EXPONENTIAL INEQUALITIES FOR EMPIRICAL UNBOUNDED CONTEXT TREES

ANTONIO GALVES AND FLORENCIA G. LEONARDI

Abstract. In this paper we obtain exponential 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.

1. Introduction

In this paper we present an exponential bound for the rate of convergence of the algorithm Context for the 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 general setting. Variable memory models were first introduced in the information theory literature by Rissanen (1983) as a universal system for data compression. Originally called by Rissanen (1983) *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)* (Bühlmann and Wyner; 1999).

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 called this relevant part of each past a *context*. 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

Date: June 6, 2007.

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 project *Stochastic modeling of speech* (grant number 475177/2004-5). AG is partially supported by a CNPq fellowship (grant 308656/2005-9) and FGL is partially supported by a FAPESP fellowship (grant 06/56980-0) .

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 (1983) 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 (1983) to Galves et al. (2006), passing by Ron et al. (1996) and Bühlmann and Wyner (1999), 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 *gain* function. A branch is pruned if the gain 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 (1983) 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 (1999) proved the weak consistency of the algorithm also in the finite case without assuming a prior 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 gain function is defined using the log likelihood ratio test to compare to 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 Galves et al. (2006), using exponential inequalities for the estimated transition probabilites 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 (2003) 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 (2006) 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. (1995). With exception of Weinberger et al. (1995), 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. (1995) 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. (2006) extends the original weak consistency result by Rissanen (1983) to the unbounded case. Assuming weaker hypothesis than Ferrari and Wyner (2003) they showed that the on-line estimation of the context function decreases as the inverse of the sample size. A different estimation procedure, inspired by Csiszár and Talata (2006), was adopted by Leonardi (2007) using a penalized likelihood algorithm to estimate the context tree. This paper proves that the estimated context tree truncated at any fixed height approximates the real truncated tree at a rate that decreases faster than the inverse of an exponential function of the penalizing term. Therefore, even with the largest possible penalizing term the obtained upper bound is summable but decreases sub exponentially fast. The main technical ingredient in this paper is an extension to the unbounded case of the exponential inequalities presented in Galves et al. (2006).

In the present paper we apply the exponential inequality approach presented in Galves et al. (2006) and extended in Leonardi (2007) to obtain an exponential upper bound for the algorithm Context in the case of unbounded probabilistic context trees. 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 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 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

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

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 Csiszár and Talata (2006) 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 $\{X_t \colon t \in \mathbb{Z}\}$ 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 $\{X_t \colon t \in \mathbb{Z}\}$ 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

$$\mathbb{P}(X_0 = a \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) \,=\, p(a|w), \quad \text{for all } a \in A, \tag{2.4}$$

and no suffix of w satisfies this equation.

Definition 2.5. We say that the process $\{X_t \colon t \in \mathbb{Z}\}$ is *compatible* with the probabilistic context tree $(\mathcal{T}, \bar{p})$ if and only if

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

In the unbounded case, the compactness of $A^{\mathbb{Z}}$ assures that there is at least one stationary stochastic chain compatible with a probabilistic context tree. The uniqueness requires further conditions, such as the ones presented in Fernández and Galves (2002).

Definition 2.6. A probabilistic context tree $(\mathcal{T}, p)$ is of *type A* if it satisfies the following conditions

(1) *Weakly non-nullness*, that is

$$\sum_{a\in A} \inf_{w\in\mathcal{T}} p(a|w) > 0;$$

(2) *Continuity*, that is

$$\beta_k \to 0 \text{ when } k \to \infty,$$

where the sequence $\{\beta_k\}_{k\in\mathbb{N}}$ is defined by

$$\beta_k := \sup\{|p(a|w) - p(a|v)| \colon a \in A, v, w \in \mathcal{T} \text{ with } w \stackrel{k}{=} v\}.$$

Here, $w \stackrel{k}{=} v$ means that there exists a sequence u, with $\ell(u) = k$ such that $u \prec w$ and $u \prec v$. The sequence $\{\beta_k\}$ is called *continuity rate*.

For a probabilistic context tree of type A with summable continuity rate, the maximal coupling argument used in Fernández and Galves (2002) implies the uniqueness of the law of the chain consistent with it. Then, we will assume here that the continuity rate is summable, that is

$$\beta := \sum_{k\in\mathbb{N}} \beta_k < +\infty. \tag{2.7}$$

Given an integer $k \geq 1$ we define

$$D_k = \min_{w\in\mathcal{T}:\ell(w)\leq k} \max_{a\in A}\{\, |p(a|w) - p(a|\text{suf}(w))|\,\}, \tag{2.8}$$

and

$$\epsilon_k = \min\{\, p(w) \colon \ell(w) \leq k \text{ and } p(w) > 0\,\}. \tag{2.9}$$

In what follows we will assume that $x_0, x_1, \ldots, x_{n-1}$ is a sample of the stationary stochastic chain $\{X_t : t \in \mathbb{Z}\}$ compatible with the probabilistic context tree $(\mathcal{T}, p)$. In this case we will say that $x_0, x_1, \ldots, x_{n-1}$ is a *realization* of $(\mathcal{T}, p)$.

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

$$N_n(w) = \sum_{t=0}^{n-\ell(w)} \mathbf{1}\{X_t^{t+\ell(w)-1} = w\}. \tag{2.10}$$

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

$$\hat{p}_n(a|w) = \frac{N_n(wa) + 1}{N_n(w\cdot) + |A|}. \tag{2.11}$$

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 2.12. Given $\delta > 0$ and $d < n$, the tree estimated with the algorithm Context is

$$\hat{\mathcal{T}}_n^{\delta,d} = \{w \in A_1^d : \Delta_n(w) > \delta \ \wedge\ \Delta_n(uw) \le \delta \ \ \forall u \in A_1^{d-\ell(w)}\},$$

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 a tree. Moreover, the way we defined $\hat{p}_n(\cdot|\cdot)$ in (2.11) associates a probability distribution to each sequence in $\hat{\mathcal{T}}_n^{\delta,d}$.

The main result in this article is the following

Theorem 2.13. *Let $(\mathcal{T}, p)$ be a probabilistic context tree with summable continuity rate and let $x_0, x_1, \ldots, x_{n-1}$ be a realization of $(\mathcal{T}, p)$. Then for any integer K and any d satisfying*

$$d > \max_{w \in \mathcal{T}|_K} \min\{\ell(v) \colon v \in \mathcal{T}, w \prec v\} \tag{2.14}$$

if $h(\mathcal{T}) > K$ or $d \ge h(\mathcal{T})$ if $h(\mathcal{T}) \le K$, for any $\delta < D_d$ and for each $n > d$ we have that

$$\mathbb{P}(\hat{\mathcal{T}}_n^{\delta,d}|_K \neq \mathcal{T}|_K) \ \le$$

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

where

$$C = \frac{1}{4e(1+\beta)}.$$

As a consequence we obtain the following strong consistency result.

Corollary 2.15. *Let $(\mathcal{T}, p)$ be a probabilistic context tree with summable continuity rate and let $x_0, x_1, \ldots, x_{n-1}$ be a realization of $(\mathcal{T}, p)$. Then for any integer K there exists a $\bar{n}$ such that, for any $n \geq \bar{n}$ we have that*

$$\hat{\mathcal{T}}_n^{\delta,d}|_K = \mathcal{T}|_K, \tag{2.16}$$

where d is given by (2.14) and δ is such that $\delta < D_d$.

3. Exponential inequalities for empirical probabilities

The main ingredient in the proof of Theorem 2.13 is a result of exponential rate of convergence for the empirical transition probabilities. This result was proven in Leonardi (2007), for a little different definition of the empirical transition probabilities given by 2.11. Here we present the proof for our specific setting. The main result in this section is the following

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

$$\mathbb{P}(\,|N_n(w) - (n-\ell(w))p(w)| \,>\, t\,) \,\leq\, e^{\frac{1}{e}} \exp\Big[\frac{-t^2 C}{(n-\ell(w))(\ell(w)+1)}\Big], \tag{3.2}$$

where

$$C = \frac{1}{4e(1+\beta)}. \tag{3.3}$$

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

Corollary 3.4. *For any finite sequence w, with $p(w) > 0$, any symbol $a \in A$ and any $t > 0$ the following inequality holds*

$$\mathbb{P}\big(|\hat{p}_n(a|w) - p(a|w)| > t\big) \leq$$

$$(|A|+1)\, e^{\frac{1}{e}} \exp\Big[-(n-\ell(w)-1)\, \frac{[t - \frac{|A|+1}{(n-\ell(w)-1)p(w)}]^2 [p(w) + \frac{|A|}{n-\ell(w)-1}]^2 C}{4|A|^2(\ell(w)+2)}\Big], \tag{3.5}$$

where C is given by (3.3).

The key property used in the proof of Theorem 3.1 is a mixture property for processes compatible with a probabilistic context tree $(\mathcal{T}, p)$ having summable continuity rate. This property is stated in the following lemma.

Lemma 3.6. *Let $\{X_t \colon t \in \mathbb{Z}\}$ be a stationary stochastic chain compatible with the probabilistic context tree $(\mathcal{T}, p)$ that has summable continuity rate. Then, for any $i \geq 1$, any $k > i$, any $j \geq 1$ and any finite sequence w_1^j, the following inequality holds*

$$\sup_{x_1^i \in A^i} |\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_1^i = x_1^i) - p(w_1^j)| \;\leq\; j\,\beta_{k-i-1}\,. \tag{3.7}$$

Proof. It is easy to see that for any $i \geq 1$,

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

where A^∞ denotes the set of all semi-infinite sequences $u_{-\infty}^0$. The reader can find a proof of the inequalities above in (Fernández and Galves; 2002, Proposition 3). Using this fact and the condition of stationarity it is sufficient to prove that for any $k \geq 0$,

$$\sup_{x \in A^\infty} |\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) - p(w_1^j)| \leq\; j\,\beta_k.$$

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

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

Then, it is enough to show that

$$\big|\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) - \mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = u_{-\infty}^{-1})\big| \leq\; j\,\beta_k. \tag{3.8}$$

We will proceed by induction on j. For $j = 1$ we have that

$$\begin{aligned}
&\left|\mathbb{P}(X_k = w_1 \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) - \mathbb{P}(X_k = w_1 \mid X_{-\infty}^{-1} = u_{-\infty}^{-1})\right| \\
&\qquad \leq \sum_{v \in A^k} \mathbb{P}(X_0^{k-1} = v) \left|\mathbb{P}(X_k = w_1 \mid X_{-\infty}^{k-1} = x_{-\infty}^{-1} v)\right. \\
&\qquad\qquad \left. - \mathbb{P}(X_k = w_1 \mid X_{-\infty}^{k-1} = u_{-\infty}^{-1} v)\right| \\
&\qquad \leq \beta_k.
\end{aligned}$$

Suppose that (3.8) is true for j. We will prove that it is also true for $j+1$. Observe that

$$\begin{aligned}
&\left|\mathbb{P}(X_k^{k+j} = w_1^{j+1} | X_{-\infty}^{-1} = x_{-\infty}^{-1}) - \mathbb{P}(X_k^{k+j} = w_1^{j+1} | X_{-\infty}^{-1} = u_{-\infty}^{-1})\right| \\
&= \left|\mathbb{P}(X_{k+j} = w_{j+1} | X_k^{k+j-1} = w_1^j, X_{-\infty}^{-1} = x_{-\infty}^{-1}) \mathbb{P}(X_k^{k+j-1} = w_1^j | X_{-\infty}^{-1} = x_{-\infty}^{-1})\right. \\
&\left. - \mathbb{P}(X_{k+j} = w_{j+1} | X_k^{k+j-1} = w_1^j, X_{-\infty}^{-1} = u_{-\infty}^{-1}) \mathbb{P}(X_k^{k+j-1} = w_1^j | X_{-\infty}^{-1} = u_{-\infty}^{-1})\right|.
\end{aligned}$$

Summing and subtracting the term

$$\mathbb{P}(X_{k+j} = w_{j+1} \mid X_k^{k+j-1} = w_1^j, X_{-\infty}^{-1} = x_{-\infty}^{-1}) \mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = u_{-\infty}^{-1})$$

we can bound above the right hand side of the last expression by

$$\begin{aligned}
&\left|\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) - \mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_{-\infty}^{-1} = u_{-\infty}^{-1})\right| \\
&\quad + \left|\mathbb{P}(X_{k+j} = w_{j+1} \mid X_k^{k+j-1} = w_1^j, X_{-\infty}^{-1} = x_{-\infty}^{-1})\right. \\
&\qquad\qquad \left. - \mathbb{P}(X_{k+j} = w_{j+1} \mid X_k^{k+j-1} = w_1^j, X_{-\infty}^{-1} = u_{-\infty}^{-1})\right| \\
&\leq j\,\beta_k + \sum_{v \in A^k} \mathbb{P}(X_0^{k-1} = v) \left|\mathbb{P}(X_{k+j} = w_{j+1} \mid X_{-\infty}^{k+j-1} = x_{-\infty}^{-1} v\, w_1^j)\right. \\
&\qquad\qquad \left. - \mathbb{P}(X_{k+j} = w_{j+1} \mid X_{-\infty}^{k+j-1} = u_{-\infty}^{-1} v\, w_1^j)\right| \\
&\leq (j+1)\,\beta_k.
\end{aligned}$$

This concludes the proof of Lemma 3.6. □

We are ready to prove Theorem 3.1. This proof uses strongly the mixture property in Lemma 3.6 and the fact that the continuity rate is summable.

Proof of Theorem 3.1. Let w be a finite sequence. Using the same argument as in Galves et al. (2006, Theorem 2.5) we have that, for any $p \geq 2$ the following inequality holds

$$\begin{aligned}
&\|N_n(w) - (n-\ell(w))p(w)\|_p \\
&\qquad \leq \Big(2p \sum_{i=1}^{n-\ell(w)} \sum_{l=i}^{n-\ell(w)} \sup_{x_0^{i-1} \in A^i} |\mathbb{P}(X_l^{l+\ell(w)} = w \,|\, X_0^i = x_0^i) - p(w)|\Big)^{\frac{1}{2}} \\
&\qquad \leq (2p(n-\ell(w))(\ell(w)+1)(1+\beta))^{\frac{1}{2}}.
\end{aligned}$$

Then, as in Galves et al. (2006) we also obtain that, for any $t > 0$,

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

where

$$C = \frac{1}{4e(1+\beta)}.$$

□

Proof of Corollary 3.4. The inequality (3.5) follows from (3.2), as explained in the sequel. As in Galves et al. (2006) we can see that

$$\begin{aligned}
\Big|\, p(a|w) - \frac{(n-\ell(w)-1)p(wa)+1}{(n-\ell(w)-1)p(w)+|A|} \,\Big| &\leq \frac{p(w)(|A|+1)}{p(w)([n-\ell(w)-1]p(w)+|A|)} \\
&\leq \frac{|A|+1}{(n-\ell(w)-1)p(w)}.
\end{aligned}$$

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

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

Let denote by $t' = t - (|A|+1)/np(w)$. Then we have that

$$\begin{aligned}
&\mathbb{P}\big(\,|\frac{N_n(wa)+1}{N_n(w\cdot)+|A|} - \frac{(n-\ell(w)-1)p(wa)+1}{(n-\ell(w)-1)p(w)+|A|}| > t'\,\big)\\
&\qquad\leq\ \mathbb{P}\big(|N_n(wa) - (n-\ell(w)-1)p(wa)| > \frac{t'}{2}([n-\ell(w)-1]p(w)+|A|)\big)\\
&\qquad\quad+\sum_{n\in A}\mathbb{P}\big(|N_n(wb) - (n-\ell(w)-1)p(wb)| > \frac{t'}{2|A|}([n-\ell(w)-1]p(w)+|A|)\big)\\
&\qquad\leq e^{\frac{1}{e}}\exp\big[-(n-\ell(w)-1)\ \frac{[t-\frac{|A|+1}{(n-\ell(w)-1)p(w)}]^2[p(w)+\frac{|A|}{n-\ell(w)-1}]^2C}{4(\ell(w)+2)}\big]\\
&\qquad\quad+|A|\,e^{\frac{1}{e}}\exp\big[-(n-\ell(w)-1)\ \frac{[t-\frac{|A|+1}{(n-\ell(w)-1)p(w)}]^2[p(w)+\frac{|A|}{n-\ell(w)-1}]^2C}{4|A|^2(\ell(w)+2)}\big]\\
&\qquad\leq\ (|A|+1)e^{\frac{1}{e}}\exp\big[-(n-\ell(w)-1)\ \frac{[t-\frac{|A|+1}{(n-\ell(w)-1)p(w)}]^2[p(w)+\frac{|A|}{n-\ell(w)-1}]^2C}{4|A|^2(\ell(w)+2)}\big],
\end{aligned}$$

where

$$C = \frac{1}{4e(1+\beta)}.$$

We obtain (3.5). □

4. Proof of the main Theorem

Proof of Theorem 2.13. Define

$$O_{n,\delta}^{K,d} = \bigcup_{\substack{w\in\mathcal{T}\\ \ell(w)<K}}\ \bigcup_{uw\in\hat{\mathcal{T}}_n^{\delta,d}}\{\Delta_n(uw)>\delta\},$$

and

$$U_{n,\delta}^{K,d} = \bigcup_{\substack{w\in\hat{\mathcal{T}}_n^{\delta,d}\\ \ell(w)<K}}\ \bigcap_{uw\in\mathcal{T}|_d}\{\Delta_n(uw)\leq\delta\}.$$

Then, if $d<n$ we have that

$$\{\hat{\mathcal{T}}_n^{\delta,d}|_K \neq \mathcal{T}|_K\} = O_{n,\delta}^{K,d}\cup U_{n,\delta}^{K,d}.$$

The result follows from a succession of lemmas.

Lemma 4.1. *For any $w \in \mathcal{T}$ with $\ell(w) < K$ and for any $uw \in \hat{\mathcal{T}}_n^{\delta,d}$ we have that*

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

where C is given by (3.3).

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 and any symbol $a \in A$ we have $p(a|w) = p(a|uw)$. Hence,

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

Using Corollary 3.4 we can bound above the right hand side of the last inequality by

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

where C is given by (3.3). □

Lemma 4.2. *For any $w \in \hat{\mathcal{T}}_n^{\delta,d}$ with $\ell(w) < K$ we have that*

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

where C is given by (3.3).

Proof. As d satisfies (2.14) we have that there exists a $u\bar{w} \in \mathcal{T}|_d$ such that $u\bar{w} \in \mathcal{T}$. Then

$$\mathbb{P}\Big(\bigcap_{uw\in\mathcal{T}|_d} \{\Delta_n(uw) \le \delta\}\Big) \;\le\; \mathbb{P}(\Delta_n(u\bar{w}) \le \delta). \tag{4.3}$$

Observe that for any $a \in A$,

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

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

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

Therefore,

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

As $\delta < D_d$ we can use Corollary 3.4 to bound above the right hand side of this inequality by

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

where C is given by (3.3). This concludes the proof. □

Now we can finish the proof of Theorem 2.13. We have that

$$\mathbb{P}(\hat{\mathcal{T}}_n^{\delta,d}|_K \ne \mathcal{T}|_K) = \mathbb{P}(O_{n,\delta}^{K,d}) + \mathbb{P}(U_{n,\delta}^{K,d}).$$

Using the definition of $O_{n,\delta}^{K,d}$ and $U_{n,\delta}^{K,d}$ we have that

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

Applying Lemma 4.1 and Lemma 4.2 we obtain the inequality

$$\begin{aligned}&\mathbb{P}(\hat{\mathcal{T}}_n^{\delta,d}|_K \ne \mathcal{T}|_K) \;\le\\ &\qquad 4\,e^{\frac{1}{e}}\,(|A|+1)\,|A|^d\exp\big[-(n-d-1)\,\frac{[\min(\frac{\delta}{2},\frac{D_d-\delta}{2}) - \frac{|A|+1}{(n-d-1)\epsilon_d}]^2[\epsilon_d + \frac{|A|}{n-d-1}]^2 C}{4|A|^2(d+2)}\big],\end{aligned}$$

where C is given by (3.3). We conclude the proof of Theorem 2.13. □

Proof of Corollary 2.15. It follows from Theorem 2.13, 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 (2.14). □

5. Acknowledgments

We thank Pierre Collet, Imre Csiszár, Nancy Garcia, Bezza Hafid, Véronique Maume-Deschamps, Jorma Rissanen and Bernard Schmitt for many discussions on the subject.

References

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

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

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

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

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

Galves, A., Maume-Deschamps, V. and Schmitt, B. (2006). Exponential inequalities for VLMC empirical trees, *Submitted.* Available at `http://math.u-bourgogne.fr/IMB/maume/articles/arbreproba.pdf` .

Leonardi, F. (2007). Rate of convergence of penalized likelihood context tree estimators, *Submitted.* arXiv: math.ST/0701810v2 .

Rissanen, J. (1983). A universal data compression system, *IEEE Trans. Inform. Theory* **29**(5): 656–664.

Ron, D., Singer, Y. and Tishby, N. (1996). The power of amnesia: Learning probabilistic automata with variable memory length, *Machine Learning* **25**(2-3): 117–149.

Weinberger, M. J., Rissanen, J. and Feder, M. (1995). A universal finite memory source., *IEEE Trans. Inform. Theory* **41**(3): 643–652.

Willems, F. M., Shtarkov, Y. M. and Tjalkens, T. J. (1995). The context-tree weighting method: basic properties, *IEEE Trans. Inform. Theory* **IT-44**: 653–664.

Antonio Galves, Florencia G. Leonardi, Instituto de Matemática e Estatística, Universidade de São Paulo, BP 66281, 05315-970 São Paulo, Brasil

E-mail address: {galves, leonardi}@ime.usp.br