BULLETIN OF THE BRAZILIAN MATHEMATICAL SOCIETY

Bull Braz Math Soc, New Series 37(4), 581-592

Markov approximation and consistent estimation of unbounded probabilistic suffix trees

Denise Duarte, Antonio Galves and Nancy L. Garcia

Abstract. We consider infinite order chains whose transition probabilities depend on a finite suffix of the past. These suffixes are of variable length and the set of the lengths of all suffix is unbounded. We assume that the probability transitions for each of these suffixes are continuous with exponential decay rate. For these chains, we prove the weak consistency of a modification of Rissanen's algorithm *Context* which estimates the length of the suffix needed to predict the next symbol, given a finite sample. This generalizes to the unbounded case the original result proved for variable length Markov chains in the seminal paper Rissanen (1983). Our basic tool is the canonical Markov approximation which enables to approximate the chain of infinite order by a sequence of variable length Markov chains of increasing order. Our proof is constructive and we present an explicit decreasing upper bound for the probability of wrong estimation of the length of the current suffix.

Keywords: probabilistic suffix trees, Markovian approximations, variable length Markov chains, algorithm Context, consistent estimation.

Mathematical subject classification: Primary: 60K99, Secondary: 60F15.

1 Introduction

Unbounded probabilistic suffix trees define an interesting family of stochastic chains of infinite order on a finite alphabet. The idea is that for each past, only a finite suffix of the past, called *context* is enough to predict the next symbol. These suffixes can be represented by a countable complete tree of finite contexts. In a probabilistic suffix tree there is a transition probability associated to each context.

The existence of an infinite order stochastic chain consistent with the probabilistic suffix tree is assured by imposing that the transition probabilities are

Received 18 April 2006.

weakly non-null and continuous, with continuity rate decaying exponentially fast.

For these chains, we prove the weak consistency of a modification of Rissanen's algorithm *Context* which estimates the context needed to predict the next symbol, given a finite sample.

Our basic tool is to approximate the chain of infinite order consistent with the unbounded probabilistic tree, by a sequence of Markov chains, generated by finite probabilistic trees of increasing height. This idea was introduced by Bressaud, Fernández and Galves (1999a), Bressaud, Fernández and Galves (1999b) and Fernández and Galves (2002).

Our proof is constructive and we present an explicit decreasing upper bound for the probability of wrongly estimating the current context. The use of the Markov approximation makes the proof simpler and, we hope, clearer.

Probabilistic suffix trees were first introduced by Rissanen (1983) in the finite case. He called his model *finitely generated source* . In his work, not only he introduces the model but also he proposes the algorithm *Context* which estimates the context needed to predict the next symbol, given a finite sample in an effective way. In his paper, there is a proof of the weak consistency of the algorithm in the case of a fixed finite tree. Here, we generalize this result to unbounded probabilistic trees for a modified version of the algorithm Context.

Recently, probabilistic suffix trees became popular in the statistics literature under the name *variable length Markov chains* coined by Bühlmann and Wyner (1999). They prove the weak consistent of a variant of the algorithm Context for finite trees without assuming a known prior on the depths of the probabilistic tree but using a bound allowed to grow with the sample size.

An extension of Bühlmann and Wyner (1999) for the unbounded case was obtained by Ferrari and Wyner (2003) using the same technical ideas. However, they impose rather obscure conditions, which in their own words "may be difficult to check". They claim it is enough to assume that the family of probability transitions is strongly non-null, i.e. the infimum for all symbols and contexts of the probability of a symbol given the context is strictly positive. This is definitively more restrictive than the weakly non-nullness property assumed by us.

A different approach to the problem was recently proposed by Csiszár and Talata (2006). They show that in the unbounded case, consistent estimation may be achieved in linear time using two penalized log-likelihood maximization procedures, namely the Bayesian Information and the Minimum Description Length criteria.

Probabilistic suffix trees have been recently used by several authors to model

scientific data coming from many different domain such as linguistics, genomics and music, see Begleiter, El-Yaniv and Yona (2004), Bejerano and Yona (2001), Leonardi (2006) among others.

This paper is organized as follows. Section 2 presents the definitions, notation and the statement of the theorem. The proof of the theorem, as well as the Markovian approximation, are presented in Section 3. In Section 4 we discuss the reason why we could not use Rissanen's original result.

2 Notations, definitions and result

Let $\mathcal{A}$ be a finite alphabet. This will be the state space of all the chains considered in this paper. We will use the shorthand notation w_m^n to denote the string $(w_m, \ldots, w_n)$ of symbols in the alphabet $\mathcal{A}$. The length of this string will be denoted by $|w_m^n| = n - m + 1$.

Definition 2.1. *A countable subset τ of $\cup_{k=1}^{\infty} \mathcal{A}^{\{-k,\ldots,-1\}}$ is a* **complete tree with finite branches** *if it satisfies the following conditions.*

- **Suffix property**. *For no $w_{-k}^{-1} \in \tau$, there exists $u_{-j}^{-1} \in \tau$ with $j < k$ such that $w_{-i} = u_{-i}$ for $i = 1, \ldots, j$.*
- **Completeness**. *τ defines a partition of $A^{\{\ldots,-2,-1\}}$. Each element of the partition coincides with the set of the sequences in $A^{\{\ldots,-2,-1\}}$ having w_{-k}^{-1} as suffix, for some $w_{-k}^{-1} \in \tau$.*

It is easy to see that the set τ can be identified with the set of leaves of a rooted tree with a countable set of finite labeled branches.

Given a finite tree, its *height* is defined as $|\tau| = \max\{|w|; w \in \tau\}$.

Definition 2.2. *A* **probabilistic suffix tree on** $\mathcal{A}$ *is an ordered pair (τ, p) such that,*

- *τ is a complete tree with finite branches; and*
- *$p = \{p(\cdot|w); w \in \tau\}$ is a family of probability transitions on $\mathcal{A}$.*

A stationary stochastic chain (X_t) is *consistent* with a probabilistic suffix tree (τ, p) if for any infinite past $x_{-\infty}^{-1}$ and any symbol $a \in \mathcal{A}$ we have

$$\mathbb{P}_p \left\{ X_0 = a \mid X_{-\infty}^{-1} = x_{-\infty}^{-1} \right\} = p\left(a \mid x_{-\ell}^{-1}\right), \tag{2.3}$$

where $x_{-\ell}^{-1}$ is the only element of τ which is a suffix of the sequence $x_{-\infty}^{-1}$. This suffix is called the *context* of the sequence $x_{-\infty}^{-1}$. The length of the context

$\ell = \ell(x_{-\infty}^{-1})$ is a function of the sequence. Observe that the suffix property implies that the set $\{\ell(X_{-\infty}^{-1}) = k\}$ is measurable with respect to the σ-algebra generated by X_{-k}^{-1}.

If $X_0, X_1, \ldots$ is a sample from a stochastic chain consistent with a probabilistic suffix tree (τ, p) we will say that $X_0, X_1, \ldots$ is a *realization* of (τ, p). We shall use the shorthand notation

$$P\left(a_1^k\right) = \mathbb{P}\left\{X_1^k = a_1^k\right\} \tag{2.4}$$

to denote the stationary probability of the cylinder defined by the finite string of symbols a_1^k.

Definition 2.5. *We say that the probabilistic suffix tree (τ, p) is* **unbounded** *if τ is countable but not finite and therefore, the function ℓ is unbounded.*

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

Definition 2.6. *A probabilistic suffix tree (τ, p) on $\mathcal{A}$ is of* **type A** *if its transition probabilities p satisfy the following conditions.*

1. **Weakly non-nullness**, *that is*

$$\sum_{a\in\mathcal{A}} \inf_{w\in\tau} p(a \mid w) > 0\,; \tag{2.7}$$

2. **Continuity**, *that is*

$$\begin{aligned}\beta(k) := \max_{a\in\mathcal{A}} \sup \Big\{ \big| p(a \mid w) - p(a \mid v) \big|, v \in \tau,\ w \in \tau \\ \text{with } w_{-k}^{-1} = v_{-k}^{-1} \Big\} \to 0\end{aligned} \tag{2.8}$$

as $k \to \infty$. We also define

$$\beta(0) = \max_{a\in\mathcal{A}} \sup \big\{ \big| p(a \mid w) - p(a \mid v) \big|, v \in \tau, w \in \tau \text{ with } w_{-1} \neq v_{-1} \big\}.$$

The sequence $\{\beta(k)\}_k \in \mathbb{N}$ is called the **continuity rate**.

For a probabilistic suffix 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.

Chains consistent with finite probabilistic suffix trees are also called *variable length Markov chains* in the literature.

We now present a simplified version of the algorithm Context introduced by Rissanen (1983) for variable length Markov chains. The goal of the algorithm is to estimate adaptively the context of the next symbol X_n given the past symbols X_0^{n-1}.

We first construct a candidate context $X_{n-k(n)}^{n-1}$ where $k(n) = C_1 \log n$ with a suitable positive constant C_1. The intuitive reason behind the choice of the upper bound length $C_1 \log n$ is the impossibility of estimating the probability of sequences of length much longer than $\log n$ based on a sample of length n. Recent versions of this fact can be found in Marton and Shields (1994), Marton and Shields (1996) and Csiszár (2002). We then shorten it according to a sequence of tests based on the likelihood ratio statistics. This is formally done as follows.

Let $X_0, X_1, \ldots, X_{n-1}$ be a sample from the finite probabilistic tree (τ, p). For any finite string w_{-j}^{-1} with $j \leq n$, we denote $N_n(w_{-j}^{-1})$ the number of occurrences of the string in the sample

$$N_n\left(w_{-j}^{-1}\right) = \sum_{t=0}^{n-j} \mathbf{1}\left\{X_t^{t+j-1} = w_{-j}^{-1}\right\}. \tag{2.9}$$

If $\sum_{b\in\mathcal{A}} N_n\left(w_{-k}^{-1}b\right) > 0$, we define the estimator of the transition probability p by

$$\hat{p}_n\left(a|w_{-k}^{-1}\right) = \frac{N_n\left(w_{-k}^{-1}a\right)}{\sum_{b\in\mathcal{A}} N_n\left(w_{-k}^{-1}b\right)} \tag{2.10}$$

where $w_{-j}^{-1}a$ denotes the string $(w_{-j}, \ldots, , w_{-1}, a)$, obtained by concatenating w_{-j}^{-1} and the symbol a. If $\sum_{b\in\mathcal{A}} N_n\left(w_{-k}^{-1}b\right) = 0$, we define $\hat{p}_n\left(a|w_{-k}^{-1}\right) = 1/|\mathcal{A}|$.

We also define

$$\Lambda_n(i, w) = -2 \sum_{w_{-i}\in\mathcal{A}} \sum_{a\in\mathcal{A}} N_n\left(w_{-i}^{-1}a\right) \log\left[\frac{\hat{p}_n\left(a|w_{-i}^{-1}\right)}{\hat{p}_n\left(a|w_{-i+1}^{-1}\right)}\right]. \tag{2.11}$$

Notice that $\Lambda_n(i, w)$ is the log-likelihood ratio statistic for testing the consistency of the sample with a probabilistic suffix tree (τ, p) against the alternative that it is consistent with (τ', p') where τ and τ' differ only by one set of sibling nodes branching from w_{-i+1}^{-1}.

We now define the length of the estimated current context $\hat{\ell}$ as

$$\hat{\ell}\left(X_0^{n-1}\right) = \max\left\{i = 2, \ldots, k(n) : \Lambda_n\left(i, X_{n-k(n)}^{n-1}\right) > C_2 \log n\right\}, \qquad (2.12)$$

where C_2 is any positive constant.

Using a random upper bound for length of the candidate context, instead of $k(n) = C_1 \log n$ Rissanen (1983) proved the following result.

Theorem. *Given a realization $X_0, \ldots, X_{n-1}$ of a probabilistic suffix tree (τ, p) with finite height, then*

$$\mathbb{P}\left\{\hat{\ell}\left(X_0^{n-1}\right) \neq \ell\left(X_0^{n-1}\right)\right\} \longrightarrow 0 \qquad (2.13)$$

as $n \to \infty$.

Unfortunately, Rissanen's definition of the candidate context and the corresponding proof of the result only applies to the case of a fixed finite probabilistic suffix tree. Using our definition together with the canonical Markov approximation, we can extend Rissanen's result for unbounded probabilistic suffix tree. This is our main result.

Theorem 1. *Let $X_0, X_2, \ldots, X_{n-1}$ be a sample from a type A unbounded probabilistic suffix tree (τ, p) with continuity rate $\beta(j) \leq f(j)\exp\{-j\}$, with $f(j) \to 0$ as $j \to \infty$. Then, for any choice of positive constants C_1 and C_2 in the definition* (2.12)*, there exist positive constants C and D such that*

$$\mathbb{P}\left\{\hat{\ell}\left(X_0^{n-1}\right) \neq \ell\left(X_0^{n-1}\right)\right\} \leq C_1 \log n(n^{-C_2} + D/n) + Cf(C_1 \log n).$$

3 Proof of Theorem 1

We will use the canonical approximation of the chain of infinite order consistent with (τ, p) introduced by Fernández and Galves (2002). We start by adapting their definitions and theorem to the framework of probabilistic suffix trees.

Definition 3.1. *The* **canonical Markov approximation of order** *k of a chain $(X_t)_{t\in\mathbb{Z}}$ is the Markov chain of order k, $X^{[k]} = (X_t^{[k]})_{t\in\mathbb{Z}}$ having as transition probabilities,*

$$p^{[k]}\left(a \mid x_{-k}^{-1}\right) := \mathbb{P}\left\{X_0 = a \mid X_{-k}^{-1} = x_{-k}^{-1}\right\} \qquad (3.2)$$

for all $k \geq 1$ and all $a \in \mathcal{A}$ and $x_{-k}^{-1} \in \mathcal{A}^k$.

Notice that, when (X_t) is consistent with a probabilistic suffix tree (τ, p), then $(X_t^{[k]})$ is consistent with a finite probabilistic suffix tree $(\tau^{[k]}, p^{[k]})$ where

$$\tau^{[k]} = \{w \in \tau; |w| \leq k\} \cup \left\{w_{-k}^{-1}; w \in \tau, |w| \geq k\right\}. \tag{3.3}$$

Observe also, that for contexts $w \in \tau$ which length does not exceed k, we have

$$p^{[k]}(a \mid w) = p(a \mid w) .$$

However, for sequences w_{-k}^{-1} which are internal nodes of τ, there is no easy explicit formula expressing $p^{[k]}(\cdot|w_{-k}^{-1})$ in terms of the family $\{p(\cdot|v), v \in \tau\}$.

The main result of Fernández and Galves (2002) that will be crucial in the proof of Theorem 1 can be stated as follows.

Theorem. *Let $(X_t)_{t\in\mathbb{Z}}$ be a chain consistent with a type A probabilistic suffix tree (τ, p) with summable continuity rate, and let $(X_t^{[k]})$ be its canonical Markov approximation of order k. Then there exists a coupling between (X_t) and $(X_t^{[k]})$ and a constant $C > 0$ such that*

$$\mathbb{P}\left\{X_0 \neq X_0^{[k]}\right\} \leq C\beta(k) . \tag{3.4}$$

From now on, we will always assume that (τ, p) is of type A with summable continuity rates $\beta(\cdot)$ and $(\tau^{[k]}, p^{[k]})$ is its canonical Markov approximation of order k.

The proof of Theorem 1 will follow from the following lemma together with a control on the error of the Markov approximation.

Lemma 3.5. *For any choice of positive constants C_1 and C_2 used in the definition of $\hat{\ell}$, we have*

$$\mathbb{P}\left\{\hat{\ell}(X_0^{[k]}, \ldots, X_{n-1}^{[k]}) \neq \ell(X_0^{[k]}, \ldots, X_{n-1}^{[k]})\right\} \leq k(n)\left(n^{-C_2} + D/n\right) \tag{3.6}$$

where $k = k(n) = C_1 \log n$.

Proof. We know that for fixed (i, w), under the null hypothesis, the statistic $\Lambda_n(i, w)$, given by (2.11), has asymptotically chi-square distribution with $|\mathcal{A}|-1$ degrees of freedom (see, for example, van der Vaart (1998)). We recall that, for each (i, w) the null hypothesis (H_0^i) is that the true context is w_{-i+1}^{-1}.

Since we are going to perform a sequence of $k(n)$ sequential tests where $k(n) \to \infty$ as n diverges, we need to control the error in the chi-square approximation. For this, we use a well-known asymptotic expansion for the distribution of $\Lambda_n(i, w)$ due to Hayakawa (1997) which implies that

$$\mathbb{P}\left\{\Lambda_n(i, w) \leq x \mid H_0^i\right\} = \mathbb{P}\left\{\chi^2 \leq x\right\} + D/n , \tag{3.7}$$

where D is a positive constant and χ^2 is random variable with distribution chi-square with $|\mathcal{A}| - 1$ degrees of freedom.

Therefore, it is immediate that

$$\mathbb{P}\{\Lambda_n(i, w) > C_2 \log n\} \leq e^{-C_2 \log n} + D/n\,.$$

By (2.12), in order to find $\hat{\ell}\left(X_0^{n-1}\right)$ we have to perform at most $k(n)$ tests. We want to give an upper bound for the overall probability of type I error in a sequence of $k(n)$ sequential tests. An upper bound is given by the Bonferroni inequality, which in our case can be written as

$$\mathbb{P}\left(\bigcup_{i=2}^{k(n)}\{\Lambda_n(i, w) > C_2 \log n\} \mid H_0^i\right) \leq \sum_{i=2}^{k(n)} \mathbb{P}\left\{\Lambda_n(i, w) > C_2 \log n \mid H_0^i\right\}.$$

This last term is bounded above by $C_1 \log n(n^{-C_2} + D/n)$. This concludes the proof. □

We are finally ready to prove Theorem 1.

Let $\left(\tau^{[k]}, p^{[k]}\right)$ be the canonical Markov approximation of order k of (τ, p). Take $k = k(n) = C_1 \log(n)$. Then,

$$\mathbb{P}\left\{\hat{\ell}\left(X_0^{n-1}\right) \neq \ell\left(X_0^{n-1}\right)\right\}$$

$$\leq \mathbb{P}\left\{\hat{\ell}\left(X_0^{n-1}\right) \neq \ell\left(X_0^{n-1}\right), X_i = X_i^{[k]}, 1 \leq i \leq n\right\} + \mathbb{P}\left(\bigcup_{i=1}^{n}\left\{X_i \neq X_i^{[k]}\right\}\right).$$

The first term equals to

$$\mathbb{P}\left\{\hat{\ell}\left(X_0^{[k]}, \ldots, X_{n-1}^{[k]}\right) \neq \ell\left(X_0^{[k]}, \ldots, X_{n-1}^{[k]}\right),\ X_i = X_i^{[k]}, i = 1, \ldots, n\right\}.$$

Using Lemma 3.5 this last expression can be bounded by

$$\mathbb{P}\left\{\hat{\ell}\left(X_0^{[k]}, \ldots, X_{n-1}^{[k]}\right) \neq \ell\left(X_0^{[k]}, \ldots, X_{n-1}^{[k]}\right)\right\} \leq n^{-C_2} + D/n\,. \tag{3.8}$$

Inequality (3.4) provides a bound for the second term

$$\mathbb{P}\left(\bigcup_{i=1}^{n}\left\{X_i \neq X_i^{[k]}\right\}\right) \leq n\,C\,\beta(k(n))\,, \tag{3.9}$$

where C is a suitable positive constant independent of $k(n)$.

Since we took $k(n) = C_1 \log(n)$ and by hypothesis $\beta(k) \leq f(k) \exp\{-k\}$, the result follows immediately from inequalities (3.8) and (3.9). □

4 Discussion

The way the algorithm Context is introduced in Rissanen (1983) is slightly different. He first constructed a candidate context $X_{n-M(n)}^{n-1}$ where $M(n)$ is a random length defined as follows

$$M(n) \;=\; \min\left\{ i = 0, 1, \ldots, \lfloor C_1 \log n \rfloor : N_n\left(X_{n-i}^{n-1}\right) \;>\; \frac{C_2\, n}{\sqrt{\log n}} \right\}, \quad (4.1)$$

where C_1 and C_2 are arbitrary positive constants. In the case the set is empty we take $M(n) = 0$. Then, the length of the estimated current context $\hat{\ell}$ is estimated as we did, using (2.12).

Imposing that the length of the candidate context is bounded above by $M(n)$ is a technical condition used by Rissanen to obtain the following upper bound which appears in his proof of (2.13). Rissanen writes it as

$$\begin{aligned} &\mathbb{P}\left\{\hat{\ell}\left(X_0^{n-1}\right) \neq \ell\left(X_0^{n-1}\right)\right\} \\ &\leq \mathbb{P}\left\{\hat{\ell}\left(X_0^{n-1}\right) \neq \ell\left(X_0^{n-1}\right) \mid N_n\left(X^{n-1}_{n-\ell\left(X_0^{n-1}\right)}\right) > \frac{C_2 n}{\sqrt{\log n}}\right\} \\ &\mathbb{P}\left\{N_n\left(X^{n-1}_{n-\ell\left(X_0^{n-1}\right)}\right) > \frac{C_2 n}{\sqrt{\log n}}\right\} + \mathbb{P}\left\{\bigcup_{w\in\tau}\left\{N_n(w) \leq \frac{C_2 n}{\sqrt{\log n}}\right\}\right\}. \end{aligned} \quad (4.2)$$

The point here is that Rissanen (1983) does not use the fact that the law of Λ_n converges to chi-square distribution as we did. Instead of that, Rissanen provides the following explicit upper bound for the conditional probability in the right-hand side of (4.2).

$$\begin{aligned} &\mathbb{P}\left\{\hat{\ell}\left(X_0^{n-1}\right) \neq \ell\left(X_0^{n-1}\right) \mid N_n\left(X^{n-1}_{n-\ell\left(X_0^{n-1}\right)}\right) > \frac{C_2 n}{\sqrt{\log n}}\right\} \\ &\leq C_1 \log n\, e^{-C_2'\sqrt{\log n}}, \end{aligned} \quad (4.3)$$

where C_1, C_2 and C_2' are positive constants independent of the height of the probabilistic suffix tree (τ, p).

With respect to the second term he only observes that, by ergodicity, for each $w \in \tau$ we have

$$\mathbb{P}\left\{N_n(w) \leq \frac{C_2 n}{\sqrt{\log n}}\right\} \longrightarrow 0 \quad (4.4)$$

as $n \to \infty$. Since τ is finite the convergence in (4.2) implies the desired result.

In the case of unbounded trees, (4.2) is not enough to assure the result. Now we need an explicit upper bound for

$$\mathbb{P}\left\{ N_n^{[k(n)]}(w) \leq \frac{C_2 n}{\sqrt{\log n}} \right\},$$

where $k(n)$ is the height of the Markov approximation estimated with the sample of size n. The height $k(n)$ diverges with n and to assure that the limit in (4.2) is really zero, using Rissanen's estimation we need to take $k(n) = c \log\log(n)$ instead of $k(n) = c \log(n)$.

The fact that $k(n)$ increases very slowly has a consequence on the quality of the Markov approximation. If $k(n) = c \log\log(n)$, then to assure that the upper bound (3.9) vanishes as n diverges, we must assume that the continuity rate of the chain decreases with a super exponential rate $\beta(k) \leq \exp\{-\exp ck\}\}$.

Our alternative approach, using directly the chi-square approximation works assuming only that $\beta(k)$ decreases exponentially fast. And this together with the canonical Markov approximation provides a very simple proof for the result in case of type A unbounded probabilistic suffix tree with continuity rate decreasing exponentially fast.

Acknowledgments. This paper 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). We acknowledge partial support of CNPq grants 301301/79 (AG) and 301054/93-2 (NLG). Many thanks to Silvia Ferrari and Francisco Cribari for helpful advice concerning the chi-square approximation. We thank the anonymous referee for the comments and corrections that improved the presentation of the paper.

References

[1] R. Begleiter, R. El-Yaniv and G. Yona, On prediction using variable order Markov models. *J. Artificial Intelligence Res.*, **22** (2004), 385–421 (electronic).

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

[3] X. Bressaud, R. Fernández and A. Galves, Speed of $\bar{d}$-convergence for Markov approximations of chains with complete connections: a coupling approach. *Stoch. Proc and Appl.*, **83** (1999a), 127–38.

[4] X. Bressaud, R. Fernández and A. Galvez, Decay of correlations for non Hölderian dynamics: a coupling approach. *Elect. J. Prob.*, (1999b). (http://www.math.washington.edu/∼ejpecp/).

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

[6] I. Csiszár, Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory*, **48**(6) (2002), 1616–1628. Special issue on Shannon theory: perspective, trends and applications.

[7] I. Csiszár and Z. Talata, Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Trans. Infor. Theory*, (2006).

[8] R. Fernández and A. Galves, Markov approximations of chains of infinite order. *Bull. Braz. Math. Soc. (N.S.)*, **33**(3) (2002), 295–306. Fifth Brazilian School in Probability (Ubatuba, 2001).

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

[10] T. Hayakawa, The likelihood ratio criterion and the asymptotic expansion of its distribution. *Ann. Inst. Statist. Math.*, **29**(3) (1977), 359–378.

[11] F.G. Leonardi, A generalization of the pst algorithm: modeling the sparse nature of protein sequences. *Bioinformatics*, **22**(7) (2006): Bioinformatics Advance Access published online on March 9, 2006 Bioinformatics, doi:10.1093/ bioinformatics/ btl088 (electronic).

[12] K. Marton and P.C. Shields, Entropy and the consistent estimation of joint distributions. *Ann. Probab.*, **22**(2) (1994), 960–977.

[13] K. Marton and P.C. Shields, Correction: "Entropy and the consistent estimation of joint distributions" [Ann. Probab., **22**(2) (1994), 960–977; MR1288138 (95g:94004)], *Ann. Probab.*, **24**(1) (1996), 541–545.

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

[15] A.W. van der Vaart, *Asymptotic statistics*, Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge.

Denise Duarte
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais
31270-901 Belo Horizonte
BRAZIL

E-mail: denisedsma@yahoo.com.br

Antonio Galves
Instituto de Matemática e Estatística
Universidade de São Paulo
Caixa Postal 66281
05315-970 São Paulo
BRAZIL

E-mail: galves@ime.usp.br

Nancy L. Garcia
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas
Caixa Postal 6065
13081-970 Campinas
BRAZIL

E-mail: nancy@ime.unicamp.br