9-2003

Estimation of General Stationary Processes by Variable Length Markov Chains

Fiorenzo Ferrari

Abraham J. Wyner
University of Pennsylvania

Recommended Citation

Ferrari, F., & Wyner, A. J. (2003). Estimation of General Stationary Processes by Variable Length Markov Chains. *Scandinavian Journal of Statistics, 30* (3), 459-480. http://dx.doi.org/10.1111/1467-9469.00342

Estimation of General Stationary Processes by Variable Length Markov Chains

Abstract

We develop new results about a sieve methodology for estimation of minimal state spaces and probability laws in the class of stationary categorical processes. We first consider finite categorical spaces. By using a sieve approximation with variable length Markov chains of increasing order, we carry out asymptotically correct estimates by an adapted version of the Context Algorithm (see Rissanen (1983)). It thereby yields a nice graphical tree representation for the potentially infinite dimensional minimal state space of the data generating process. This procedure is also consistent for increasing size countable categorical spaces. Finally, we show similar results for real-valued general stationary processes by using a quantization procedure based on the distribution function.

Keywords

context algorithm, sieve approximation, state space estimation, strong mixing sequence, time series, tree model representation

Disciplines

Statistics and Probability

Estimation of General Stationary Processes by Variable Length Markov Chains

by

Fiorenzo Ferrari

Research Report No. 88
October 1999

Seminar für Statistik
Eidgenössische Technische Hochschule (ETH)
CH-8092 Zürich
Switzerland

Estimation of General Stationary Processes by Variable Length Markov Chains

Fiorenzo Ferrari

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

October 1999

Abstract We develop new results about a sieve methodology for estimation of minimal state spaces and probability laws in the class of stationary categorical processes. We first consider finite categorical spaces. By using a sieve approximation with variable length Markov chains of increasing order, we carry out asymptotically correct estimates by an adapted version of the Context Algorithm (see Rissanen (1983)). It thereby yields a nice graphical tree representation for the potentially infinite dimensional minimal state space of the data generating process. This procedure is also consistent for increasing size countable categorical spaces. Finally, we show similar results for real-valued general stationary processes by using a quantization procedure based on the distribution function.

AMS 1991 subject classifications. Primary 60J10; secondary 62G20, 62M10.
Key words and phrases. Context Algorithm, sieve approximation, time series, state space estimation, tree model representation, strong mixing sequence.

1 Introduction

The assumption that a sequence of data belongs to a certain type of models, helps to better understand the features of the analyzed realizations and allows in particular to predict possible developments of the underlying process. On the other hand, a fixed model almost never corresponds to reality. The method of sieves (Grenander (1981)) combines the advantages of a model but allows model-misspecification for any finite sample size. It only requires that in the limit, as sample size tends to infinity, some basic assumptions such as stationarity hold.

For the estimation of general stationary processes, we propose the method of sieves with variable length Markov chains of increasing order. These models are still Markovian of potentially high order, but with a sparse memory having some states lumped together. In favourable cases, for example when the process has a memory which tends to certain "directions", this yields a drastic reduction in the number of parameters to be estimated without restricting necessarily to short memories.

The advantage of the presented method in comparison to the use of full Markov chains is higher efficiency for estimation. For a full Markov chain of order d taking values in a finite categorical space $\mathcal{X}$, the number of free parameters is $|\mathcal{X}|^d (|\mathcal{X}| - 1)$ ($|\mathcal{X}|$ = cardinality of $\mathcal{X}$), which is already very big for moderate values of d. Estimation is therefore very poor in many practical applications with only moderate values of d. Since the dimension of the models in the class of full Markov chains grows exponentially in the order d, their structure is not so flexible as in the case of variable length Markov chains. Consequently, as described in Remark 7, variable length Markov chain approximation is often naturally linked to an increasing order $d = d_n$ which is polynomial in the sample size n, whereas full Markov chains typically use an approximation of order $d_n = \mathcal{O}(\log(n))$. The idea of sieve approximation is better understood thanks to a nice graphical representation. This uses trees, which grow downwards and whose branches stand for the relevant history of the underlying process (see Subsection 2.2).

For general stationary processes taking values in a finite categorical space, the probability distribution and the minimal state space, i.e. the relevant memory for future outcomes, are approximated by that of variable length Markov chains of increasing order. For the latter the estimation is performed by using an adapted version of the Context Algorithm (see Rissanen (1983),Weinberger, Rissanen and Feder (1995) and Bühlmann and Wyner (1999)), whose main operations are local decision between two possible states.

If the minimal state space has finite length (the underlying process is thus a variable length Markov chain), then the Context Algorithm consistently finds the right model (see also Weinberger, Rissanen and Feder (1995) and Bühlmann and Wyner (1999)). The most important new result in our article is given for the estimate of the memory of a process, whose order is infinite; in this case, the Context Algorithm selects automatically variable length Markov chains whose orders grow to infinity for increasing sample size. This new development guarantees broader perspectives: the adaptation of models to data is now possible without necessarily assuming finite minimal state spaces. Similar results are shown to hold also for increasing size categorical spaces. The operation of the Context Algorithm can hence be also interpreted as a model selection in the class of variable length Markov chains. Attacking this problem with conventional criteria, such as AIC or BIC, is computationally infeasible.

For real-valued general stationary processes, we present a quantization procedure based on the distribution function, which partitions $\mathbb{R}$ into a countable union of disjoint intervals. Since for the quantization becoming finer, the quantized process takes value in a categorical space with increasing alphabet, we can apply the above proceedings to achieve consistent estimates.

Our results have potential impact to a variety of applications: to mention a few, modelling of categorical time-series (for example DNA sequences, see Bühlmann and Wyner (1999), Braun and Müller (1998), quantization of nonlinear stationary real-valued time-series (see Section 4.2) and sieve-bootstrapping stationary categorical time-series.

In the first section we define the variable length Markov chains on a finite categorical space and give a tree representation of their minimal state space, which will be useful in the second section, when describing a version of the Context Algorithm proposed by Bühlmann and Wyner (1999). Theoretical results about consistent estimation of the minimal state space and the probability distribution of general stationary processes are given in the third section, whether for countable or uncountable spaces, the latter treated with a quantization procedure. A small simulation experiment is also given there. The last section contains all the proofs.

2 Variable Length Markov Chains

2.1 Definition

Let $\mathcal{X}$ be a finite categorical space and $(X_t)_{t\in\mathbb{Z}}$ an $\mathcal{X}$-valued stationary Markov chain of finite order p. We denote by P the probability distribution of $(X_t)_{t\in\mathbb{Z}}$ on $\mathcal{X}^{\mathbb{Z}}$ and use the notation

$$\begin{aligned}&P(x_a^b) := \mathbb{P}_P[X_a^b = x_a^b]\,,\\ &P(x_b|x_a^{b-1}) := \mathbb{P}_P[X_b = x_b|X_a^{b-1} = x_a^{b-1}]\,,\ \text{for } x_a^b \in \mathcal{X}^{b-a+1}\,,\end{aligned}$$

where in general for $a, b \in \mathbb{Z}\cup\{-\infty,\infty\}$, $a < b$, $x_a^b := x_b, x_{b-1}, ..., x_a$. Thus, $(X_t)_{t\in\mathbb{Z}}$ is specified by

$$P(x_1|x_{-p+1}^0)\,,\ \text{for } x_1 \in \mathcal{X} \text{ and } x_{-p+1}^0 \in \mathcal{X}^p.$$

Without loss of generality we concentrate on the random variable X_1, since by stationarity, the transition probabilities are time-homogeneous. The random variable X_1 might not necessarily be influenced by its full history x_{-p+1}^0. Therefore, it is important to distinguish between relevant and irrelevant states in the infinite past and then lump irrelevant states together yielding a possibly parsimonious Markov chain. Formalizing this idea leads to the concept of *variable length Markov chains.*

Definition 1 *Let $\mathcal{X}$ be a finite categorical space and $(X_t)_{t\in\mathbb{Z}}$ an $\mathcal{X}$-valued stationary process.*

(i) The projection function

$$c : \mathcal{X}^\infty \longrightarrow \bigcup_{i=0}^{\infty}\mathcal{X}^i \quad (\mathcal{X}^0 = \emptyset)\,, \quad x_{-\infty}^0 \longmapsto c(x_{-\infty}^0) = x_{-\ell+1}^0\,,$$

where

$$\ell = \ell(x_{-\infty}^0) := \min\{p : P(x_1|x_{-\infty}^0) = P(x_1|x_{-p+1}^0)\,, \forall x_1 \in \mathcal{X}\}\,,$$

is called the context function of the process $(X_t)_{t\in\mathbb{Z}}$.

(ii) The elements of the set $\{c(x_{-\infty}^0) : x_{-\infty}^0 \in \mathcal{X}^\infty\}$ *are called contexts of the process* $(X_t)_{t\in\mathbb{Z}}$.

The name *context* derives from the fact, that now the random variable X_1 does no more depend on the full history x_{-p+1}^0, as in the case of a Markov chain of order p, but only on some pieces of variable length $\ell(\cdot)$ from the infinite past $x_{-\infty}^0$.
From Definition 1 we see that the context length $\ell(\cdot)$ and the context function $c(\cdot)$ are equivalent, because $c(\cdot)$ is a projection function and $\ell(x_{-\infty}^0) = \left|c(x_{-\infty}^0)\right|$, $\forall x_{-\infty}^0 \in \mathcal{X}^\infty$.

Definition 2 *Let* $\mathcal{X}$ *be a finite categorical space and* $(X_t)_{t\in\mathbb{Z}}$ *an* $\mathcal{X}$*-valued stationary process with context function* $c(\cdot)$*. The smallest integer* d*, such that*

$$\left|c(x_{-\infty}^0)\right| = \ell(x_{-\infty}^0) \leq d\,, \quad \forall x_{-\infty}^0 \in \mathcal{X}^\infty,$$

is called the order of the context function. If $d < \infty$*, then* $(X_t)_{t\in\mathbb{Z}}$ *is called stationary variable length Markov chain (VLMC) of order* d*.*

Obviously, a VLMC of order d can be embedded in a Markov chain of order d, however with a memory of variable length $\ell(\cdot) \leq d$. The case $\ell(\cdot) \equiv 0$ coincides with an independent, stationary process. If $c(x_{-\infty}^0) = x_{-d+1}^0$, $\forall x_{-\infty}^0 \in \mathcal{X}^\infty$, then $(X_t)_{t\in\mathbb{Z}}$ is a full Markov chain of order d. Since there is a large variety of context functions of order d with different structures (particularly of sparse type), VLMC's of order d build a more flexible class of processes than full Markov chains of order d, and they better face the curse of dimensionality.

2.2 Tree Representation

Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary VLMC of order d with context function $c(\cdot)$ and probability distribution P on $\mathcal{X}^\mathbb{Z}$. Because of stationarity, P is completely specified by its transitions probabilities: $P(x_1|c(x_{-\infty}^0))\,,\ x_{-\infty}^1 \in \mathcal{X}^\infty$, which themselves are functions of the values of the context function $c(\cdot)$. The latter are thus the minimal state space of the process $(X_t)_{t\in\mathbb{Z}}$.
For better insight of a VLMC, it is convenient to adopt a tree representation for $c(\cdot)$. This will also be useful later, when fitting a VLMC to general stationary processes.
A tree is a directed graph composed by nodes and edges. We consider for our purposes trees, which grow downwards. The root, i.e. the node on top, is connected to any other node by means of exactly one branch (or path). From every internal node there originate at most $|\mathcal{X}|$ edges. The branches connecting the root with the final nodes represent the values of the context function $c(\cdot)$. The following example should clarify our objective to use a tree representation for the context function.

EXAMPLE 1 Let $\mathcal{X} = \{0, 1, 2\}$ and $(X_t)_{t\in\mathbb{Z}}$ be an $\mathcal{X}$-valued VLMC of order 2 with context function $c(\cdot)$ given by

$$c(x_{-\infty}^{0}) = \begin{cases} 00, \text{ if } x_{-1}^{0} = 00\,,\ x_{-\infty}^{2} \text{ arbitrary} \\ 01, \text{ if } x_{-1}^{0} = 01\,,\ x_{-\infty}^{2} \text{ arbitrary} \\ 02, \text{ if } x_{-1}^{0} = 02\,,\ x_{-\infty}^{2} \text{ arbitrary} \\ \ 1, \text{ if } x_0 = 1\,,\ x_{-\infty}^{1} \text{ arbitrary} \\ \ 2, \text{ if } x_0 = 2,\ x_{-1} \in \{1, 2\}\,,\ x_{-\infty}^{2} \text{ arbitrary} \\ 20, \text{ if } x_{-1}^{0} = 20\,,\ x_{-\infty}^{2} \text{ arbitrary.} \end{cases}$$

The minimal state space is represented by the tree in Figure 1. For instance, the branch most on the left stands for the context $c(x_{-\infty}^{0}) = 00$.

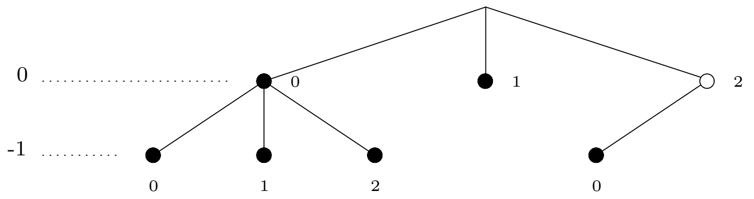

Figure 1: Minimal state space for Example 1.

To represent a VLMC we do not necessarily need a full tree, i.e. a tree with exactly $|\mathcal{X}|$ edges growing down from every internal node (this would correspond to a full Markov chain), but in many cases it suffices to use a sparse tree (which is one of the important advantages of the concept of a VLMC). It is also important to note, that there are two different types of nodes in the tree representation of a VLMC, indicated with black and white (see Example 1), which give rise to the next definition.

DEFINITION 3 *Let $\mathcal{X}$ be a finite categorical space and $(X_t)_{t\in\mathbb{Z}}$ an $\mathcal{X}$-valued stationary variable length Markov chain of order d with context function $c(\cdot)$.*

(i) The tree representation of

$$\tau := \{w : w = c(x_{-\infty}^{0}), x_{-\infty}^{0} \in \mathcal{X}^{\infty}\}$$

is called the ($|\mathcal{X}|$-ary) context tree of the process $(X_t)_{t\in\mathbb{Z}}$.

(ii) The tree representation of

$$\tau^t := \{w : w \in \tau \text{ and } wu \notin \tau, \forall u \in \mathcal{X}\}$$

is called the terminal ($|\mathcal{X}|$-ary) context tree of the process $(X_t)_{t\in\mathbb{Z}}$.

Example 1 (Continued) The context tree of the process $(X_t)_{t\in\mathbb{Z}}$ is given by

$$\tau = \{00,\ 01,\ 02,\ 1,\ 2,\ 20\}\,,$$

which are all nodes of the tree in Figure 1, and the terminal context tree by

$$\tau^t = \{00,\ 01,\ 02,\ 1,\ 20\}\,,$$

which consists of all the black nodes in Figure 1 only.

The context tree τ is the minimal state space of a VLMC with context function $c(\cdot)$. It is clear from Definition 3, that we can reconstruct the context function $c(\cdot)$ from either the context tree τ or the terminal context tree τ^t, and vice versa. The notion of terminal context tree will be useful in Section 3, when formulating an algorithm to estimate the context tree.

3 The Context Algorithm

Let $\mathcal{X}$ be a finite categorical space and $(X_t)_{t\in\mathbb{Z}}$ a stationary process taking values in $\mathcal{X}$. Given realizations $X_1, ..., X_n$, the aim is to find a good estimate of both the underlying context function $c(\cdot)$, which can be of infinite order, and the probability distribution P of $(X_t)_{t\in\mathbb{Z}}$. An adapted version of the Context Algorithm (see Bühlmann and Wyner (1999), Rissanen (1983)) can be used for addressing this problem.
Let $n_w := n - |w| + 1$. We denote by

$$N(w) = \sum_{t=1}^{n_w} 1_{\{X_t^{t+|w|-1}=w\}}\,, \quad w \in \bigcup_{i=1}^{\infty} \mathcal{X}^i \tag{1}$$

the number of occurrences of the string w in the data sequence X_1^n. Let

$$\hat{P}(w) = \frac{N(w)}{n}\,, \quad \hat{P}(u|w) = \frac{N(uw)}{N(w)}\,, \quad w, u \in \bigcup_{i=1}^{\infty} \mathcal{X}^i. \tag{2}$$

Asymptotically $\hat{P}(w)$ possess the same features as the more correct $N(w)/(n - |w| + 1)$, since n_w is of the same order as n. We have opted for $\hat{P}(w)$ for simplicity in the definition of $\hat{P}(u|w)$.
The operation of the Context Algorithm takes place in three steps. Starting from a predetermined initial terminal context tree for the data $X_1, ..., X_n$, we prune its branches, until the past history, represented by the latter, becomes relevant. The condition for pruning (see 3) is based on the Kullback-Leibler distance, which is defined by

$$D(P, Q) := \sum_{x\in\mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)\,,$$

where P, Q are probability measures on the categorical space $\mathcal{X}$.

The Context Algorithm

Step 1

Fit to the data $X_1, ..., X_n$ the terminal context tree $\tau^t_{(0)}$, whose (terminal) nodes have been observed at least twice in X_1^n.

Step 2

Let $wu = x^0_{-\ell+1}$, with $u = x_{-\ell+1}$ and $w = x^0_{-\ell+2}$, be a terminal node of $\tau^t_{(0)}$ (if $\ell = 1$, the pruned version is the empty branch, i.e. the root node). Prune $wu = x^0_{-\ell+1}$ to $w = x^0_{-\ell+2}$, if

$$\Delta_{wu} := D(\hat{P}(\cdot|wu), \hat{P}(\cdot|w))N(wu) < K_n \tag{3}$$

where

$$K_n \sim C\log(n)\,,\; C > 2\,|\mathcal{X}| + 3. \tag{4}$$

Construct in this way the terminal context tree $\tau^t_{(1)}$.

Step 3

Repeat Step 2 with $\tau^t_{(i)}$ instead of $\tau^t_{(i-1)}$ $(i = 1, ...)$ until no more pruning is possible. Denote the final obtained terminal context tree by $\hat{\tau}^t$ and the corresponding context function by $\hat{c}(\cdot)$.

The underlying context function $c(\cdot)$ is hence estimated by means of $\hat{c}(\cdot)$, while the estimate of the transition probability $P(x_1|c(x^0_{-\infty}))$ is given by $\hat{P}(x_1|\hat{c}(x^0_{-\infty}))$.

Remark 1 The initial terminal context tree $\tau^t_{(0)}$ in Step 1 is constructed on the basis of at least two occurrences of every terminal node in the data sequence. This is reasonable in practice. Asymptotic properties of the algorithm remain unchanged, when replacing the number two by any other finite number. The order of testing the terminal nodes of the terminal context tree $\tau^t_{(i)}$ in Step 2 is irrelevant.

Remark 2 The Context Algorithm prunes branches wu to w, for which the estimated transition probabilities $\hat{P}(\cdot|wu)$ are close (in Kullback-Leibler sense) to $\hat{P}(\cdot|w)$. It is also possible to interpret the Context Algorithm as multiple likelihood ratio test, with an acceptance region $[0, \log(n)]$ for the pruned tree (see Bühlmann and Wyner (1999), Remark 3.1).

Remark 3 The L_1-distance $||P - Q||_1$ between P and Q is defined by

$$||P - Q||_1 := \sum_{x \in \mathcal{X}} |P(x) - Q(x)| \ ,$$

and it is well-known that (see Cover and Thomas (1991))

$$D(P, Q) \geq \frac{1}{2} ||P - Q||_1^2.$$

One can show that consistency of the Context Algorithm is still valid when replacing the Kullback-Leibler distance by the squared L_1-distance; note, that the constant in the cut-off value K_n has then to satisfy $C > 4\,|\mathcal{X}| + 6$.

Remark 4 The cut-off value $K_n \sim C \log(n)$, $C > 2\,|\mathcal{X}| + 3$ for the pruning decision in Step 2 is specified by asymptotic considerations (see the proof of Theorem 1). The condition on C comes from Lemma 4. An estimation of C has been given in Bühlmann (1998b). The cut-off value can be interpreted as a stepwise $(1 - \alpha)$-quantile with $\alpha = \alpha_n \longrightarrow 0$ $(n \to \infty)$. The necessity for α_n converging to zero is explained in Rissanen (1989).

Remark 5 This adapted version of the Context Algorithm makes no a-priori restriction on the length of the contexts of the process, such as $\ell(\cdot) = |c(\cdot)| \leq \log(n)/\log(|\mathcal{X}|)$ employed in Weinberger, Rissanen and Feder (1995), which can be a severe restriction in practical applications. However we will see in the next section, that to prove consistency for the estimate of the context function, we assume some milder condition for $|c(\cdot)|$ (see assumption (A3)).

4 Consistency

4.1 Processes with Values in a Finite Categorical Space

Let $\mathcal{X}$ be a finite categorical space and $(X_t)_{t \in \mathbb{Z}}$ a general stationary $\mathcal{X}$-valued process with probability distribution P (defined on $\mathcal{X}^{\mathbb{Z}}$). For such processes the order of the context function $c(\cdot)$ may be infinite, since we do not assume the process $(X_t)_{t \in \mathbb{Z}}$ to be a VLMC and hence to have finite order. To prove consistency for the estimate of the context function $c(\cdot)$ (given by the Context Algorithm) we approximate $c(\cdot)$ by a sequence of context functions $(c_n(\cdot))_{n \in \mathbb{N}}$, corresponding to VLMC's of increasing order and then show, that the event $\{\hat{c}_n(\cdot) = c_n(\cdot)\}$ has asymptotically probability 1. This implies that we approximate in a reasonable sense general stationary processes by VLMC's.

Definition 4 *The truncated context function $c_n(\cdot)$, $n \in \mathbb{N}$, is defined by*

$$c_n(x_{-\infty}^0) := \begin{cases} x_{-d_n+1}^0\,, & \text{if } \left|c(x_{-\infty}^0)\right| \geq d_n\,,\ (d_n)_{n\in\mathbb{N}} \text{ an increasing sequence,} \\ c(x_{-\infty}^0)\,, & \text{otherwise.} \end{cases}$$

Therefore, the branches of the context tree τ (corresponding to $c(\cdot)$), which are too long, particularly longer than d_n, are cut off. This allows us to define a finite context tree τ_n, $n \in \mathbb{N}$, by means of the context function $c_n(\cdot)$. With the Context Algorithm we then estimate the truncated context function $c_n(\cdot)$ by $\hat{c}_n(\cdot)$ (resp. the context tree $\tau_n(\cdot)$ by $\hat{\tau}_n(\cdot)$) and consequently the probability distribution P by $\hat{P}_{\hat{c}_n}$, being an estimated VLMC with context function $\hat{c}_n(\cdot)$.

We make the following assumptions:

(A1) $(X_t)_{t\in\mathbb{Z}} \sim P$ is geometrically α-mixing with α-mixing coefficients $(\alpha(i))_{i\in\mathbb{N}}$ satisfying

$$\alpha(i) \leq C_\alpha \nu^i\,, \text{ for some constants } C_\alpha > 0 \text{ and } \nu \in (0,1).$$

(A2) For some $\gamma \in (0,1)$, some $\sigma \in (0,1)$ and some $\theta > 0$, for all n sufficiently large,

$$\Gamma_n := \min_{w\in\tau_n^t} P(w) \geq \frac{1}{n^\gamma}\,,$$

$$\Upsilon_n := \min_{wu\in\tau_n^t, u\in\mathcal{X}} \|P(\cdot|wu) - P(\cdot|w)\|_1 \geq \Big(\frac{\log(n)^{1+\theta}}{(n\Gamma_n^{(1-\sigma)/2})^{1-\sigma}}\Big)^{1/2}.$$

(A3) The order d_n of the context tree τ_n satisfies, for all n sufficiently large,

$$d_n \leq n^\delta\,, \text{ for some } \delta \in (0,1)\,,$$

such that

$$[(n_w)^\sigma] - d_n \geq n^\lambda\,, \text{ for some } \lambda > 0\,, \text{ and all } w \in \tau_n\,,$$

where $n_w := n - |w| + 1$ and $[x] := \max\{\ell \in \mathbb{N} : \ell \leq x\}$.

(A4) For the minimal transition probabilities, for all n sufficiently large,

$$\min_{x\in\mathcal{X}, w\in\tau_n} P(x|w) \geq \frac{1}{n}\,.$$

The assumptions (A2)-(A4) are all probabilistic conditions about the sparseness of the terminal context tree τ_n^t.

Remark 6 Because of the first condition in assumption (A2), the cardinality of the terminal context tree τ_n^t is bounded by

$$\left|\tau_n^t\right| \leq \frac{1}{\Gamma_n} \leq n^\gamma.$$

The second condition in assumption (A2) is measuring relevance of terminal nodes in comparison with their ancestors.

Remark 7 Assumption (A3) is about the maximal growth rate for the approximating order of the context tree. Consider a full Markov chain of order d_n. Then, $|\tau_n| = |\mathcal{X}|^{d_n}$, which by assumption (A2) is required to be smaller than n^γ. Thus, the assumption becomes

$$d_n \leq \frac{\gamma}{\log(|\mathcal{X}|)} \log(n).$$

Hence, the choice of δ in the interval $(0,1)$ from (A3) is without restrictions. With VLMC's we can also treat models with a memory growing only in certain directions with d_n of polynomial order less than 1. This is a big advantage of VLMC's in comparison with full Markov chains.

The power of the Context Algorithm is shown in the next two theorems, which state, that the estimate of the minimal state space and of the probability distribution of general stationary processes is asymptotically correct. According to Theorem 1, the Context Algorithm selects asymptotically the right final-dimensional model components and increases model complexity for infinite dimensional components. This cannot be achieved by more traditional selection criterion such as AIC or BIC due to the extremely large number of possible submodels.

Theorem 1 *Under the assumptions (A1)-(A4),*

$$\mathbb{P}[\hat{\tau}_n = \tau_n] \longrightarrow 1 \quad (n \to \infty).$$

Since knowledge of the context function $c_n(\cdot)$ (or of the terminal context tree τ_n^t) is equivalent to knowing the context tree τ_n, the assertion of Theorem 1 can be restated as

$$\mathbb{P}[\hat{c}_n(\cdot) = c_n(\cdot)] \longrightarrow 1 \quad (\text{or } \mathbb{P}[\hat{\tau}_n^t = \tau_n^t] \longrightarrow 1) \quad (n \to \infty).$$

Theorem 2 *Under the assumptions (A1)-(A4),*

(i) $\displaystyle\sup_{x \in \mathcal{X}, w \in \tau_n} \left|\hat{P}_{\hat{c}_n}(x|w) - P(x|w)\right| = o_P(1)$,

(ii) $\hat{P}_{\hat{c}_n}(x_1^r) \xrightarrow{\mathbb{P}} P(x_1^r) \quad (n \to \infty)$, $\forall\, x_1^r \in \mathcal{X}^r$, $\forall\, r \in \mathbb{N}$.

Example 2 We consider the threshold-AR(1) process $(Y_t)_{t\in\mathbb{Z}}$ defined by

$$\begin{aligned} Y_t &= m(Y_{t-1}) + \varepsilon_t \\ m(x) &= \phi x 1_{\{x>0\}} , \end{aligned}$$

where $\phi \in \mathbb{R}$ with $|\phi| < 1$, $1_{\{\cdot\}}$ the indicator function and $\varepsilon_t \, iid \sim F$ with F having a density with respect to Lebesgue measure. The process $(Y_t)_{t\in\mathbb{Z}}$ is stationary and α-mixing (see Doukhan (1994)). We then construct the process $(X_t)_{t\in\mathbb{Z}}$, defined on $\mathcal{X} = \{0, 1\}$ by

$$X_t := 1_{\{Y_t>0\}}.$$

The context function of $(X_t)_{t\in\mathbb{Z}}$ is given by

$$c(x^0_{-\infty}) = x^0_{-h} ,$$

where $h = \min\{k : x_{-k} = 0 \text{ and } x^0_{-k+1} = 1 \cdots 1\}$ is depending on $x^0_{-\infty}$. Thus, the context tree τ of $(X_t)_{t\in\mathbb{Z}}$ grows up to the infinity as shown in Figure 2.
To support empirically the consistency of the estimate of the context tree τ we have simulated six series of data with $\varepsilon_t \, iid \sim \mathcal{N}(0, 1)$ and $\phi = 0.85$, once with n=1000 and once with n=10000. Then from each of the six series we have opted for the two most representative. The choice of the cut-off value K in the Context Algorithm is purely subjective: for n=1000 we used $\chi^2_{1;0.965}/2$ and for n=10000 $\chi^2_{1;0.995}/2$.
Whereas with 1000 data the estimates of τ show still small deviations from the right structure (see Figure 3), those with 10000 data are almost perfect (see Figure 4).

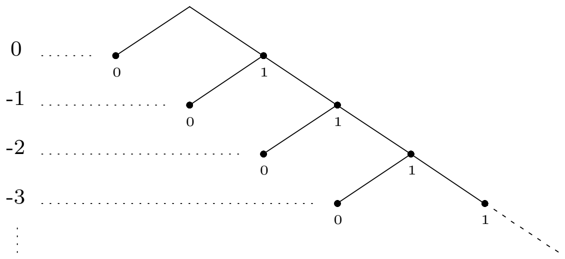

Figure 2: Context tree for Example 2.

4.2 Processes with Values in an Increasing Size Categorical Space

We now consider the case of an increasing size categorical space and without loss of generality denote it by $\mathcal{X}_n = \{0, 1, ..., \mathcal{M}_n - 1\}$, $n \in \mathbb{N}$. To estimate the minimal state space and the probability distribution of general stationary processes with values in $\mathcal{X}_n$, we can make use of the same ideas developed in Section 3. For the cut-off constant $C = C_n$ of the Context Algorithm we have now $C > 2\,|\mathcal{X}_n| + 3$ which increases with at least the same order as $|\mathcal{X}_n|$.

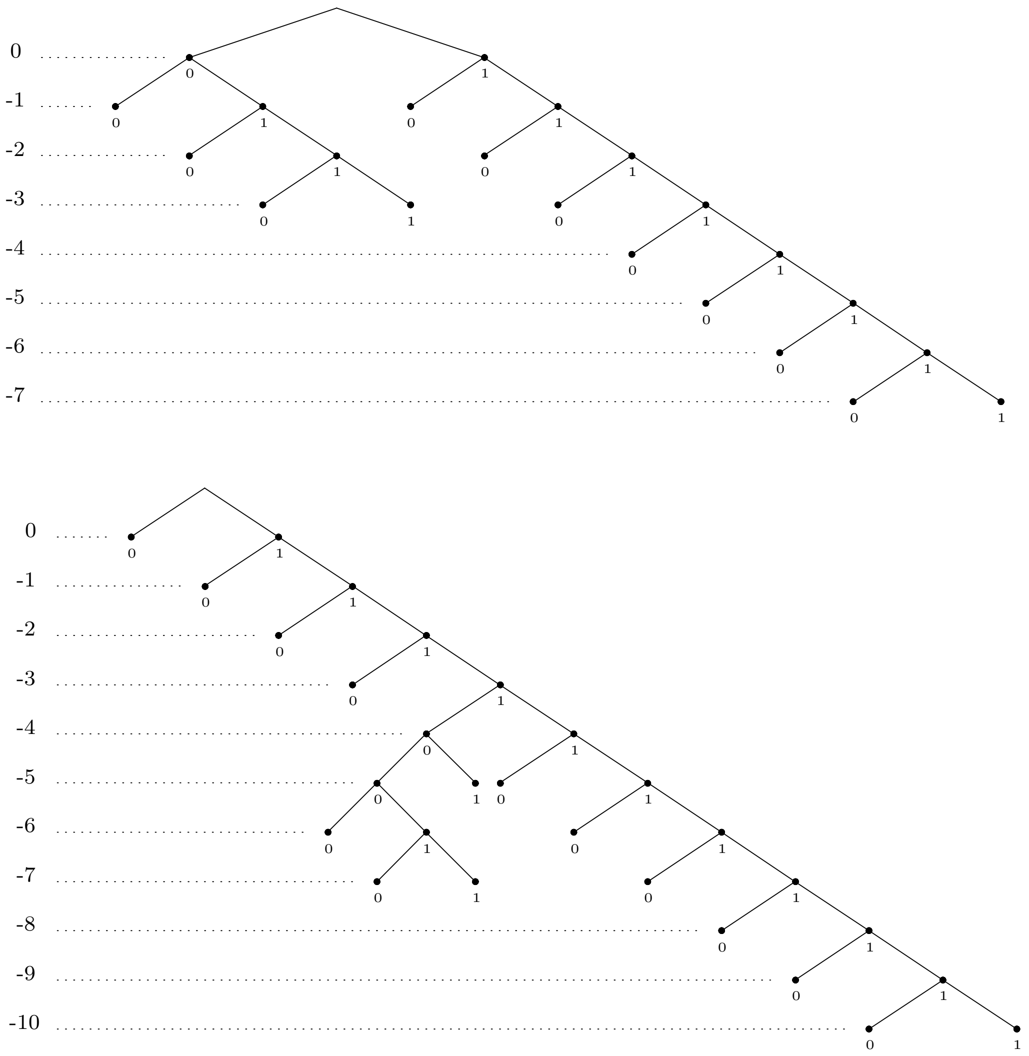

Figure 3: Estimates of the context tree for Example 2 with 1000 data.

We make the further assumption:

(A5) The cardinality of $\mathcal{X}_n$ satisfies for all n sufficiently large

$$|\mathcal{X}_n| \leq \log(n)^{1+\mu}\ ,\ \text{for some } \mu > 0. \tag{5}$$

Corollary 1 *Under the assumptions (A1)-(A5),*

$$\mathbb{P}[\hat{\tau}_n = \tau_n] \longrightarrow 1 \quad (n \to \infty).$$

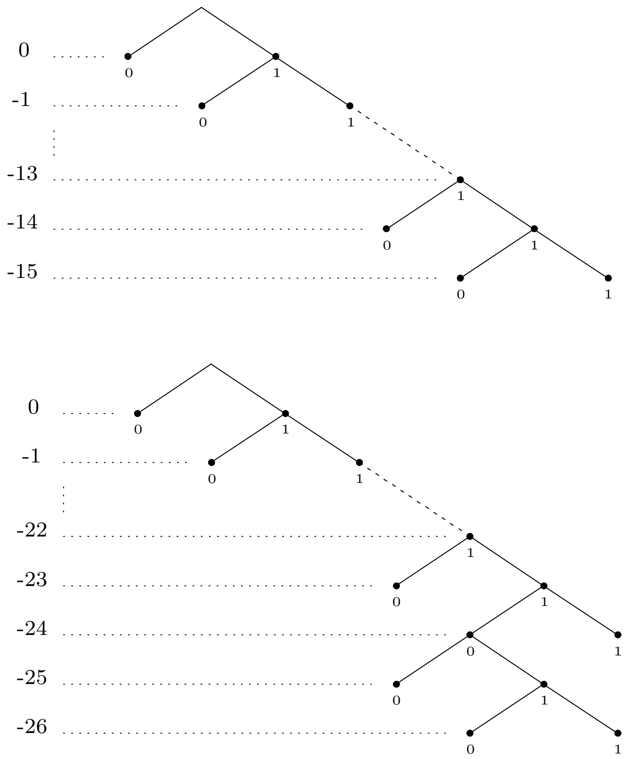

Figure 4: Estimates of the context tree for Example 2 with 10000 data.

Corollary 2 *Under the assumptions (A1)-(A5),*

(i) $\sup_{x\in\mathcal{X}_n, w\in\tau_n} \left|\hat{P}_{\hat{c}_n}(x|w) - P(x|w)\right| = o_P(1)$,

(ii) $\hat{P}_{\hat{c}_n}(x_1^r) \xrightarrow{\mathbb{P}} P(x_1^r) \quad (n \to \infty)\,, \; \forall\, x_1^r \in (\mathcal{X}_n)^r\,, \; \forall\, r \in \mathbb{N}.$

An important application of the above corollaries occurs when quantizing real-valued processes. Let $(Y_t)_{t\in\mathbb{Z}}$ be a process with values in $\mathbb{R}$. We consider quantizers

$$\begin{array}{rccl} Q_n: & \mathbb{R} & \longrightarrow & \mathcal{X}_n = \{0, 1, ..., \mathcal{M}_n - 1\} \\ & y \in \mathrm{I}_{x,n} & \longmapsto & Q_n(y) = x \end{array}$$

where $(\mathrm{I}_{x,n})_{x\in\mathcal{X}_n}$ is a partition of $\mathbb{R}$ into disjoint sets, i.e.

$$\dot{\bigcup_{x\in\mathcal{X}_n}} \mathrm{I}_{x,n} = \mathbb{R} \quad \text{and} \quad \mathrm{I}_{x_1,n} \cap \mathrm{I}_{x_2,n} = \emptyset \text{ for } x_1 \neq x_2.$$

The process $(X_t)_{t\in\mathbb{Z}} := (Q_n(Y_t))_{t\in\mathbb{Z}}$ is then a process with values in $\mathcal{X}_n$.

A specific example of a quantizer is as follows. Assume $(Y_t)_{t\in\mathbb{Z}}$ to be an $\mathbb{R}$-valued stationary, geometrically α-mixing process with mixing coefficients $(\alpha_Y(i))_{i\in\mathbb{N}}$ and one-dimensional continuous cumulative distribution function F. The assumption on the distribution function F is only made to avoid discontinuity problems. We define the quantized process $(X_t)_{t\in\mathbb{Z}}$ on $\mathcal{X}_n = \{0, 1, ..., \mathcal{M}_n - 1\}$ (and hence the quantizer Q_n) by

$$X_t := \begin{cases} 0\,, & -\infty < Y_t \le F^{-1}(\frac{1}{\mathcal{M}_n}) \\ x\,, & F^{-1}(\frac{x}{\mathcal{M}_n}) < Y_t \le F^{-1}(\frac{x+1}{\mathcal{M}_n})\,,\ x = 1, ..., \mathcal{M}_n - 2 \\ \mathcal{M}_n - 1\,, & F^{-1}(\frac{\mathcal{M}_n - 1}{\mathcal{M}_n}) < Y_t < \infty. \end{cases} \tag{6}$$

The process $(X_t)_{t\in\mathbb{Z}}$ is hence $\mathcal{X}_n$-valued, stationary and geometrically α-mixing with coefficients $(\alpha_X(i))_{i\in\mathbb{N}}$ bounded by those of the process $(Y_t)_{t\in\mathbb{Z}}$. By means of the Context Algorithm we estimate the probability distribution P_n of $(X_t)_{t\in\mathbb{Z}}$ on $(\mathcal{X}_n)^{\mathbb{Z}}$ by $\hat{P}_{n,\hat{c}_n}$. We define an estimate of F by

$$\hat{F}_n(y) := \sum_{Q_n(y)\le y} \hat{P}_{n,\hat{c}_n} \circ Q_n(y)\,,\ y \in \mathbb{R}$$

and those of the finite-dimensional distributions $F^{(r)}$, $r \in \mathbb{N}$, of $(Y_t)_{t\in\mathbb{Z}}$ by

$$\hat{F}_n^{(r)}(y_1^r) := \sum_{Q_n(y_1^r)\le y_1^r} \hat{P}_{n,\hat{c}_n} \circ Q_n(y_1^r)\,,\ y \in \mathbb{R}$$
$$Q_n(y_1^r) := (Q_n(y_1), ..., Q_n(y_r))\,,\ r \in \mathbb{N}.$$

where the summation range is (defined) componentwise.

Corollary 3 *Under the assumptions (A1)-(A5),*

$$\hat{F}_n^{(r)}(y_1^r) \xrightarrow{\mathbb{P}} F^{(r)}(y_1^r) \quad (n \to \infty)\,,\ \forall\, y_1^r \in \mathbb{R}^r\,,\ \forall\, r \in \mathbb{N}.$$

The quantizer in (6), but replacing F by the empirical one-dimensional distribution $\hat{F}$ and approximating the quantized process by a VLMC, has been successfully applied in practical problems (see Bühlmann (1998a)). Our Corollaries 1-3 justify this procedure on a theoretical basis.

5 Proofs

To prove Theorem 1 we make use of some ideas developed in Bühlmann and Wyner (1999) and apply an exponential inequality for α-mixing processes.

Proof of Theorem 1 The error event $E_n = \{\hat{\tau}_n \neq \tau_n\}$ for sample size n for the context tree τ_n can be decomposed into the disjoint union of the under- and the overestimation events U_n and O_n, where

$$U_n = \{\exists\ w \in \hat{\tau}_n \text{ with } wu \in \tau_n \text{ and } wu \notin \hat{\tau}_n,\ \text{ for some } u \in \bigcup_{i=1}^{\infty} \mathcal{X}^i\}$$
$$O_n = \{\exists\ w \in \tau_n \text{ with } wu \in \hat{\tau}_n \text{ and } wu \notin \tau_n,\ \text{ for some } u \in \bigcup_{i=1}^{\infty} \mathcal{X}^i\}$$

Therefore, we can bound the error of estimating the underlying context tree by separately treating the under- and the overestimation. Let us first bound the underestimation event U_n.

Lemma 1 *Under the assumptions (A1)-(A3),*

$$\mathbb{P}[U_n] = \mathcal{O}(\exp(-D\log(n)^{1+\theta})) \tag{7}$$

for some constant $D > 0$ and θ as in assumption (A2).

Proof We define a sequence $(\rho_n)_{n\in\mathbb{N}}$ by $\rho_n := n\Gamma_n^{(1-\sigma)/2}$ and then using the event

$$H_n = \{N(w) \geq \rho_n \text{ for every } w \in \tau_n^t\}$$

we partition U_n. It follows, that

$$\mathbb{P}[U_n] \leq \mathbb{P}[U_n \cap H_n] + \mathbb{P}[H_n^c].$$

To bound $\mathbb{P}[U_n \cap H_n]$ we apply the same techniques used in Bühlmann and Wyner (1999) and reformulate for our case Lemma 5.1 and Lemma 5.2. First, from Bühlmann and Wyner (1999) (see (5.1)-(5.4)) we have

$$\begin{aligned}
&\mathbb{P}\big[U_n \cap H_n\big] \\
&\quad\leq \sum_{wu\in\tau_n^t, u\in\mathcal{X}} \sum_{k=\rho_n}^{n_{wu}} \sum_{j=k}^{n_w} \mathbb{P}\big[D(\hat{P}(\cdot|wu)||\hat{P}(\cdot|w)) < C\log(n)/k, N(wu) = k, N(w) = j\big] \\
&\quad\leq |\mathcal{X}| \Bigg(\sum_{wu\in\tau_n^t, u\in\mathcal{X}} \sum_{k=\rho_n}^{n_{wu}} \sum_{j=k}^{n_w} \Big(\sup_{x\in\mathcal{X}} \mathbb{P}\Big[\Big|\hat{P}(x|wu) - P(x|wu)\Big|^2 \geq a_n(k), N(wu) = k\Big] \\
&\qquad\qquad + \sup_{x\in\mathcal{X}} \mathbb{P}\Big[\Big|\hat{P}(x|w) - P(x|w)\Big|^2 \geq a_n(k), N(w) = j\Big]\Big)\Bigg) \\
&\quad\leq |\mathcal{X}|\,\big|\tau_n^t\big|\, n^2 \Big(\sup_{x\in\mathcal{X}} \mathbb{P}\Big[\Big|\hat{P}(x|wu) - P(x|wu)\Big|^2 \geq a_n(k), N(wu) = k\Big] \\
&\qquad\qquad + \sup_{x\in\mathcal{X}} \mathbb{P}\Big[\Big|\hat{P}(x|w) - P(x|w)\Big|^2 \geq a_n(k), N(w) = j\Big]\Big)
\end{aligned} \tag{8}$$

where

$$a_n(k) = \Big(\frac{\Upsilon_n}{2} - \sqrt{\frac{C\log n}{k}}\Big)^2. \tag{9}$$

Note that for n sufficiently large

$$\min_{k\geq\rho_n} a_n(k) \geq \frac{\log(n)^{1+\theta}}{\rho_n^{1-\sigma}} \ . \tag{10}$$

We treat the two last summands in (8) in the same manner denoting with v either wu or w. Let $p = P(x|v)$ and $\hat{p} = \hat{P}(x|v)$. In order to find an upper probabilistic bound for the event

$$\{|\hat{p} - p|^2 \geq a_n(k), N(v) = k\}$$

consider the extension of $X_1, ..., X_n$ to the infinite sequence $(X_t)_{t\in\mathbb{N}}$ and define $I_i(v)$ as the time of the i^{th} occurrence of v in $(X_t)_{t\in\mathbb{N}}$. Then let $Z_i = X_{I_i+1}$ be the symbol that occurs after the i^{th} occurrence of v in $(X_t)_{t\in\mathbb{N}}$. The stochastic process $(Z_i)_{i\in\mathbb{N}}$ is stationary and α-mixing with mixing coefficients bounded by the same bound as the α-mixing coefficients of $(X_t)_{t\in\mathbb{N}}$. Define $Y_i = 1_{\{Z_i=x\}}$ and observe, that

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

and consequently

$$\mathbb{P}[|\hat{p} - p|^2 > a_n(k), N(v) = k] \leq \mathbb{P}\Big[\Big|\sum_{i=1}^{k} \frac{Y_i}{k} - p\Big|^2 > a_n(k)\Big]. \tag{11}$$

Lemma 2 *Let $(Y_i)_{i\in\mathbb{N}}$ with $\mathbb{E}[Y_i] = p$ be the above defined process and $a_n(k)$ be as in (9). Then under the assumptions (A1)-(A2), for $k \geq \rho_n$ and for all n sufficiently large*

$$\sup_{0<p<1} \mathbb{P}\Big[\Big|\sum_{i=1}^{k} \frac{Y_i}{k} - p\Big|^2 > a_n(k)\Big] \leq 4\exp\big(-\frac{1}{16}\log(n)^{1+\theta}\big) + \frac{11\sqrt{5}C_\alpha}{\nu^{\tilde{\sigma}}} n^{\frac{(5-\sigma)(1-\sigma)}{4}} \nu^{\tilde{\sigma}n} \ ,$$

for $\tilde{\sigma} = \sigma(1-\sigma)(1-\frac{\gamma}{2}(1-\sigma))$, C_α as in assumption (A1) and σ, θ as in assumption (A2).

Proof The process $(X_t)_{t\in\mathbb{Z}}$ has α-mixing coefficients $\alpha(j) \leq C_\alpha \nu^j$, $\nu \in (0,1)$, and the same bound applies also for the α-mixing coefficients of the process $(Y_i)_{i\in\mathbb{N}}$. Since $(Y_i)_{i\in\mathbb{N}}$ is a zero-mean real-valued process with $|Y_i| \leq 1$, for all $i \in \mathbb{N}$, we get from Theorem 1.3, Chapter 1.4 in Bosq (1996)

$$\begin{aligned}\sup_{0<p<1} \mathbb{P}\Big[\Big|\sum_{i=1}^{k} \frac{Y_i}{k} - p\Big|^2 > a_n(k)\Big] &= \sup_{0<p<1} \mathbb{P}\Big[\Big|\sum_{i=1}^{k}(Y_i - p)\Big| > k\sqrt{a_n(k)}\Big] \\ &\leq 4\exp\Big(-\frac{1}{8}q a_n(k)\Big) + 22\Big(1 + \frac{4}{\sqrt{a_n(k)}}\Big)^{1/2} q\alpha([k/2q]). \end{aligned} \tag{12}$$

From inequality (10) by choosing $q := [k^{1-\sigma}/2]$ we obtain for $k \geq \rho_n$ and for all n sufficiently large

$$qa_n(k) \geq \frac{1}{2}\rho_n^{1-\sigma}\frac{\log(n)^{1+\theta}}{\rho_n^{1-\sigma}} = \frac{1}{2}\log(n)^{1+\theta}.$$

For the second summand in the inequality (12) we have by (10) and by $\rho_n \leq n^{1-\sigma}$

$$\Big(1 + \frac{4}{\sqrt{a_n(k)}}\Big) \leq 1 + 4\rho_n^{\frac{1-\sigma}{2}} \leq 5n^{\frac{(1-\sigma)^2}{2}}.$$

Since $q = [k^{1-\sigma}/2] \leq k^{1-\sigma}/2 \leq n^{1-\sigma}/2$ and $\rho_n = (n\Gamma_n^{(1-\sigma)/2})^{1-\sigma} \geq n^{(1-\sigma)(1-\frac{\gamma}{2}(1-\sigma))}$ we get

$$\begin{aligned} q \cdot \alpha([k/2q]) &\leq \frac{1}{2}n^{1-\sigma}\alpha([\rho_n^{\sigma}]) \leq \frac{1}{2}n^{1-\sigma}\alpha([n^{(1-\sigma)(1-\frac{\gamma}{2}(1-\sigma))\sigma}]) \\ &\leq \frac{C_\alpha}{2\nu^{(1-\sigma)(1-\frac{\gamma}{2}(1-\sigma))\sigma}}n^{1-\sigma}\nu^{\sigma(1-\sigma)(1-\frac{\gamma}{2}(1-\sigma))n}. \end{aligned}$$

The assertion of the lemma follows then immediately. □

A direct application of Lemma 2 to the above inequality (8) proves, that for $k, j \geq \rho_n$ and for all n sufficiently large

$$\begin{aligned} \mathbb{P}[U_n \cap H_n] &\leq 2\,|\mathcal{X}|\,n^{2+\gamma}(4\exp(-\frac{1}{16}\log(n)^{1+\theta}) + \frac{11\sqrt{5}C_\alpha}{\nu^{\tilde{\sigma}}}n^{\frac{(5-\sigma)(1-\sigma)}{4}}\nu^{\tilde{\sigma}n}) \\ &= \mathcal{O}(\exp(-D_1\log(n)^{1+\theta}))\ ,\ \text{for some constant } D_1 > 0. \end{aligned}$$

The next step is to find a bound for $\mathbb{P}[H_n^c]$. First of all note, that since

$$\mathbb{E}[N(w)] = \sum_{t=1}^{n_w} P(w) \geq n_w\Gamma_n \geq n_w\Gamma_n^{(1-\sigma)/2} \tag{13}$$

and $\rho_n < \frac{1}{2}n_w\Gamma_n^{(1-\sigma)/2}$, for all n sufficiently large, we have

$$\begin{aligned} \mathbb{P}[H_n^c] &\leq \sum_{w\in\tau_n^t}\mathbb{P}[N(w) < \rho_n] = \sum_{w\in\tau_n^t}\mathbb{P}[N(w) - \mathbb{E}[N(w)] < \rho_n - \mathbb{E}[N(w)]] \\ &\leq \sum_{w\in\tau_n^t}\mathbb{P}[N(w) - \mathbb{E}[N(w)] < -\frac{1}{2}n_w\Gamma_n^{(1-\sigma)/2}] \\ &\leq \sum_{w\in\tau_n^t}\mathbb{P}[|N(w) - \mathbb{E}[N(w)]| > \frac{1}{2}n_w\Gamma_n^{(1-\sigma)/2}] \\ &\leq \left|\tau_n^t\right| \sup_{w\in\tau_n^t}\mathbb{P}[|N(w) - \mathbb{E}[N(w)]| > \frac{1}{2}n_w\Gamma_n^{(1-\sigma)/2}] \qquad (14) \end{aligned}$$

Lemma 3 *Under the assumptions (A1)-(A3), for all n sufficiently large*

$$\sup_{w\in\tau_n}\mathbb{P}\Big[\,|N(w)-\mathbb{E}[N(w)]|>\frac{1}{2}n_w\Gamma_n^{(1-\sigma)/2}\Big]\le 4\exp(-\frac{1}{128}n^{(1-\sigma)(1-\gamma)})+33C_\alpha\nu n^{(1-\sigma)(1+\frac{\gamma}{4})}\nu^{n^\lambda}\ ,$$

for C_α as in assumption (A1), σ,γ as in assumption (A2) and λ as in assumption (A3).

Proof For $t\le n_w$ and $w\in\tau_n$ we define $W_t:=1_{\{X_t^{t+|w|-1}=w\}}-P_n(w)$. The process $(W_t)_{t\in\mathbb{Z}}$ has mean zero with $|W_t|\le 1$, for all $t\in\mathbb{Z}$. We have

$$N(w)-\mathbb{E}[N(w)]=\sum_{t=1}^{n_w}W_t.$$

Note that for the α-mixing coefficients $(\alpha_W(i))_{i\in\mathbb{N}}$ of $(W_t)_{t\in\mathbb{Z}}$ we obtain

$$\alpha_W(i)\le\begin{cases}\alpha(i-|w|+1)\ ,\ \text{if } i\ge|w|\\ \\ 1\ ,\ \text{if } i<|w|\end{cases}\tag{15}$$

where $(\alpha(i))_{i\in\mathbb{Z}}$ are the α-mixing coefficients of $(X_t)_{t\in\mathbb{Z}}$. From Theorem 1.3, Chapter 1.4 in Bosq (1996) we get for $q:=[(n_w)^{1-\sigma}/2]$, σ as in assumption (A2) and $w\in\tau_n$

$$\begin{aligned}&\mathbb{P}\Big[\,|N(w)-\mathbb{E}[N(w)]|>\frac{1}{2}n_w\Gamma_n^{(1-\sigma)/2}\Big]=\mathbb{P}\Big[\Big|\sum_{t=1}^{n_w}W_t\Big|>\frac{1}{2}n_w\Gamma_n^{(1-\sigma)/2}\Big]\\&\le 4\exp\big(-\frac{1}{32}q\Gamma_n^{1-\sigma}\big)+22\big(1+8\Gamma_n^{-(1-\sigma)/2}\big)^{1/2}q\alpha_W\big([(n_w)/2q]\big).\end{aligned}$$

Because of $q\ge(n_w-1)^{1-\sigma}/2$ and assumption (A2), we have for all n sufficiently large

$$q\Gamma_n^{1-\sigma}\ge\frac{1}{2}\frac{(n_w-1)^{1-\sigma}}{n^{\gamma(1-\sigma)}}=\frac{1}{2}\big(\frac{n_w-1}{n}\big)^{1-\sigma}n^{(1-\gamma)(1-\sigma)}\ge\frac{1}{4}n^{(1-\gamma)(1-\sigma)}\tag{16}$$

and

$$\big(1+8\Gamma_n^{-(1-\sigma)/2}\big)^{1/2}\le\big(9\Gamma_n^{-(1-\sigma)/2}\big)^{1/2}\le 3n^{\frac{(1-\sigma)\gamma}{4}}.\tag{17}$$

For the other part of the second summand in inequality (16), because of $q\le n^{1-\sigma}/2$, assumption (A3) and (15),

$$\begin{aligned}q\alpha_W([(n_w)/2q])\ &\le\ \frac{1}{2}n^{1-\sigma}\alpha_W([(n_w)^\sigma])\le\frac{1}{2}n^{1-\sigma}\alpha([(n_w)^\sigma]-|w|+1)\\&\le\ \frac{C_\alpha\nu}{2}n^{1-\sigma}\nu^{([(n_w)^\sigma]-|w|)}\le\frac{C_\alpha\nu}{2}n^{1-\sigma}\nu^{n^\lambda}.\end{aligned}\tag{18}$$

Since the upper bounds of the inequalities (16), (17) and (18) do not depend on w, the assertion of the lemma then follows immediately. □

From inequality (14) we obtain

$$\mathbb{P}[H_n^c] \leq n^{\gamma}(4\exp(-\frac{1}{128}n^{(1-\sigma)(1-\gamma)}) + 33C_{\alpha}\nu n^{(1-\sigma)(1+\frac{\gamma}{4})}\nu^{n^{\lambda}}) = \mathcal{O}(\exp(-D_2 n^{\xi}))\ , \quad (19)$$

for some constants $D_2 > 0$ and $0 < \xi < \min(\lambda, (1-\sigma)(1-\gamma))$.

This completes the proof of Lemma 1. □

In order to find a bound for the overestimation event O_n, we use the same method as in Bühlmann and Wyner (1999).

Lemma 4 *Under the assumptions (A1)-(A4), for all n sufficiently large*

$$\mathbb{P}[O_n] \leq |\mathcal{X}| \cdot n^{-r}\ , \quad (20)$$

where $r := C - 2\,|\mathcal{X}| - 3 > 0$ and C is the cut-off constant of the Context Algorithm.

Proof We define the event $O_n(swv)$ for $s \in \tau_n$, $w \in \bigcup_{i=1}^{\infty} \mathcal{X}^i$ and $v \in \mathcal{X}$ with $swv \notin \tau_n$ by

$$O_n(swv) = \{\Delta_{swv} \geq C\log(n),\ N(swv) \geq 2\}.$$

Then from Lemma 5.3 in Bühlmann and Wyner (1999), using also assumption (A4) we have

$$\mathbb{P}[O_n(swv)] \leq n^{-C+2|\mathcal{X}|+1}\mathbb{P}[sw \in \hat{\tau}_{(0)}].$$

Now let L denote the number of sequences occurring at least twice in the sequence $X_1, ..., X_n$. Since

$$\begin{aligned} \mathbb{P}[O_n] &\leq \sum_{swv}\mathbb{P}[O_n(swv)] \leq n^{-C+2|\mathcal{X}|+1} \cdot \sum_{swv}\mathbb{P}[sw \in \hat{\tau}_{(0)}] \\ &\leq n^{-C+2|\mathcal{X}|+1}\,|\mathcal{X}|\,\mathbb{E}[L] \leq |\mathcal{X}| \cdot n^{-C+2|\mathcal{X}|+3} \end{aligned}$$

and $C > 2\,|\mathcal{X}| + 3$ (see (4)), the assertion of the lemma follows then immediately. □

By Lemma 1 and Lemma 4, we complete the proof of Theorem 1. □

Proof of Theorem 2 (i) For $\varepsilon > 0$, let $G_n := \{\left|\hat{P}_{\hat{c}_n}(x|w) - P(x|w)\right| > \varepsilon\}$. Then, by means of the event $\tilde{H}_n = \{N(w) \geq \rho_n = n\Gamma_n^{(1-\sigma)/2}$ for every $w \in \tau_n\}$, we partition G_n and get

$$\mathbb{P}[G_n] \leq \mathbb{P}[G_n \cap \tilde{H}_n] + \mathbb{P}[\tilde{H}_n^c]. \quad (21)$$

Using the same arguments as in (14), we have

$$\mathbb{P}[\tilde{H}_n^c] \leq |\tau_n| \sup_{w \in \tau_n^t} \mathbb{P}[|N(w) - \mathbb{E}[N(w)]| > \frac{1}{2}n_w\Gamma_n^{(1-\sigma)/2}].$$

Since $|\tau_n| \leq \left|\tau_n^t\right| d_n \leq n^2$, from Lemma 3 follows

$$\mathbb{P}[\tilde{H}_n^c] = \mathcal{O}(\exp(-D_2 n^{\xi}))\,,$$

for some constants $D_2 > 0$ and $0 < \xi < \min(\lambda, (1-\sigma)(1-\gamma))$. Now, note that

$$\mathbb{P}\big[G_n \cap \tilde{H}_n\big] \leq \sum_{k=\rho_n}^{n} \mathbb{P}\Big[\Big|\hat{P}_{\hat{c}_n}(x|w) - P(x|w)\Big| > \varepsilon, N(w) = k\Big].$$

For all n sufficiently large, we have $\varepsilon > a_n(k)$ ($a_n(k)$ as in (9)) and thus by means of (11) and Lemma 2 we obtain

$$\begin{aligned}
\mathbb{P}\big[G_n \cap \tilde{H}_n\big] &\leq \sum_{k=\rho_n}^{n} \mathbb{P}\Big[\Big|\hat{P}_{\hat{c}_n}(x|w) - P(x|w)\Big| > \sqrt{a_n(k)}, N(w) = k\Big] \\
&\leq \sum_{k=\rho_n}^{n} \mathbb{P}\Big[\Big|\sum_{i=1}^{k} \frac{Y_i}{k} - p\Big| > \sqrt{a_n(k)}\Big] \\
&\leq (n-\rho_n) \sup_{0<p<1} \mathbb{P}\Big[\Big|\sum_{i=1}^{k} \frac{Y_i}{k} - p\Big| > \sqrt{a_n(k)}\Big] \\
&= \mathcal{O}(\exp(-D_3 \log(n)^{(1+\theta)}))\,,
\end{aligned}$$

for some constant $D_3 > 0$ and θ as in assumption (A2).

(ii) Follows from part (i). □

Proof of Corollary 1 We follow the proof of Theorem 1 and note that for the underestimation event U_n the found bound still holds (Lemma 1 with a constant $\widetilde{D}$ different from D). For the overestimation event O_n we have from Lemma 4

$$\mathbb{P}[O_n] \leq |\mathcal{X}_n| \cdot n^{-C+2|\mathcal{X}_n|+3}.$$

Because of assumption (A5) and $C > 2|\mathcal{X}_n| + 3$ the assertion of Corollary 1 follows immediately. □

Proof of Corollary 2 See the proof of Theorem 2. □

Proof of Corollary 3 Follows immediately from the assertion of Corollary 2, since F is continuous and the quantization becomes finer. □

Acknowledgments I am very grateful to Peter Bühlmann for having introduced me to the field of variable length Markov chains. With many suggestions he contributed to great improvements in the presented results. I also thank Michele Marcionelli for computational aid.

References

Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes.* Lecture Notes in Statistics 110, Springer.

Braun, J. and Müller, H.-G. (1998). Statistical Methods for DNA Sequence Segmentation. Statistical Science 13, 142-162.

Bühlmann, P. and Wyner, A. (1999). Variable Length Markov Chains. The Annals of Statistics 27, 480-513.

Bühlmann, P. (1998a). Dynamic Adaptive Partitioning for Nonlinear Time Series. Research Report 84, Seminar für Statistik, ETH Zürich.

Bühlmann, P. (1998b). Model Selection for Variable Length Markov Chains and Tuning the Context Algorithm. Research Report 72, Seminar für Statistik, ETH Zürich.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* John Wiley & Sons.

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

Grenander, U. (1981). *Abstract Inference.* John Wiley & Sons.

Rissanen, J. (1983). A Universal Data Compression System. IEEE Transactions in Information Theory 29, 656-664.

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

Weinberger, M., Rissanen, J. and Feder, M. (1995). A Universal Finite Memory Source. IEEE Transactions in Information Theory 41, 643-652.

Seminar für Statistik
ETH Zürich
CH-8092 Zürich
E-mail: fiore@stat.math.ethz.ch