Progress in Probability, Vol. 60, 257–269

Exponential Inequalities for Empirical Unbounded Context Trees

Antonio Galves and Florencia Leonardi

Abstract. In this paper we obtain non-uniform exponential upper bounds for the rate of convergence of a version of the algorithm Context, when the underlying tree is not necessarily bounded. The algorithm Context is a well-known tool to estimate the context tree of a Variable Length Markov Chain. As a consequence of the exponential bounds we obtain a strong consistency result. We generalize in this way several previous results in the field.

Mathematics Subject Classification (2000). 62M09, 60G99.

Keywords. Variable memory processes, unbounded context trees, algorithm Context.

1. Introduction

In this paper we present an exponential bound for the rate of convergence of the algorithm Context for a class of unbounded variable memory models, taking values on a finite alphabet A. From this it follows a strong consistency result for the algorithm Context in this setting. Variable memory models were first introduced in the information theory literature by Rissanen [11] as a universal system for data compression. Originally called by Rissanen *finite memory source* or *probabilistic tree*, this class of models recently became popular in the statistics literature under the name of *Variable Length Markov Chains (VLMC)* [1].

The idea behind the notion of variable memory models is that the probabilistic definition of each symbol only depends on a finite part of the past and the length of this relevant portion is a function of the past itself. Following Rissanen we call

This work is part of PRONEX/FAPESP's project *Stochastic behavior, critical phenomena and rhythmic pattern identification in natural languages* (grant number 03/09930-9) and CNPq's projects *Stochastic modeling of speech* (grant number 475177/2004-5) and *Rhythmic patterns, prosodic domains and probabilistic modeling in Portuguese Corpora* (grant number 485999/2007-2). AG is partially supported by a CNPq fellowship (grant 308656/2005-9) and FL is supported by a FAPESP fellowship (grant 06/56980-0).

context the minimal relevant part of each past. The set of all contexts satisfies the suffix property which means that no context is a proper suffix of another context. This property allows to represent the set of all contexts as a rooted labeled tree. With this representation the process is described by the tree of all contexts and a associated family of probability measures on A, indexed by the tree of contexts. Given a context, its associated probability measure gives the probability of the next symbol for any past having this context as a suffix. From now on the pair composed by the context tree and the associated family of probability measures will be called *probabilistic context tree*.

Rissanen not only introduced the notion of variable memory models but he also introduced the algorithm Context to estimate the probabilistic context tree. The way the algorithm Context works can be summarized as follows. Given a sample produced by a chain with variable memory, we start with a maximal tree of candidate contexts for the sample. The branches of this first tree are then pruned until we obtain a minimal tree of contexts well adapted to the sample. We associate to each context an estimated probability transition defined as the proportion of time the context appears in the sample followed by each one of the symbols in the alphabet. From Rissanen [11] to Galves et al. [10], passing by Ron et al. [12] and Bühlmann and Wyner [1], several variants of the algorithm Context have been presented in the literature. In all the variants the decision to prune a branch is taken by considering a *cost* function. A branch is pruned if the cost function assumes a value smaller than a given threshold. The estimated context tree is the smallest tree satisfying this condition. The estimated family of probability transitions is the one associated to the minimal tree of contexts.

In his seminal paper Rissanen proved the weak consistency of the algorithm Context in the case where the contexts have a bounded length, i.e., where the tree of contexts is finite. Bühlmann and Wyner [1] proved the weak consistency of the algorithm also in the finite case without assuming a priori known bound on the maximal length of the memory, but using a bound allowed to grow with the size of the sample. In both papers the cost function is defined using the log likelihood ratio test to compare two candidate trees and the main ingredient of the consistency proofs was the chi-square approximation to the log likelihood ratio test for Markov chains of fixed order. A different way to prove the consistency in the finite case was introduced in [10], using exponential inequalities for the estimated transition probabilities associated to the candidate contexts. As a consequence they obtain an exponential upper bound for the rate of convergence of their variant of the algorithm Context.

The unbounded case as far as we know was first considered by Ferrari and Wyner [8] who also proved a weak consistency result for the algorithm Context in this more general setting. The unbounded case was also considered by Csiszár and Talata [3] who introduced a different approach for the estimation of the probabilistic context tree using the Bayesian Information Criterion (BIC) as well as the Minimum Description Length Principle (MDL). We refer the reader to this last paper for a nice description of other approaches and results in this field, including

the context tree maximizing algorithm by Willems et al. [14]. With exception of Weinberger et al. [13], the issue of the rate of convergence of the algorithm estimating the probabilistic context tree was not addressed in the literature until recently. Weinberger et al. proved in the bounded case that the probability that the estimated tree differs from the finite context tree generating the sample is summable as a function of the sample size. Duarte et al. in [6] extends the original weak consistency result by Rissanen [11] to the unbounded case. Assuming weaker hypothesis than [8], they showed that the on-line estimation of the context function decreases as the inverse of the sample size.

In the present paper we generalize the exponential inequality approach presented in [10] to obtain an exponential upper bound for the algorithm Context in the case of unbounded probabilistic context trees. Under suitable conditions, we prove that the truncated estimated context tree converges exponentially fast to the tree generating the sample, truncated at the same level. This improves all results known until now.

The paper is organized as follows. In section 2 we give the definitions and state the main results. Section 3 is devoted to the proof of an exponential bound for conditional probabilities, for unbounded probabilistic context trees. In section 4 we apply this exponential bound to estimate the rate of convergence of our version of the algorithm Context and to prove its consistency.

2. Definitions and results

In what follows A will represent a finite alphabet of size $|A|$. Given two integers $m \leq n$, we will denote by w_m^n the sequence $(w_m, \ldots, w_n)$ of symbols in A. The length of the sequence w_m^n is denoted by $\ell(w_m^n)$ and is defined by $\ell(w_m^n) = n-m+1$. Any sequence w_m^n with $m > n$ represents the empty string and is denoted by λ. The length of the empty string is $\ell(\lambda) = 0$.

Given two finite sequences w and v, we will denote by vw the sequence of length $\ell(v) + \ell(w)$ obtained by concatenating the two strings. In particular, $\lambda w = w\lambda = w$. The concatenation of sequences is also extended to the case in which v denotes a semi-infinite sequence, that is $v = v_{-\infty}^{-1}$.

We say that the sequence s is a *suffix* of the sequence w if there exists a sequence u, with $\ell(u) \geq 1$, such that $w = us$. In this case we write $s \prec w$. When $s \prec w$ or $s = w$ we write $s \preceq w$. Given a sequence w we denote by $\mathrm{suf}(w)$ the largest suffix of w.

In the sequel A^j will denote the set of all sequences of length j over A and A^* represents the set of all finite sequences, that is

$$A^* = \bigcup_{j=1}^{\infty} A^j.$$

Definition 2.1. A countable subset $\mathcal{T}$ of A^* is a *tree* if no sequence $s \in \mathcal{T}$ is a suffix of another sequence $w \in \mathcal{T}$. This property is called the *suffix property*.

We define the *height* of the tree $\mathcal{T}$ as

$$h(\mathcal{T}) = \sup\{\ell(w) : w \in \mathcal{T}\}.$$

In the case $h(\mathcal{T}) < +\infty$ it follows that $\mathcal{T}$ has a finite number of sequences. In this case we say that $\mathcal{T}$ is *bounded* and we will denote by $|\mathcal{T}|$ the number of sequences in $\mathcal{T}$. On the other hand, if $h(\mathcal{T}) = +\infty$ then $\mathcal{T}$ has a countable number of sequences. In this case we say that the tree $\mathcal{T}$ is *unbounded*.

Given a tree $\mathcal{T}$ and an integer K we will denote by $\mathcal{T}|_K$ the tree $\mathcal{T}$ *truncated* to level K, that is

$$\mathcal{T}|_K = \{w \in \mathcal{T} \colon \ell(w) \le K\} \cup \{w \colon \ell(w) = K \text{ and } w \prec u, \text{ for some } u \in \mathcal{T}\}.$$

We will say that a tree is *irreducible* if no sequence can be replaced by a suffix without violating the suffix property. This notion was introduced in [3] and generalizes the concept of complete tree.

Definition 2.2. A *probabilistic context tree over* A is an ordered pair $(\mathcal{T}, p)$ such that

1. $\mathcal{T}$ is an irreducible tree;
2. $p = \{p(\cdot|w); w \in \mathcal{T}\}$ is a family of transition probabilities over A.

Consider a stationary stochastic chain $(X_t)_{t\in\mathbb{Z}}$ over A. Given a sequence $w \in A^j$ we denote by

$$p(w) \;=\; \mathbb{P}(X_1^j = w)$$

the stationary probability of the cylinder defined by the sequence w. If $p(w) > 0$ we write

$$p(a|w) \;=\; \mathbb{P}(X_0 = a \mid X_{-j}^{-1} = w)\,.$$

Definition 2.3. A sequence $w \in A^j$ is a *context* for the process (X_t) if $p(w) > 0$ and for any semi-infinite sequence $x_{-\infty}^{-1}$ such that w is a suffix of $x_{-\infty}^{-1}$ we have that

$$\mathbb{P}(X_0 = a \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) \;=\; p(a|w), \quad \text{for all } a \in A,$$

and no suffix of w satisfies this equation.

Definition 2.4. We say that the process (X_t) is *compatible* with the probabilistic context tree $(\mathcal{T}, \bar{p})$ if the following conditions are satisfied

1. $w \in \mathcal{T}$ if and only if w is a context for the process (X_t).
2. For any $w \in \mathcal{T}$ and any $a \in A$, $\bar{p}(a|w) = \mathbb{P}(X_0 = a \mid X_{-\ell(w)}^{-1} = w)$.

Define the sequence $(\alpha_k)_{k\in\mathbb{N}}$ as

$$\alpha_0 := \sum_{a\in A} \inf_{w\in\mathcal{T}} \{\, p(a|w) \,\},$$

$$\alpha_k := \inf_{u\in A^k} \sum_{a\in A} \inf_{w\in\mathcal{T}, w\succ u} \{\, p(a|w) \,\}.$$

From now on we will assume that the probabilistic context tree $(\mathcal{T}, p)$ satisfies the following assumptions.

Assumption 2.5. Non-nullness, that is $\inf_{w\in\mathcal{T}}\{p(a|w)\} > 0$ for any $a \in A$.

Assumption 2.6. Summability of the sequence $(1-\alpha_k), k \geq 0$. In this case denote by

$$\alpha := \sum_{k\in\mathbb{N}}(1-\alpha_k) \ < +\infty.$$

For a probabilistic context tree satisfying Assumptions 2.5 and 2.6, the maximal coupling argument used in [7], or alternatively the perfect simulation scheme presented in [2], imply the uniqueness of the law of the chain compatible with it.

Given an integer $k \geq 1$ we define

$$\mathcal{C}_k = \{u \in \mathcal{T}|_k : p(a|u) \neq p(a|\mathrm{suf}(u)) \text{ for some } a \in A\}$$

and

$$D_k = \min_{u\in\mathcal{C}_k}\max_{a\in A}\{\,|p(a|u) - p(a|\mathrm{suf}(u))|\,\}.$$

We denote by

$$\epsilon_k = \min\{\,p(w)\colon \ell(w) \leq k \text{ and } p(w) > 0\,\}.$$

In what follows we will assume that $x_0, x_1, \ldots, x_{n-1}$ is a sample of the stationary stochastic chain (X_t) compatible with the probabilistic context tree $(\mathcal{T}, p)$.

For any finite string w with $\ell(w) \leq n$, we denote by $N_n(w)$ the number of occurrences of the string in the sample; that is

$$N_n(w) = \sum_{t=0}^{n-\ell(w)} \mathbf{1}\{X_t^{t+\ell(w)-1} = w\}.$$

For any element $a \in A$, the empirical transition probability $\hat{p}_n(a|w)$ is defined by

$$\hat{p}_n(a|w) = \frac{N_n(wa)+1}{N_n(w\cdot)+|A|}. \tag{2.7}$$

where

$$N_n(w\cdot) = \sum_{b\in A} N_n(wb)\,.$$

This definition of $\hat{p}_n(a|w)$ is convenient because it is asymptotically equivalent to $\frac{N_n(wa)}{N_n(w\cdot)}$ and it avoids an extra definition in the case $N_n(w\cdot) = 0$.

A variant of Rissanen's *Algorithm Context* is defined as follows. First of all, let us define for any finite string $w \in A^*$:

$$\Delta_n(w) = \max_{a\in A}|\hat{p}_n(a|w) - \hat{p}_n(a|\mathrm{suf}(w))|.$$

The $\Delta_n(w)$ operator computes a distance between the empirical transition probabilities associated to the sequence w and the one associated to the sequence $\mathrm{suf}(w)$.

Definition 2.8. Given $\delta > 0$ and $d < n$, the tree estimated with the Algorithm Context is

$$\hat{\mathcal{T}}_n^{\delta,d} = \{w \in A_1^d : N_n(aw\cdot) > 0, \Delta_n(a\,\mathrm{suf}(w)) > \delta \text{ for some } a \in A \text{ and}$$
$$\Delta_n(uw) \leq \delta \text{ for all } u \in A_1^{d-\ell(w)} \text{ with } N_n(uw\cdot) \geq 1\},$$

where A_1^r denotes the set of all sequences of length at most r. In the case $\ell(w) = d$ we have $A_1^{d-\ell(w)} = \emptyset$.

It is easy to see that $\hat{\mathcal{T}}_n^{\delta,d}$ is an irreducible tree. Moreover, the way we defined $\hat{p}_n(\cdot|\cdot)$ in (2.7) associates a probability distribution to each sequence in $\hat{\mathcal{T}}_n^{\delta,d}$.

The main result in this article is the following

Theorem 2.9. *Let $(\mathcal{T}, p)$ be a probabilistic context tree satisfying Assumptions* 2.5 *and* 2.6 *and let (X_t) be a stationary stochastic chain compatible with $(\mathcal{T}, p)$. Then for any integer K, any d satisfying*

$$d > \max_{u \notin \mathcal{T}, \ell(u) \leq K} \min\{k : \exists w \in \mathcal{C}_k, w \succ u\}, \tag{2.10}$$

any $\delta < D_d$ and any

$$n > \frac{2(|A|+1)}{\min(\delta, D_d - \delta)\epsilon_d} + d$$

we have that

$$\mathbb{P}(\hat{\mathcal{T}}_n^{\delta,d}|_K \neq \mathcal{T}|_K) \;\leq\; 4\, e^{\frac{1}{e}}\, |A|^{d+2} \exp\left[-(n-d)\, \frac{[\min(\frac{\delta}{2}, \frac{D_d-\delta}{2}) - \frac{|A|+1}{(n-d)\epsilon_d}]^2 \epsilon_d^2 C}{4|A|^2(d+1)}\right],$$

where

$$C = \frac{\alpha_0}{8e(\alpha+\alpha_0)}.$$

As a consequence we obtain the following strong consistency result.

Corollary 2.11. *Under the conditions of Theorem* 2.9 *we have*

$$\hat{\mathcal{T}}_n^{\delta,d}|_K = \mathcal{T}|_K,$$

eventually almost surely as $n \to +\infty$.

3. Exponential inequalities for empirical probabilities

The main ingredient in the proof of Theorem 2.9 is the following exponential upper bound

Theorem 3.1. *For any finite sequence w, any symbol $a \in A$ and any $t > 0$ the following inequality holds*

$$\mathbb{P}(\,|N_n(wa) - (n-\ell(w))p(wa)| \;>\; t\,) \;\leq\; e^{\frac{1}{e}} \exp\left[\frac{-t^2 C}{(n-\ell(w))\ell(wa)}\right],$$

where

$$C = \frac{\alpha_0}{8e(\alpha+\alpha_0)}. \tag{3.2}$$

As a direct consequence of Theorem 3.1 we obtain the following corollary.

Corollary 3.3. *For any finite sequence w with $p(w) > 0$, any symbol $a \in A$, any $t > 0$ and any $n > \frac{|A|+1}{tp(w)} + \ell(w)$ the following inequality holds*

$$\mathbb{P}\big(|\hat{p}_n(a|w) - p(a|w)| > t\big) \leq 2\,|A|\,e^{\frac{1}{e}} \exp\left[-(n-\ell(w))\frac{[t - \frac{|A|+1}{(n-\ell(w))p(w)}]^2 p(w)^2 C}{4|A|^2 \ell(wa)}\right],$$

where C is given by (3.2).

To prove Theorem 3.1 we need a mixture property for processes compatible with a probabilistic context tree $(\mathcal{T}, p)$ satisfying Assumptions 2.5 and 2.6. This is the content of the following lemma.

Lemma 3.4. *Let (X_t) be a stationary stochastic chain compatible with the probabilistic context tree $(\mathcal{T}, p)$ satisfying Assumptions* 2.5 *and* 2.6*. Then, there exists a summable sequence $\{\rho_l\}_{l\in\mathbb{N}}$, satisfying*

$$\sum\nolimits_{l\in\mathbb{N}} \rho_l \;\leq\; 1 + \frac{2\alpha}{\alpha_0}, \tag{3.5}$$

such that for any $i \geq 1$, any $k > i$, any $j \geq 1$ and any finite sequence w_1^j, the following inequality holds

$$\sup_{x_1^i \in A^i} |\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_1^i = x_1^i) - p(w_1^j)| \;\leq\; \sum\nolimits_{l=0}^{j-1} \rho_{k-i-1+l}\,. \tag{3.6}$$

Proof. First note that

$$\inf_{u\in A^\infty} \mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^i = u_{-\infty}^0 x_1^i) \;\leq\; \mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_1^i = x_1^i) \leq \sup_{u\in A^\infty} \mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^i = u_{-\infty}^0 x_1^i).$$

where A^∞ denotes the set of all semi-infinite sequences $u_{-\infty}^0$. The reader can find a proof of the inequalities above in [7, Proposition 3]. Using this fact and the condition of stationarity it is sufficient to prove that for any $k \geq 0$,

$$\sup_{x\in A^\infty} |\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) - p(w_1^j)| \leq \sum\nolimits_{l=0}^{j-1} \rho_{k+l}.$$

Note that for all pasts $x_{-\infty}^{-1}$ we have

$$\begin{aligned}
&\big|\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) - p(w_1^j)\big| \\
&\qquad = \Big|\int_{u\in A^\infty} \big[\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) \\
&\qquad\qquad - \mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = u_{-\infty}^{-1})\big] dp(u)\Big| \\
&\qquad \leq \int_{u\in A^\infty} \big|\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) \\
&\qquad\qquad - \mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = u_{-\infty}^{-1})\big|\, dp(u).
\end{aligned}$$

Therefore, applying the loss of memory property proved in [2, Corollary 4.1] we have that

$$\left|\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) - \mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = u_{-\infty}^{-1})\right| \leq \sum_{l=0}^{j-1} \rho_{k+l},$$

where ρ_m is defined as the probability of return to the origin at time m of the Markov chain on $\mathbb{N}$ starting at time zero at the origin and having transition probabilities

$$p(x,y) = \begin{cases} \alpha_x, & \text{if y = x+1,} \\ 1-\alpha_x, & \text{if y=0,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

This concludes the proof of (3.6). To prove (3.5), let(Z_n) be the Markov chain with probability transitions given by (3.7). By definition, we have

$$\begin{aligned} \prod_{l\geq 1}(1-\rho_l) &= \prod_{l\geq 1}\sum_{j=1}^{l} \mathbb{P}(Z_l = j | Z_{l-1} = j-1)\mathbb{P}(Z_{l-1} = j-1) \\ &\geq \prod_{l\geq 1} \alpha_{l-1} \prod_{i=0}^{l-2} \alpha_i \prod_{l\geq 0} \alpha_l^2. \end{aligned}$$

From this, using the inequality $x \leq -\ln(1-x) \leq \frac{x}{1-c}$ which holds for any $x \in (-1, c]$, it follows that

$$\sum_{l\geq 1} \rho_l \leq \sum_{l\geq 0} \log \alpha_l \leq \sum_{l\geq 0} \frac{1-\alpha_l}{\alpha_0}.$$

This concludes the proof of the lemma. □

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Let w be a finite sequence and a any symbol in A. Define the random variables

$$U_j = \mathbf{1}\{X_j^{j+\ell(w)} = wa\} - p(wa),$$

for $j = 0, \ldots, n-\ell(wa)$. Then, using [4, Proposition 4] we have that, for any $p \geq 2$

$$\begin{aligned} &\|N_n(wa) - (n-\ell(w))p(wa)\|_p \\ &\leq \left(2p \sum_{i=0}^{n-\ell(wa)} \sum_{k=i}^{n-\ell(wa)} \| \mathbb{E}(U_k \mid U_0, \ldots, U_i) \|_\infty\right)^{\frac{1}{2}} \\ &\leq \left(2p \sum_{i=0}^{n-\ell(wa)} \sum_{k=i}^{n-\ell(wa)} \sup_{u\in A^{i+\ell(wa)}} |\mathbb{P}(X_k^{k+\ell(w)} = wa \mid X_0^{i+\ell(w)} = u) - p(wa)|\right)^{\frac{1}{2}} \\ &\leq \left(2p\,\ell(wa)(n-\ell(w))\frac{2(\alpha+\alpha_0)}{\alpha_0}\right)^{\frac{1}{2}}. \end{aligned}$$

Then, as in [5, Proposition 5] we also obtain that, for any $t > 0$,

$$\mathbb{P}(|N_n(wa) - (n - \ell(w))p(wa)| > t) \;\le\; e^{\frac{1}{e}} \exp\Big[\frac{-t^2 C}{(n-\ell(w))\ell(wa)}\Big],$$

where

$$C = \frac{\alpha_0}{8e(\alpha + \alpha_0)}. \qquad \square$$

Proof of Corollary 3.3. First observe that

$$\Big| p(a|w) - \frac{(n-\ell(w))p(wa)+1}{(n-\ell(w))p(w)+|A|} \Big| \;\le\; \frac{|A|+1}{(n-\ell(w))p(w)}.$$

Then, for all $n \ge (|A|+1)/tp(w) + \ell(w)$ we have that

$$\begin{aligned}
&\mathbb{P}\big(\big|\hat{p}_n(a|w) - p(a|w)\big| > t \big) \\
&\qquad \le\; \mathbb{P}\Big(\Big| \frac{N_n(wa)+1}{N_n(w\cdot)+|A|} - \frac{(n-\ell(w))p(wa)+1}{(n-\ell(w))p(w)+|A|} \Big| > t - \frac{|A|+1}{(n-\ell(w))p(w)} \Big)
\end{aligned}$$

Denote by $t' = t - (|A|+1)/(n-\ell(w))p(w)$. Then

$$\begin{aligned}
&\mathbb{P}\Big(\Big| \frac{N_n(wa)+1}{N_n(w\cdot)+|A|} - \frac{(n-\ell(w))p(wa)+1}{(n-\ell(w))p(w)+|A|} \Big| > t' \Big) \\
&\qquad \le\; \mathbb{P}\big(\big| N_n(wa) - (n-\ell(w))p(wa) \big| > \frac{t'}{2} [(n-\ell(w))p(w) + |A|] \big) \\
&\qquad\quad + \sum_{b \in A} \mathbb{P}\big(\big| N_n(wb) - (n-\ell(w))p(wb) \big| > \frac{t'}{2|A|} [(n-\ell(w))p(w) + |A|] \big).
\end{aligned}$$

Now, we can apply Theorem 3.1 to bound above the last sum by

$$2\,|A|\, e^{\frac{1}{e}} \exp\left[-(n-\ell(w)) \, \frac{[t - \frac{|A|+1}{(n-\ell(w))p(w)}]^2 p(w)^2 C}{4|A|^2 \ell(wa)} \right],$$

where

$$C = \frac{\alpha_0}{8e(\alpha + \alpha_0)}.$$

This finishes the proof of the corollary. $\square$

4. Proof of the main results

Proof of Theorem 2.9. Define

$$O_n^{\delta,d} = \bigcup_{\substack{w \in \mathcal{T} \\ \ell(w) < K}} \bigcup_{uw \in \hat{T}_n^{\delta,d}} \{\Delta_n(uw) > \delta\}, \quad \text{and} \quad U_n^{\delta,d} = \bigcup_{\substack{w \in \hat{T}_n^{\delta,d} \\ \ell(w) < K}} \bigcap_{uw \in \mathcal{T}|_d} \{\Delta_n(uw) \le \delta\}.$$

Then, if $d < n$ we have that

$$\{\hat{\mathcal{T}}_n^{\delta,d}|_K \neq \mathcal{T}|_K\} = O_n^{\delta,d} \cup U_n^{\delta,d}.$$

The result follows from a succession of lemmas.

Lemma 4.1. *For any $n > \frac{2(|A|+1)}{\delta\epsilon_d} + d$, for any $w \in \mathcal{T}$ with $\ell(w) < K$ and for any $uw \in \hat{\mathcal{T}}_n^{\delta,d}$ we have that*

$$\mathbb{P}(\Delta_n(uw) > \delta) \;\leq\; 4\,|A|^2\, e^{\frac{1}{e}} \exp\big[-(n-d)\,\frac{[\frac{\delta}{2} - \frac{|A|+1}{(n-d)\epsilon_d}]^2\epsilon_d^2 C}{4|A|^2(d+1)}\big],$$

where C is given by (3.2).

Proof. Recall that

$$\Delta_n(uw) = \max_{a\in A} |\hat{p}_n(a|uw) - \hat{p}_n(a|\mathrm{suf}(uw))|.$$

Note that the fact $w \in \mathcal{T}$ implies that for any finite sequence u with $p(u) > 0$ and any symbol $a \in A$ we have $p(a|w) = p(a|uw)$. Hence,

$$\begin{aligned}\mathbb{P}(\Delta_n(uw) > \delta) \;\leq\; \sum_{a\in A} \big[\,&\mathbb{P}\big(|\hat{p}_n(a|w) - p(a|w)| > \frac{\delta}{2}\big) \\ &+ \mathbb{P}\big(|\hat{p}_n(a|uw) - p(a|uw)| > \frac{\delta}{2}\big)\big].\end{aligned}$$

Using Corollary 3.3 we can bound above the right-hand side of the last inequality by

$$4\,|A|^2\, e^{\frac{1}{e}} \exp\left[-(n-d)\,\frac{[\frac{\delta}{2} - \frac{|A|+1}{(n-d)\epsilon_d}]^2\epsilon_d^2 C}{4|A|^2(d+1)}\right],$$

where C is given by (3.2). □

Lemma 4.2. *For any $n > \frac{2(|A|+1)}{(D_d-\delta)\epsilon_d} + d$ and for any $w \in \hat{\mathcal{T}}_n^{\delta,d}$ with $\ell(w) < K$ we have that*

$$\mathbb{P}\Big(\bigcap_{uw\in\mathcal{T}|_d} \{\Delta_n(uw) \leq \delta\}\Big) \leq 4\,|A|\, e^{\frac{1}{e}} \exp\left[-(n-d)\,\frac{[\frac{D_d-\delta}{2} - \frac{|A|+1}{(n-d)\epsilon_d}]^2\epsilon_d^2 C}{4|A|^2(d+1)}\right],$$

where C is given by (3.2).

Proof. As d satisfies (2.10) there exists $u\bar{w} \in \mathcal{T}|_d$ such that $p(a|u\bar{w} \neq p(a|\mathrm{suf}\,(u\bar{w}))$ for some $a \in A$. Then

$$\mathbb{P}\Big(\bigcap_{uw\in\mathcal{T}|_d} \{\Delta_n(uw) \leq \delta\}\Big) \;\leq\; \mathbb{P}(\Delta_n(u\bar{w}) \leq \delta).$$

Observe that for any $a \in A$,

$$\begin{aligned}|\hat{p}_n(a|\mathrm{suf}(u\bar{w})) - \hat{p}_n(a|u\bar{w})| \;\geq\; &|p(a|\mathrm{suf}(u\bar{w})) - p(a|u\bar{w})| \\ &- |\hat{p}_n(a|\mathrm{suf}(u\bar{w})) - p(a|\mathrm{suf}(u\bar{w}))| - |\hat{p}_n(a|u\bar{w}) - p(a|u\bar{w})|.\end{aligned}$$

Hence, we have that for any $a \in A$

$$\Delta_n(u\bar{w}) \;\geq\; D_d - |\hat{p}_n(a|\mathrm{suf}(u\bar{w})) - p(a|\mathrm{suf}(u\bar{w}))| - |\hat{p}_n(a|u\bar{w}) - p(a|u\bar{w})|.$$

Therefore,

$$\begin{aligned}\mathbb{P}(\Delta_n(u\bar{w}) \leq \delta) \quad &\leq \quad \mathbb{P}\Big(\bigcap_{a\in A} \{\, |\hat{p}_n(a|\mathrm{suf}(u\bar{w})) - p(a|\mathrm{suf}(u\bar{w}))| \geq \frac{D_d - \delta}{2} \,\} \Big) \\ &\quad + \mathbb{P}\Big(\bigcap_{a\in A} \{\, |\hat{p}_n(a|u\bar{w}) - p(a|u\bar{w})| \geq \frac{D_d - \delta}{2} \,\} \Big).\end{aligned}$$

As $\delta < D_d$ and $n > \frac{2(|A|+1)}{(D_d-\delta)\epsilon_d} + d$ we can use Corollary 3.3 to bound above the right-hand side of this inequality by

$$4\,|A|\,e^{\frac{1}{e}} \exp\left[-(n-d)\, \frac{[\frac{D_d-\delta}{2} - \frac{|A|+1}{(n-d)\epsilon_d}]^2 \epsilon_d^2 C}{4|A|^2(d+1)} \right],$$

where C is given by (3.2). This concludes the proof of the lemma. □

Now we can finish the proof of Theorem 2.9. We have that

$$\mathbb{P}(\hat{\mathcal{T}}_n^{\delta,d}|_K \neq \mathcal{T}|_K) = \mathbb{P}(O_n^{\delta,d}) + \mathbb{P}(U_n^{\delta,d}).$$

Using the definition of $O_n^{\delta,d}$ and $U_n^{\delta,d}$ we have that

$$\mathbb{P}(\hat{\mathcal{T}}_n^{\delta,d}|_K \neq \mathcal{T}|_K) \leq \sum_{\substack{w\in\mathcal{T} \\ \ell(w)<K}} \sum_{uw\in\hat{\mathcal{T}}_n^{\delta,d}} \mathbb{P}(\Delta_n(uw) > \delta) + \sum_{\substack{w\in\hat{\mathcal{T}}_n^{\delta,d} \\ \ell(w)<K}} \mathbb{P}\Big(\bigcap_{uw\in\mathcal{T}|_d} \Delta_n(uw) \leq \delta \Big).$$

Applying Lemma 4.1 and Lemma 4.2 we can bound above the last expression by

$$\mathbb{P}(\hat{\mathcal{T}}_n^{\delta,d}|_K \neq \mathcal{T}|_K) \;\leq\; 4\,e^{\frac{1}{e}}\,|A|^{d+2} \exp\left[-(n-d)\, \frac{[\min(\frac{\delta}{2}, \frac{D_d-\delta}{2}) - \frac{|A|+1}{(n-d)\epsilon_d}]^2 \epsilon_d^2 C}{4|A|^2(d+1)} \right],$$

where C is given by (3.2). We conclude the proof of Theorem 2.9. □

Proof of Corollary 2.11. It follows from Theorem 2.9, using the first Borel-Cantelli Lemma and the fact that the bounds for the error estimation of the context tree are summable in n for a fixed d satisfying (2.10) and $\delta < D_d$. □

5. Final remarks

The present paper presents an upper bound for the rate of convergence of a version of the algorithm Context, for unbounded context trees. This generalizes previous results obtained in [10] for the case of bounded variable memory processes. We obtain an exponential bound for the probability of incorrect estimation of the truncated context tree, when the estimator is given by Definition (2.8). Note that the definition of the context tree estimator depends on the parameter δ, and this parameter appears in the exponent of the upper bound. To assure the consistency

of the estimator we need to choose a δ sufficiently small, depending on the transition probabilities of the process. Therefore, our estimator is not universal, in the sense that for any fixed δ it fails to be consistent for any process having $D_d < \delta$. The same happens with the parameter d. In order to choose δ and d not depending on the process, we can allow these parameters to be a function of n, in such a way δ_n goes to zero and d_n goes to $+\infty$ as n diverges. When we do this, we loose the exponential property of the upper bound.

As an anonymous referee has pointed out, Finesso et al. [9] proved that in the simpler case of estimating the order of a Markov chain, it is not possible to obtain pure exponential bounds for the overestimation event with a universal estimator. The above discussion illustrates this fact.

Acknowledgments

We thank Pierre Collet, Imre Csiszár, Nancy Garcia, Aurélien Garivier, Bezza Hafidi, Véronique Maume-Deschamps, Eric Moulines, Jorma Rissanen and Bernard Schmitt for many discussions on the subject. We also thank an anonymous referee that attracted our attention to the interesting paper [9].

References

[1] P. Bühlmann and A. J. Wyner. Variable length Markov chains. *Ann. Statist.*, 27:480–513, 1999.

[2] F. Comets, R. Fernández, and P. Ferrari. Processes with long memory: Regenerative construction and perfect simulation. *Ann. Appl. Probab.*, 12(3):921–943, 2002.

[3] I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory*, 52(3):1007–1016, 2006.

[4] J. Dedecker and P. Doukhan. A new covariance inequality and applications. *Stochastic Process. Appl.*, 106(1):63–80, 2003.

[5] J. Dedecker and C. Prieur. New dependence coefficients. examples and applications to statistics. *Probab. Theory Related Fields*, 132:203–236, 2005.

[6] D. Duarte, A. Galves, and N.L. Garcia. Markov approximation and consistent estimation of unbounded probabilistic suffix trees. *Bull. Braz. Math. Soc.*, 37(4):581–592, 2006.

[7] R. Fernández and A. Galves. Markov approximations of chains of infinite order. *Bull. Braz. Math. Soc.*, 33(3):295–306, 2002.

[8] F. Ferrari and A. Wyner. Estimation of general stationary processes by variable length Markov chains. *Scand. J. Statist.*, 30(3):459–480, 2003.

[9] L. Finesso, C-C. Liu, and P. Narayan. The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory*, 42(5):1488–1497, 1996.

[10] A. Galves, V. Maume-Deschamps, and B. Schmitt. Exponential inequalities for VLMC empirical trees. *ESAIM Prob. Stat.* (*accepted*), 2006.

[11] J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 29(5):656–664, 1983.
[12] D. Ron, Y. Singer, and N. Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25(2-3):117–149, 1996.
[13] M.J. Weinberger, J. Rissanen, and M. Feder. A universal finite memory source. *IEEE Trans. Inform. Theory*, 41(3):643–652, 1995.
[14] F.M. Willems, Y.M. Shtarkov, and T.J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Trans. Inform. Theory*, IT-44:653–664, 1995.

Antonio Galves and Florencia Leonardi
Instituto de Matemática e Estatística
Universidade de São Paulo
BP 66281, 05315-970
São Paulo, Brasil
e-mail: galves@ime.usp.br

e-mail: leonardi@ime.usp.br