|U_n^* - \theta_n^*\|, \qquad 1 \le i \le w, 1 \le j \le v.\end{aligned} \tag{5.23}$$

A first-order Taylor expansion, (5.23), Lemma 5.5(iii) and the boundedness of $n\Sigma_n^* = O_P(1)$ imply

$$\begin{aligned}\sup_{x \in \mathbb{R}^w} \Big|\mathbb{P}^*\big[n^{1/2}(T_n^* - g(\theta_n^*)) \le x\big] - \mathbb{P}\big[n^{1/2}\Sigma_n^{1/2} Dg(\theta_n) Z \le x\big]\Big| \\ = o_P(1), \qquad n \to \infty,\end{aligned} \tag{5.24}$$

where $Z \sim \mathscr{N}_v(0, I)$.

By (5.16) and (5.24) we complete the proof of Theorem 4.1. □

Acknowledgments. We thank Itai Zuckerman for carrying out the computations. We also acknowledge interesting general comments from P. Bickel, H.-R. Künsch, R. Olshen and T. Speed and specific suggestions from F. Ferrari, the Editor and the referees.

REFERENCES

Bickel, P. J., Götze, F. and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statist. Sinica* **7** 1–32.

Braun, J. V. and Müller, H.-G. (1998). Statistical methods for DNA sequence. *Statist. Sci.* **13** 142–162.

Breiman, L.; Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Brillinger, D. R. (1995). Trend analysis: binary-valued and point cases. *Stochastic Hydrology and Hydraulics* **9** 207–213.

Bühlmann, P. (1999). Model selection for variable length Markov chains and tuning the context algorithm. *Ann. Inst. Statist. Math.* To appear.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York.

Doukhan, P. (1994). *Mixing Properties and Examples. Lecture Notes in Statist.* **85**. Springer, Berlin.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** 1–26.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, Berlin.

Feder, M.; Merhav, N. and Gutman, M. (1992). Universal prediction of individual sequences. *IEEE Trans. Inform. Theory* **IT-38** 1258–1270.

Guttorp, P. (1995). *Stochastic Modeling of Scientific Data*. Chapman and Hall, London.

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

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241.

PRUM, B.; RODOLPHE, F. and DE TURCKHEIM, E. (1995). Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *J. Roy. Statist. Soc. Ser. B* **57** 205–220.

RAFTERY, A. and TAVARÉ, S. (1994). Estimation and modelling repeated patterns in high-order Markov chains with the mixture transition distribution model. *Appl. Statist.* **43** 179–199.

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

RISSANEN, J. (1983). A universal data compression system. *IEEE Trans. Inform. Theory* **IT-29** 656–664.

RISSANEN, J. (1986). Complexity of strings in the class of Markov sources. *IEEE Trans. Inform. Theory* **IT-32** 526–532.

RISSANEN, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.

RITOV, Y. and BICKEL, P. J. (1990). Achieving information bounds in non- and semiparametric models. *Ann. Statist.* **18** 925–938.

WEINBERGER, M. J. and FEDER, M. (1994). Predictive stochastic complexity and model estimation for finite-state processes. *J. Statist. Plann. Inference* **39** 353–372.

WEINBERGER, M. J., LEMPEL, A. and ZIV, J. (1992). A sequential algorithm for the universal coding of finite memory sources. *IEEE Trans. Inform. Theory* **IT-38** 1002–1014.

WEINBERGER, M. J., RISSANEN, J. and FEDER, M. (1995). A universal finite memory source. *IEEE Trans. Inform. Theory* **IT-41** 643–652.

WITHERS, C. S. (1981). Central limit theorems for dependent variables **I**. *Z. Wahrsch. Verw. Gebiete* **57** 509–534 [Corrigendum (1983) **63** 555.]

SEMINAR FÜR STATISTIK
ETH ZÜRICH
CH-8092 ZÜRICH
SWITZERLAND
E-MAIL: buhlmann@stat.math.ethz.ch

DEPARTMENT OF STATISTICS
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
E-MAIL: ajw@compstat.wharton.upenn.edu