ESAIM: PS
January 2008, Vol. 12, p. 219–229
DOI: 10.1051/ps:2007035

ESAIM: Probability and Statistics
www.esaim-ps.org

EXPONENTIAL INEQUALITIES FOR VLMC EMPIRICAL TREES *

Antonio Galves[1], Véronique Maume-Deschamps[2] and Bernard Schmitt[2]

Abstract. A seminal paper by Rissanen, published in 1983, introduced the class of Variable Length Markov Chains and the algorithm Context which estimates the probabilistic tree generating the chain. Even if the subject was recently considered in several papers, the central question of the rate of convergence of the algorithm remained open. This is the question we address here. We provide an exponential upper bound for the probability of incorrect estimation of the probabilistic tree, as a function of the size of the sample. As a consequence we prove the almost sure consistency of the algorithm Context. We also derive exponential upper bounds for type I errors and for the probability of underestimation of the context tree. The constants appearing in the bounds are all explicit and obtained in a constructive way.

Mathematics Subject Classification. 62M05, 60G99.

Received October 22, 2006. Revised June 10, 2007.

1. Introduction

Variable Length Markov Chains were first introduced in the information theory literature by Rissanen [11] under the name of *finite memory source* or *probabilistic tree*. More recently this class of models became quite popular in the statistics literature under the name of *Variable Length Markov Chains (VLMC)* (Bühlman and Wyner [2]).

VLMC is a flexible class of Markov chains in which the part of the past which is relevant to predict the next symbol has a variable length depending on the observed past values. These relevant parts of the past are called *contexts*. The set of contexts define a partition of the past and can be represented by a tree.

The notion of context makes VLMC models parsimonious, with less parameters to estimate than the traditional approach to Markov chains in which the number of parameters grows exponentially with the length of the fixed past. Rissanen [6] introduced VLMC models as an universal tool to perform data compression. Recently, it has been used as a flexible class of processes to model scientific data, for instance in genomics to classify proteins [1, 9].

Keywords and phrases. Variable Length Markov Chain, context tree, algorithm context, weak dependance.

* *This work has been partially supported by the program ACI-NIM* DynamicAl *and by CNPq grant 301301/79 (A.G.) ans is part of Pronex/FAPESP's Project* Stochastic behavior, critical phenomena and rhythmic pattern identification in natural languages *(grant number 03/09930-9), FAPESP-CNRS Project* Probabilistic phonlogy of rhythm *and CNPq Project* Stochastic modeling of speech *(grant 475177/2004-5).*

[1] Instituto de Matemática e Estatística, Universidade de São Paulo, BP 66281, 05315-970 São Paulo, Brasil; galves@ime.usp.br

[2] Institut de Mathématiques de Bourgogne, BP 47870, 21078 Dijon cedex France; {vmaume;schmittb}@u-bourgogne.fr

Rissanen [11] not only introduced the notion of VLMC, but he also introduced the algorithm Context to estimate the tree of contexts of the VLMC as well as the family of probability transitions generating the VLMC.

The way the algorithm Context works can be summarized as follows. Given a sample produced by a VLMC, 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. Given a candidate tree of contexts we associate to each context a probability transition which is estimated as the proportion of time in the sample the context is followed by this symbol. 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.

Several recent papers addressed the question of the estimation of the tree of contexts of the VLMC as well as its associated set of probability transitions, either using variants of the algorithm Context (*cf.* [2,11] for the bounded case, and Ferrari and Wyner [8] for infinite trees), the BIC (Csiszar and Talata [4]), or other algorithms (*cf.* [12,13]). However the central question of the rate of convergence of the estimation algorithm remained open. This is the question we address here.

The central result of this paper is an exponential upper bound for the probability of error estimation of the finite probabilistic tree defining a Variable Length Markov Chain. As a consequence we prove the almost sure consistency of the algorithm Context. We also derive exponential upper bounds for type I errors and for the probability of underestimation of the context tree.

Our proofs are inspired by recent exponential upper bounds obtained by Dedecker and Doukhan [5], Dedecker and Prieur [6] and Maume-Deschamps [10]. In the first paper, an upper bound for the L^p norm of sums of weakly dependant variables is obtained. This upper bound is used in the second paper to obtain an exponential inequality for sums of weakly dependent random variables. Finally, this exponential inequality leads to an estimation of conditional probabilities in the third paper. VLMC satisfy the τ-dependence condition as defined in [6], then exponential inequalities follows. Nevertheless, the present paper takes advantage of the VLMC structure to obtain a more specific version of the estimation of conditional probabilities. This makes possible to control the probability of error of the version of the algorithm Context we consider here. Our proofs are constructive and the constants appearing in the bounds are all explicit and computable.

This paper is organized as follows. In Section 2, we state the definitions and our results. Section 3 is devoted to the proof of an exponential bound for conditional probabilities. In Section 4, we apply this exponential bound to estimate the rate of convergence of the algorithm Context and the corollaries.

2. Definitions and results

Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary process taking values on a finite alphabet A. Given two time indices $m<n$ we shall use the short hand notation x_m^n to denote the string $(x_m,\dots,x_n)$. The process is called a *Variable Length Markov Chain (VLMC)* if for any past $x_{-\infty}^{-1}$ there exists an index $k=k(x_{-\infty}^{-1})$ such that for any $a\in A$ we have

$$\mathbb{P}(X_0=a \mid X_{-\infty}^{-1}=x_{-\infty}^{-1})=\mathbb{P}(X_0=a \mid X_{-k}^{-1}=x_{-k}^{-1}). \tag{2.1}$$

Following Rissanen [11], we call any suffix string x_{-k}^{-1} satisfying (2.1) a *context.*

We will use the shorthand notation

$$p(a \mid x_{-k}^{-1})=\mathbb{P}(X_0=a \mid X_{-k}^{-1}=x_{-k}^{-1})\text{ and}$$

$$p(x_{-k}^{-1})=\mathbb{P}(\,X_{-k}^{-1}=x_{-k}^{-1})\,.$$

Let us call τ the set of all contexts associated to the VLMC $(X_n)_{n\in\mathbb{Z}}$. We will assume that τ is minimal. This means that for each context $x_{-k}^{-1} \in \tau$, there is no $1 \leq j < k$ such that

$$p(a \mid x_{-j}^{-1}) = p(a \mid x_{-k}^{-1})\,,$$

for any $a \in A$. This minimality condition is equivalent to the *suffix property* which means that no context is a suffix of any other context. The suffix property implies that τ can be represented as a *tree* in which contexts are identified by the leaves of the tree. The tree of contexts τ defines a partition of the pasts accepted by the process $(X_n)_{n\in\mathbb{Z}}$.

The tree of contexts τ together with a family p of conditional probabilities on A, indexed by the leaves of τ, will be called a *probabilistic tree.*

In what follows we shall always assume that the tree τ is *finite*. This means that for any past $x_{-\infty}^{-1}$ the index $k = k(x_{-\infty}^{-1})$ appearing in (2.1) is bounded.

From now on the length of a finite string w will be denoted $|w|$. Given two finite strings w and v, we will denote by vw the string with length $|w| + |v|$ obtained by concatenating the two strings. We will also denote by $h(\tau)$ the height of the tree τ, *i.e.*

$$h(\tau) = \max\{|w|; w \in \tau\}.$$

Given a finite sample $(X_1, \ldots, X_n)$ and a finite string w, with $|w| \leq n$, we define the counting random variables $N_n(w)$ as the number of times w appears in the sample

$$N_n(w) = \sum_{m=1}^{m=n-|w|+1} \mathbf{1}\{X_m^{m+|w|-1} = w\}.$$

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

$$\widehat{p}_n(a|w) = \frac{N_n(wa) + 1}{N_n(w) + |A|}\cdot \tag{2.2}$$

This definition of $\widehat{p}_n(a|w)$ is convenient because it is asymptotically equivalent to $\frac{N_n(wa)}{N_n(w)}$ and it avoids an extra definition in the case $N_n(w) = 0$. To estimate the probabilistic tree producing a sample $X_1, \ldots, X_n$ we define the *gain* function Δ as follows. For any pair of finite strings u and w, define the random function

$$\Delta_n(u, w) = \max_{a\in A} |\hat{p}_n(a|w) - \hat{p}_n(a|uw)|.$$

For any $\ell \geq 1$ and any $\delta > 0$, we consider the complete tree of height ℓ and prune it in such way that all nodes w of the resulting tree satisfy:

$$\max_{a\in A} \Delta_n(a, w) > \delta.$$

Denote $\widehat{\tau}_n(\ell, \delta)$ the resulting tree.

We associate to each $w \in \widehat{\tau}_n(\ell, \delta)$ a probability $\widehat{p_n}(\cdot|w)$ on A using Definition (2.2).

If there is no danger of confusion we will omit to mention ℓ and δ in the notation $\widehat{\tau}_n(\ell, \delta)$. This construction of the empirical probabilistic tree $(\widehat{\tau}_n, \widehat{p_n})$ is a variant of Rissanen's algorithm Context.

Let us introduce some quantities associated to the family of probability transitions p which will be useful in what follows. First of all, we extend the conditional probabilities p to suffixes of the contexts in a natural way. Namely, if $w = (w_{-1}, \ldots, w_{-k}) \in \tau$, for all $i = 1, \ldots, k-1$, let

$$p(a|w_{-i}^{-1}) = \mathbb{P}(X_0 = a|X_{-i}^{-1} = w_{-i}^{-1})\,,$$

where the probabilities appearing in the formula are those of the stationary VLMC corresponding to (τ, p). Let us introduce some extra notations

$$D(\tau) = \min_{w\in\tau} \min_{i=1,\ldots,\ |w|-1} \max_{a\in A} |p(a|w) - p(a|w_{-i}^{-1})|,$$

$$\rho = \min_{\substack{a\in A\\ w\in\tau}} p(a|w),$$

$$p_{\min} = \min_{w\in\tau} p(w),$$

$$\begin{cases} \theta = \min\limits_{(u,v)\in\tau} \left[\sum\limits_{a\in A} \min(p(a|u), p(a|v))\right] \\ \gamma = \left(1-\theta^{h(\tau)}\right)^{\frac{1}{h(\tau)}} \\ \beta = (1-\gamma). \end{cases} \tag{2.3}$$

The quantity θ is called the Dobrushin's coefficient of $(X_n)_{n\in\mathbb{Z}}$. Observe that τ finite implies that $p_{\min} > 0$.

We will assume the following assumption.

Positivity Assumption $\rho > 0$.

This implies that the process is aperiodic and irreducible and that $0 < \theta$. Obviously θ is always bounded above by 1. Moreover it only takes the value 1 in the trivial case in which the process is a sequence of independent random variables which will not be considered here. Our main result is the following theorem.

Theorem 2.1. *For all $\ell \geq h(\tau)$, there exists $\bar{n}(\ell)$, such that for all $n \geq \bar{n}$ the following inequality holds*

$$\mathbb{P}\left(\widehat{\tau}_n(\ell) = \tau\right) \geq 1 - 4e^{\frac{1}{e}} \left(|A|^{\ell} \exp\left[-n\frac{\beta p_{\min}^2 \rho^{\ell-h(\tau)}}{4e} \left(\frac{\delta}{2} - \frac{|A|+1}{np_{\min}}\right)^2 \right] \right.$$
$$\left. + |A|^{h(\tau)} \exp\left[-n\frac{\beta p_{\min}^2}{4e} \left(\frac{D(\tau)-\delta}{2} - \frac{|A|+1}{np_{\min}} \rho^{\ell-h(\tau)} \right)^2 \right] \right).$$

A first consequence of the theorem is the following almost sure consistency result for the empirical probability trees.

Corollary 2.2. *Let us assume that the sample $(X_1, \ldots, X_n)$ has been produced by a VLMC whose probabilistic tree is (τ, p_τ). Let $\kappa = (\kappa_n)_{n\geq 1}$ be any sequence such that*

$$\sum_{n\geq 1} \exp[-\kappa_n^2 n \frac{\beta p_{\min}^2}{4e}] < \infty.$$

Then for almost all infinite samples $X_1, X_2, \ldots$ there exists $\bar{n} = \bar{n}(X_1, X_2, \ldots)$, such that for all $n \geq \bar{n}$ we have

- *$\widehat{\tau}_n = \tau$ and*
- $\sup_{w\in\tau} \max_{a\in A} |\widehat{p}_n(a|w) - p(a|w)| < \kappa_n$.

Where $(\hat{\tau}_n, \widehat{p}_n)$ is the probabilistic tree constructed with $\ell = h(\tau)$.
This shows that our variant of the context algorithm is consistent almost surely.
A convenient choice for κ_n could be $\frac{1}{n^\alpha}$, $0 < \alpha < \frac{1}{2}$.

Another application of the theorem are the following exponential upper bounds for type I errors and for the probability of underestimation of the context tree.

In the first case, we want to test the *null assumption* that the sample was generated by a VLMC corresponding to the probabilistic tree (τ, p). To estimate the probability of type I error, we take the control parameter $\ell = h(\tau)$. The result is the following

Corollary 2.3. *Let us assume that the sample $(X_1, \ldots, X_n, \ldots)$ has been produced by a VLMC whose probabilistic tree is (τ, p_τ). Then the following inequality holds*

$$\mathbb{P}\left(\{\widehat{\tau}_n \neq \tau\ \} \cup \{\widehat{\tau}_n = \tau\,,\ \sup_{w\in\tau} \max_{a\in A} |\widehat{p}_n(a|w) - p_\tau(a|w)| > t\} \right) \leq \\ e^{\frac{1}{e}} \left(8|A|^{h(\tau)} \exp[-nk_1] + 2|A| \exp\left[-\left(t - \frac{|A|+1}{np_{\min}} \right)^2 nk_2 \right] \right)$$

where

$$k_1 = \frac{\beta p_{\min}^2}{8e} \max\left([\delta/2 - |A|/(np_{\min})]^2 \rho^{h(\tau)-1}, \left(\frac{D(\tau)-\delta}{2} - \frac{|A|+1}{np_{\min}} \right)^2 \right)$$

and

$$k_2 = \frac{\beta p_{\min}^2}{8e}\cdot$$

Corollary 2.3 provides an exponential upper bound for the error of type I, *i.e.* the probability of erroneously rejecting the null hypothesis that the sample was generated by the probabilistic tree (τ, p).

When testing in the class of probabilistic trees, the most serious mistake consists in erroneously accepting that the contexts are smaller than they really are. The following lemma provides an exponential upper bound for this type of error that we call *underestimate error*.

Given two trees of context τ and τ', we shall write $\tau \leq \tau'$ if all $w \in \tau$ is a suffix of some $w' = uw \in \tau'$. This defines an order on trees, which is not complete. Obviously, if $\tau < \tau'$, there is at least one $w \in \tau$ which is a strict suffix $w' = uw \in \tau'$, $|u| \geq 1$.

For $0 < \theta_0 < 1$, $0 < \beta_0 < 1$, $0 < p_0 < 1$ and $\ell \in \mathbb{N}$ define

$$\mathcal{T}(D_0, \beta_0, p_0, \ell) = \{(\tau, p), /\ h(\tau) \leq \ell,\ D(\tau) \geq D_0,\ \beta(\tau) \geq \beta_0,\ p_{\min}(\tau) \geq p_0^\ell\}.$$

Corollary 2.4. *Let us suppose that the sample $X_1, \ldots, X_n$ has been generated by a probabilistic tree belonging to the class $\mathcal{T}(D_0, \beta_0, p_0, \ell)$. The following holds*

$$\sup_{(\tau,p)\in\mathcal{T}(D_0,\beta_0,p_0,\ell)} \mathbb{P}(\hat{\tau}_n < \tau) \leq 4e^{\frac{1}{e}} |A|^\ell \exp\left[-\frac{\beta_0 p_0^{2\ell} n}{8e} \left(\frac{D_0 - \delta}{2} - \frac{|A|+1}{np_0^\ell} \right)^2 \right].$$

The main ingredient of the proof are the following exponential inequalities.

Theorem 2.5. *For any finite sequence a_0^j and any $t > 0$ the following inequality holds*

$$\mathbb{P}(|N_n(a_0^j) - np(a_0^j)| > t) \leq e^{\frac{1}{e}} \exp\left[-\beta \frac{t^2 p_{\min}}{2enp(a_0^j)} \right].$$

Moreover, for any $n > (|A|+1)/tp(a_0^j)$ we have

$$\mathbb{P}\left(|\hat{p}_n(a_j|a_0^{j-1}) - p(a_j|a_0^{j-1})| > t \right) \leq 2 \cdot e^{\frac{1}{e}} \exp\left(-\frac{\left(t - \frac{|A|+1}{np(a_0^{j-1})} \right)^2 np_{\min} p(a_0^{j-1}) \beta}{8e} \right). \tag{2.4}$$

3. Proof of Theorem 2.5

The proof of Theorem 2.5 follows the strategy developed in Maume-Deschamps [10] together with the following classical mixing property of Markov chains.

Lemma 3.1. *Let $(X_i)_{i\in\mathbb{Z}}$ be a VLMC satisfying the Positivity Assumption $\rho > 0$. Let τ be the related tree whose height equals $h < \infty$. For any $n \geq 0$, for any $m \geq 0$, any word u_{-k}^{-1} in τ and any b_0^m finite sequence, the following inequality holds*

$$|\mathbb{P}(X_n^{n+m} = b_0^m | X_{-k}^{-1} = u_{-k}^{-1}) - p(b_0^m)| \leq \frac{p(b_0^m)}{p_{\min}}\gamma^n. \tag{3.1}$$

Recall that γ and $p_{\min}$ have been defined by (2.3).

Proof. We will use the method of maximal coupling as presented in [7]. The VLMC process $(X_i)_{i\in\mathbb{Z}}$ being a Markov chain of order h, we consider the Markov process of order 1 taking values in A^h: $(Y_n)_{n\in\mathbb{Z}}$, $Y_n = (X_{nh}, \ldots, X_{(n+1)h-1})$. For any $n \in \mathbb{Z}$, $(a_h^{2h-1}) \in A^h$, $(b_0^{h-1}) \in A^h$,

$$\begin{aligned}&\mathbb{P}(Y_1 = a_h^{2h-1} | Y_0 = b_0^{h-1})\\ &= \prod_{j=0}^{h-1} \mathbb{P}(X_j = a_j | X_{j+1}^{2h-1} = a_{j+1}^{2h-1}).\end{aligned}$$

If we denote θ_h the Dobrushin's coefficient of $(Y_n)_{n\in\mathbb{Z}}$, a simple computation gives $\theta_h \geq \theta^h$.

The irreducibility of $(X_n)_{n\in\mathbb{Z}}$ induces irreducibility of $(Y_n)_{n\in\mathbb{Z}}$. The condition $\rho > 0$ implies that $(X_n)_{n\in\mathbb{Z}}$ is aperiodic of degree 1 and $(Y_n)_{n\in\mathbb{Z}}$ is also aperiodic of degree 1. It remains to apply Theorem 3.2.3 in Ferrari and Galves [7] for the process $(Y_n)_{n\in\mathbb{Z}}$. We obtain : $\forall (u_{nh}^{(n+1)h-1}) \in A^h$, $\forall (v_0^{h-1}) \in A^h$, $\forall n \geq 0$,

$$|\mathbb{P}(Y_n = u_{nh}^{(n+1)h-1} | Y_0 = v_0^{h-1}) - \mu(u_{nh}^{(n+1)h-1})| \leq (1 - \theta_h)^n \tag{3.2}$$

where μ is the invariant probability measure of $(Y_n)_{n\in\mathbb{Z}}$ on A^h. Of course, μ coincides with the probability that we have denoted p. We can reformulate (3.2) in the following way: $\forall (u_{nh}^{(n+1)h-1}) \in A^h$, $\forall (v_0^{h-1}) \in \tau$, $\forall n \geq 0$,

$$|\mathbb{P}(X_{nh}^{(n+1)h-1} = u_{nh}^{(n+1)h-1} | X_0^{h-1} = v_0^{h-1}) - p(u_{nh}^{(n+1)h-1})| \leq (1 - \theta^h)^n. \tag{3.3}$$

We are now able to prove the announced mixing property. Take $a_0^{h-1} \in A^h$ and any finite sequence b_0^m and any $n \geq 0$, denote $n_0 = [\frac{n}{k}]$. Using the fact that $(X_n)_{n\in\mathbb{Z}}$ is a Markov chain of order h:

$$\begin{aligned}&\mathbb{P}(X_{-h}^{-1} = a_{-h}^{-1} \cap X_n^{n+m} = b_0^m)\\ &= \mathbb{P}(X_{-h}^{-1} = a_{-h}^{-1}) \sum_{z_{-h}^{-1}\in A^h} \mathbb{P}(X_n^{n+m} = b_0^m | Y_{n_0} = z_{-h}^{-1}) \cdot \mathbb{P}(Y_{n_0} = z_{-h}^{-1} | Y_0 = a_{-h}^{-1})\\ &= \mathbb{P}(Y_0 = a_{-h}^{-1}) \Bigg(\mathbb{P}(X_n^{n+m} = b_0^m) + \sum_{z_{-h}^{-1}\in A^h} \mathbb{P}(X_n^{n+m} = b_0^m | Y_{n_0} = z_{-h}^{-1}) \cdot\\ &\quad \big[\mathbb{P}(Y_{n_0} = z_{-h}^{-1} | Y_0 = a_{-h}^{-1}) - p(z_{-h}^{-1})\big] \Bigg).\end{aligned}$$

Using (3.2), we get :

$$\begin{aligned}&|\mathbb{P}(Y_0 = a_{-h}^{-1} \cap X_n^{n+m} = b_0^m) - \mathbb{P}(Y_0 = a_{-h}^{-1})\mathbb{P}(X_n^{n+m} = b_0^m)|\\ &\leq \frac{\mathbb{P}(Y_0 = a_{-h}^{-1})}{p_{\min}} \cdot \mathbb{P}(X_n^{n+m} = b_0^m)(1 - \theta^h)^{[\frac{n}{h}]}.\end{aligned} \tag{3.4}$$

Then, (3.1) follows from (3.4) by dividing by $\mathbb{P}(Y_0 = a_{-h}^0)$ and observing that the conditional probabilities of the VLMC $(X_n)_{n\in\mathbb{Z}}$ by the event $\{Y_0 = a_{-h}^0\}$ are the same as those conditioned by the event $\{X_{-k}^{-1} = a_{-k}^{-1}\}$ with $a_{-k}^{-1} \in \tau$ (see formulae (2.1)). □

Corollary 3.2. *Let $(X_i)_{i\in\mathbb{Z}}$ be a VLMC satisfying the Positivity Assumption $\rho > 0$. Let τ be the related tree whose height equals $h < \infty$. For any $n \geq 0$, for any $m \geq 0$, any finite words u_{-j}^{-1}, b_0^m, inequality (3.1) holds.*

Proof. We only have to prove (3.1) for a word u_{-k}^{-1} which is a strict suffix of a branch of τ. We denote by $\tau_{u_{-j}^{-1}}$ the set of branches of τ admitting u_{-k}^{-1} as a strict suffix. These are words w such that : $wu_{-k}^{-1} \in \tau$. Then :

$$\begin{aligned}
&|\mathbb{P}(X_n^{n+m} = b_0^m \cap X_{-j}^{-1} = u_{-j}^{-1}) - \mathbb{P}(X_n^{n+m} = b_0^m)\mathbb{P}(X_{-j}^{-1} = u_{-j}^{-1})| \\
= \;& |\sum_{w\in\tau_{u_{-j}^{-1}}} \mathbb{P}(X_n^{n+m} = b_0^m \cap X_{-j-1}^{-j-|w|} = w \cap X_{-j}^{-1} = u_{-j}^{-1}) \\
& -\mathbb{P}(X_n^{n+m} = b_0^m)\mathbb{P}(X_{-j-1}^{-j-|w|} = w \cap X_{-j}^{-1} = u_{-j}^{-1})| \\
\leq \;& \frac{1}{p_{\min}}|\sum_{w\in\tau_{u_{-j}^{-1}}} \mathbb{P}(X_n^{n+m} = b_0^m)\cdot\mathbb{P}(X_{-j-|w|}^{-j-1} = w \cap X_{-j}^{-1} = u_{-j}^{-1})(1-\theta^h)^{[\frac{n}{h}]}| \\
= \;& \frac{1}{p_{\min}}(1-\theta^h)^{[\frac{n}{h}]}\mathbb{P}(X_n^{n+m} = b_0^m)\cdot\mathbb{P}(X_{-j}^{-1} = u_{-j}^{-1}).
\end{aligned}$$

We conclude the proof by dividing by $\mathbb{P}(X_{-j}^{-1} = u_{-j}^{-1})$. □

Proof of Theorem 2.5. It is similar to Corollary 1.3 and Theorem 1.5 in Maume-Deschamps [10]. It is based on an inequality by Dedecker-Doukhan (2003)[5], Proposition 4.

In our setting, let $a_0^j = w\cdot u$ with $w \in \tau$ and u a finite sequence. Let $Y_i = \mathbf{1}_{\{X_i^{i+j-1} = a_0^j\}} - p(a_0^j)$ and $\mathcal{M}_i$ be the σ-algebra generated by $X_0, \ldots, X_{i-1}$. From the definition of the counting function N_n, following Dedecker-Doukhan [5], Prop. 4), we have for any $q \geq 2$,

$$\begin{aligned}
\|N_n(a_0^j) - np(a_0^j)\|_q \;&\leq\; \left(2q\sum_{i=1}^{n-j-1}\max_{i\leq\ell\leq n}\|Y_i\sum_{k=i}^{\ell}\mathbb{E}(Y_k|\mathcal{M}_i)\|_{\frac{q}{2}}\right)^{\frac{1}{2}} \\
&\leq\; \left(2q\sum_{i=1}^{n-j-1}\max_{i\leq\ell\leq n}\|Y_i\|_{\frac{q}{2}}\sum_{k=i}^{n-j-1}\|\mathbb{E}(Y_k|\mathcal{M}_i)\|_\infty\right)^{\frac{1}{2}} \\
&\leq\; \left(2q\sum_{i=1}^{n-j-1}\sum_{k=i}^{n-j-1}\sup_{x_0^{i-1}\in A^i}|\mathbb{P}(X_k^{k+j} = a_0^j|X_0^i = x_0^i) - p(a_0^j)|\right)^{\frac{1}{2}} \\
&\leq\; \left(n\frac{2qp(a_0^j)}{p_{\min}\beta}\right)^{\frac{1}{2}}.
\end{aligned}$$

The last inequality follows from (3.4), where the parameter β was defined in (2.3).

Then, we follow Dedecker-Prieur (2005)[6] and we get that for all $t > 0$,

$$\mathbb{P}(|N_n(a_0^j) - np(a_0^j)| > t) \leq \mathrm{e}^{\frac{1}{e}}\exp\left(\frac{-t^2\beta p_{\min}}{2enp(a_0^j)}\right). \tag{3.5}$$

A similar estimation may be found in I. Csiszár [3], but our parameters and our proof are different.

Inequality (2.4) follows from (3.5) with an improvement of the proof of Theorem 1.5 in Maume-Deschamps [10]. Let us sketch it.
First of all, remark that

$$\frac{N_n(a_0^j)+1}{N_n(a_0^{j-1})+|A|}\leq 1.$$

Now,

$$\begin{aligned}
&\mathbb{P}\left(|\hat{p}_n(a_j|a_0^{j-1})-\frac{np(a_0^j)+1}{np(a_0^{j-1})+|A|}|>t\right)\\
&\leq \mathbb{P}\left(|N_n(a_0^j)-np(a_0^j)|>\frac{t}{2}(np(a_0^{j-1})+|A|)\right)\\
&+\mathbb{P}\left(|N_n(a_0^{j-1})-np(a_0^{j-1})|>\frac{t}{2}(np(a_0^{j-1})+|A|)\right)\\
&\leq 2e^{\frac{1}{e}}\exp\left[-\frac{\beta p_{\min}}{8enp(a_0^{j-1})}t^2(np(a_0^{j-1})+|A|)^2\right]
\end{aligned}$$

using (3.5). To conclude the proof, it is enough to observe that

$$\begin{aligned}
\left|p(a_j|a_0^{j-1})-\frac{np(a_0^j)+1}{np(a_0^{j-1})+|A|}\right| &\leq \frac{p(a_0^{j-1})(|A|+1)}{p(a_0^{j-1})(np(a_0^{j-1})+|A|)}\\
&\leq \frac{|A|+1}{n}p(a_0^{j-1}.
\end{aligned}$$

We get (2.4). □

4. Proof of the main theorem

Proof of the main theorem. First of all we need to bound above the probability of selecting a context which is too long. Given two strings u and w define $O_n(u,\ w)$ as the event in which $\Delta_n(u,w)>\delta$. The event in which the algorithm selects a context longer than the real one is

$$O_n=\bigcup_{\substack{w\in\tau\\ uw\in\hat{\tau}_n}}O_n(u,\ w).$$

We also need to estimate the probability of selecting a context shorter than the real one. This event is represented by

$$U_n(u,\ w)=\{\Delta_n(u,w)\leq\delta\}\ ,\ U_n=\bigcup_{\substack{uw\in\tau\\ w\in\hat{\tau}_n}}U_n(u,\ w).$$

If $n>\ell$ then

$$\{\hat{\tau}_n\neq\tau\}=O_n\cup U_n.$$

The proof of the theorem follows from a succession of lemmas.

Lemma 4.1. *For any $w\in\tau$, $uw\in\hat{\tau}_n$ and $\ell\geq h(\tau)$ and for any*

$$n\geq\frac{2|A|}{\delta p_{\min}\rho^{\ell-h(\tau)}},$$

we have

$$\mathbb{P}(O_n(u,\ w)) \leq 4e^{\frac{1}{e}}|A| \exp\left[-\frac{\beta n p_{\min}^2 \rho^{\ell-1}}{8e}\left(\frac{\delta}{2} - \frac{|A|}{np_{\min}}\right)^2\right].$$

Proof. Recall that

$$\Delta_n(u,w) = \max_{a\in A} |\hat{p}_n(a|w) - \hat{p}_n(a|uw)|.$$

Remark that $w \in \tau$ then for all finite sequence u and $a \in A$ we have $p(a|w) = p(a|uw)$. Therefore,

$$\begin{aligned}\mathbb{P}(\Delta_n(u,w) \quad > \quad \delta) \leq \sum_{a\in A} \mathbb{P}\left(|\hat{p}_n(a|w) - p(a|w)| > \frac{\delta}{2}\right) \\ +\mathbb{P}\left(|\hat{p}_n(a|uw) - p(a|uw)| > \frac{\delta}{2}\right).\end{aligned}$$

Using Theorem 2.5 we bound above the right hand side of this inequality by

$$4e^{\frac{1}{e}}|A| \exp\left[-\frac{\beta n p(uw) p_{\min}}{8e}\left(\frac{\delta}{2} - \frac{|A|}{np_{\min}}\rho^{\ell-h(\tau)}\right)^2\right].$$

Recall that, by definition, a complete branch of $\hat{\tau}_n$ has its length bounded above by ℓ. Thus $p(uw) \geq p_{\min}\rho^{\ell-h(\tau)}$. This concludes the proof. □

Lemma 4.2. *For any $uw \in \tau$, for any*

$$n \geq \frac{2|A|}{p_{\min}(D(\tau) - \delta)},$$

and $w \in \hat{\tau}_n$ we have

$$\mathbb{P}(U_n(u,\ w)) \leq 4e^{\frac{1}{e}} \exp\left[-\frac{\beta n p_{\min}^2}{8e}\left[\frac{D(\tau)-\delta}{2} - \frac{|A|}{np_{\min}}\right]^2\right].$$

Proof. We start by observing that for any $a \in A$,

$$\begin{aligned}|\hat{p}_n(a|w) - \hat{p}_n(a|uw)| \quad &\geq \quad |p(a|w) - p(a|uw)| - |\hat{p}_n(a|w) - p(a|w)| \\ &\quad +|\hat{p}_n(a|uw) - p(a|uw)|].\end{aligned}$$

I follows that for any $a \in A$ we have

$$\Delta_n(u,w) \ \geq \ D(\tau) - |\hat{p}_n(a|w) - p(a|w)| - |\hat{p}_n(a|uw) - p(a|uw)|.$$

Therefore,

$$\begin{aligned}\mathbb{P}(\Delta_n(u,w) \leq \delta) \quad &\leq \quad \mathbb{P}\left(\forall a \in A,\ |\hat{p}_n(a|w) - p(a|w)| \geq \frac{D(\tau)-\delta}{2}\right) \\ &\quad +\mathbb{P}\left(\forall a \in A,\ |\hat{p}_n(a|uw) - p(a|uw)| \geq \frac{D(\tau)-\delta}{2}\right).\end{aligned}$$

Observing that $p(uw) \geq p_{\min}$, the result now follows from Theorem 2.5. □

We can now conclude the proof of the main theorem. We have

$$\mathbb{P}(\widehat{\tau}_n \neq \tau) = \mathbb{P}(O_n) + \mathbb{P}(U_n).$$

It follows from the definition that

$$\mathbb{P}(O_n) \leq \sum_{\substack{w \in \tau \\ uw \in \hat{\tau}_n}} \mathbb{P}(O_n(u,\ w)).$$

Using Lemma 4.1 we obtain the inequality

$$\mathbb{P}(O_n) \leq 4e^{\frac{1}{e}}|A|^{\ell} \exp\left[-\frac{\beta n p_{\min}^2 \rho^{\ell-h(\tau)}}{8e}\left(\frac{\delta}{2} - \frac{|A|}{n p_{\min} \rho^{\ell-h(\tau)}}\right)^2\right].$$

Using Lemma 4.2 we obtain the bounds

$$\mathbb{P}(U_n) \leq 4e^{\frac{1}{e}}|A|^{h(\tau)} \exp\left[-\frac{\beta n p_{\min}^2}{8e}\left[\frac{D(\tau)-\delta}{2} - \frac{|A|}{n p_{\min}}\right]^2\right]. \tag{4.1}$$

To conclude the proof, it suffices to sum these two terms. □

Proof of Corollary 2.2. Using Theorem 2.5, we have that

$$\begin{aligned}&\mathbb{P}(\sup_{w\in\tau}\max_{a\in A}|\widehat{p}_n(a|w) - p(a|w)| > \kappa_n) \\ &\quad\leq 2e^{\frac{1}{e}}|A| \exp\left[-\left(\kappa_n - \frac{|A|}{n p_{\min}}\right)^2 n \frac{\beta p_{\min}^2}{8e}\right].\end{aligned}$$

Now, we use Theorem 2.1 with $\ell = h(\tau)$ and $\delta = (1-\varepsilon)D(\tau)$. The Borel-Cantelli lemma implies the following almost sure result

$$\begin{aligned}&\mathbb{P}\left(\widehat{\tau}_n = \tau,\ \sup_{w\in\tau}\max_{a\in A}|\widehat{p}_n(a|w) - p(a|w)| < \kappa_n,\ \forall n \geq \widetilde{n}\right) \\ &\quad\geq\ 1 - Ce^{-L\tilde{n}} + Ke^{\frac{1}{e}}|A| \sum_{n\geq\tilde{n}} \exp\left[-\kappa_n^2 n \frac{\beta p_{\min}^2}{8e}\right],\end{aligned}$$

where K and C are suitable positive constants. □

Proof of Corollary 2.3. We take $\ell = h(\tau)$ and apply Theorem 2.1 together with equation (2.4). □

Proof of Corollary 2.4. This is an immediate application of equation (4.1). □

As a final remark, we observe that Theorem 2.1 could be extended by taking ℓ increasing with n in a suitable way. An example of this appears in Corollary 2.4 where we could take $\ell = O(\ln n)$. In the same way, δ may be decreasing with n in a suitable way. This appears in Corollary 2.3 where we could take $\delta = O(n^{-\alpha})$ with $0 < \alpha < 1/2$.

Acknowledgements. Many thanks to Pierre Collet whose careful reading of a preliminary version of this paper helped us making it hopefully more clear. We are also grateful to the referee for many interesting questions that remain open and are a source for a further work.

References

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

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

[3] I. Csiszár, Large-scale typicality of Markov sample paths and consistency of MDL order estimators. Special issue on Shannon theory: perspective, trends, and applications. *IEEE Trans. Inform. Theory* **48** (2002) 1616–1628.

[4] I. Csiszár and Z. Talata, *Context tree estimation for not necessarily finite memory processes via BIC and MDL*, manuscript (2005).

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

[6] J. Dedecker and C. Prieur, New dependence coefficients. Examples and applications to statistics. *Prob. Theory Relat. Fields* **132** (2005) 203–236.

[7] P. Ferrari and A. Galves, *Coupling and regeneration for stochastic processes.* Notes for a minicourse presented in XIII Escuela Venezolana de Matematicas. Can be downloaded from `www.ime.usp.br/pablo/book/abstract.html` (2000).

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

[9] F. Leonardi and A. Galves, Sequence Motif identification and protein classification using probabilistic trees. *Lect. Notes Comput. Sci.* **3594** (2005) 190–193.

[10] V. Maume-Deschamps, Exponential inequalities and estimation of conditional probabilities in Dependence in probability and statistics, *Lect. Notes in Stat.*, Vol. **187**, P. Bertail, P. Doukhan and P. Soulier Eds. Springer (2006).

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

[12] T.J. Tjalkens and F.M.J.F. Willems, Implementing the context-tree weighting method: arithmetic coding. Recent advances in interdisciplinary mathematics (Portland, ME, 1997). *J. Combin. Inform. System Sci.* **25** (2000) 49-58.

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