Variable Length Markov Chains

Peter Bühlmann * and Abraham J. Wyner †
University of California
Berkeley

January 1997

Abstract

In an information theoretical set-up, Rissanen (1983) has proposed the algorithm 'context' for data compression. Using his idea we introduce a new sub-class of stationary, possibly sparse, Markov chains (context models), whose dimension is allowed to grow with increasing sample size. Asymptotically, this new class covers infinite dimensional models.

Proposing a modification of Rissanen's algorithm (context algorithm), we show how such context models can be selected and fitted in a data-driven way. From this we gain several grounds: an excellent exploratory tool, a novel universal resampling procedure for categorical time series (context bootstrap) and a nonparametric prediction machine.

We prove a novel consistency result for our data-driven context algorithm in an asymptotically infinite dimensional setting and also show the asymptotic validity of the context bootstrap for a broad range of situations.

The computations can be done recursively and are very fast. A simulation study for the context bootstrap completes our exposition.

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

Short title: Variable Length Markov Chain

*Research supported in part by the Swiss National Science Foundation. Part of the work has been done while visiting the University of Heidelberg, Germany.

†Research supported by grant NSF DMS 9508933.

1 Introduction

One of the most nonparametric models for a stationary process $\{X_t\}_{t\in\mathbb{Z}}$ assuming no particular underlying mechanistic system is maybe a full Markov chain of finite order. The only implicit assumption made is about the finite memory of the process. Probabilistically a nice model, such full Markov chains can become very hard to estimate. Even when the process $\{X_t\}_{t\in\mathbb{Z}}$ takes only values in a finite space, these models run very soon into the curse of dimensionality. This corresponds to an explosion in the number of parameters yielding highly variable estimates.

Trying to avoid the curse of dimensionality and still achieving a substantial reduction of the complexity in the data, we make use of an existing method in computer science and information theory. The algorithm 'context', proposed by Rissanen (1983) and further developed by Weinberger et al. (1995) has been designed for compression of (dependent) finite state sequences. The idea in a Markovian set-up is to lump irrelevant states together, resulting in a sometimes huge reduction of the number of parameters.

We first redefine here what we call the context model as some sort of finite state, possibly sparse Markov chain. Then, we consider families of such context models which change with sample size. We allow a growth in the model dimension as sample size increases, thus extending to the class of infinite dimensional models. The range of applications of such models is very broad, as examples we mention genetics with DNA sequencing, cf. Prum et al. (1995), seismology with Mercalli intensities, cf. Brillinger (1994), and finance with modeling extreme events, cf. Bühlmann (1996).

Given data, we propose to fit a context model by lumping irrelevant states together in a fully adaptive way. We modify Rissanen's algorithm 'context' which is related to some kind of hierarchical backward model selection. Fitted context models can be used as an excellent exploratory tool for the dynamics of a categorical time series. Finite state Markov chains can be represented by trees, where every branch corresponds to a history of times $t-1,\ldots$ attached with the probabilities for moving on to time t. Sparse Markov chains are then represented by unbalanced trees and a fitted context model yields the structure of such a tree.

We give an entirely new consistency result, showing that our modified context algorithm is consistent for estimating the true underlying model whose dimension also grows with sample size. This is in some sense an analog of a convergence rate result. As an important consequence, our result implies a balance between over- and under-estimation of the true model, the effect of these miss-estimations becoming eventually negligible. This corresponds to the well known bias-variance tradeoff.

We also make use of the general consistency result described above to propose a novel resampling scheme, the context bootstrap. We prove asymptotic validity of the context bootstrap for a whole class of estimators and argue why such a scheme works under very general conditions. The context bootstrap is tailored for categorical time series and offers an alternative to the blockwise bootstrap, which has been proposed by Künsch (1989) for the case of general stationary observations. In particular, the context bootstrap has a nice probabilistic interpretation and enjoys the advantage of being applicable as a simple plug-in rule. Based on results from our simulation study we have with the context bootstrap a new universally well working resampling tool for categorical time series which usually outperforms the blockwise bootstrap.

The paper is organized as follows. In section 2 we give the definition of a context model. In section 3 we describe the process of fitting such models and state the general consistency result of finding the true underlying model. In section 4 we describe the context bootstrap, state results about asymptotic validity thereof and present results from a simulation study. In particular, a comparison with the blockwise bootstrap is included. In section 5 we give the proofs.

2 Context models as variable length Markov chains

In the sequel, we denote by $x_i^j = x_j, x_{j-1}, \ldots, x_i$ $(i < j,\ i, j \in \mathbb{Z} \cup \{-\infty, \infty\})$ a string written in reverse 'time'. We usually denote by capital letters X random variables and by small letters x fixed deterministic values. We follow here the ideas of Weinberger et al. (1995) and define what we call the context model. As a starting point, consider $(X_t)_{t\in\mathbb{Z}}$, being a stationary Markov chain of finite order k with values in a finite space $\mathcal{X}$. Thus,

$$\mathbb{P}[X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0] = \mathbb{P}[X_1 = x_1 | X_{-k+1}^0 = x_{-k+1}^0], \text{ for all } x_{-\infty}^1. \tag{2.1}$$

Such full Markov chains are very hard to estimate since they involve $|\mathcal{X}|^k(|\mathcal{X}| - 1)$ free parameters. For example, if $|\mathcal{X}| = 5$ and $k = 5$, the number of free parameters is 12500, which is prohibitive! To get less complex models, the idea is to lump irrelevant states in the history X_{-k+1}^0 in formula (2.1) together, resulting in a sparse Markov chain.

For a time point $t \in \mathbb{Z}$, typically only some values from the infinite history $X_{-\infty}^{t-1}$ of the variable X_t are relevant. This relevant history can be thought as a *context* for the actual variable X_t. 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 actual values $X_{-\infty}^{t-1}$. In other words, we might have for the variable X_t a context of length 1 and for $X_{t'}$ a context of length 5. We can formalize this as follows.

Definition 2.1 *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process with values $X_t \in \mathcal{X}$, $|\mathcal{X}| < \infty$. Denote by $c : \mathcal{X}^\infty \to \mathcal{X}^\infty$ a (projection) function which maps*

$$c : x_{-\infty}^0 \mapsto x_{-\ell+1}^0, \text{ where } \ell \text{ is defined by}$$
$$\ell = \min\{k; \mathbb{P}[X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0] = \mathbb{P}[X_1 = x_1 | X_{-k+1}^0 = x_{-k+1}^0] \text{ for all } x_1 \in \mathcal{X}\}$$
($\ell = 0$ corresponds to independence).

Then, $c(.)$ 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.

By the projection structure of the context function $c(.)$, the context-length $\ell(.) = |c(.)|$ determines $c(.)$ and vice-versa; here $|.|$ denotes the cardinality of a tuple. The definition of ℓ implicitly reflects the fact that the context-length of a variable x_t is $\ell = |c(x_{-\infty}^{t-1})| = \ell(x_{-\infty}^{t-1})$, depending on the history $x_{-\infty}^{t-1}$.

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

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

Then $c(.)$ is called a context function of order k, and $(X_t)_{t\in\mathbb{Z}}$ is called a stationary context model of order k.

Clearly, a context model of order k is a Markov chain of order k, now having a *memory of variable length* ℓ. By requiring stationarity, a context model is thus completely specified by its transition probabilities,

$$p(x_1|c(x_{-\infty}^0)) = \mathbb{P}[X_1 = x_1|c(X_{-\infty}^0) = c(x_{-\infty}^0)],\ x_{-\infty}^1 \in \mathcal{X}^\infty.$$

Our framework, given below in section 2.1, will be such that the order $k = k_n$ of a context model is allowed to grow with sample size n. In retrospect, we could define a context function $c(.) : \mathcal{X}^k \to \mathcal{X}^k$, since there is no functional dependence of the function $c(x_{-\infty}^0)$ on a variable x_{-k+1-m} $(m > 0)$. We sometimes use the definition on $\mathcal{X}^\infty$ and sometimes on $\mathcal{X}^k$. The context function projects the k-th (or infinite) order history x_{-k+1}^0 into $\mathcal{X}^k$. Often the range space of the context function $c(.)$ is not the full space $\mathcal{X}^k$, but also not the empty space. If the context function $c(.)$ of order k is the full projection $x_{-k+1}^0 \mapsto x_{-k+1}^0$ for all x_{-k+1}^0, the context model is a full Markov chain of order k. The class of context functions of length k is rich enough to obtain a broad class of Markov chains, including special sparse types given by the notion of a short context. In particular, some context functions $c(.)$ would yield a substantial reduction in the number of parameters compared to a full Markov chain of the same order as the context function.

2.1 Family of context models

Sometimes it is appropriate to assume an underlying probability distribution (model) P_n on $\mathcal{X}^\infty$ which changes with sample size n. This means, we have a finite realization $X_{1,n}, \ldots, X_{n,n}$ from P_n, where P_n is the distribution of a stationary context model of finite order k_n as given in Definition 2.2. Such a model is also called 'moving truth'. We allow $k_n \to \infty$ $(n \to \infty)$, the rate of increase not being too fast. This then incorporates models of unbounded dimensions and infinite dimensional models in the limit. The precise description is as follows.

Let $\mathcal{P}$ be a class of probability distributions on $\mathcal{X}^\infty$, corresponding to stationary context models of finite order,

$$\mathcal{P} = \{P;\ (X_t)_{t\in\mathbb{Z}} \sim P,\ (X_t)_{t\in\mathbb{Z}} \text{ defined as in Definition 2.2 with order } k < \infty\}. \quad (2.2)$$

The 'moving truth' model then reads,

$$\begin{aligned} &X_{1,n}, \ldots X_{n,n} \text{ a finite realization of } P_n, \\ &P_n \in \mathcal{P},\ \mathcal{P} \text{ as in (2.2)},\ n \in \mathbb{N}. \end{aligned} \quad (2.3)$$

In order to learn consistently about P_n from $X_{1,n}, \ldots, X_{n,n}$, we need additional restrictions for the sequence $(P_n)_{n\in\mathbb{N}}$, see assumptions (A1)-(A3) in section 3.2. But we do not necessarily assume a 'limiting truth' $\lim_{n\to\infty} P_n = P$ (where the limit would have to be defined first). Consistent estimation of a sequence $(P_n)_{n\in\mathbb{N}} \in \mathcal{P}$ (or consistent learning) in a 'moving truth' model is defined as follows. Let $d(.,.)$ be a metric on $\mathcal{P}$. An estimate $\hat{P}_n$, based on a realization $X_{1,n}, \ldots, X_{n,n}$ from P_n, is called d-consistent for P_n, if for any $\varepsilon > 0$, there exists an $n_0 = n_0(\varepsilon)$, such that

$$\mathbb{P}[d(\hat{P}_n, P_n) < \varepsilon] > 1 - \varepsilon \text{ for all } n \geq n_0.$$

2.2 Representation as context trees

In order to explain our procedure for adaptively selecting and fitting a context model, it is most convenient to represent a context function, and hence the set of relevant histories of a context model, 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 $|\mathcal{X}|$ offsprings. Then, each value of a context function $c(.) : \mathcal{X}^k \to \mathcal{X}^k$ can be represented as a branch (or final node) of such a tree. The context $w = c(x_{-k+1}^0)$ is represented by a branch, whose sub-branch on the top is determined by x_0, the next sub-branch by x_{-1} and so on, and the final sub-branch by $x_{-\ell(x_0,\ldots,x_{-k+1})+1}$.

Example 2.1 $|\mathcal{X}| = 2$, $k = 3$.
The function

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

can be represented by the tree $\tau_{c(.)}$,

A 'growing to the left' sub-branch represents the symbol 0 *and vice versa for the symbol* 1.

Note that context trees do not have to be complete, i.e., every internal does not need to have exactly $|\mathcal{X}|$ offsprings (when $|\mathcal{X}| > 2$).

Definition 2.3 *Let* $c(.)$ *be a context function of a stationary context model of order* k. *The* ($|\mathcal{X}|$*-ary) context tree* τ *and final node context tree* τ^f *are defined as*

$$\tau = \tau_c = \{w; w = c(x_{-k+1}^0),\ x_{-k+1}^0 \in \mathcal{X}^k\},$$
$$\tau^f = \tau_c^f = \{w; w \in \tau_c \text{ and } wu \notin \tau_c \text{ for all } u \in \mathcal{X}\}.$$

Definition 2.3 says that only final nodes in the tree representation τ are considered as elements of the final node context tree τ^f. Clearly, we can reconstruct the context function $c(.)$ from τ_c or τ_c^f. An internal node with $b < |\mathcal{X}|$ offsprings can be implicitly thought to be complete by adding one complementary offspring, lumping the $|\mathcal{X}| - b$ non-present nodes together.

3 Context algorithm and its consistency

Given data $X_{1,n}, \ldots, X_{n,n}$ as in (2.3), the aim is to find the underlying context function $c_n(.)$ and an estimate of P_n. We will attack and solve this problem in a purely nonparametric way, incorporating ideas from data compression as given by Weinberger et al. (1995). It is exactly this nonparametric character which makes our data driven algorithm an excellent exploratory tool and attractive for resampling, see section 4.

3.1 Context algorithm

We describe now the algorithm for the aim mentioned above. In the sequel we always make the convention that quantities involving time indices $\notin \{1, \ldots, n\}$ equal zero (or are irrelevant). Let

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

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

$$\hat{p}(w) = N(w)/n, \ \hat{p}(u|w) = \frac{N(uw)}{N(w)}, \ w, u \in \mathcal{X}^\infty, uw = (\ldots, u_2, u_1, \ldots, w_2, w_1). \tag{3.2}$$

The algorithm below constructs the estimated context tree $\hat{\tau}$ to be the biggest context tree such that

$$\Delta_{wu} = \sum_{x \in \mathcal{X}} \hat{p}(x|wu) \log(\frac{\hat{p}(x|wu)}{\hat{p}(x|w)}) N(wu) \geq K \log(n) \text{ for all } wu \in \hat{\tau}^f \tag{3.3}$$

with $K > 2|\mathcal{X}| + 3$.

Step 1 Given data $X_1, \ldots, X_n$ taking values in a finite space $\mathcal{X}$, fit a maximal $|\mathcal{X}|$-ary context tree, i.e., search for the context function $c_{max}(.)$ with final node context tree representation τ_{max}^f, where τ_{max}^f is the biggest tree such that every element (final node) in τ_{max}^f has been observed at least twice in the data. This can be formalized as follows:

$$w \in \tau_{max}^f \text{ implies } N(w) \geq 2,$$
$$\tau_{max}^f \supseteq \tau^f, \text{ where } w \in \tau^f \text{ implies } N(w) \geq 2.$$

($\tau_1 \subseteq \tau_2$ means: $w \in \tau_1 \Rightarrow wu \in \tau_2$ for some $u \in \cup_{m=0}^{\infty} \mathcal{X}^m$ ($\mathcal{X}^0 = \emptyset$)). Set $\tau_{(0)}^f = \tau_{max}^f$.

Step 2 Examine every element (final node) of $\tau_{(0)}^f$ as follows (the order of examining is irrelevant, see Remark 3.2). Let $c(.)$ be the corresponding context function to $\tau_{(0)}^f$ and let

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

be an element (final node) of $\tau^{(0)}$, which we compare with its pruned version $w = X^0_{-\ell+2}$ (if $\ell = 1$, the pruned version is the empty branch, i.e., the root node).
Replace the context $wu = X^0_{-\ell+1}$ by $w = X^0_{-\ell+2}$ if

$$\Delta_{wu} = \sum_{x \in \mathcal{X}} \hat{p}(x|wu) \log(\frac{\hat{p}(x|wu)}{\hat{p}(x|w)}) N(wu) < K \log(n),$$

with $K > 2|\mathcal{X}| + 3$ and $\hat{p}(.)$ and $\hat{p}(.|.)$ as defined in (3.2). Decision about pruning for every final node in $\tau^f_{(0)}$ yields a (possibly) smaller tree $\tau_{(1)} \subseteq \tau^f_{(0)}$. Let

$$\tau^f_{(1)} = \{w;\ w \in \tau_{(1)} \text{ and } wu \notin \tau_{(1)} \text{ for all } u \in \mathcal{X}\}.$$

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

Step 4 If interested in probability sources, estimate $p(x_1|c(x^0_{-\infty})) = \mathbb{P}[X_1 = x_1|c(X^0_{-\infty}) = c(x^0_{-\infty})]$ by $\hat{p}(x_1|\hat{c}(x^0_{-\infty}))$, where $\hat{p}(.|.)$ is defined as in (3.1).

Remark 3.1. The pruning in the context algorithm can be viewed as some sort of hierarchical backward selection. Dependence on some values further back in the history should be weaker, so that deep nodes in the tree are considered, in a hierarchical way, to be less relevant. This hierarchic structure is a clear distinction to the CART algorithm (Breiman et al., 1984), where the tree architecture has no built in time structure.

Remark 3.2. It does not matter which final node wu in Step 2 is examined first, second and so on. This relates to the orthogonal decomposition in analysis of variance, where the order to test various effects does not matter. Here, for every tree $\tau_{(i)}$ the order of testing the final nodes is irrelevant, constituting a semi-orthogonality.

Remark 3.3. The pruning decision in Step 2 can be related to the Kullback-Leibler distance and to the likelihood ratio test. By definition,

$$\begin{aligned} \Delta_{wu} &= \sum_{x \in \mathcal{X}} \hat{p}(x|wu) \log(\frac{\hat{p}(x|wu)}{\hat{p}(x|w)}) N(wu) \\ &= D(\hat{p}(.|wu)||\hat{p}(.|w)N(wu), \end{aligned} \tag{3.4}$$

where $N(wu)$ is defined in (3.1) and $D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log(P(x)/Q(x))$ is the Kullback-Leibler distance between two probability measures P and Q on $\mathcal{X}$.
Denote the estimated likelihood function, based on context function $c(.)$ by

$$\hat{\mathbb{P}}_c(X_1^n) = \prod_{t=1}^{n} \hat{p}(X_t|c(X^{t-1}_{-\infty})), \tag{3.5}$$

where $\hat{p}(X_t|c(X^{t-1}_{-\infty}))$ is defined in (3.2).
Denote by $c(.)$ the context function of a non-pruned context tree and by $c'(.)$ the context function of the sub–tree, pruned at one final node $wu = x^0_{-\ell+1}$ 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 only the term remains at the node considered for pruning. One gets

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

Formula (3.6) says that our pruning criterion is nothing else than a likelihood ratio test, but now with a large acceptance region $[0, K\log(n)]$ for the pruned (sub–)tree. The large acceptance region takes care about the multiple test problem, our algorithm can be viewed as doing very many likelihood ratio tests.

Remark 3.4. The cut-off value $K\log(n)$ in Step 2 for the pruning decision is chosen by an asymptotic consideration. Clearly, by the interpretation as likelihood ratio tests, small cut-off values will result in larger context trees and overfitting occurs. It is an open question which cut-off yields a procedure, being optimal in some (still to be defined) sense. Since the likelihood ratio statistic in (3.6) satisfies

$$2\log(\frac{\hat{\mathbb{P}}_c(X_1^n)}{\hat{\mathbb{P}}_{c'}(X_1^n)}) \xrightarrow{d} \chi^2_{d-1} \ (n \to \infty),$$

the cut-off value can be interpreted as a (stepwise) $1-\alpha$ quantile divided by 2 of the appropriate χ^2 distribution $\chi^2_{d-1;1-\alpha}/2$. The level α would typically be chosen to be small. For an automatic selection of the cut-off value one could try to minimize a measure for model complexity such as AIC,

$$-2\log(\hat{\mathbb{P}}_{\hat{c}}(X_1^n)) + 2(\text{number of parameters}),$$

where $\log(\hat{\mathbb{P}}_{\hat{c}}(X_1^n))$ is the log-likelihood of the data with respect to an estimated context model with context function $\hat{c}$.

3.2 Consistency

We give two results, the first one dealing with consistency for finding the structure of a context model, the second one yielding d-consistency as defined in section 2.1.

We consider a sequence of context models $(P_n)_{n\in\mathbb{N}}$, $P_n \in \mathcal{P}$ as defined in (2.2). Every context model P_n is specified by its context function $c_n(.)$ or equivalently its context tree τ_n and the transition probabilities $\{p_n(.|w); w \in \tau_n\}$. With a slight abuse of notation, we write for any $v = (v_m, \ldots, v_1) \in \mathcal{X}^m$, $P_n(v)$ instead of $P_n \circ \pi^{-1}_{1,\ldots,m}(v)$ with π being the coordinate function, see (3.7). We also denote by $P_n(x|v) = P_n(xv)/P_n(v)$ for $x, v \in \mathcal{X}^\infty$. Under the assumption (A1) below, the transition probabilities $\{p_n(.|w); w \in \tau_n\}$ generate the unique stationary probability measure P_n on $\mathcal{X}^\infty$. Thus, for a context $w \in \tau_n$, $P_n(.|w) = p_n(.|w)$. We make the following assumptions.

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

$$\sup_{n\in\mathbb{N}} \sup_{v,w,w'} |p^{(r)}_{Z_n}(v,w) - p^{(r)}_{Z_n}(v,w')| < 1-2\kappa, \text{ for some } \kappa > 0,$$

where $p^{(r)}_{Z_n}(v,w) = \mathbb{P}[Z_{r,n} = v | Z_{0,n} = w]$ denotes the r-step transition kernel of the state process $Z_{t,n} = c(X^{t,n}_{0,n} x_0^\infty)$, $x_0^\infty = x_0, x_0, \ldots$ $(t \in \mathbb{N}_0)$ with $(X_{t,n})_{t\in\mathbb{Z}} \sim P_n$.

The definition of $Z_{t,n}$ reflects our implicit assumption here that the initial state is padded with elements $x_0 \in \mathcal{X}$, i.e., $Z_{0,n} = w$ means $Z_{0,n} = wx_0^\infty$ so that the next states $Z_{t,n}$ $(t > 0)$ are uniquely determined by the transition probabilities $\{p_n(.|w);\ w \in \tau_n\}$.

(A2) Let $b_n = \min_{w\in\tau_n} P_n(w)$ and $\epsilon_n = \min_{wu\in\tau_n, u\in\mathcal{X}} \|P_n(.|wu) - P_n(.|w)\|_1$. Then, for some $\delta > 0$,

$$b_n \geq \log(n)^{3+\delta}/n,$$
$$\epsilon_n \geq 4(\frac{2K\log(n)}{nb_n})^{1/2}.$$

(A3) The minimal transition probabilities satisfy

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

Remark 3.5. The assumption about transition kernels in (A1) is related to the ergodicity coefficient for stationary Markov processes, cf. Iosifescu and Theodorescu (1969) and Rajarshi (1990).

Remark 3.6. By remark 3.5, the stationary probabilities are $\pi_n(w) = P_n(w),\ w \in \tau_n$. Thus, assumption (A2) about the minimum stationary probability bounds the size of the context tree as $|\tau_n| \leq b_n^{-1} \leq n/\log(n)^{2+\delta}$, which is the order of the number of parameters in the model.

Remark 3.7. For distinguishing a context wu form its parent node w in the context tree, assumption (A2) also guarantees a minimal L_1 distance between the relevant conditional distributions.

Theorem 3.1 *Consider data $X_{1,n}, \ldots, X_{n,n}$ as in (2.3), where $c_n(.)$ denotes the context function of model P_n, satisfying (A1)-(A3). Let $\hat{p}(.|.)$ be defined as in (3.2) and $\hat{c}(.)$ the estimate in Step 2 of the context algorithm. Then,*

(i) $\lim_{n\to\infty} \mathbb{P}[\hat{c}(.) = c_n(.)] = 1$, *or equivalently* $\lim_{n\to\infty} \mathbb{P}[\hat{\tau} = \tau_n] = 1$,

(ii) $\sup_{x_{-\infty}^1 \in \mathcal{X}^\infty} |\hat{p}(x_1|\hat{c}(x_{-\infty}^0)) - P_n(x_1|c_n(x_{-\infty}^0))| = o_P(1)\ (n \to \infty)$.

A proof of Theorem 3.1 is given in section 5. For d-consistency, we use the metric for probability measures P, Q on $\mathcal{X}^\infty$,

$$d(P,Q) = \sum_{m=1}^{\infty} 2^{-m} d_m(P \circ \pi_{1,\ldots,m}^{-1}, Q \circ \pi_{1,\ldots,m}^{-1}),$$
$$d_m(P \circ \pi_{1,\ldots,m}^{-1}, Q \circ \pi_{1,\ldots,m}^{-1}) = \sup_{x_1^m \in \mathcal{X}^m} |P(x_1^m) - Q(x_1^m)|, \tag{3.7}$$

where $\pi_{1,\ldots,m} : x \mapsto x_1, \ldots, x_m,\ x \in \mathcal{X}^\infty$.

Theorem 3.2 *Consider data $X_{1,n}, \ldots, X_{n,n}$ as in (2.3) with P_n satisfying (A1)-(A3). Then,*

(i) for $\hat{p}(.|.)$ as in (3.2) and $\hat{c}(.)$ the estimate in Step 2 of the context algorithm,

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

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

Remark 3.8. The measure $\hat{P}_n$ is with high probability geometrically ϕ-mixing, see Lemma 5.5. This, together with the d-consistency allows to reconstruct the probability law of a broad class of measurable functions of the true moving P_n, see Theorem 4.1.

Although Definition 2.2 only includes finite spaces with $|\mathcal{X}| < \infty$, our theoretical framework is flexible enough to allow also spaces $\mathcal{X}_n$ with $|\mathcal{X}_n| \to \infty$ as $n \to \infty$. We do not need to specify or bound the speed at which $|\mathcal{X}_n| \to \infty$, note that assumptions (A1)-(A3) are getting more restrictive when $|\mathcal{X}|$ is getting larger. Theorems 3.1 and 3.2 remain true for such generalizations. The growing alphabet case $|\mathcal{X}_n| \to \infty$ $(n \to \infty)$ is interesting when fitting context models to real valued stationary time series: first, the data would be quantized and then, a context model would be fitted on the quantized data. Obviously, the quantization should depend on the sample size n, getting finer as $n \to \infty$.

4 The context bootstrap

Theorem 3.2 indicates, that the estimate $\hat{P}_n$ of P_n can be used for resampling. Given observations $X_{1,n}, \ldots, X_{n,n}$ which take values in a finite space $\mathcal{X}$, we fit a context model as described in section 3.1 and simulate from it to obtain $X_1^*, \ldots, X_n^*$, now being the bootstrap sample of interest. In this case, our proposal will be a bootstrap for categorical time series which has a wide range of applications.

Since the context algorithm is nonparametric, our context bootstrap for categorical time series inherits the nonparametric property and offers an attractive and often more accurate alternative to the model free blockwise bootstrap, which has been proposed by Künsch (1989). We proceed as follows.

Step 1 Fit a context model as described in section 3.1, yielding a stationary probability measure $\hat{P}_n$ on $\mathcal{X}^{\infty}$, see Theorem 3.2.

Step 2 Draw a finite realization

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

The variables $X_1^*, \ldots, X_n^*$ are called the context bootstrap sample, they are nothing else than one random sample from the fitted context model. In practice, one would choose some starting values, generate a longer random sample via the estimated transition probabilities $\hat{p}(x_1|\hat{c}(x_{-\infty}^{0}))$ in Theorem 3.1 and then use the last n elements of such a longer sample as our bootstrap sample. By doing this, we avoid nonstationarity of a simulated Markov chain, due to starting values. Of course, one could also draw bootstrap samples of size $m \neq n$, cf. Bickel et al. (1994), but such generalizations are not the scope of this paper.

Given an estimator $T_n = T_n(X_{1,n}, \ldots, X_{n,n})$, which is a measurable function of $X_{1,n}, \ldots, X_{n,n}$, the bootstrapped estimator is defined by the plug-in rule $T_n^* = T_n(X_1^*, \ldots, X_n^*)$. Quantities induced by the resampling in Step 2 are denoted by an asterisk *.

4.1 Consistency of the context bootstrap

We present here an asymptotic result which justifies the use of the context bootstrap as defined in section 4. Such an asymptotic justification can only be given for a certain class of estimators T_n, our goal is to establish a consistency result for smooth functions of means. We will also discuss informally why the context bootstrap should work in the more general context of empirical processes, without giving the exact arguments.

We assume that we have observations $X_{1,n}, \ldots, X_{n,n} \in \mathcal{X}$ from a family of context models as given in (2.3). Consider the class of estimators, being smooth functions of means,

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

Examples of such estimators include estimation of transition probabilities in finite state 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 et al. (1995). We usually make the following assumption.

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

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

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 $d(.,.)$ defined in (3.7). Generally, P is not a context model anymore. It is then sufficient to assume

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

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

Remark 4.2. The function f is bounded, since $|\mathcal{X}| < \infty$.

The following Theorem justifies the context bootstrap for smooth functions of means.

Theorem 4.1 *Let $X_{1,n}, \ldots, X_{n,n}$ be as in (2.3) with P_n satisfying (A1) and (A2). Assume also that (B1) holds. Let the context bootstrap be defined as in section 4 and denote by $\theta_n^* = \mathbb{E}^*[f((X^*)_1^m)]$. Then,*

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

The proof of Theorem 4.1 is given in section 5.

4.1.1 General empirical processes

We point out that our results can probably be generalized to consistency of the context bootstrap for general empirical processes. As we will see in section 5, Lemma 5.5, the context bootstrap for categorical time series will satisfy a ϕ-mixing property with exponentially decaying mixing coefficients. This key result could then be used to show the consistency of the context bootstrap for general empirical processes indexed by Vapnik-Cervonenkis subgraph classes or by bracketing function classes with a weak condition on the bracketing number. Such results should be consequences of the general results about empirical processes in Arcones and Yu (1994) and Andrews and Pollard (1994). A route for general results for some bootstrap in time series, satisfying a mixing condition, has been given in Bickel and Bühlmann (1995). We do not give the precise arguments for such straightforward extensions.

These extensions are useful for studying the consistency of estimators

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

being 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, cf. Fahrmeir and Tutz (1994).

Consistency in empirical process theory would then imply that the context bootstrap works for the probabilistic core part of statistical procedures as given in (4.2). A sufficient technical condition for consistency in this class (4.2) would be compact differentiability of the functional T at the true underlying distribution, cf. Gill (1989).

4.2 Simulations

We study here the context bootstrap for variance estimation in various cases by simulation. We represent the models by context trees and equip final nodes with tuples, describing the transition probabilities. A tuple $(i_0, \ldots, i_{|\mathcal{X}|-1})$ corresponds to $p(j|w) = i_j / \sum_{j=0}^{|\mathcal{X}|-1} i_j$, $j \in \{0, \ldots, |\mathcal{X}| - 1\}$ (without loss of generality we let $\mathcal{X} = \{0, \ldots, |\mathcal{X}| - 1\}$).

We consider the following models:

(M1) Full binary Markov chain of order 3.

(M2) Full 4-ary Markov chain of order 2.

(M3) Semi-sparse binary context model of order 5.

(M4) Semi-sparse 4-ary context model of order 3.

(M5) Sparse binary context model of order 8.

(M6) Sparse 4-ary context model of order 4.

As sample sizes, we choose $n = 1000$ and $n = 2000$. We consider one statistic for the binary models (M1), (M3), (M5) and one for the 4-ary models (M2), (M4), (M6).

(S1) $T_n = \hat{p}_n(1|0) = N_n(1,0)/N_n(0)$ for binary models,

(S2) $T_n = N_n(1,3,3)$, the frequency of the word $(x_3, x_2, x_1) = (1,3,3)$, for 4-ary models.

The variance estimates are

$$\hat{\sigma}_n^2 = nVar^*(\hat{p}_n^*(1|0)) \text{ for } nVar(\hat{p}_n(1|0)),$$
$$\hat{\sigma}_n^2 = Var^*(N_n^*(1,3,3)) \text{ for } Var(N_n(1,3,3)),$$

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

Our moment estimates are based on 200 simulations over the different models, the results are given in Table 4.1 and 4.2. The true value of $nVar(T_n)$ (for (S1)) and of $Var(T_n)$ (for (S2)) is denoted by σ_n^2, computed over 1000 simulations. The relative mean square error is given by $RMSE(\hat{\sigma}_n^2) = \mathbb{E}|\hat{\sigma}_n^2 - \sigma_n^2|^2/\sigma_n^4$ and an estimated standard error thereof is given in parentheses. Instead of using cut-off values $K\log(n)$ in Step 2 of the context algorithm, we tried different cut-off values according to the $\chi^2/2$-quantiles,

	σ_n^2	$\mathbb{E}[\hat{\sigma}_n^2] - \sigma_n^2$	$Var(\hat{\sigma}_n^2)$	$RMSE(\hat{\sigma}_n^2)$
(M1,S1) 95%	0.8093	-0.0353	0.0203	0.0328 (0.0035)
(M1,S1) 98%	0.8093	-0.1038	0.0242	0.0534 (0.0040)
(M1,S1) 99.9%	0.8093	-0.2593	0.0158	0.1268 (0.0038)
(M3,S1) 95%	0.6688	-0.0222	0.0132	0.0306 (0.0029)
(M3,S1) 98%	0.6688	-0.0477	0.0138	0.0359 (0.0030)
(M3,S1) 99.9%	0.6688	-0.1702	0.0057	0.0776 (0.0027)
(M5,S1) 95%	0.5277	0.0072	0.0058	0.0210 (0.0027)
(M5,S1) 98%	0.5277	-0.0046	0.0020	0.0072 (0.0006)
(M5,S1) 99.9%	0.5277	0.0034	0.0015	0.0054 (0.0005)
(M2,S2) 95%	14.450	-0.5131	9.145	0.0451 (0.0051)
(M2,S2) 98%	14.450	0.1237	5.772	0.0277 (0.0027)
(M2,S2) 99.9%	14.450	-0.0321	3.785	0.0181 (0.0019)
(M4,S2) 95%	14.101	-0.2880	6.353	0.0324 (0.0044)
(M4,S2) 98%	14.101	-0.4388	5.543	0.0288 (0.0042)
(M4,S2) 99.9%	14.101	-0.4692	2.849	0.0154 (0.0014)
(M6,S2) 95%	11.201	-0.0179	4.756	0.0379 (0.0043)
(M6,S2) 98%	11.201	-0.0737	3.129	0.0250 (0.0029)
(M6,S2) 99.9%	11.201	-0.2893	2.008	0.0167 (0.0026)

Table 4.1: Context bootstrap variance estimates, sample size $n = 1000$.

	σ_n^2	$\mathbb{E}[\hat{\sigma}_n^2] - \sigma_n^2$	$Var(\hat{\sigma}_n^2)$	$RMSE(\hat{\sigma}_n^2)$
(M1,S1) 95%	0.8205	-0.0113	0.0146	0.0219 (0.0021)
(M1,S1) 98%	0.8205	-0.0248	0.0103	0.0162 (0.0018)
(M1,S1) 99.9%	0.8205	-0.1412	0.0241	0.0654 (0.0048)
(M3,S1) 95%	0.6670	-0.0048	0.0061	0.0138 (0.0013)
(M3,S1) 98%	0.6670	-0.0330	0.0066	0.0173 (0.0016)
(M3,S1) 99.9%	0.6670	-0.0876	0.0109	0.0418 (0.0033)
(M5,S1) 95%	0.5183	0.0065	0.0028	0.0108 (0.0012)
(M5,S1) 98%	0.5183	0.0024	0.0018	0.0070 (0.0008)
(M5,S1) 99.9%	0.5183	0.0086	0.0012	0.0048 (0.0004)
(M2,S2) 95%	12.854	0.1541	4.907	0.0257 (0.0026)
(M2,S2) 98%	12.854	0.4105	4.430	0.0240 (0.0026)
(M2,S2) 99.9%	12.854	0.8552	3.169	0.0203 (0.0025)
(M4,S2) 95%	14.653	-0.6567	4.252	0.0218 (0.0022)
(M4,S2) 98%	14.653	-0.6301	2.643	0.0142 (0.0013)
(M4,S2) 99.9%	14.653	-0.9512	1.691	0.0121 (0.0011)
(M6,S2) 95%	11.506	-0.0814	3.980	0.0301 (0.0036)
(M6,S2) 98%	11.506	-0.2734	1.854	0.0146 (0.0015)
(M6,S2) 99.9%	11.506	-0.4738	1.374	0.0121 (0.0016)

Table 4.2: Context bootstrap variance estimates, sample size $n = 2000$.

	σ_n^2	$\mathbb{E}[\hat{\sigma}_n^2] - \sigma_n^2$	$Var(\hat{\sigma}_n^2)$	$RMSE(\hat{\sigma}_n^2)$
(M5,S1) $\ell = 10$	0.5277	0.1075	0.0065	0.0650 (0.0053)
(M5,S1) $\ell = 20$	0.5277	0.0578	0.0102	0.0486 (0.0066)
(M5,S1) $\ell = 30$	0.5277	0.0296	0.0141	0.0537 (0.0065)

Table 4.3: Blockwise bootstrap variance estimates, sample size $n = 1000$.

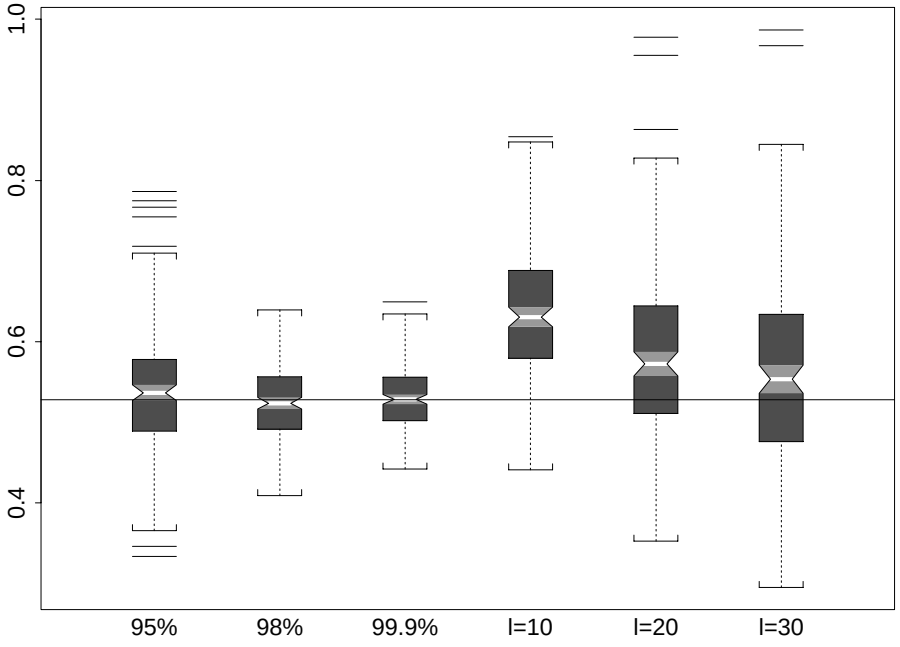

Figure 4.1: Boxplots of bootstrap variance estimates for case (M5,S1). Context bootstrap estimates are denoted with their $\chi_1^2/2$-quantiles as cut-off values, blockwise bootstrap estimates are denoted with their blocklengths ℓ, the line denotes the true variance.

see Remark 3.4. For the binary models (M1), (M3), (M5) we used the cut-offs $\chi^2_{1;\alpha}/2$, denoted in short by $\alpha 100\%$; for the 4-ary models (M2), (M4), (M6) we used the cut-offs $\chi^2_{3;\alpha}/2$, denoted in short by $\alpha 100\%$.

The results are promising in that the relative mean square error is most often smaller than 5%. Though there are some exceptions, we found that often 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.

For comparison, we also tried the blockwise bootstrap (Künsch, 1989) in the case (M5,S1) for sample size $n = 1000$ with different blocklengths b, see Table 4.3. A graphical representation of this comparison is given in Figure 4.2. The blockwise bootstrap exhibits a serious bias and a large variability, both in accordance with the asymptotic behavior for different blocksizes b: the bias decreases, whereas the variance increases with growing b. The context bootstrap is far better for this sparse context model (M5).

5 Proofs

We first recall and introduce some useful notation. We usually denote by $w, u, v \in \mathcal{X}^\infty$ (possibly finite) sequences, written in reverse time: $w = (\ldots, w_2, w_1)$. Sometimes we look at the (finite) sequence $wu = (\ldots, u_2, u_1, \ldots, w_2, w_1) \in \mathcal{X}^\infty$ $(w, u \in \mathcal{X}^\infty)$. Transition probabilities (outcome distributions) in a context tree τ are indexed by $w \in \tau$: $p_w(.) = p(.|w)$. We also denote by $P_w(x) = P(xw)/P(w)$ for general $w \in \mathcal{X}^\infty$, $x \in \mathcal{X}$ (w not necessarily a context in τ) and P a stationary probability measure on $\mathcal{X}^\infty$. In the new notation, a context model is completely specified by the context tree τ and the set $\{p_w(.); w \in \tau\}$. Estimated transition probabilities are denoted by $\hat{p}_w(x) = N(xw)/N(w)$, $N(.)$ as defined in (3.1). We recall that for any context $w = w'u$ $(u \in \mathcal{X})$ we have defined $\Delta_w = D(\hat{p}_{w'u}||\hat{p}_{w'})N(w)$. When looking at a sequence $(P_n)_{n\in\mathbb{N}}$ of context models, we sometimes drop the index n.

Proof of Theorem 3.1.
We define first the events of under- and overestimation for sample size n,

$$U_n = \{\text{there exists } w \in \hat{\tau} \text{ with } wu \in \tau_n \text{ and } wu \notin \hat{\tau}^f \ (u \in \mathcal{X}^\infty)\}$$
$$O_n = \{\text{there exists } wu \in \hat{\tau} \ (u \in \mathcal{X}^\infty) \text{ with } w \in \tau_n^f \text{ and } wu \notin \tau_n^f\},$$

where τ^f denotes the final node context tree corresponding to τ, see Definition 2.3. Note that by formula (3.3) we can also characterize U_n and O_n in terms of the pruning criterion $\Delta_{wu} < K \log(n)$. 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 $\delta > 0$ hold. Then for any $0 < r < 1/3$,*

$$\mathbb{P}[U_n] = O(n^{-\log(n) r\delta}) \ (n \to \infty).$$

Proof: We partition the underestimation event U_n using the the event

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

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,

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

It is well known, c.f. Cover and Thomas (1991), that the divergence can be lower bounded by the 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 $A = \{x \in \mathcal{X} : \hat{p}_{wu}(x) > \hat{p}_w(x)\}$. Therefore,

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

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 $\gamma_n^2(k) = \frac{K\log(n)}{k}$ and $\hat{p}_{wu}(x) = a$, $\hat{p}_w(x) = b$, $p_{wu}(x) = r$ and $p_w(x) = s$, where $x \in \mathcal{X}$. 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 $r - s > \epsilon_n$. Now if $b < s$, then $|a-b| < \gamma_n(k)$ implies that $|a-r| > \epsilon_n - \gamma_n(k)$. Furthermore, if $b > r$, then it must be that $|s-b| > \epsilon_n$. Now if $s \leq b \leq r$ then either $s \leq b < s + \frac{r-s}{2}$, in which case $|r-a| > \frac{\epsilon_n}{2} - \gamma_n(k)$ or $r - \frac{r-s}{2} \leq b \leq r$, in which case $|s-b| > \frac{\epsilon_n}{2}$. Taken together we have proved that if $|\hat{p}_{wu}(x) - \hat{p}_w(x)| < \gamma_n(k)$, then either $|\hat{p}_{wu}(x) - p_{wu}(x)| > \frac{\epsilon_n}{2} - \gamma_n(k)$ or $|\hat{p}_w(x) - p_w(x)| > \frac{\epsilon_n}{2} - \gamma_n(k)$. Thus, when applied to (5.2), we have proved that for

$$a_n(k) = (\frac{\epsilon_n}{2} - \gamma_n(k))^2, \quad (5.3)$$

it must be that

$$\begin{aligned}&\mathbb{P}[D(\hat{P}_{wu}||\hat{P}_w) < \frac{K\log(n)}{k}, N(wu) = k, N(w) = j] \\ \leq\ &\mathbb{P}[\sum_{x \in A} |\hat{p}_{wu}(x) - p_{wu}(x)| \geq a_n(k)^{1/2}, N(wu) = k] \\ +\ &\mathbb{P}[\sum_{x \in A} |\hat{p}_w(x) - p_w(x)| \geq a_n(k)^{1/2}, N(w) = j] \\ \leq\ &|\mathcal{X}| \max_{x \in \mathcal{X}} \mathbb{P}[|\hat{p}_{wu}(x) - p_{wu}(x)| \geq a_n(k)^{1/2}, N(wu) = k] \\ +\ &|\mathcal{X}| \max_{x \in \mathcal{X}} \mathbb{P}[|\hat{p}_w(x) - p_w(x)| \geq a_n(k)^{1/2}, N(w) = j]. \quad (5.4)\end{aligned}$$

Since $k \geq \rho_n$, is must be that $\gamma_n(k)^2 \leq \frac{K\log(n)}{\rho_n}$. Thus it follows that for $\epsilon_n \geq 4\sqrt{\frac{K\log(n)}{\rho_n}}$,

$$\min_{k \geq \rho_n} a_n(k) = \min_{k \geq \rho_n} (\frac{\epsilon_n}{2} - \gamma_n(k))^2 \geq \frac{K\log(n)}{\rho_n}.$$

Also, we will now choose $\rho_n = b_n n/2 \geq \log(n)^{3+\delta}/2$ and note that $k a_n(k) \geq K\log(n)$ for $k \geq \rho_n$.

We treat the two cases on the RHS 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 $X_{-\infty}^{\infty}$. Define,

$$I_i = \{\text{the time of the } i^{th}\text{occurrence of } v \text{ in } X_{-\infty}^{\infty}\}.$$

Then let

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

The sequence Z_1^{∞} is a stationary ϕ-mixing sequence with mixing coefficients bounded by the same bound as the original sequence $X_{-\infty}^{\infty}$. The marginal probability distribution of Z_1 on $\mathcal{X}$ is equal to P_v. Let $Y_i = 1_{[Z_i = x]}$. Now observe that

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

Thus, we have established the upper bound,

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

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

Lemma 5.1 *Let Y_1^{∞} with $E[Y_1] = p$ be defined as above and $a_n(k)$ as in (5.3). Assume the conditions (A1) and (A2) with $\delta > 0$. Then, for $k \geq \rho_n = b_n n/2$,*

$$\sup_{0<p<1} \mathbb{P}[|\sum_{i=1}^{k} \frac{Y_i}{k} - p|^2 > a_n(k)] \leq 2exp(3\sqrt{e})exp(-D\log(n)^{1+\delta/3}),$$

$D > 0$ a constant depending on the mixing rate.

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 $(Y_i)_{i\in\mathbb{N}}$. Thus, by applying Proposition 2 from Doukhan (1994, Ch.1.4.2) with $\sigma^2 = 8BM\log(k)\sqrt{K\log(n)}/\sqrt{k}$, $M \geq -1/\log(1-2\kappa)$ and $B^{-1} = 8(1 + 4\sum_{j=1}^{\infty}(1-2\kappa)^j)$, we get

$$\mathbb{P}[|\sum_{i=1}^{k} \frac{Y_i}{k} - p|^2 > a_n(k)] \leq 2exp(3\sqrt{e})exp(-const.\frac{(\log(n)k)^{1/2}}{\log(k)}).$$

Now using that $k \geq \rho_n \geq \log(n)^{3+\delta}/2$, the result follows. □

Denote by $M_n = 2exp(3\sqrt{e})exp(-D\log(n)^{1+\delta/3})$. A straightforward application of Lemma 5.1 to equation (5.5) proves that for $k, j \geq \rho_n$,

$$\max_{x\in\mathcal{X}} \mathbb{P}[(\hat{p}_{wu}(x) - p_{wu}(x))^2 \geq a_n(k), N(wu) = k] \leq M_n,$$
$$\max_{x\in\mathcal{X}} \mathbb{P}[(\hat{p}_{w}(x) - p_{w}(x))^2 \geq a_n(k), N(w) = j\} \leq M_n.$$

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

$$\mathbb{P}[U_n \cap D_n] \leq |\mathcal{X}| \sum_{wu \in \tau_n} \sum_{k=\rho_n}^{n} \sum_{j=k}^{n} [M_n + M_n]. \tag{5.6}$$

By Remark 6.1, (5.6) and assumption (A2),

$$\begin{aligned} & \mathbb{P}[U_n \cap D_n] \leq const.|\tau_n| n^2 M_n \\ \leq \; & const. b_n^{-1} n^2 exp(-const. \log(n)^{1+\delta/3}) \leq const. n^{-\log(n) r\delta}, \; 0 < r < 1/3. \end{aligned} \tag{5.7}$$

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

$$\begin{aligned} & \mathbb{P}[D_n^c] \leq \sum_{w \in \tau_n} \mathbb{P}[N(w) < \rho_n] = \sum_{w \in \tau_n} \mathbb{P}[N(w) - n\pi_n(w) < \rho_n - n\pi_n(w)] \\ \leq \; & \sum_{w \in \tau_n} \mathbb{P}[N(w) - n\pi_n(w) < -b_n n/2] \leq \sum_{w \in \tau_n} \mathbb{P}[|N(w) - E[N(w)]| > b_n n/2]. \end{aligned}$$

We bound this quantity by the following exponential inequality.

Lemma 5.2 *Assume that (A1) and (A2) with $\delta > 0$ hold. Then, for any $0 < r < 1$,*

$$\max_{w \in \tau_n} \mathbb{P}[|N(w) - \mathbb{E}[N(w)]| \geq b_n n/2] \leq n^{-\log(n)(1+r\delta)}(1 + o(1)).$$

Proof: Since $w \in \tau_n$ we can write

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

By assumption (A1), $(Z_{t,n})_{t \in \mathbb{Z}}$ is ϕ-mixing with mixing coefficients bounded by $\sup_{n \in \mathbb{N}} \phi_n(j) \leq (1 - 2\kappa)^j$, cf. Rajarshi (1990). Thus, we can apply Proposition 2 in Doukhan (1994, Ch. 1.4.2) with $\sigma^2 = 8BM \log(n) b_n n/(2n)$, $M \geq -1/\log(1 - 2\kappa)$ and $B^{-1} = 8(1 + 4\sum_{j=1}^{\infty}(1 - 2\kappa)^j)$. This then yields

$$\max_{w \in \tau_n} \mathbb{P}[|N(w) - \mathbb{E}[N(w)]| \geq b_n n/2] \leq 2exp(3e^{1/2}) exp(-const. \frac{b_n n}{\log(n)}).$$

By using assumption (A2) about b_n we complete the proof. □

By Lemma 5.2

$$\mathbb{P}[D_n^c] = O(b_n^{-1} n^{-\log(n)(1+r\delta)}) = O(n^{-\log(n)}).$$

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 $w = w'u \in \hat{\tau}$ $(u \in \mathcal{X}^\infty)$ with $w' \in \tau_n$ and $w \notin \tau_n^f\}$. For a sequence w to be an element of $\hat{\tau}$, it is necessary that $N(w) > 1$ and $\Delta_w \geq K \log(n)$. Now Weinberger et. al. (1995) establish for any $w = w'u$ $(w' \in \tau_n, \; u \in \mathcal{X}^\infty)$,

$$\mathbb{P}[\Delta_w \geq K \log(n+1)] \leq (n+1)^{2a}(n+1)^{-K}.$$

Here, $a = |\mathcal{X}|$. In their algorithm, an overestimation event can only occur at any string w if $|w| \leq \frac{\log(n)}{\log(a)}$. Thus they establish that

$$\mathbb{P}[O_n] \leq \sum_{|w| \leq \frac{\log(n)}{\log(a)}} (n+1)^{-K+2a} \leq n^{-K+2a+1}.$$

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

It is possible to prove a stronger result, eliminating the need for a length restriction. We just give an outline of such a proof.

Lemma 5.3 *Let swv be any possible string with $s \in \tau_n, w \in \mathcal{X}^\infty \cup \emptyset$ and $v \in \mathcal{X}$. Let $O_n(swv) = \{\Delta_{swv} \geq K\log(n), N(swv) > 1\}$. Denote by $p_{\min}(n) = \min_{x \in \mathcal{X}, w \in \tau_n} p_w(x)$ and by $\hat{\tau}_{max}$ the maximal context tree in Step 1 of the context algorithm. Then, under the assumptions (A1)-(A2),*

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

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,

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

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,

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

Therefore, since $K > 2a + 3$ 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 $\tilde{\Delta}_{wu} = \|\hat{p}_w(.) - \hat{p}_{wu}(.)\|_1^2$ and define $\tilde{O}_n = \{\text{there exists } w = w'u \ (\ w' \in \tau_n^f,\ u \in \mathcal{X}^\infty), \text{ such that } \tilde{\Delta}_w \geq K\log(n) \text{ and } N(w) > 1\}$.

Theorem 5.3 *Under the assumptions (A1)-(A2),*

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

Proof: The main step in establishing the overestimation bound is practically identical to the underestimation step in the proof of Theorem 5.1. □

Proof of Lemma 5.3. Let $s \in \tau_n$ be a context and $su = swv$ with $w \in \mathcal{X}^\infty \cup \emptyset$ and $v \in \mathcal{X}$. Our aim is to bound the probability of overestimation at su. We begin by recalling several inequalities and definitions from Weinberger et. al. (1995). First, we fix a sequence x_1^n, being a realization from P_n. We can determine a probability law given by $Q_{su}(y_1^n|x_1^n)$ (on the set of sequences of length n), defined as follows:

$$\begin{aligned}\log(Q_{su}(y_1^n|x_1^n)) &= R_{sw}(y_1^n|S_s) + \sum_{x\in\mathcal{X}}\sum_{b\neq v} N_{y_1^n}(x|swb)\log(\hat{P}_{x_1^n}(x|sw)) \\ &\quad + \sum_{x\in\mathcal{X}} N_{y_1^n}(x|su)\log(\hat{P}_{x_1^n}(x|su)).\end{aligned}$$

where $R_{sw}(y_1^n|S_s)$, defined formally in Weinberger et al. (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 $x \in \mathcal{X}$. If $\Delta_{x_1^n}(swv) > K\log(n)$, it follows from (A9) in Weinberger et. al. (1995), that

$$P_n(\sigma_{x_1^n}) \leq Q_{su}(\sigma_{x_1^n}|x_1^n)n^{-K}. \tag{5.8}$$

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

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

If $b_0 = v$ then we define

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

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

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

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

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

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

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

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

$$\mathbb{P}[O_n(swv)] = P_n(y_1^n; \Delta_{y_1^n}(swv) > K\log(n))\frac{1}{p_{\min}(n)} \leq P_n(sw \in \hat{\tau}_{max})n^{-K+2a}.$$

□

Theorem 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 3.1 (i): partition with the set D_n and use Lemma 5.1 and 5.2. □

Theorem 3.2 (i) follows from the more general formula (5.16) and Theorem 3.2 (ii) is an immediate consequence of Theorem 3.1 (ii). □

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

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

and denote by $\Sigma_n = Cov[U_n]$ 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 $n \geq n_0$.

(ii) for $Z \sim \mathcal{N}_v(0, I)$,

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

Proof: For every $n \in \mathbb{N}$, the process $(X_{t,n})_{t\in\mathbb{Z}}$ is ϕ_n-mixing with mixing coefficient

$$\phi_n(k) = \sup\{\mathbb{P}[A] - \mathbb{P}[A\cap B]/\mathbb{P}[B];\ A \in \mathcal{F}_{-\infty,n}^{0,n},\ B \in \mathcal{F}_{k,n}^{\infty,n},\ \mathbb{P}[B] \neq 0\},$$

where the σ-fields are $\mathcal{F}_{a,n}^{b,n} = \sigma(\{X_{a,n}^{b,n}\}),\ a < b$.
By assumption (A1), the mixing coefficients are bounded by

$$\sup_{n\in\mathbb{N}} \phi_n(k) \leq (1-2\kappa)^k, \tag{5.9}$$

cf. Rajarshi (1990, Lem. 2.1).
Bounding covariances in terms of mixing coefficients, cf. Doukhan (1994), and using the bound in (5.9) implies for $i, j \in \{1, \ldots, v\}$,

$$(n-m+1)(\Sigma_n)_{i,j} = \sum_{k=-n+1}^{n-1} Cov(f_i(X_0^{m-1}), f_j(X_k^{k+m-1})) + O(n^{-1}). \tag{5.10}$$

Hence, assertion (i) follows from the assumption in (B1).
Assertion (i), assumption (B1) and (5.10) allow us to write

$$\Sigma_n^{-1/2} = n^{1/2},_n,\ \sup_{n\in\mathbb{N}}\max_{1\leq i,j\leq v} |(,_n)_{i,j}| < \infty. \tag{5.11}$$

Now write

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

where $\tilde{f}_n(X_t^{t+m-1}) = , _n(f(X_t^{t+m-1}) - \theta_n)$.
By construction and (5.11),

$$\begin{aligned} &\mathbb{E}[\tilde{f}_n(X_t^{t+m-1})] = 0, \ t \in \mathbb{Z}, \\ &Var(n^{-1/2} \sum_{t=1}^{n-m+1} \tilde{f}_n(X_t^{t+m-1})) \to 1 \ (n \to \infty), \\ &\sup_{n \in \mathbb{N}} \mathbb{E}|\tilde{f}_n(X_t^{t+m-1})|^2 < \infty, \ t \in \mathbb{Z}. \end{aligned} \tag{5.12}$$

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 in Vol. 63) are easily verified by invoking the mixing bound in (5.9) and (5.12). Thus,

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

and assertion (ii) follows by Polya's Theorem. □.

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

$$n^{1/2}(T_n - g(\theta_n)) = n^{1/2} Dg(\tilde{\theta}_n)(U_n - \theta_n), \tag{5.13}$$

where $Dg(\theta) = [\frac{\partial g_i(u)}{\partial u_j}]_{i,j}, \ (1 \le i \le w, \ 1 \le j \le v)$ and $\|\tilde{\theta}_n - \theta_n\| \le \|U_n - \theta_n\|$.
By (5.11) and Lemma 5.4 (ii), $U_n - \theta_n = o_P(1)$, so that

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

This, together with (5.13), the boundedness of $n^{1/2}\Sigma_n^{1/2}$ (use (5.10)) and Lemma 5.4 (ii) implies

$$\sup_{x \in \mathbb{R}^w} |\mathbb{P}[n^{1/2}(T_n - g(\theta_n)) \le x] - \mathbb{P}[n^{1/2}\Sigma_n^{1/2} Dg(\theta_n) Z \le x]| = o(1) \ (n \to \infty), \tag{5.14}$$

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

We are going now to show the bootstrap analog of (5.14). We first establish a mixing property for the bootstrap process $(X_t^*)_{t \in \mathbb{Z}}$. Note that the distribution of $(X_t^*)_{t \in \mathbb{Z}}$ depends again on the sample size n. We define

$$\phi_n^*(k) = \sup\{|\mathbb{P}^*[A] - \mathbb{P}^*[A \cap B]/\mathbb{P}^*[B]|; A \in \mathcal{F}_{-\infty,0}^*, \ B \in \mathcal{F}_{k,\infty}^*, \ \mathbb{P}^*[B] \ne 0\},$$

where $\mathcal{F}_{a,b}^* = \sigma(\{(X^*)_a^b\}), \ a < b$.
The next result establishes the mixing property for the bootstrap process $(X_t^*)_{t \in \mathbb{Z}}$.

Lemma 5.5 *Consider data $X_{1,n}, \ldots, X_{n,n}$ from $P_n \in \mathcal{P}$ as in (2.3), satisfying (A1) and the context bootstrap is defined as in section 4. Then,*

$$\mathbb{P}[\phi_n^*(k) \le (1-\kappa)^k \text{ for all } k \in \mathbb{N}_0] \to 1 \ (n \to \infty).$$

In particular, the bound for the mixing coefficients $\phi_n^*(k)$ is non-random and the same for all $n \in \mathbb{N}$.

Proof: The r-step transition kernel $p_Z^{(r)}(v,w) = \mathbb{P}[Z_r = v | Z_0 = w]$ $(r \ge 1)$ for the state process $Z_t = c(X_0^t x_0^\infty)$ $(t > 0)$ of a context model $(X_t)_{t\in\mathbb{Z}}$ can be characterized by $p(.|.)$ and $c(.)$, i.e.,

$$T(v|w; r, p(.|.), c(.)) = p_Z^{(r)}(v,w) = \sum_{x_1^{r-1} \in \mathcal{X}^{r-1},\ c(x_r \ldots x_1 w x_0^\infty) = v} \prod_{i=0}^{r-1} p(x_{r-i} | c(x_1^{r-i-1} w x_0^\infty)) \quad (5.15)$$

For every $n \in \mathbb{N}$, the bootstrap process $(X_t^*)_{t\in\mathbb{Z}}$ is a context model. We consider its r-step transition kernel for the states $P^{*(r)}(v,w) = \mathbb{P}^*[Z_r^* = v | Z_0^* = w]$ $(r \ge 1)$, where $Z_t^* = \hat{c}((X^*)_0^t x_0^\infty)$ $(t \in \mathbb{N}_0)$ is the bootstrap state process. This transition is characterized by

$$T(v|w; r, \hat{p}(.|.), \hat{c}(.)) = P^{*(r)}(v,w) \ (r \ge 1).$$

We now obtain an analogon of (A1) for the bootstrap. We have,

$$\begin{aligned}
& |T(v|w; r, \hat{p}(.|.), \hat{c}(.)) - T(v|w'; r, \hat{p}(.|.), \hat{c}(.))| \\
\le\ & |T(v|w; r, p(.|.), c(.)) - T(v|w'; r, p(.|.), c(.))| + |T(v|w; r, \hat{p}(.|.), c(.)) - T(v|w; r, p(.|.), c(.))| \\
+\ & |T(v|w'; r, \hat{p}(.|.), c(.)) - T(v|w'; r, p(.|.), c(.))| + 2 1_{[\hat{c}(y) \ne c(y) \text{ for some } y \in \mathcal{X}^\infty]}.
\end{aligned}$$

We now invoke (A1) for $T(.|.; r, p(.|.), c(.))$ about the true underlying process. For the other terms we use the finiteness of r and $\mathcal{X}$, together with (5.15) and Theorem 3.1. We then obtain,

$$\begin{aligned}
\sup_{v,w,w'} |P^{*(r)}(v,w) - P^{*(r)}(v,w')| &= \sup_{v,w,w'} |T(v|w; r, \hat{p}(.|.), \hat{c}(.)) - T(v|w'; r, \hat{p}(.|.), \hat{c}(.))| \\
&\le 1 - 2\kappa + o_P(1),
\end{aligned}$$

so that

$$\sup_{v,w,w'} |P^{*(r)}(v,w) - P^{*(r)}(v,w')| < 1 - \kappa \text{ in probability.} \quad (5.16)$$

Thus, we can restrict ourselves to sets $A_n = \{\omega;\ \sup |P^*(v,w) - P^*(v,w')|(\omega) < 1 - \kappa \text{ and } \hat{c}(.;\omega) = c(.)\}$, where the sup is over the set as in (5.16) and ω is an element of the underlying probability space. By construction of $(X_t^*)_{t\in\mathbb{Z}}$ we conclude that $(X_t^*)_{t\in\mathbb{Z}}$ is ϕ-mixing on A_n with mixing coefficients bounded by

$$\phi_n^*(k) \le (1-\kappa)^k \text{ for all } k \in \mathbb{N}_0 \text{ on the set } A_n,$$

cf. Rajarshi (1990, Lem. 2.1).
But by (5.16), $\mathbb{P}[A_n] \to 1$ as $n \to \infty$, which completes the proof. □

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

Lemma 5.6 *Assume the conditions of Theorem 4.1. Then,*

(i) $n(\Sigma_n^* - \Sigma_n)_{i,j} = o_P(1)$ $(n \to \infty)$, $i, j = 1, \ldots, v$.

(ii) $\lim_{n\to\infty} \mathbb{P}[n\Sigma_n^*$ *is positive definite*$] = 1$.

(iii) for $Z \sim \mathcal{N}_v(0, I)$,

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

Proof: For any $i, j \in \{1, \ldots, v\}$,

$$n(\Sigma_n^*)_{i,j} = \sum_{k=-n+m}^{n-m} Cov^*(f_i((X^*)_0^{m-1}), f_j((X^*)_k^{k+m-1}))(1 - \frac{|k|}{n-m+1})$$

$$= \sum_{k=-M}^{M} Cov^*(f_i((X^*)_0^{m-1}), f_j((X^*)_k^{k+m-1}))(1 - \frac{|k|}{n-m+1}) + \Delta_{n,M}, \qquad (5.17)$$

where M is a finite constant.
By well known bounds of covariances in terms of mixing coefficients, cf. Doukhan (1994),

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

Therefore by Lemma 5.5,

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

By Theorem 3.2 (ii),

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

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

$$|\sum_{k=-M}^{M} Cov^*(f_i((X^*)_0^{m-1}), f_j((X^*)_k^{k+m-1})) - \sum_{k=-M}^{M} Cov(f_i(X_0^{m-1}), f_j(X_k^{k+m-1}))|$$

$$= o_P(1) \ (n \to \infty). \qquad (5.20)$$

By the geometric ϕ-mixing property of $(X_t)_{t\in\mathbb{Z}}$, see (5.9), and the boundedness of f,

$$|\sum_{k=-M}^{M} Cov(f_i(X_0^{m-1}), f_j(X_k^{k+m-1})) - \sum_{k=-\infty}^{\infty} Cov(f_i(X_0^{m-1}), f_j(X_k^{k+m-1}))|$$

$$= o(1) \ (M \to \infty). \qquad (5.21)$$

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 Lemma 5.4 (ii); we now invoke the mixing bound in Lemma 5.5 and use (i). □

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

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

and hence by the continuous differentiability of g,

$$[Dg(\tilde{\theta}_n^*) - Dg(\theta_n)]_{i,j} = o_P(1) \text{ for } \|\tilde{\theta}_n^* - \theta^*\| \le \|U_n^* - \theta^*\|,\ (1 \le i \le w,\ 1 \le j \le v). \tag{5.23}$$

A first order Taylor expansion, (5.23), Lemma 5.6 (iii) and the boundedness of $n\Sigma_n^* = O_P(1)$ imply

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

where $Z \sim \mathcal{N}_v(0, I)$.
By (5.14) and (5.24) we complete the proof of Theorem 4.1. □

Acknowledgments. We thank Itai Zukerman for carrying out the computations.

References

[1] Andrews, D.W.K. and Pollard, D. (1994). An introduction to functional central limit theorems for dependent stochastic processes. Internat. Statist. Rev. **62** 119-132.

[2] Arcones, M.A. and Yu, B. (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences. J. Theoret. Probab. **7** 47-71.

[3] Bickel, P.J. and Bühlmann, P. (1995). Mixing property for some sieve bootstrap in time series and functional central limit theorems. Tech Rep. 447, Dept. Statist., University of California, Berkeley.

[4] Bickel, P.J., Götze, F. and van Zwet, W.R. (1994). To appear in Statistica Sinica.

[5] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone. C.J. (1984). Classification and Regression Trees. Wadsworth, Belmont.

[6] Brillinger, D.R. (1994). Examples of scientific problems and data analyses in demography, neurophysiology, and seismology. J. Comp. and Graph. Statistics **3** 1-22.

[7] Bühlmann, P. (1996). Extreme events from return–volume process: a discretization approach for complexity reduction. To appear in Applied Financial Economics.

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

[9] Doukhan, P. (1994). Mixing. Properties and Examples. Lect. Notes in Stat. **85**. Springer, New York.

[10] Fahrmeir, L. and Tutz, G. (1994). Multivariate Statistical Modelling Based in Generalized Linear Models. Springer, New York.

[11] Gill, R.D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1). Scand. J. Statist. **16** 97-128.

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

[13] Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. Ann. Statist. **17** 1217-1241.

[14] Prum, B., Rodolphe, F. and deTurckheim, E. (1995). Finding words with unexpected frequencies in deoxyribonucleic acid sequences. J. Roy. Statist. Soc B **57** 205-220.

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

[16] Rissanen, J.J. (1983). A universal data compression system. IEEE Trans. Inform. Theory **IT-29** 656-664.

[17] Weinberger, M.J., Rissanen, J.J. and Feder, M. (1995). A universal finite memory source. IEEE Trans. Inform. Theory **IT-41** 643-652.

[18] Withers, C.S. (1981). Central limit theorems for dependent variables I. Z. Wahrsch. verw. Gebiete **57** 509-534 (Corr: **63** p555).

Department of Statistics
University of California
367 Evans Hall #3860
Berkeley, CA 94720-3860