Estimation of General Stationary Processes by Variable Length Markov Chains

FIORENZO FERRARI
BSI SA Lugano

ABRAHAM WYNER
University of Pennsylvania

ABSTRACT. We develop new results about a sieve methodology for the estimation of minimal state spaces and probability laws in the class of stationary processes defined on finite categorical spaces. Using a sieve approximation with variable length Markov chains of increasing order, we show that an adapted version of the Context algorithm yields asymptotically correct estimates for the minimal state space and for the underlying probability distribution. As a side product, the method of sieves yields a nice graphical tree representation for the potentially infinite dimensional minimal state space of the data generating process, which is very useful for exploration of the memory.

Key words: Context algorithm, sieve approximation, state space estimation, strong mixing sequence, time series, tree model representation.

Running Heading: Estimation by VLMC's

1 Introduction

The assumption that a sequence of data is generated by 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, described in Grenander (1981), combines the advantages of a model, but allows model-misspecification for any finite sample size. In principle, 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 with values in a categorical space, we propose the method of sieves with variable length Markov chains (VLMC's) 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, e.g. when the process has a memory which tends to certain directions, i.e. a sparse memory with a few but typically long states, 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 quite small 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 VLMC's. In fact inequality (5) shows that full Markov chains typically use an approximation of logarithmic order, whereas VLMC approximation is often naturally linked to an increasing order $d = d_n$, which is polynomial in the sample size n, as specified in condition (i) of Assumption 2.

The sieve approximation with VLMC's can be graphically represented by using the so-called context trees, which are rooted trees growing downwards, whose branches represent the relevant history (memory) of the underlying process. Since we do not assume finite memory for the processes to be estimated, the tree representation for the underlying memory may be of infinite order, with some branches growing to infinity. Our approximation uses a sequence of VLMC's of increasing finite order d_n, whose tree representation is a sequence of increasing trees (growing to the true tree as sample size tends to infinity), which are truncated versions of the underlying true tree.

The probability distribution and the minimal state space of general stationary processes taking values in a finite categorical space are thus approximated by those of VLMC's of increasing order d_n, whose estimation is performed using an adapted version of the Context algorithm, for which we refer to Rissanen (1983), Weinberger *et al.* (1995) and Bühlmann & Wyner (1999). The main operations of the algorithm are local decisions between two possible states, lumping them together whenever their corresponding transition probabilities are similar.

If the minimal state space is finite (the underlying process is thus a VLMC), Weinberger *et al.* (1995) and Bühlmann & Wyner (1999) proved that the Context algorithm consistently finds the right model.

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

VLMC's, 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. The operation of the Context algorithm can be also interpreted as a very difficult model selection. Attacking this problem with conventional criteria, such as AIC or BIC, is computationally infeasible. This indicates the importance of the Context algorithm and hence the need to better understand its properties.

As pointed out in remark 7, similar results can be shown to hold for more general stationary processes defined on an increasing size categorical space.

Our results have potential impact to a variety of applications: to mention a few, modeling of categorical time series, e.g. DNA sequences (Bühlmann & Wyner, 1999), (Braun & Müller, 1998) or protein families (Bejerano & Yona, 2001), quantization of nonlinear stationary real-valued time series (Bühlmann, 1999) and sieve-bootstrapping stationary categorical time series (Bühlmann, 2002).

The paper is organized as follows: in Section 2 we define VLMC's on finite categorical spaces and give a tree representation of their minimal state space, which will be useful in Section 3, when describing a version of the Context algorithm proposed by Bühlmann & Wyner (1999). Theoretical results about consistent estimation of the minimal state space and the probability distribution of general stationary processes on finite categorical state spaces are given Section 4. The last section contains all the proofs.

2 Variable length Markov chains

2.1 Definition

Let $\mathcal{X}$ be a finite categorical space, $|\mathcal{X}|$ its cardinality, and $\{X_t : t \in \mathbb{Z}\}$ an $\mathcal{X}$-valued stationary irreducible 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

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

where in general for $a, b \in \overline{\mathbb{Z}}$, $a < b$, x_a^b denotes the sequence $x_b, x_{b-1}, \ldots, x_a$. 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 past and then lump irrelevant states together yielding a possibly parsimonious Markov chain. Formalizing this idea leads to the notion of variable length Markov chain.

In information theory, this model is known as finite memory (or tree) source (Weinberger *et al.*, 1995) and in the field of machine learning as probabilistic suffix automaton

(Ron *et al.*, 1996).

Consider now a process with infinite dependence, but which can be approximated uniformly by a Markov chain. By this we mean that the conditional probabilities $P(x_1|x_{-\infty}^0)$ are continuous functions of $x_{-\infty}^0$ with respect to the product topology, or equivalently for any $\varepsilon > 0$ there exists a $p \in \mathbb{N}$ such that

$$\left|P(x_1|x_{-\infty}^0) - P(x_1|x_{-p}^0 xxx\cdots)\right| < \varepsilon\,,$$

for any $x \in \mathcal{X}$, for all $x_1 \in \mathcal{X}$ and for all $x_{-\infty}^0 \in \mathcal{X}^\infty$.

Definition 1
Let $\mathcal{X}$ be a finite categorical space and $\{X_t : t \in \mathbb{Z}\}$ an $\mathcal{X}$-valued stationary process with continuous conditional probability distribution P.

(i) The projection function

$$\begin{array}{llll} c: & \mathcal{X}^\infty & \longrightarrow & \bigcup_{i=0}^{\infty} \mathcal{X}^i \cup \mathcal{X}^{\mathbb{N}} \quad (\mathcal{X}^0 = \emptyset) \\ & x_{-\infty}^0 & \longmapsto & c(x_{-\infty}^0) = x_{-\ell+1}^0\,, \end{array}$$

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

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

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

A VLMC of order d can be embedded in a Markov chain of order d, however with a memory of variable length $\ell(\cdot) \le d$. 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.

Because of stationarity and irreducibility, the probability distribution P of a VLMC is completely specified by the 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 a VLMC.

2.2 Tree representation

For better insight into the structure of the context function of a VLMC defined on a finite categorical space $\mathcal{X}$, it is convenient to adopt a tree representation. This will also be useful later, when fitting VLMC's to general stationary processes.

The context function $c(\cdot)$ can be represented as a complete tree τ^{com}, where every node has $|\mathcal{X}|$ edges. We consider for our purposes directed rooted trees growing downwards, whose root (the node on top) is connected to any other node by means of exactly one branch (or path). The branches connecting the root with the final nodes represent the values of the context function.

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 with context function $c(\cdot)$. The tree τ^{com} with branches $\{w : w = c(x_{-\infty}^0), x_{-\infty}^0 \in \mathcal{X}^\infty\}$ is called the ($|\mathcal{X}|$-ary) *complete context tree* of the process $\{X_t : t \in \mathbb{Z}\}$.

We allow for additional more parsimonious structure of the context tree. At most $|\mathcal{X}| - 1$ terminal nodes having the same ancestor will be lumped together to one new terminal node if the conditional distributions are the same.

Definition 4
The potentially lumped tree, denoted by τ, is called the *context tree.*

Example 1. Let $\mathcal{X} = \{0, 1, 2\}$ and consider an $\mathcal{X}$-valued VLMC of order $d = 2$ with context function given by

$$c(x_{-\infty}^0) = \begin{cases} 00, \text{ if } x_{-1}^0 = 00\,,\ x_{-\infty}^{-2} \text{ arbitrary} \\ 01, \text{ if } x_{-1}^0 = 01\,,\ x_{-\infty}^{-2} \text{ arbitrary} \\ 02, \text{ if } x_{-1}^0 = 02\,,\ x_{-\infty}^{-2} \text{ arbitrary} \\ \ \ 1, \text{ if } x_0 = 1\,,\ x_{-\infty}^{1} \text{ arbitrary} \\ 20, \text{ if } x_{-1}^0 = 20\,,\ x_{-\infty}^{-2} \text{ arbitrary} \\ 21, \text{ if } x_{-1}^0 = 21\,,\ x_{-\infty}^{-2} \text{ arbitrary} \\ 22, \text{ if } x_{-1}^0 = 22\,,\ x_{-\infty}^{-2} \text{ arbitrary} \end{cases}$$

As additional structure we assume $P(x_1|21) = P(x_1|22) \neq P(x_1|20)$. We then lump the terminal nodes 21 and 22 together to one new terminal node which is represented as an *internal* node. We thus have states or contexts which correspond to terminal nodes, drawn in black, and we have states or contexts which correspond to internal nodes, drawn in white.

(Figure 1 here)

The context tree consists of the states $\tau = \{00, 01, 02, 1, 2, 20\}$, which are all nodes of the context tree in Figure 1.

It will often be useful to distinguish between the whole context tree τ and its terminal nodes $\tau^t = \{w : w \in \tau \text{ and } wu \notin \tau, \forall u \in \mathcal{X}\}$.

Example 1 (continued). The set of terminal nodes is given by $\tau^t = \{00,\ 01,\ 02,\ 1,\ 20\}$, which consists of all the black leaves (terminal nodes) in Figure 1 only.

The minimal state space of a VLMC is thus given by the context tree τ. The associated transition probabilities are then $P(x|w) = \mathbb{P}\left[X_1 = x | c(X_{-\infty}^0)\right]\ ,\ x \in \mathcal{X}\ ,\ w \in \tau$, with $c(X_{-\infty}^0)$ belonging to $w \in \tau$.

Example 1 (continued). The following transition probabilities are in action:

$$P(x|w) = \begin{cases} \pi_1(x), \text{ if } x_{-1}^0 = 00\,,\ x_{-\infty}^{-2} \text{ arbitrary} \\ \pi_2(x), \text{ if } x_{-1}^0 = 01\,,\ x_{-\infty}^{-2} \text{ arbitrary} \\ \pi_3(x), \text{ if } x_{-1}^0 = 02\,,\ x_{-\infty}^{-2} \text{ arbitrary} \\ \pi_4(x), \text{ if } x_0 = 1\,,\ x_{-\infty}^{1} \text{ arbitrary} \\ \pi_5(x), \text{ if } x_{-1}^0 = 20\,,\ x_{-\infty}^{-2} \text{ arbitrary} \\ \pi_6(x), \text{ if } x_0 = 2\,, x_{-1} \in \{1,2\}\,,\ x_{-\infty}^{-2} \text{ arbitrary} \end{cases}$$

where $\pi_i(\cdot)$, $i = 1, \ldots, 6$, are 3×1 probability vectors, whose components sum up to one.

Remark 1. It is worth pointing out that knowledge of a context tree τ implies knowledge of the complete context tree τ^{com} and thus of the context function $c(\cdot)$. Going vice versa, $c(\cdot)$ implies τ^{com}, but generally we cannot infer from this to τ.

3 The Context algorithm

Let $\mathcal{X}$ be a finite categorical space and $\{X_t : t \in \mathbb{Z}\}$ an $\mathcal{X}$-valued general stationary process with distribution function P, context function $c(\cdot)$ and context tree τ.

Consider only one realization which is a sequence $x_1, \ldots, x_n$. Our purpose is to find good estimates of both the underlying context function (and therefore of the context tree), which can be of infinite order, and the probability distribution. A computationally feasible method is given by an adapted version of the Context algorithm, introduced by Rissanen (1983) and recently re-proposed by Bühlmann & Wyner (1999).

Let $n_w = n - |w| + 1$, where $w \in \bigcup_{i=1}^n \mathcal{X}^i$ and $|w|$ is the number of elements of w. We denote by $N_{x_1^n}(w)$ the number of occurrences of the substring w in the reverse data string x_1^n. Hence,

$$N_{x_1^n}(w) = \sum_{t=1}^{n_w} 1_{\{x_t^{t+|w|-1} = w\}}\,, \quad w \in \bigcup_{i=1}^{n} \mathcal{X}^i\,.$$

Furthermore we define

$$\widehat{P}_{x_1^n}(w) = \frac{N_{x_1^n}(w)}{n}\,, \quad \widehat{P}_{x_1^n}(u|w) = \frac{N_{x_1^n}(uw)}{N_{x_1^{n-1}}(w)}\,, \quad u \in \mathcal{X},\ w \in \bigcup_{i=1}^{n-1} \mathcal{X}^i\,.$$

The estimator $\widehat{P}_{x_1^n}(w) = N_{x_1^n}(w)/n$ possesses asymptotically the same features as the unbiased version $\widehat{P}_{x_1^n}(w) = N_{x_1^n}(w)/(n - |w| + 1)$, since n_w is of the same order as n. We have opted for $\widehat{P}_{x_1^n}(w) = N_{x_1^n}(w)/n$ for simplicity in the definition of $\widehat{P}_{x_1^n}(u|w)$.

The operation of the Context algorithm takes place in three steps. Starting from a predetermined initial maximal context tree for the sequence $x_1, \ldots, x_n$, we prune its branches until the remaining states are all relevant for the development of the underlying process. The condition for pruning, stated in inequality (2), is based on the Kullback-Leibler information $D\left(\cdot,\cdot\right)$, which is defined by

$$D\left(P,Q\right)=\mathbb{E}_P\left[\log\left(\frac{P(X)}{Q(X)}\right)\right]=\sum_{x\in\mathcal{X}^m}P(x)\log\left(\frac{P(x)}{Q(x)}\right)\,, \tag{1}$$

where P and Q are m-dimensional ($1\leq m\leq\infty$) probability measures defined on the same categorical space $\mathcal{X}^m$.

The Context algorithm

Step 1. Build the maximal context tree corresponding to the sequence $x_1,\ldots,x_n$. This is the maximal context tree $\tau_{(0)}$, whose terminal nodes w satisfy $N_{x_1^n}(w)\geq 2$, i.e they have been observed at least twice in the reverse data string x_1^n.

Step 2. Let $wu=x_{-\ell+1}^0$, with $u=x_{-\ell+1}$ and $w=x_{-\ell+2}^0$, be a terminal node of $\tau_{(0)}$. Prune $wu=x_{-\ell+1}^0$ to $w=x_{-\ell+2}^0$, if

$$\Delta_{wu}(x_1^n)=D\left(\widehat{P}_{x_1^n}(\cdot|wu),\widehat{P}_{x_1^n}(\cdot|w)\right)N_{x_1^n}(wu)<K_n\,, \tag{2}$$

where

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

Repeat the pruning procedure for all terminal nodes wu of $\tau_{(0)}$ and build in this way the context tree $\tau_{(1)}$. If $\ell=1$, the pruned version is the root node. If an internal node $w\in\tau_{(1)}$ is lacking in edges, then the context tree $\tau_{(1)}$ is completed: to every non-complete internal node w one additional state u is attached, such that wu represents the needed (possibly lumped together) terminal nodes. The corresponding estimate $\widehat{P}_{x_1^n}(\cdot|wu)$ of the transition probability is set equal to $\widehat{P}_{x_1^n}(\cdot|w)$.

Step 3. Repeat Step 2 with $\tau_{(i)}$ instead of $\tau_{(i-1)}$ $(i=1,2,\ldots)$ until no more pruning is possible. Denote the final obtained context tree by $\widehat{\tau}_n$ and the corresponding context function by $\widehat{c}_n(\cdot)$.

The underlying context function $c(\cdot)$ is estimated by $\widehat{c}_n(\cdot)$, and the transition probability $P(x_1|c(x_{-\infty}^0))$ by $\widehat{P}_{\widehat{c}_n}(x_1|x_{-\infty}^0):=\widehat{P}_{x_1^n}(x_1|\widehat{c}_n(x_{-\infty}^0))$.

Remark 2. The initial maximal context tree $\tau_{(0)}$ 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, since it allows to start the pruning procedure with a large tree. 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 context tree $\tau_{(i)}$ in Step 2 is irrelevant.

Remark 3. The Context algorithm can be viewed as a multiple likelihood ratio test with rejection region (K_n,∞) for the null hypothesis $P(\cdot|wu)=P(\cdot|w)$.

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, as can be seen in the proof of Theorem 1). The condition on C is established by asymptotic considerations in Lemma 5, whereas estimation of C is discussed in Bühlmann (2000). 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 memory of the process, such as $\ell(\cdot) = |c(\cdot)| \leq \log(n)/\log(|\mathcal{X}|)$ employed in Weinberger *et al.* (1995), which can be a severe restriction in practical applications. However we will see in the next section, that in order to prove consistency for the estimate of the context function, we assume for $|c(\cdot)|$ the milder condition (i) of Assumption 2.

Remark 6. Ron *et al.* (1996) propose a learning algorithm (called Learn-PSA) for model selection (estimation) in the class of probabilistic suffix automata. This algorithm constructs the model forwardly: starting from the root node, it adds nodes, if some probabilistic conditions are satisfied. On the contrary, the Context algorithm works backwardly, pruning an initial maximal context tree. This backward pruning is analogous to the one in the tree-structured CART algorithm (Breiman *et al.*, 1984). It usually perform better than the more simple forward construction.

4 Consistency

We consider a discrete-time general stationary process $\{X_t : t \in \mathbb{Z}\}$ defined on a finite categorical space $\mathcal{X}$, with continuous conditional probability distribution P.

For this kind of processes the context function $c(\cdot)$ (or the context tree τ, respectively) needs not to be of finite order. To prove consistency for the estimate of $c(\cdot)$ (given by the Context algorithm), we approximate $c(\cdot)$ by a sequence of context functions corresponding to VLMC's of increasing finite order. This means that we approximate in a reasonable sense general stationary processes by VLMC's. Such approximations with increasing-dimensional parametric models as sample size n grows, are also known as method of sieves, as illustrated in Grenander (1981).

Definition 5
Let $\mathcal{X}$ be a finite categorical space and $\{X_t : t \in \mathbb{Z}\}$ an $\mathcal{X}$-valued general stationary process with context function $c(\cdot)$ and context tree τ. Let $\{d_n : n \in \mathbb{N}\}$ be a sequence increasing to infinity. The sequence $\{c_n(\cdot) : n \in \mathbb{N}\}$ of *truncated context functions* corresponding to $c(\cdot)$ is defined by

$$c_n(\cdot) = \begin{cases} x^0_{-d_n+1}\,, \text{ if } \left|c(x^0_{-\infty})\right| \geq d_n \\ \\ c(x^0_{-\infty})\,, \text{ otherwise}\,. \end{cases}$$

We assume here and in the sequel of Section 4, that an internal node $v \in \tau$ has been completed to one node $w \in \tau$ representing all the lumped states. Hence, τ is a complete tree where every (possibly lumped) state is a terminal node in τ.

Example 1 (continued). The completed context tree would look as in Figure 2.

(Figure 2 here)

The white ellipse represents the completion of the context tree, with the terminal nodes 21 and 22 lumped together.

Definition 6
The sequence $\{\tau_n : n \in \mathbb{N}\}$ of *truncated context trees* corresponding to τ is defined by

$$\tau_n = \Big\{ w \in \tau : |w| \leq d_n \Big\} \cup \Big\{ w \in \mathcal{X}^{d_n} : wu \in \tau,\, u \in \bigcup_{i=1}^{\infty} \mathcal{X}^i \Big\} .$$

Furthermore let $\widehat{c}_n^{tr}(\cdot)$ be the truncation of the estimated context function $\widehat{c}_n$ at level d_n, as described in Definition 5, and $\widehat{\tau}_n^{tr}$ be the truncation of the estimated context tree $\widehat{\tau}_n$ at level d_n, as described in Definition 6.

Definition 7
Let $\{X_t : t \in \mathbb{Z}\}$ be a stationary process with probability distribution P. Let $\sigma_a^b = \sigma(X_a, \ldots, X_b)$ be the σ-algebra generated in the time interval $[a, b]$. The process $\{X_t : t \in \mathbb{Z}\}$ is *α-mixing* (or *strong mixing*), if

$$\alpha(k) = \sup_{A \in \sigma_{-\infty}^0,\, B \in \sigma_k^{\infty}} |P(A)P(B) - P(A \cap B)| \longrightarrow 0 \quad (k \to \infty) .$$

We make the following assumptions:

Assumption 1. The process $\{X_t : t \in \mathbb{Z}\}$ is stationary and geometrically α-mixing with α-mixing coefficients $\{\alpha(k) : k \in \mathbb{N}\}$ satisfying $\alpha(k) \leq C_\alpha \nu^k$, for some constants $C_\alpha > 0$ and $\nu \in (0, 1)$.

Assumption 2. The sequence of truncated context functions $\{c_n(\cdot) : n \in \mathbb{N}\}$ is determined by an increasing sequence $\{d_n : n \in \mathbb{N}\}$, such that

(i) For all n sufficiently large, $d_n \leq n^\delta$, for some $\delta \in (0, \sigma)$, where $\sigma \in (0, 1)$ is specified in condition (ii).

(ii) For some $\theta > 0$, some $\sigma \in (0, 1)$ and some $\gamma \in (0, (1 - \sigma)/2)$, for all n sufficiently large,

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

and

$$\Upsilon_n = \min_{wu \in \tau_n, u \in \mathcal{X}} \|P(\cdot|wu) - P(\cdot|w)\|_1$$

satisfies

$$\Upsilon_n^2 \geq \frac{\log(n)^{1+\theta}}{\left(n \Gamma_n^{(1-\sigma)/2}\right)^{1-\sigma}} .$$

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

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

The conditions in Assumption 2 are all probabilistic conditions about the sparseness and the growth rate of the truncated context tree τ_n. They may be hard to check, but have some intuitive meanings, which we now discuss.

Condition (i) is about the maximal growth rate for the approximating order of the context tree. Consider a Markov chain of order d_n. Then, $|\tau_n| = |\mathcal{X}|^{d_n}$. Because of the first inequality in condition (ii), the cardinality of the context tree τ_n is bounded by

$$|\tau_n| \leq \frac{1}{\Gamma_n} \leq n^\gamma \, . \tag{4}$$

Thus, in this case, condition (i) becomes

$$d_n \leq \frac{\gamma}{\log(|\mathcal{X}|)} \log(n) \, . \tag{5}$$

Hence, the choice of δ in the interval $(0, \sigma)$ is without restrictions. With VLMC's we can also treat models with a memory growing only in certain directions (e.g. a sparse memory represented by a context tree, which has only a few but typically long branches, as described in Example 2 and Example 3, with a growth rate d_n of polynomial order $\delta \in (0, \sigma)$. This is a big advantage of VLMC's in comparison with Markov chains.

The second inequality in condition (ii) is measuring relevance of terminal nodes in comparison with their ancestors and ensures that they are not too close to each other. Without this condition the Context algorithm could not distinguish between them.

As pointed out in Bühlmann & Wyner (1999), for general stationary processes with finite memory (being therefore VLMC's), it suffices to assume

$$\min_{x \in \mathcal{X}, w \in \tau} P(x|w) > 0 \, . \tag{6}$$

Condition (6) implies Assumptions 1 and 2 directly.

The power of the Context algorithm is shown in the next Theorem 1 and Theorem 2. The first theorem states that for general stationary processes the Context algorithm produces context trees, whose truncated versions consistently estimate the underlying truncated context tree. The second theorem asserts that the transition probabilities, given finite (possibly unbounded) contexts, are also consistently estimated. According to Theorem 1, the Context algorithm selects asymptotically the correct finite dimensional model components. This cannot be achieved by more traditional selection criterion such as AIC or BIC due to the extremely large number of possible sub-models.

This work is so far the only attempt to model a general stationary process with a sequence of VLMC's of increasing order.

Bühlmann & Wyner (1999), using the same Context algorithm, showed consistency for the moving truths model, namely a sequence of VLMC's with a memory allowed to depend on the sample size. In their seminal work, Weinberger *et al.* (1995) obtained consistent estimates for fixed order VLMC's, imposing in their version of the Context algorithm a search between models with bounded context length, which can nevertheless slowly grow with the sample size. Ron *et al.* (1996) by means of their learning algorithm Learn-PSA, already discussed in Remark 6), and Willems *et al.* (1995) using their Context Tree Weighting algorithm achieve convergence (in Kullback-Leibler distance) of the estimated to the true underlying probability distribution for fixed finite order VLMC's. Their algorithms do not propose a model, but are very useful for prediction.

Theorem 1
Under the Assumptions 1 and 2,

$$\mathbb{P}\left[\widehat{\tau}_n^{tr} = \tau_n\right] \longrightarrow 1 \quad (n \to \infty)\,.$$

Theorem 2
Under the Assumptions 1 and 2,

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

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

Assertion (ii) of Theorem 2 also holds for the non-truncated version of the estimate $\widehat{c}_n(\cdot)$ of the context function.

Example 2. We consider the threshold first order autoregressive process $\{Y_t : t \in \mathbb{Z}\}$ defined by $Y_t = \phi Y_{t-1} 1_{\{Y_{t-1}>0\}} + Z_t$, where $|\phi| < 1$, $1_{\{\cdot\}}$ is the indicator function and $\{Z_t : t \in \mathbb{Z}\}$ is a sequence of independent and identically distributed random variables (the innovation process), having a density with respect to the Lebesgue measure and a finite first absolute moment. The stochastic process $\{Y_t : t \in \mathbb{Z}\}$ is stationary and geometrically α-mixing, as described in Example 3, Chapter 2.4 in Doukhan (1994).

We define the categorical process $\{X_t : t \in \mathbb{Z}\}$ on $\mathcal{X} = \{0, 1\}$ by $X_t = 1_{\{Y_t>0\}}$. This is also stationary and geometrically α-mixing (with the same bound for the α-mixing coefficients). The context function of $\{X_t : t \in \mathbb{Z}\}$ is given by

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

where $h = h(X_{-\infty}^0) = \min\{k : X_{-k} = 0 \text{ and } X_{-k+1}^0 = 1, \ldots, 1\}$. Whenever a state $X_{-k} = 0$ occurs in the past $X_{-\infty}^0$, then $X_{-\infty}^{-k-1}$ becomes irrelevant for the future state X_1. This process has an infinitely long context function $c(\cdot)$, whose corresponding context tree τ is represented in Figure 3, and is therefore not a VLMC.

(Figure 3 here)

For this Example 2 and in general, the verification of the conditions stated in Assumption 2 is very difficult.

Example 3. Let $\{X_t : t \geq 0\}$ be a stationary binary process with initial probability given by $\mathbb{P}[X_0 = 0] = \mathbb{P}[X_0 = 1] = 1/2$, and where the times $\{T_i : i \geq 1\}$ between switches of

X_t from 1 to 0 or 0 to 1 are independent and identically distributed. $\{X_t : t \geq 0\}$ is thus a stationary alternating renewal process. We assume

$$\mathbb{P}[T = j] = c_1 \rho_1^j + c_2 \rho_2^j \, , \; 0 < \rho_2 < \rho_1 < 1 \, . \tag{7}$$

and let $\mu = \mathbb{E}[T]$. The same results hold if we add to (7) a remainder term of the order $o(\rho_2^j)$. The context tree of the process $\{X_t : t \geq 0\}$ is represented in Figure 4. Whenever in the history a state different from the preceding states occurs, then the later past does not play any role for the future development of the process. We will prove in the Appendix that this example satisfies our Assumptions 1 and 2.

(Figure 4 here)

Remark 7. The minimal state space and the probability distribution of general stationary processes defined on increasing size categorical spaces $\mathcal{X}_n$, $n \in \mathbb{N}$, can be also consistently estimated with the approximation by VLMC's. The only further assumption is a slowly enough growth for the cardinality of $\mathcal{X}_n$, namely $|\mathcal{X}_n| = \mathcal{O}(\log(n)^{1+r})$, for some $r > 0$. Such a result allows us to show consistency for real-valued stationary processes using the quantization procedure explained in Bühlmann (1999), where also some practical applications are given, and Ferrari (2002).

5 Conclusions

We have presented new consistency results about estimation with variable length Markov chains (VLMC's) in the class of general stationary processes defined on finite categorical spaces. We have proposed a sieve approximation with VLMC's of increasing order, based on the so-called context function, which is a function describing the memory of a process. We have proved that with this methodology the Context algorithm yields estimates for the minimal state space and for the underlying probability distribution, which are asymptotically correct, for very general stationary processes with possibly infinite memory.

This is the first work, where a general stationary process is modeled with VLMC's of increasing order. The underlying process is neither assumed to be of fixed order nor to have a memory depending on the sample size.

For the exploration of the potentially infinitely long memory of a general stationary process, we also obtain an interesting graphical tree representation for the underlying minimal state space. By means of two illustrative examples we show the power of our methodology. Also for processes defined on increasing size categorical spaces or with real values, our method can be successfully applied. The presented sieve approximation finds application in many fields, such as modeling or sieve-bootstrapping stationary categorical time series or quantization of nonlinear stationary real-valued time series.

Acknowledgements

The first author largely carried out this research when he was at the Department of Mathematics at the ETH in Zurich. We are very grateful to Peter Bühlmann. With many suggestions he contributed to great improvements in the presented results. We thank

Hansruedi Künsch for some technical refinements. Our gratitude goes also to Michele Marcionelli for computational aid. The comments of a referee led to a better presentation of the contents.

References

Asmussen, S. (1987). *Applied probabilities and queues.* Wiley, New York.

Bejerano, G. & Yona, G. (2001). Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics* **17**, 23-43.

Bosq, D. (1996). *Nonparametric statistics for stochastic processes.* Lecture Notes in Statist. **110**.

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

Breiman, L., Friedman, J., Olshen, R. and Stone, C.J. (1984). *Classification and regression trees.* Wadsworth, Belmont, California.

Bühlmann, P. (1999). Dynamic adaptive partitioning for nonlinear time series. *Biometrika* **86**, 555-571.

Bühlmann, P. (2000). Model selection for variable length Markov chains and tuning the Context algorithm. *Ann. Inst. Statist. Math.* **52**, 287-315.

Bühlmann, P. (2002). Sieve bootstrap with variable length Markov chains for stationary categorical time series (with discussion). *J. Amer. Statist. Assoc.* **97**, 443-471.

Bühlmann, P. & Wyner, A. (1999). Variable length Markov chains. *Ann. Statist.* **27**, 480-513.

Cover, T. & Thomas, J. (1991). *Elements of information theory.* Wiley, New York.

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

Ferrari, F. (2002). *Variable length Markov chains and dynamic combination of models.* Ph.D. Thesis (No. 14503). ETH Zurich.

Grenander, U. (1981). *Abstract inference.* Wiley, New York.

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

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry.* World Scientific, Singapore.

Ron, D., Singer, Y. & Tishby, N. (1996). The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning* **25**, 117-149.

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

Willems, F.M.J., Shtarkov, Y.M. & Tjalkens, T.J. (1995). The context tree weighting method: basic properties. *IEEE Trans. Inform. Theory* **41**, 653-664.

Fiorenzo Ferrari, Analysis and Strategies, BSI SA, via Peri 23, 6900 Lugano, Switzerland
E-mail: fiorenzo.ferrari@bsi.ch

Appendix

Proof of Theorem 1. We essentially follow the proof of Theorem 3.1 in Bühlmann & Wyner (1999). We rewrite for our case Lemma 5.1 and Lemma 5.2, applying an exponential inequality for α-mixing processes, and also Lemma 5.3.

Let $x_1, \ldots, x_n$ be realizations of $\{X_t : t \in \mathbb{Z}\}$. The error event $E_n = \{\widehat{\tau}_n^{tr} \neq \tau_n\}$ 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

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

Therefore, we can bound the error in the estimation of the underlying context tree by separately treating the under- and the overestimation.

Lemma 1
Under the Assumptions 1 and 2 (without condition (iii)),

$$\mathbb{P}\left[U_n\right] = \mathcal{O}\left(\exp\left(-D\log(n)^{1+\theta}\right)\right) \tag{8}$$

for some constant $D > 0$ and θ as in Assumption 2.

Proof. We define the sequence $\{\rho_n : n \in \mathbb{N}\}$ by $\rho_n = n\Gamma_n/2$ and the event H_n by

$$H_n = \{N_{x_1^n}(w) \geq \rho_n \text{ for every } w \in \tau_n\}.$$

Since $U_n = U_n \cap (H_n \cup H_n^c) \subseteq (U_n \cap H_n) \cup H_n^c$, it follows that

$$\mathbb{P}\left[U_n\right] \leq \mathbb{P}\left[U_n \cap H_n\right] + \mathbb{P}\left[H_n^c\right]. \tag{9}$$

Now, we separately bound the two probabilities on the right side of inequality (9). The event $U_n \cap H_n$ is the underestimation event for branches observed at least ρ_n times in the reversed sequence X_1^n. Thus,

$$\mathbb{P}\left[U_n \cap H_n\right] \leq \sum_{wu\in\tau_n, u\in\mathcal{X}} \sum_{k=\rho_n}^{n_{wu}} \sum_{j=k}^{n_w} \mathbb{P}\Big[D\left(\widehat{P}_{x_1^n}(\cdot|wu), \widehat{P}_{x_1^n}(\cdot|w)\right) < C\frac{\log(n)}{k}, \\ N_{x_1^n}(wu) = k, N_{x_1^n}(w) = j\Big]. \tag{10}$$

The Kullback-Leibler information between two probability distributions is lower bounded by the half squared L_1-distance (Cover & Thomas, 1991). This, with the statements in Bühlmann & Wyner (1999), pp. 502-503, leads to the bound

$$\begin{aligned}\mathbb{P}&\left[D\left(\widehat{P}_{x_1^n}(\cdot|wu), \widehat{P}_{x_1^n}(\cdot|w)\right) < C\frac{\log(n)}{k}, N_{x_1^n}(wu) = k, N_{x_1^n}(w) = j\right] \\ &\leq |\mathcal{X}| \sup_{x\in\mathcal{X}} \left(\mathbb{P}\left[\left|\widehat{P}_{x_1^n}(x|wu) - P(x|wu)\right|^2 > a_n(k), N_{x_1^n}(wu) = k\right]\right. \\ &\qquad \left. + \mathbb{P}\left[\left|\widehat{P}_{x_1^n}(x|w) - P(x|w)\right|^2 > a_n(k), N_{x_1^n}(w) = j\right]\right),\end{aligned} \tag{11}$$

with

$$a_n(k) = \left(\frac{\Upsilon_n}{2} - \sqrt{\frac{C \log n}{k}} \right)^2 . \tag{12}$$

For n sufficiently large, because of condition (ii) of Assumption 2, we have

$$\min_{k \geq \rho_n} a_n(k) \geq \frac{\log(n)^{1+\theta}}{\rho_n^{1-\sigma}} . \tag{13}$$

This bound is not the same as those given in Bühlmann & Wyner (1999). The difference is due to condition (ii) of Assumption 2, which in our case must be modified with an exponent $(1-\sigma)$, $\sigma \in (0,1)$, in the denominator. This is the price we have to pay, when estimating general stationary processes. The two summands in (11) are now handled in the same manner denoting with v either wu or w. Let $p = P(x|v)$ and $\widehat{p} = \widehat{P}_{x_1^n}(x|v)$. In order to find an upper probabilistic bound for the event

$$\left\{ |\widehat{p} - p|^2 > a_n(k), N_{x_1^n}(v) = k \right\} ,$$

consider the extension of $X_1, \ldots, X_n$ 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 $\{X_t : t \in \mathbb{N}\}$. Let $Z_i = X_{I_i+1}$ be the symbol that occurs after the i^{th} occurrence of v. 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

$$\left\{ \left| \sum_{i=1}^{N_{x_1^n}(v)} \frac{Y_i}{N_{x_1^n}(v)} - p \right|^2 > a_n(k), N_{x_1^n}(v) = k \right\} \subseteq \left\{ \left| \sum_{i=1}^{k} \frac{Y_i}{k} - p \right|^2 > a_n(k) \right\} ,$$

and consequently

$$\mathbb{P}\left[|\widehat{p} - p|^2 > a_n(k), N_{x_1^n}(v) = k \right] \leq \mathbb{P}\left[\left| \sum_{i=1}^{k} \frac{Y_i}{k} - p \right|^2 > a_n(k) \right] . \tag{14}$$

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 (12). Then under the Assumptions 1 and 2 (without condition (iii)), for $k \geq \rho_n$ and for all n sufficiently large

$$\sup_{0<p<1} \mathbb{P}\left[\left| \sum_{i=1}^{k} \frac{Y_i}{k} - p \right|^2 > a_n(k) \right] \leq 4 \exp\left(-B_1 \log(n)^{1+\theta} \right) + 11\sqrt{5} C_\alpha n^{5(1-\sigma)/4} \nu^{n^{(1-\gamma)\sigma/2}} ,$$

for some constant $B_1 > 0$, C_α as in Assumption 1 and σ, θ, γ as in Assumption 2.

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 $(Y_i - p)_{i\in\mathbb{N}}$ is a zero-mean real-valued process with $|Y_i - p| \le 1$, for all $i \in \mathbb{N}$, we get from Theorem 1.3, Chapter 1.4 in Bosq (1996) for each integer q in $1, \dots, \lfloor k/2 \rfloor$

$$\begin{aligned} \sup_{0<p<1} \mathbb{P}\left[\left|\sum_{i=1}^{k} \frac{Y_i}{k} - p\right|^2 > a_n(k)\right] &= \sup_{0<p<1} \mathbb{P}\left[\left|\sum_{i=1}^{k} (Y_i - p)\right| > k\sqrt{a_n(k)}\right] \\ &\le 4\exp\left(-\frac{1}{8} q a_n(k)\right) + 22\sqrt{1 + 4\frac{1}{\sqrt{a_n(k)}}} \cdot q\alpha\left(\left\lfloor \frac{k}{2q} \right\rfloor\right) . \end{aligned} \tag{15}$$

From inequality (13), by setting $q = \lfloor k^{1-\sigma}/2 \rfloor$, we obtain for $k \ge \rho_n$ and for all n sufficiently large

$$q a_n(k) \ge \left\lfloor \frac{\rho_n^{1-\sigma}}{2} \right\rfloor \frac{\log(n)^{1+\theta}}{\rho_n^{1-\sigma}} .$$

The sequence $\lfloor \rho_n^{1-\sigma}/2 \rfloor / \rho_n^{1-\sigma}$ tends increasingly to $1/2$ ($n \to \infty$), and thus there exists a positive constant B_1, such that for all n sufficiently large $\lfloor \rho_n^{1-\sigma}/2 \rfloor / \rho_n^{1-\sigma} \ge B_1$. This leads therefore to $q a_n(k) \ge B_1 \log(n)^{1+\theta}$.

For the first term in the second summand of inequality (15), we have again by (13) and by $\rho_n = n\Gamma_n/2 \le n$, that $1 + 4/a_n(k)^{1/2} \le 1 + 4\rho_n^{(1-\sigma)/2} \le 5n^{(1-\sigma)/2}$. For the second term, since $q = \lfloor k^{1-\sigma}/2 \rfloor \le k^{1-\sigma}/2 \le n^{1-\sigma}/2$, and, by condition (ii) of Assumption 2, $\rho_n = n\Gamma_n/2 \ge n^{1-\gamma}/2$, we get for $k \ge \rho_n$ and for all n sufficiently large

$$\begin{aligned} q\alpha\left(\left\lfloor \frac{k}{2q} \right\rfloor\right) &\le \frac{1}{2} n^{1-\sigma} \alpha\left(\lfloor \rho_n^{\sigma} \rfloor\right) \le \frac{1}{2} n^{1-\sigma} \alpha\left(\left\lfloor \left(\frac{1}{2} n^{1-\gamma}\right)^{\sigma} \right\rfloor\right) \\ &\le \frac{1}{2} n^{1-\sigma} \alpha\left(n^{(1-\gamma)\sigma/2}\right) \le \frac{1}{2} n^{1-\sigma} C_\alpha \nu^{n^{(1-\gamma)\sigma/2}} . \end{aligned}$$

The assertion of the lemma follows then immediately.

A direct application of Lemma 2 to the above inequality (11) proves, that for all n sufficiently large

$$\begin{aligned} &\mathbb{P}\left[U_n \cap H_n\right] \\ &\quad \le 2\left|\mathcal{X}\right| \sum_{wu \in \tau_n, u \in \mathcal{X}} \sum_{k=\rho_n}^{n_{wu}} \sum_{j=k}^{n_w} \left(4\exp\left(-B_1 \log(n)^{1+\theta}\right) + 11\sqrt{5} C_\alpha n^{5(1-\sigma)/4} \nu^{n^{(1-\gamma)\sigma/2}}\right) \\ &\quad \le 2\left|\mathcal{X}\right| \left|\tau_n\right| n^2 \left(4\exp\left(-B_1 \log(n)^{1+\theta}\right) + 11\sqrt{5} C_\alpha n^{5(1-\sigma)/4} \nu^{n^{(1-\gamma)\sigma/2}}\right) . \end{aligned}$$

The cardinality of the context tree τ_n is bounded by n^γ with $\gamma \in (0, (1-\sigma)/2)$, as explained in inequality (4). In consequence

$$\mathbb{P}\left[U_n \cap H_n\right] = \mathcal{O}\left(\exp\left(-D_1 \log(n)^{1+\theta}\right)\right) ,$$

for some constant $D_1 > 0$ and θ as in Assumption 2.

The next step is to find a bound for $\mathbb{P}\left[H_n^c\right]$. This is the probability that all terminal nodes have been observed less than ρ_n times in the reversed sequence X_1^n. First of all note, that for $w \in \tau_n$ holds

$$\mathbb{E}\left[N_{x_1^n}(w)\right] = \mathbb{E}\left[\sum_{t=1}^{n_w} 1_{\{X_t^{t+|w|-1}=w\}}\right] = \sum_{t=1}^{n_w} P(w) \geq n_w \Gamma_n \,. \tag{16}$$

For all n sufficiently large we have then

$$\begin{aligned}
\mathbb{P}\left[H_n^c\right] &\leq \sum_{w\in\tau_n} \mathbb{P}\left[N_{x_1^n}(w) < \rho_n\right] = \sum_{w\in\tau_n} \mathbb{P}\left[N_{x_1^n}(w) - \mathbb{E}\left[N_{x_1^n}(w)\right] < \rho_n - \mathbb{E}\left[N_{x_1^n}(w)\right]\right] \\
&= \sum_{w\in\tau_n} \mathbb{P}\left[N_{x_1^n}(w) - \mathbb{E}\left[N_{x_1^n}(w)\right] < \frac{1}{2} n\Gamma_n - n_w\Gamma_n\right] \\
&\leq \sum_{w\in\tau_n} \mathbb{P}\left[N_{x_1^n}(w) - \mathbb{E}\left[N_{x_1^n}(w)\right] < -\frac{1}{3} n_w\Gamma_n\right] \\
&\leq \sum_{w\in\tau_n} \mathbb{P}\left[\left|N_{x_1^n}(w) - \mathbb{E}\left[N_{x_1^n}(w)\right]\right| > \frac{1}{3} n_w\Gamma_n\right] \\
&\leq |\tau_n| \sup_{w\in\tau_n} \mathbb{P}\left[\left|N_{x_1^n}(w) - \mathbb{E}\left[N_{x_1^n}(w)\right]\right| > \frac{1}{3} n_w\Gamma_n\right] \,.
\end{aligned} \tag{17}$$

Lemma 3
Under the Assumptions 1 and 2 (without condition (iii)), for all n sufficiently large

$$\begin{aligned}
&\sup_{w\in\tau_n} \mathbb{P}\left[\left|N_{x_1^n}(w) - \mathbb{E}\left[N_{x_1^n}(w)\right]\right| > \frac{1}{3} n_w\Gamma_n\right] \\
&\qquad\qquad \leq 4\exp\left(-B_2 n^{1-\sigma-2\gamma}\right) + 11\sqrt{13} C_\alpha \nu n^{1-\sigma+\gamma/2} \nu^{n^{\sigma/2}} \,,
\end{aligned}$$

for some constant $B_2 > 0$, C_α as in Assumption 1 and δ, σ, γ as in Assumption 2.

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

$$N_{x_1^n}(w) - \mathbb{E}\left[N_{x_1^n}(w)\right] = \sum_{t=1}^{n_w} W_t.$$

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

$$\alpha_W(i) \leq \begin{cases} \alpha(i - |w| + 1)\,, \text{ if } i \geq |w| \\ \\ 1\,, \text{ if } i < |w| \end{cases} \tag{18}$$

where $\{\alpha(i) : i \in \mathbb{N}\}$ 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 each integer q in $1, \ldots, \lfloor n_w/2 \rfloor$, σ as in Assumption

2 and $w \in \tau_n$

$$\mathbb{P}\left[\left|N_{x_1^n}(w) - \mathbb{E}\left[N_{x_1^n}(w)\right]\right| > \frac{1}{3} n_w \Gamma_n\right] = \mathbb{P}\left[\left|\sum_{t=1}^{n_w} W_t\right| > \frac{1}{3} n_w \Gamma_n\right]$$
$$\leq 4 \exp\left(-\frac{1}{72} q \Gamma_n^2\right) + 22 \sqrt{1 + 12 \frac{1}{\Gamma_n}} \cdot q \alpha_W \left(\left\lfloor \frac{n_w}{2q} \right\rfloor\right) . \qquad (19)$$

By setting $q = \left\lfloor n_w^{1-\sigma}/2 \right\rfloor$ and by condition (ii) of Assumption 2, we get for all n sufficiently large

$$q \Gamma_n^2 \geq \frac{\left\lfloor n_w^{1-\sigma} \right\rfloor}{n^{2\gamma}} . \qquad (20)$$

The sequence $\left\lfloor n_w^{1-\sigma}/2 \right\rfloor / n^{1-\sigma}$ tends increasingly to $1/2$ ($n \to \infty$), and therefore there exists a positive constant B_2, such that for all n sufficiently large

$$q \Gamma_n^2 \geq B_2 n^{1-\sigma-2\gamma} . \qquad (21)$$

Since by condition (ii) of Assumption 2, $1/\Gamma_n \leq n^{\gamma}$, for all n sufficiently large

$$1 + 12 \frac{1}{\Gamma_n} \leq 13 \frac{1}{\Gamma_n} \leq 13 n^{\gamma} . \qquad (22)$$

For the other part of the second summand in inequality (19), because of $q \leq n_w^{1-\sigma}/2 \leq n^{1-\sigma}/2$ and (18), we have

$$\begin{aligned} q \alpha_W \left(\left\lfloor \frac{n_w}{2q} \right\rfloor\right) &\leq \frac{1}{2} n^{1-\sigma} \alpha_W(\lfloor n_w^{\sigma} \rfloor) \leq \frac{1}{2} n^{1-\sigma} \alpha(\lfloor n_w^{\sigma} \rfloor - |w| + 1) \\ &\leq \frac{1}{2} C_\alpha \nu n^{1-\sigma} \nu^{(\lfloor n_w^{\sigma} \rfloor - |w|)} . \end{aligned}$$

Now, for $w \in \tau_n$ we have $|w| \leq d_n$ and by condition (i) of Assumption 2, for $\delta \in (0, \sigma)$, $d_n \leq n^{\delta}$. Because $\delta < \sigma$, for all n sufficiently large

$$\lfloor n_w^{\sigma} \rfloor - |w| \geq n^{\sigma/2} . \qquad (23)$$

From (23) follows

$$q \alpha_W \left(\frac{\lfloor n_w \rfloor}{2q}\right) \leq \frac{1}{2} C_\alpha \nu n^{1-\sigma} \nu^{n^{\sigma/2}} . \qquad (24)$$

Since the upper bounds of the inequalities (21), (22) and (24) do not depend on w, the assertion of the lemma follows immediately.

From inequality (17) we obtain

$$\mathbb{P}\left[H_n^c\right] \leq n^{\gamma} \left(4 \exp\left(-B_2 n^{1-\sigma-2\gamma}\right) + 11 \sqrt{13} C_\alpha \nu n^{1-\sigma+\gamma/2} \nu^{n^{\sigma/2}}\right) .$$

Therefore

$$\mathbb{P}\left[H_n^c\right] \; = \; \mathcal{O}\left(\exp\left(-D_2 n^{\xi}\right)\right) ,$$

for some constants $D_2 > 0$, $0 < \xi < \min(1 - \sigma - 2\gamma, \sigma/2)$ and σ, γ as in Assumption 2.

This completes the proof of Lemma 1.

Our next step is to prove that the overestimation of the truncated context tree τ_n by $\widehat{\tau}_n^{tr}$ is increasingly unlikely as the sample size n tends to infinity.

Lemma 4
Let swv be a string with $s \in \tau_n$, $w \in \bigcup_{m=0}^{\infty} \mathcal{X}^m$, $v \in \mathcal{X}$ and $swv \notin \tau_n$. Let $O_n(swv) = \{\Delta_{swv}(x_1^n) \geq C\log(n)\,, N_{x_1^n}(swv) \geq 2\}$. Under the Assumption 1 and condition (iii) of Assumption 2, for all n sufficiently large

$$\mathbb{P}\left[O_n(swv)\right] \leq \frac{\mathbb{P}\left[sw \in \tau_{(0)}\right]}{P_{min}(n)} n^{-C+2|\mathcal{X}|}\,,$$

where $\tau_{(0)}$ is the initial maximal context tree in Step 1 of the Context algorithm.

Proof. We focus our attention on an arbitrary node $s \in \tau_n$. Suppose that u is any string of letters denoted by $su = swv$ with $w \in \cup_{m=0}^{\infty}\mathcal{X}^m$, $v \in \mathcal{X}$ and $swv \notin \tau_n$. That is wv is the extension of the node s to su. The terminal letter of the string u is the letter v. Overestimation of the context s would be the erroneous inclusion of su in $\widehat{\tau}_n^{tr}$. Let $l = |sw|$ be the length of the string sw. We begin by letting the sequence x_1^n be a realization from P. Now, for each string su and each x_1^n our goal is to define a probability law on sequences y_1^n in $\mathcal{X}^n$. To that end, recalling Weinberger *et al.* (1995), first let

$$R_{sw}(y_1^n) \;=\; \sum_{i:\; y_{i-l+1}^i \neq sw} \log\left(P(y_{i+1}|y_1^i)\right)\,,$$

where $\log\left(P(y_1|y_1^0)\right)$ stands for $\log P(y_1)$. Here, $P(y_{i+1}|y_1^i)$ is the true conditional probability of observing the symbol y_{i+1} in the *full* context y_1^i. As in Weinberger *et al.* (1995), we can determine a probability law given by $Q_{su}(y_1^n|x_1^n)$ defined as follows:

$$\begin{aligned}\log\left(Q_{su}(y_1^n|x_1^n)\right) = R_{sw}(y_1^n) + \sum_{x\in\mathcal{X}}\sum_{b\neq v} N_{y_1^n}(xswb)\log\left(\widehat{P}_{x_1^n}(x|sw)\right) \\ + \sum_{x\in\mathcal{X}} N_{y_1^n}(xsu)\log\left(\widehat{P}_{x_1^n}(x|sw)\right)\,.\end{aligned}$$

Thus $Q_{su}(\cdot|x_1^n)$ is a collection of well defined probability measures on sequences of length n indexed by contexts su and sequences x_1^n. An important observation is that for any sequence y_1^n with $N_{y_1^n}(sw) = 0$, i.e. sw does not occur in y_1^n, the Q_{su} probability of y_1^n is the same as the P probability. Strictly speaking every sequence x_1^n and every su indexes a probability measure $Q_{su}(\cdot|x_1^n)$. It is easy to see from the definition that, if z_1^n is a sequence in $\mathcal{X}^n$, then $Q_{su}(\cdot|z_1^n) = Q_{su}(\cdot|x_1^n)$ provided that $N_{z_1^n}(xsw) = N_{x_1^n}(xsw)$ and $N_{z_1^n}(xswv) = N_{x_1^n}(xswv)$. Thus we can establish equivalence classes of sequences whose Q_{su} measures are identical. To this end, 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 $x \in \mathcal{X}$. This implies that

$$\widehat{P}_{x_1^n}(x|sw) = \widehat{P}_{y_1^n}(x|sw)\,,$$

and also that

$$\widehat{P}_{x_1^n}(x|swv) = \widehat{P}_{y_1^n}(x|swv)\,,$$

provided that $y_1^n \in \sigma_{x_1^n}$. At this point we are in a position to consider an overestimation event. Recalling inequality (2), this will occur for any sequence x_1^n such that $\Delta_{su}(x_1^n) \geq K_n = C\log(n)$. From the definition of $\Delta_{su}(x_1^n)$ we know that this means that

$$D\left(\widehat{P}_{x_1^n}(\cdot|swv), \widehat{P}_{x_1^n}(\cdot|sw)\right) N_{x_1^n}(swv) \geq C\log(n)\,,$$

and by explicitly writing the Kullback-Leibler distance specified in (1), we have

$$N_{x_1^n}(swv) \sum_{x\in\mathcal{X}} \widehat{P}_{x_1^n}(x|swv) \log\left(\frac{\widehat{P}_{x_1^n}(x|swv)}{\widehat{P}_{x_1^n}(x|sw)}\right) \geq C\log(n)\,.$$

Collecting the terms, we obtain

$$\sum_{x\in\mathcal{X}} N_{x_1^n}(xswv) \log\left(\widehat{P}_{x_1^n}(x|swv)\right) \geq C\log(n) + \sum_{x\in\mathcal{X}} N_{x_1^n}(xswv) \log\left(\widehat{P}_{x_1^n}(x|sw)\right)\,.$$

Now it follows for x_1^n with $\Delta_{su}(x_1^n) \geq C\log(n)$ and for $y_1^n \in \sigma_{x_1^n}$ that

$$\begin{aligned}
\log\left(Q_{su}(y_1^n|x_1^n)\right) &\geq R_{sw}(y_1^n) + \sum_{x\in\mathcal{X}}\sum_{b\neq v} N_{y_1^n}(xswb)\log\left(\widehat{P}_{y_1^n}(x|sw)\right) \\
&\quad + C\log(n) + \sum_{x\in\mathcal{X}} N_{y_1^n}(xsu)\log\left(\widehat{P}_{y_1^n}(x|sw)\right) \\
&= R_{sw}(y_1^n) + \sum_{x\in\mathcal{X}} N_{y_1^n}(xsw)\log\left(\widehat{P}_{y_1^n}(x|sw)\right) + C\log(n)\,.
\end{aligned}$$

Continuing, we add and subtract the difference between $R_{sw}(y_1^n)$ and $\log\left(P(y_1^n)\right)$. Thus,

$$\begin{aligned}
\log\left(Q_{su}(y_1^n|x_1^n)\right) &\geq R_{sw}(y_1^n) + \sum_{i:\ y_{i-l+1}^i = sw} \log\left(P(y_{i+1}|y_1^i)\right) \\
&\quad - \sum_{i:\ y_{i-l+1}^i = sw} \log\left(P(y_{i+1}|y_1^i)\right) \\
&\quad + \sum_{x\in\mathcal{X}} N_{y_1^n}(xsw)\log\left(\widehat{P}_{y_1^n}(x|sw)\right) + C\log(n) \\
&= \log\left(P(y_1^n)\right) + \sum_{i:\ y_{i-l+1}^i = sw} \log\left(\frac{\widehat{P}_{y_1^n}(y_{i+1}|sw)}{P(y_{i+1}|y_1^i)}\right) + C\log(n)\,.
\end{aligned}$$

Since $\widehat{P}_{y_1^n}(y_{i+1}|sw)$ is the maximum likelihood conditional distribution given sw and y_1^n, it follows that $\widehat{P}_{y_1^n}(y_{i+1}|sw) \geq P(y_{i+1}|sw)$. Thus we have that

$$\log\left(Q_{su}(y_1^n|x_1^n)\right) \geq \log\left(P(y_1^n)\right) + \sum_{i:\ y_{i-l+1}^i = sw} \log\left(\frac{P(y_{i+1}|sw)}{P(y_{i+1}|y_1^i)}\right) + C\log(n)\,.$$

For every $s \in \tau_n$, we have $P(y_{i+1}|sw) = P(y_{i+1}|y_1^i)$, for every i, such that $y_{i-l+1}^i = sw$. Applying this fact we obtain the inequality

$$\log\left(Q_{su}(y_1^n|x_1^n)\right) \geq \log\left(P(y_1^n)\right) + C\log(n)\,.$$

By exponentiating both sides we have this upper bound for $P(y_1^n)$:

$$P(y_1^n) \leq Q_{su}(y_1^n|x_1^n)n^{-C}\,.$$

Since this inequality holds for every $y_1^n \in \sigma_{x_1^n}$, for every x_1^n with $\Delta_{su}(x_1^n) \geq C\log(n)$, we have that

$$P(\sigma_{x_1^n}) \leq Q_{su}(\sigma_{x_1^n}|x_1^n)n^{-C}\,.$$

Now observe that for any su, the event that $\Delta_{su}(x_1^n) > C\log(n)$ is contained in the event that string su occurs at least twice in x_1^n. Thus

$$Q_{su}\left(\Delta_{su}(x_1^n) \geq C\log(n)\right) \leq Q_{su}(y_1^n : N_{su}(y_1^n) > 1|x_1^n)\,.$$

Applying this above we have

$$P(\sigma_{x_1^n}) \leq Q_{su}\left(y_1^n : N_{su}(y_1^n) > 1|x_1^n\right)n^{-C}\,. \tag{25}$$

In order to complete the proof we require a uniform bound on $Q_{su}(y_1^n : N_{su}(y_1^n) > 1|x_1^n)$ (uniformity is with respect to all choices of su and all choices of x_1^n). To find a bound, we construct a new probability distribution Q' on the set of sequences of length n. Our goal for Q' is that it should place almost the same probability on the events $\{y_1^n : N_{su}(y_1^n) > 1\}$ that P places on these events. To that end, for every sequence y_1^n, suppose that sw first occurs at index i and let x_0 be the symbol that occurs after this first occurrence. 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

$$\log\left(Q'_{su}(y_1^n|x_1^n)\right) = \log\left(Q_{su}(y_1^n|x_1^n)\right) + \log\left(P(x_0|y_1^i)\right) - \log\left(\widehat{P}_{x_1^n}(x_0|sw)\right)\,,$$

whereas if $b_0 = v$, we define

$$\log\left(Q'_{su}(y_1^n|x_1^n)\right) = \log\left(Q_{su}(y_1^n|x_1^n)\right) + \log\left(P(x_0|y_1^i)\right) - \log\left(\widehat{P}_{x_1^n}(x_0|swv)\right)\,.$$

The upshot of this construction is to adjust Q_{su} so that it remains identical to P on the longest prefix of y_1^n that contains only 1 occurrence of sw. That is, for all sequences y_1^n such that $N_{sw}(y_1^n) < 2$ it follows that $P(y_1^n) = Q'_{su}(y_1^n|x_1^n)$. It also follows from the definition of Q' that

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

From here we can deduce our uniform upper bound, namely that

$$Q_{su}(y_1^n : N_{su}(y_1^n) > 1|x_1^n) \leq Q'_{su}(y_1^n : N_{su}(y_1^n) > 1|x_1^n)\frac{1}{P_{min}(n)}\,.$$

Substituting this bound into (25) we have, for all x_1^n with $\Delta_{su}(x_1^n) \geq C\log(n)$, that

$$P(\sigma_{x_1^n}) \leq n^{-C}Q'(y_1^n : N_{su}(y_1^n) > 1|x_1^n)\frac{1}{P_{min}(n)}\,.$$

Using the fact that for all x_1^n and all su we know that Q'_{su} and P each attach the same probability to the set of sequences that contain at least two occurrences of su we have

$$P(\sigma_{x_1^n}) \leq n^{-C} P(x_1^n : N_{su}(x_1^n) > 1) \frac{1}{P_{min}(n)} \,.$$

Now observe that since the equivalence classes are defined by numbers of counts, then there can be at most $n^{2|\mathcal{X}|}$ distinct classes $\sigma_{x_1^n}$. Thus we have that

$$\mathbb{P}\left[O_n(su)\right] \leq n^{2|\mathcal{X}|} n^{-C} P(x_1^n : N_{su}(x_1^n) > 1) \frac{1}{P_{min}(n)} \,.$$

Since $\tau_{(0)}$ is defined to be the set of strings that have appeared at least twice in x_1^n, it follows that

$$\mathbb{P}\left[O_n(su)\right] \leq n^{-C+2|\mathcal{X}|} \mathbb{P}\left[su \in \tau_{(0)}\right] \frac{1}{P_{min}(n)} \,.$$

This completes the proof of the lemma.

Lemma 5
Under the Assumptions 1 and 2, for all n sufficiently large

$$\mathbb{P}\left[O_n\right] \leq |\mathcal{X}| \, n^{-\lambda} \,, \tag{26}$$

for some constant $\lambda > 0$.

Proof. Applying Lemma 4, using also condition (iii) of Assumption 2, we obtain the inequality

$$\mathbb{P}\left[O_n\right] \leq \sum_{swv} \mathbb{P}\left[O_n(swv)\right] \leq n^{-C+2|\mathcal{X}|+1} \sum_{swv} \mathbb{P}\left[sw \in \tau_{(0)}\right] \,.$$

Now denote by L the number of subsequences occurring at least twice in the reversed sequence X_1^n. It holds $\mathbb{E}\left[L\right] \leq n^2$, and thus we have that

$$\sum_{swv} \mathbb{P}\left[sw \in \tau_{(0)}\right] \leq |\mathcal{X}| \, \mathbb{E}\left[\sum_{sw} 1_{\{sw \text{ occurs at least twice in } X_1^n\}}\right] \leq |\mathcal{X}| \, \mathbb{E}\left[L\right] \leq |\mathcal{X}| \, n^2 \,.$$

Since from inequality (3), $C > 2\,|\mathcal{X}| + 3$, the assertion of the lemma follows.

By Lemma 1 and Lemma 5, we complete the proof of Theorem 1.

Proof of Theorem 2. (i) We essentially follow the same strategy adopted to bound the underestimation event in Theorem 1. Let $\{\rho_n = n\Gamma_n/2 : n \in \mathbb{N}\}$ and $H_n = \{N_{x_1^n}(w) \geq \rho_n$ for every $w \in \tau_n\}$. Partitioning G_n with H_n leads to $\mathbb{P}\left[G_n\right] \leq \mathbb{P}\left[G_n \cap H_n\right] + \mathbb{P}\left[H_n^c\right]$. Now observe that

$$\mathbb{P}\left[G_n \cap H_n\right] \leq \sum_{k=\rho_n}^{n} \mathbb{P}\left[\left|\widehat{P}_{\widehat{C}_n^{tr}}(x|w) - P(x|w)\right| > \varepsilon, N_{x_1^n}(w) = k\right] \,.$$

For all n sufficiently large, we have $\varepsilon > \sqrt{a_n(k)}$, with $a_n(k)$ as in (12), and thus

$$\mathbb{P}\left[G_n \cap H_n\right] \leq \sum_{k=\rho_n}^{n} \mathbb{P}\left[\left|\widehat{P}_{\widehat{c}_n^{tr}}(x|w) - P(x|w)\right| > \sqrt{a_n(k)}, N_{x_1^n}(w) = k\right] .$$

We now define $A_n = \left\{w \in \tau_n : \widehat{c}_n^{tr}(w) = c_n(w)\right\}$. It follows

$$\begin{aligned}\mathbb{P}\left[\left|\widehat{P}_{\widehat{c}_n^{tr}}(x|w) - P(x|w)\right| > \sqrt{a_n(k)}, N_{x_1^n}(w) = k\right] \ \leq \ \mathbb{P}\left[A_n^c\right] + \\ \mathbb{P}\left[\left|\widehat{P}_{\widehat{c}_n^{tr}}(x|w) - P(x|w)\right| > \sqrt{a_n(k)}, N_{x_1^n}(w) = k, \widehat{c}_n^{tr}(w) = c_n(w)\right] .\end{aligned}$$

Because of Lemma 1 and Lemma 5,

$$\mathbb{P}\left[A_n^c\right] = \mathcal{O}\left(n^{-\lambda}\right) ,$$

for some constant $\lambda > 0$. Using (14) and Lemma 2 we obtain

$$\begin{aligned}&\mathbb{P}\left[\left|\widehat{P}_{\widehat{c}_n^{tr}}(x|w) - P(x|w)\right| > \sqrt{a_n(k)}, N_{x_1^n}(w) = k, \widehat{c}_n^{tr}(w) = c_n(w)\right] \\ &\qquad \leq \sup_{0<p<1} \mathbb{P}\left[\left|\sum_{i=1}^{k} \frac{Y_i}{k} - p\right| > \sqrt{a_n(k)}\right] = \mathcal{O}\left(\exp\left(-D_4 \log(n)^{1+\theta}\right)\right) ,\end{aligned}$$

for some constant $D_4 > 0$ and θ as in Assumption 2. The assertion of the theorem is now immediate.
(ii) Follows from part (i).

Check of Assumptions 1 and 2 for Example 3. In order to verify the Assumption 1, we use the following inequality, for which we refer to Proposition 1, Chapter 1.1 in Doukhan (1994),

$$\alpha(k) \leq \mathbb{E}\left[\sup_{B \in \sigma_k^\infty} \left|P(B|\sigma_{-\infty}^0) - P(B)\right|\right] .$$

We can estimate the difference on the right with a coupling argument: we construct two processes $\{X_t : t \geq 0\}$ and $\{X'_t : t \geq 0\}$ on the same probability space such that $X_t = X'_t$ for all t greater or equal than some random time $\tau < \infty$. The first process $\{X_t : t \geq 0\}$ is our stationary alternating renewal process, and the second process $\{X'_t : t \geq 0\}$ has the distribution $P(\cdot|\sigma_{-\infty}^0)$. This means that $\{X'_t : t \geq 0\}$ is also an alternating renewal process, but the value X'_0 is fixed and the time until the first renewal after zero has a different distribution, depending on $\sigma_{-\infty}^0$. Then we have for any $B \in \sigma_k^\infty$

$$\left|P(B) - P(B|\sigma_{-\infty}^0)\right| = \left|\mathbb{P}\left[\{X_t : t \geq k\} \in B\right] - \mathbb{P}\left[\{X'_t : t \geq k\} \in B\right]\right| \leq \mathbb{P}\left[\tau > k\right] ,$$

compare (2.3) and (2.4), Chapter VI in Asmussen (1987). Hence it is sufficient to show that such a coupling exists with $\mathbb{E}\left[\exp(\epsilon\tau)\right] < \infty$ for some $\epsilon > 0$. This can be achieved with the same arguments as in Theorem 2.3 and Lemma 2.5, Chapter VI in Asmussen

(1987). Since we are in discrete time, some simplifications occur. It is sufficient to assume that the distribution of T is non-lattice and has exponential tails.

In order to deal with the alternating processes, we first construct the time points of switches from zero to one (which form an ordinary renewal process). In a second step, we finally determine the other switches.

To show that the conditions stated in Assumption 2 hold, we first prove a lemma.

Lemma 6
For the process $\{X_t : t \geq 0\}$ described in Example 3, for $k \geq 1$,

$$\begin{aligned}\mathbb{P}\left[X_t = X_{t-1} = \ldots = X_{t-k} = 1\right] &= \mathbb{P}\left[X_t = X_{t-1} = \ldots = X_{t-k} = 0\right] \\ &= \frac{1}{2\mu}\left(\frac{c_1}{(1-\rho_1)^2}\rho_1^{k+1} + \frac{c_2}{(1-\rho_2)^2}\rho_2^{k+1}\right) .\end{aligned}$$

Proof. Using the Renewal Theorem, stated in Theorem 4.3 in Asmussen (1987), we obtain

$$\begin{aligned}\mathbb{P}\left[X_t = X_{t-1} = \ldots = X_{t-k} = 1\right] &= \frac{1}{2}\,\mathbb{P}\left[\text{Excess} > k\right] \\ &= \frac{1}{2\mu}\sum_{j=k}^{\infty}\mathbb{P}\left[T > j\right] = \frac{1}{2\mu}\sum_{j=k+1}^{\infty}\mathbb{P}\left[T \geq j\right] .\end{aligned}$$

By means of the summation formula for geometric series we have

$$\mathbb{P}\left[T \geq k\right] = \sum_{j=k}^{\infty}\mathbb{P}\left[T = j\right] = \frac{c_1}{1-\rho_1}\rho_1^k + \frac{c_2}{1-\rho_2}\rho_2^k .$$

Using the same formula once again, the lemma follows.

For what follows we also need to calculate

$$\begin{aligned}&\mathbb{P}\left[X_t = X_{t-1} = \ldots = X_{t-k+1} = 1, X_{t-k} = 0\right] \\ &\qquad = \mathbb{P}\left[X_t = X_{t-1} = \ldots = X_{t-k+1} = 0, X_{t-k} = 1\right] .\end{aligned}$$

But this can be written as

$$\begin{aligned}&\mathbb{P}\left[X_t = X_{t-1} = \ldots = X_{t-k+1} = 1, X_{t-k} = 0\right] \\ &\qquad = \mathbb{P}\left[X_t = \ldots = X_{t-k+1} = 1\right] - \mathbb{P}\left[X_t = \ldots = X_{t-k} = 1\right] ,\end{aligned}$$

and therefore from Lemma 6 we have

$$\mathbb{P}\left[X_t = X_{t-1} = \ldots = X_{t-k+1} = 1, X_{t-k} = 0\right] = \frac{1}{2\mu}\left(\frac{c_1}{1-\rho_1}\rho_1^k + \frac{c_2}{1-\rho_2}\rho_2^k\right) . \qquad (27)$$

Now, for the first inequality in condition (ii) of Assumption 2, from Lemma 6 and (27) it follows immediately, that for a constant $C_\Gamma > 0$,

$$\Gamma_n \geq C_\Gamma\,\rho_1^{d_n} , \qquad (28)$$

because

$$\tau_n = \{\underbrace{1 \cdots 1}_{d_n}, \underbrace{0 \cdots 0}_{d_n}, \underbrace{1 \cdots 10}_{k}, \underbrace{0 \cdots 01}_{k}\}\,,$$

with $2 \leq k \leq d_n$. Consequently, the sequence $\{d_n : n \in \mathbb{N}\}$ has to be of logarithmic order.

For the second inequality in condition (ii) of Assumption 2, for a constant $C_\Upsilon > 0$, holds

$$\Upsilon_n \geq C_\Upsilon \left(\frac{\rho_2}{\rho_1}\right)^{d_n} . \tag{29}$$

To prove this, first note that for $wu \in \tau_n$ and $u \in \{0, 1\}$ holds $\|P(\cdot|wu) - P(\cdot|w)\|_1 = 2\,|P(1|wu) - P(1|w)|$. For simplicity we concentrate on $wu = 1 \cdots 1$ (of length d_n), but the same arguments also apply for the other states in τ_n. In the following calculation, by wu (resp. w) we mean the sequence $1 \cdots 1$ of length d_n (resp. $d_n - 1$). We have

$$\begin{aligned}
&|P(1|wu) - P(1|w)| \\
&\quad = |\mathbb{P}\left[X_t = 1|X_{t-1} = \ldots = X_{t-d_n} = 1\right] - \mathbb{P}\left[X_t = 1|X_{t-1} = \ldots = X_{t-d_n+1} = 1\right]| \\
&\quad = \left|\frac{\mathbb{P}\left[X_t = \ldots = X_{t-d_n} = 1\right]}{\mathbb{P}\left[X_t = \ldots = X_{t-d_n+1} = 1\right]} - \frac{\mathbb{P}\left[X_t = \ldots = X_{t-d_n+1} = 1\right]}{\mathbb{P}\left[X_t = \ldots = X_{t-d_n+1} = 2\right]}\right|
\end{aligned}$$

Using Lemma 6, we obtain

$$|P(1|wu) - P(1|w)| = \rho_1 \left(\frac{1 + c_3 q^{d_n+1}}{1 + c_3 q^{d_n}} - \frac{1 + c_3 q^{d_n}}{1 + c_3 q^{d_n-1}}\right) ,$$

where $q = \rho_2/\rho_1$ and $c_3 = (c_2(1-\rho_1)^2)/(c_1(1-\rho_2)^2)$. This can now be rewritten as

$$\begin{aligned}
&|P(1|wu) - P(1|w)| \\
&\quad = \rho_1\big((1 + c_3 q^{d_n+1})(1 - c_3 q^{d_n}) - (1 + c_3 q^{d_n})(1 - c_3 q^{d_n-1}) + \mathcal{O}\left(q^{2d_n}\right)\big) \\
&\quad = \rho_1 c_3 (1-q)^2 q^{d_n-1} + o\left(q^{d_n}\right)
\end{aligned}$$

and thus finally

$$|P(1|wu) - P(1|w)| = c_3 \left(\frac{1}{\rho_2}\right) (\rho_1 - \rho_2)^2 \left(\frac{\rho_2}{\rho_1}\right)^{d_n} + o\left(q^{d_n}\right) .$$

From inequalities (28) and (29) follows

$$\Upsilon_n^2 \Gamma_n^{(1-\sigma)^2/2} \geq C_\Upsilon^2 C_\Gamma^{(1-\sigma)^2/2} \left(\rho_1^{(1-\sigma)^2/2} \left(\frac{\rho_2}{\rho_1}\right)^2\right)^{d_n} . \tag{30}$$

The term on the right side of inequality (30) has to be greater than $\log(n)^{1+\theta}/n^{1-\sigma}$. With

$$d_n = C \log(n)\,, \quad \text{for a positive constant } C \text{ small enough}\,,$$

this condition, and also condition (ii) of Assumption 2 are satisfied.

Because $P_{min}(n)$ is greater than a positive constant, the condition (iii) of Assumption 2 is obvious.

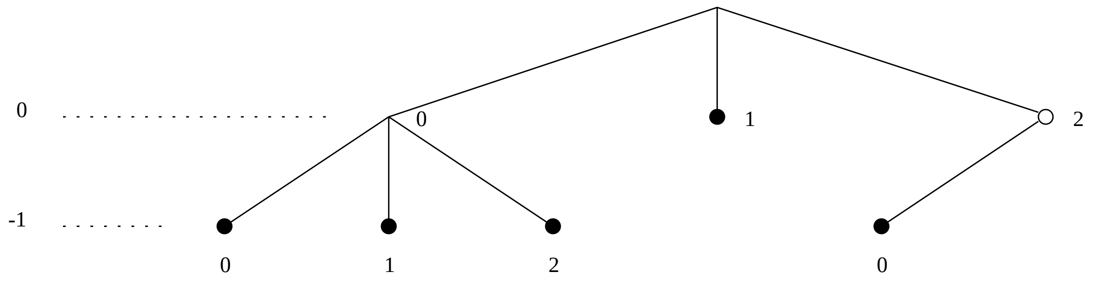

Figure 1: Context tree for Example 1.

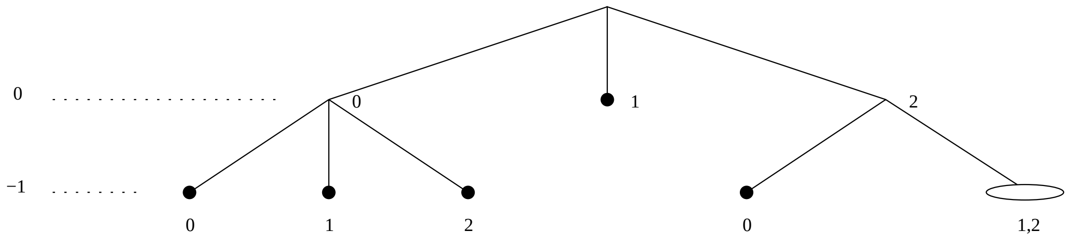

Figure 2: Complete context tree for Example 1.

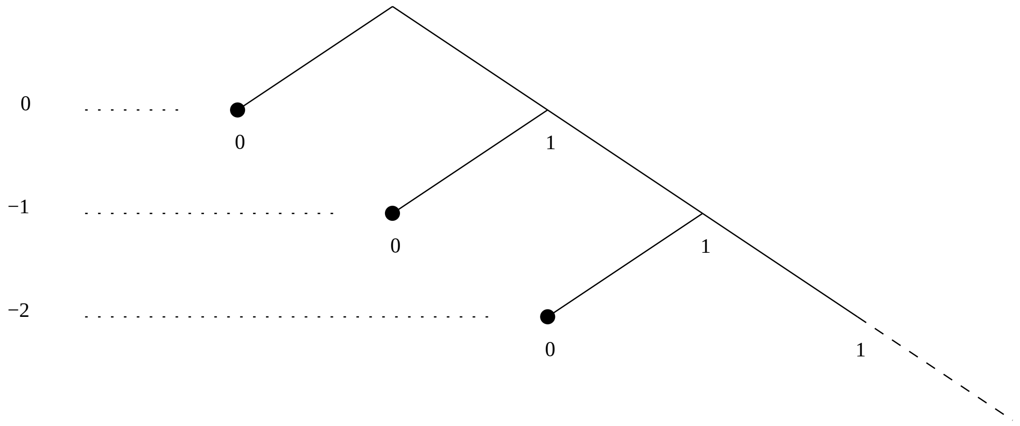

Figure 3: Context tree for Example 2.

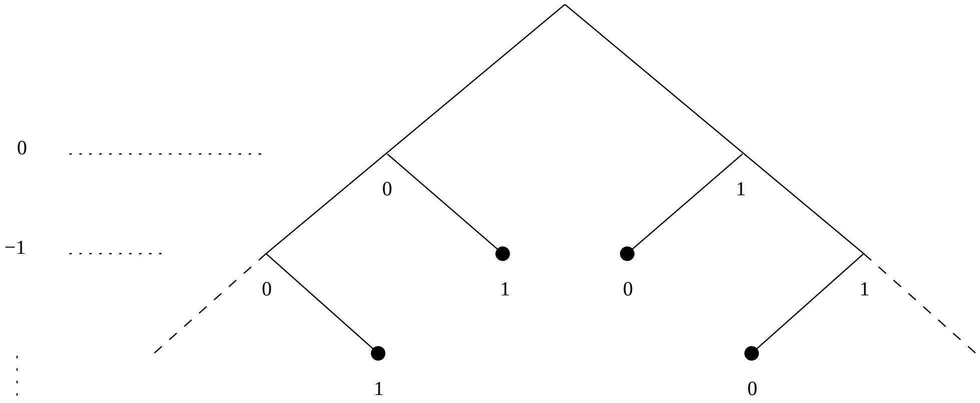

Figure 4: Context tree for Example 3.