Brazilian Journal of Probability and Statistics
2010, Vol. 24, No. 2, 321–336
DOI: 10.1214/09-BJPS033

Some upper bounds for the rate of convergence of penalized likelihood context tree estimators

Florencia Leonardi

Instituto de Matemática e Estatística, Universidade de São Paulo

Abstract. We find upper bounds for the probability of underestimation and overestimation errors in penalized likelihood context tree estimation. The bounds are explicit and applies to processes of not necessarily finite memory. We allow for general penalizing terms and we give conditions over the maximal depth of the estimated trees in order to get strongly consistent estimates. This generalizes previous results obtained in the case of estimation of the order of a Markov chain.

1 Introduction

In this paper we obtain an exponential upper bound for the underestimation of the *context tree* of a variable memory process by penalized likelihood (PL) criteria and a subexponential upper bound for the overestimation event. Our result applies to processes of not necessarily finite memory that satisfies some continuity requirements, generalizing the bound obtained in Dorea and Zhao (2006) for the estimation of the order of a Markov chain by similar methods (EDC criterion).

The concept of context tree was first introduced by Rissanen (1983) to denote the minimum set of sequences that are necessary to predict the next symbol in a finite memory stochastic chain. A particular case of context tree is the set of all sequences of length k, representing a Markov chain of order k. For that reason, context trees allow a more detailed and parsimonious representation of processes than finite order Markov chains do.

In the statistical literature, the processes allowing a context tree representation are called Variable Length Markov Chains [Bühlmann and Wyner (1999)]. This class of models has shown to be useful in real data modeling as, for example, for the case of protein classification into families [Bejerano and Yona (2001), Leonardi (2006)].

Historically, the estimation of the context tree of a process has been addressed by different versions of the algorithm *Context*, introduced by Rissanen in its seminal paper. This algorithm was proven to be weak consistent in the case of bounded memory [Bühlmann and Wyner (1999)] and also in the case of unbounded memory [Ferrari and Wyner (2003), Duarte, Galves and Garcia (2006)]. Recently, in

Key words and phrases. Context tree, penalized maximum likelihood estimation, Bayesian information criterion, rate of convergence.

Received December 2008; accepted June 2009.

Galves, Maume-Deschamps and Schmitt (2008) it was obtained an upper bound for the rate of convergence of the algorithm Context in the case of bounded memory processes. A generalization of this result to the case of unbounded memory processes was given in Galves and Leonardi (2008).

The estimation of context trees by PL criteria had not been addressed in the literature until the recent work by Csiszár and Talata (2006). The reason for that was the exponential cost of the estimation, due to the number of trees that had to be considered in order to find the optimal one. In their article, Csiszár and Talata showed that the Bayesian Information Criterion (BIC), which is a particular case of the PL estimators (using a penalizing term growing logarithmically), is strongly consistent and can be computed in linear time, using a suitable version of the Context Tree Weighting method of Willems, Shtarkov and Tjalkens [Willems, Shtarkov and Tjalkens (1995), Willems (1998)]. Their result applies to unbounded memory processes and the depth of the estimated tree is allowed to grow with the sample size as a sublogarithmic function. This last condition was proven to be unnecessary in the case of finite memory processes, as proven in Garivier (2006). An explicit bound on the rate of convergence of the PL context tree estimators had remained until now as an open question.

The paper is organized as follows. In Section 2 we introduce some definitions and state the main result. In Section 3 we present the proofs and in Section 4 we do some final remarks. Finally, the Appendix contains some results needed in our proofs and obtained elsewhere in the literature.

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 $\ell(\lambda) = 0$. In the sequel A^j will denote the set of all sequences of length j over A.

Given two sequences $v = v_j^k$ and $w = w_m^n$, we will denote by vw the sequence of length $\ell(v) + \ell(w)$ obtained by concatenating the two strings, with the symbols in v preceding the symbols in w. 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, $v = (\ldots, v_{-2}, v_{-1})$, denoted by $v = v_{-\infty}^{-1}$.

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 $s \preceq w$.

Definition 2.1. A set $\mathcal{T}$ of finite or semi-infinite sequences 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 $h(\mathcal{T}) < +\infty$ we say that $\mathcal{T}$ is *bounded* and we denote by $|\mathcal{T}|$ the number of sequences in $\mathcal{T}$. On the other hand, if $h(\mathcal{T}) = +\infty$ 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,

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

The expression $\mathrm{Int}(\mathcal{T})$ will denote the set of all sequences that are suffixes of some $u \in \mathcal{T}$, that is,

$$\mathrm{Int}(\mathcal{T}) = \{w : w \prec u \text{ for some } u \in \mathcal{T}\}.$$

We will say that a tree $\mathcal{T}$ is *irreducible* if no $w \in \mathcal{T}$ can be replaced by a suffix without violating the suffix property. On the other hand, a tree $\mathcal{T}$ will be *complete* if for every semi-infinite sequence $w_{-\infty}^{-1}$ there exists a sequence $s \in \mathcal{T}$ such that $s \preceq w_{-\infty}^{-1}$.

Consider a stationary ergodic 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

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

In the sequel we will use the simpler notation X_t for the process $\{X_t : t \in \mathbb{Z}\}$.

Definition 2.2. A sequence $w \in A^j$ is a *context* for the process X_t if it satisfies:

(a) For any semi-infinite sequence $x_{-\infty}^{-1}$ having w as a suffix

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

(b) No suffix of w satisfies (a).

An *infinite context* is a semi-infinite sequence $w_{-\infty}^{-1}$ such that none of its suffixes w_{-j}^{-1}, $j = 1, 2, \ldots$ is a context.

Definition 2.2 implies that the set of all contexts (finite or infinite) satisfies the suffix property and hence it is a tree. This tree is called *the context tree* of the process X_t and will be denoted by $\mathcal{T}_0$.

Remark 2.3. In this paper we will also consider i.i.d. processes. We assume that these processes have as context tree the set $\mathcal{T}_0 = \{\lambda\}$.

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

$$\alpha_0 := \inf_{x_{-\infty}^{-1}, a\in A} \{p(a|x_{-\infty}^{-1})\},$$

$$\alpha_k := \inf_{u\in A^k} \sum_{a\in A} \inf_{x_{-\infty}^{-1}} p(a|x_{-\infty}^{-1}u), \qquad k \geq 1.$$

It is important to note that $\{\alpha_k\}_{k\in\mathbb{N}}$ is a nondecreasing sequence such that $0 \leq \alpha_k \leq 1$ for all k. Moreover, if the tree $\mathcal{T}_0$ is bounded, that is $h(\mathcal{T}_0) = K$ for some $K \in \mathbb{N}$, then $\alpha_k = 1$ for all $k \geq K$. In the unbounded case we will assume the sequence $\{\alpha_k\}$ converges to 1 sufficiently fast. This is related to the loss of memory of a process of infinite order [see Comets, Fernández and Ferrari (2002) and Galves and Leonardi (2008) for more details].

Assumption 1. From now on we will assume the process X_t satisfies the following conditions:

1. $\alpha_0 > 0$ and
2. $\alpha := \sum_{k\in\mathbb{N}}(1-\alpha_k) < +\infty$.

The positivity assumption over α_0 implies that the context tree of the process X_t is complete, that is, any semi-infinite sequence $w_{-\infty}^{-1}$ either belongs to $\mathcal{T}_0$ or has a suffix that belongs to $\mathcal{T}_0$.

In what follows we will assume $x_1, x_2, \ldots, x_n$ is a sample of the process X_t. Let $d(n) < n$ be a function taking integer values and growing to infinity with n. This will denote the maximal height of the estimated context trees (and will be denoted simply by d). Then, given a sequence w, with $1 \leq \ell(w) \leq d$, and a symbol $a \in A$ we denote by $N_n(w,a)$ the number of occurrences of symbol a preceded by the sequence w, starting at $d+1$, that is,

$$N_n(w,a) = \sum_{t=d+1}^{n} \mathbf{1}\{x_{t-\ell(w)}^{t-1} = w, x_t = a\}.$$

On the other hand, $N_n(w)$ will denote the sum $\sum_{a\in A} N_n(w,a)$.

Definition 2.4. We will say that the tree $\mathcal{T}$ is *feasible* if it is irreducible, $h(\mathcal{T}) \leq d$, $N_n(w) \geq 1$ for all $w \in \mathcal{T}$ and any string w' with $N_n(w') \geq 1$ either belongs to $\mathcal{T}$, is a suffix of some $w \in \mathcal{T}$ or has a suffix w that belongs to $\mathcal{T}$.

We will denote by $\mathcal{F}^d(x_1^n)$ the set of all feasible trees and, given a sequence w, we will denote by $\mathcal{F}_w^d(x_1^n)$ the set of trees $\mathcal{T}'$ such that $\mathcal{T}' = \mathcal{T} \cap \{u : u \succeq w\}$, with $\mathcal{T} \in \mathcal{F}^d(x_1^n)$. Then, given a tree $\mathcal{T} \in \mathcal{F}^d(x_1^n)$ and a family of transition probabilities associated to $\mathcal{T}$, that is a set $q = \{q(a|w) : a \in A, w \in \mathcal{T}\}$, we have that the

likelihood of the sequence $x_1, \ldots, x_n$ given $\mathcal{T}$ and q (conditioned on the first d symbols $x_1, \ldots, x_d$) is given by

$$L_{(\mathcal{T},q)}(x_1^n) = \prod_{w \in \mathcal{T}} \prod_{a \in A} q(a|w)^{N_n(w,a)}. \tag{2.1}$$

We are interested in the family q of transition probabilities that maximizes $L_{(\mathcal{T},q)}(x_1^n)$ for a given context tree $\mathcal{T}$. Then, maximizing (2.1) subject to the restrictions $\sum_{a \in A} p(a|w) = 1$ for all $w \in \mathcal{T}$ we obtain that the family of maximum likelihood estimators of the transition probabilities is given by

$$\hat{p}_n(a|w) = \frac{N_n(w,a)}{N_n(w)}, \qquad w \in \mathcal{T}, a \in A.$$

Note that by Definition 2.4, as $N_n(w) \geq 1$ for any $w \in \mathcal{T}$, it is not necessary to give an extra definition of $\hat{p}_n(a|w)$ in the case $N_n(w) = 0$. Therefore, the maximum likelihood of the sequence $x_1, \ldots, x_n$ (conditioned on $x_1, \ldots, x_d$) is given by

$$\hat{\mathbb{P}}_{\mathrm{ML},\mathcal{T}}(x_1^n) = \prod_{w \in \mathcal{T}} \prod_{a \in A} \hat{p}_n(a|w)^{N_n(w,a)}. \tag{2.2}$$

Here and in the sequel we use the convention $0^0 = 1$, for example, in the case of $N_n(w,a) = 0$ in expression (2.2).

Given a sequence w, with $N_n(w) \geq 1$, we will denote by

$$\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n) = \prod_{a \in A} \hat{p}_n(a|w)^{N_n(w,a)}.$$

Hence, we have

$$\hat{\mathbb{P}}_{\mathrm{ML},\mathcal{T}}(x_1^n) = \prod_{w \in \mathcal{T}} \hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n).$$

Let $f(n)$ be any positive function such that $f(n) \to +\infty$, when $n \to +\infty$, and $n^{-1} f(n) \to 0$, when $n \to +\infty$. This function will represent the generic penalizing term of our estimator, replacing the function $\frac{|A|-1}{2} \log n$ in the classical definition of BIC [Csiszár and Talata (2006)]. A function satisfying these conditions will be called *penalizing term*.

Definition 2.5. Given a penalizing term $f(n)$, the PL context tree estimator is given by

$$\hat{\mathcal{T}}(x_1^n) = \underset{\mathcal{T} \in \mathcal{F}^d(x_1^n)}{\arg\min} \{-\log \hat{\mathbb{P}}_{\mathrm{ML},\mathcal{T}}(x_1^n) + |\mathcal{T}| f(n)\}.$$

Remark 2.6. Here and throughout the rest of the paper, we will assume the logarithm is taken to the base 2.

As can be seen, the computation of the estimated context tree using its raw definition would imply a search for the optimal tree on the set of all feasible trees. This was the biggest drawback of this approach, because the size of this set grows extremely fast as a function of the maximal height d. Fortunately, there is a way of computing the PL estimator without exploring the set of all trees, as shown by Csiszár and Talata (2006). The details of this algorithm are given in the Appendix and will be used in the proof of our main result.

Let $K \in \mathbb{N}$. Define the underestimation event with respect to the truncated tree $\mathcal{T}_0|_K$ by

$$U_n^K = \bigcup_{w \in \mathrm{Int}(\mathcal{T}_0|_K)} \{w \in \hat{\mathcal{T}}_n(x_1^n)\}$$

and the overestimation event by

$$O_n^K = \bigcup_{v \succ w \in \mathcal{T}_0, \ell(w) < K} \{v \in \hat{\mathcal{T}}_n(x_1^n)\}.$$

We are ready to present the main result in this paper. It establishes upper bounds for the probability of occurrence of the underestimation and overestimation events.

Theorem 2.7. *Let $x_1, x_2, \ldots$ be a sample of the stationary ergodic stochastic process X_t having context tree $\mathcal{T}_0$ and satisfying Assumption* 1. *For any constant $K \in \mathbb{N}$ there exist an integer n_0 and positive constants c_1, c_2, c_3 and c_4 depending on the process X_t such that*:

(a) $\mathbb{P}[U_n^K] \le c_1 |A|^K e^{-c_2(n-d)}$ *for all* $n \ge n_0$;
(b) $\mathbb{P}[O_n^K] \le c_3 |A|^{d+K} e^{-c_4 f(n) \alpha_0^{2d}/(d+1)}$ *for all* $n \ge 1$.

Corollary 2.8. *Let $f(n)$ be a penalizing term and $d(n)$ be a function such that for any constant $c > 0$,*

$$\sum_{n \in \mathbb{N}} |A|^{d(n)} \exp\left[-\frac{f(n) c^{d(n)}}{d(n)}\right] < +\infty. \tag{2.3}$$

Then, for almost every infinite sample $x_1, x_2, \ldots$ we have that $\hat{\mathcal{T}}_n(x_1^n)|_K = \mathcal{T}_0|_K$ for any n sufficiently large.

3 Proof of Theorem 2.7

Using Definition A.3 and Lemma A.5 we see that the tree in (2.5) can be written as

$$\hat{\mathcal{T}}(x_1^n) = \left\{ w \in \bigcup_{j=1}^{d} A^j : \mathcal{X}_w(x_1^n) = 0, \mathcal{X}_v(x_1^n) = 1 \text{ for all } v \prec w \right\}$$

if $\mathcal{X}_\lambda(x_1^n) = 1$ and $\hat{\mathcal{T}}(x_1^n) = \{\lambda\}$ if $\mathcal{X}_\lambda(x_1^n) = 0$. Then we have

$$U_n^K \subset \bigcup_{w \in \mathrm{Int}(\mathcal{T}_0|_K)} \{\mathcal{X}_w(x_1^n) = 0\}$$

and

$$O_n^K \subset \bigcup_{w \in \mathcal{T}_0, \ell(w) < K} \{\mathcal{X}_w(x_1^n) = 1\}.$$

To prove (a) let $w \in \mathrm{Int}(\mathcal{T}_0|_K)$. Then, as w is not a context there must exists a finite complete tree $\mathcal{T}$ such that

$$\begin{aligned}\delta_{\mathcal{T}}(w) &= \sum_{a \in A}\Big[\sum_{u \in \mathcal{T}_w} p(ua)\log p(a|u) - p(wa)\log p(a|w)\Big] \\ &= \sum_{u \in \mathcal{T}_w} D\big(p(\cdot|u) \parallel p(\cdot|w)\big) > 0,\end{aligned}$$

where $\mathcal{T}_w = \mathcal{T} \cap \{u : u \succeq w\} \neq \{w\}$ and D is the *Kullback–Leibler divergence* between the two distributions $p(\cdot|u)$ and $p(\cdot|w)$ over A (see the Appendix). Then, using Definition A.2 we have that

$$\begin{aligned}\mathbb{P}[\mathcal{X}_w(x_1^n) = 0] \le \mathbb{P}\Big[\prod_{a \in A} V_{aw}(x_1^n) \le e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n), \mathcal{T}_w \in \mathcal{F}_w^d(x_1^n)\Big] \\ + \mathbb{P}[\mathcal{T}_w \notin \mathcal{F}_w^d(x_1^n)]\end{aligned}$$

and by Lemma A.4, for any $a \in A$

$$V_{aw}(x_1^n) = \max_{\mathcal{T}' \in \mathcal{F}_{aw}^d(x_1^n)} \prod_{s \in \mathcal{T}'} e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},s}(x_1^n).$$

Notice that if $\mathcal{T}_w \in \mathcal{F}_w^d(x_1^n)$, then $\mathcal{T}_{aw} \in \mathcal{F}_{aw}^d(x_1^n)$ for all $a \in A$, because $\mathcal{T}_w \neq \{w\}$ and it is complete. Therefore,

$$\prod_{u \in \mathcal{T}_w} e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n) = \prod_{a \in A}\prod_{s \in \mathcal{T}_{aw}} e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},s}(x_1^n) \le \prod_{a \in A} V_{aw}(x_1^n)$$

and

$$\begin{aligned}&\mathbb{P}\Big[\prod_{a \in A} V_{aw}(x_1^n) \le e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n), \mathcal{T}_w \in \mathcal{F}_w^d(x_1^n)\Big] \\ &\quad \le \mathbb{P}\Big[\prod_{u \in \mathcal{T}_w} e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n) \le e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n), \mathcal{T}_w \in \mathcal{F}_w^d(x_1^n)\Big] \\ &\quad = \mathbb{P}\Big[\sum_{u \in \mathcal{T}_w} \log\hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n) - \log\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n) \le (|\mathcal{T}_w| - 1)f(n), \\ &\qquad \mathcal{T}_w \in \mathcal{F}_w^d(x_1^n)\Big].\end{aligned}$$

Dividing by $n-d$ and subtracting on both sides the term $\delta_{\mathcal{T}}(w)$ we have that, for all $n \geq n_0$ we can bound above the last expression by

$$\mathbb{P}\Big[|L_n(w)| > \frac{\delta_{\mathcal{T}}(w)}{4}, N_n(w) \geq 1\Big] + \sum_{u \in \mathcal{T}_w} \mathbb{P}\Big[|L_n(u)| > \frac{\delta_{\mathcal{T}}(w)}{4|\mathcal{T}_w|}, N_n(u) \geq 1\Big],$$

where for any finite sequence s

$$L_n(s) = \sum_{a \in A} p(sa) \log p(a|s) - \frac{N_n(s,a)}{n-d} \log \hat{p}_n(a|s)$$

and n_0 is a sufficiently large integer such that $h(\mathcal{T}_w) \leq d(n)$ and for all $n \geq n_0$,

$$\frac{|\mathcal{T}_w| f(n)}{n-d} \leq \frac{|\mathcal{T}_w| f(n_0)}{n_0 - d} < \frac{\delta_{\mathcal{T}}(w)}{2}.$$

Using part (c) of Corollary A.7 we can bound above this expression by

$$3e^{1/e}|A|^2(1+|\mathcal{T}_w|)\exp\Big[-\frac{(n-d)\min(1,(\delta_{\mathcal{T}}(w)/4|\mathcal{T}_w|)^2)\alpha_0^{2(h(\mathcal{T}_w)+1)}}{64e|A|^3(\alpha+\alpha_0)\log^2\alpha_0(h(\mathcal{T}_w)+1)}\Big], \tag{3.1}$$

by noticing that $p(w) \geq p(u) \geq \alpha_0^{h(\mathcal{T}_w)}$ for any $u \in \mathcal{T}_w$. On the other hand, using part (a) of Corollary A.7 we obtain that

$$\begin{aligned}\mathbb{P}[\mathcal{T}_w \notin \mathcal{F}_w^d(x_1^n)] &\leq \mathbb{P}\Big[\bigcup_{u \in \mathcal{T}_w} \{N_n(u) = 0\}\Big] \\ &\leq e^{1/e}|A||\mathcal{T}_w|\exp\Big[-\frac{(n-d)\alpha_0^{2h(\mathcal{T}_w)+1}}{8e(\alpha+\alpha_0)|A|^2(h(\mathcal{T}_w)+1)}\Big].\end{aligned} \tag{3.2}$$

We conclude the proof of part (a) by observing that we only have a finite number of sequences $w \in \mathrm{Int}(\mathcal{T}_0|_K)$, so we can sum (3.1) and (3.2) and take

$$c_1 = 4e^{1/e}|A|^2\Big(1 + \max_{w \in \mathrm{Int}(\mathcal{T}_0|_K)} |\mathcal{T}_w|\Big)$$

and

$$c_2 = \min_{w \in \mathrm{Int}(\mathcal{T}_0|_K)} \Big\{\frac{\min(1,(\delta_{\mathcal{T}}(w)/4|\mathcal{T}_w|)^2)\alpha_0^{2(h(\mathcal{T}_w)+1)}}{64e|A|^3(\alpha+\alpha_0)\log^2\alpha_0(h(\mathcal{T}_w)+1)}\Big\}.$$

To prove part (b) observe that for any $w \in \mathcal{T}_0$ with $\ell(w) < K$

$$\mathbb{P}[\mathcal{X}_w(x_1^n) = 1] = \mathbb{P}\Big[\prod_{a \in A} V_{aw}(x_1^n) > e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n)\Big]. \tag{3.3}$$

Using Lemma A.4 we have that

$$\prod_{a \in A} V_{aw}(x_1^n) = \prod_{u \in \mathcal{T}_w(x_1^n)} e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n).$$

Then, applying the logarithm function the probability (3.3) is equal to

$$\begin{aligned}
&\mathbb{P}\Big[\sum_{u\in\mathcal{T}_w(x_1^n)} \log e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n) > \log e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n)\Big] \\
&\quad = \mathbb{P}\Big[\log \hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n) - \sum_{u\in\mathcal{T}_w(x_1^n)} \log \hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n) \\
&\qquad < \big(1 - |\mathcal{T}_w(x_1^n)|\big) f(n)\Big].
\end{aligned} \tag{3.4}$$

We know, by the maximum likelihood estimators of the transition probabilities that

$$\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n) \geq \prod_{a\in A} p(a|w)^{N_n(w,a)}.$$

Therefore, we can bound above the right-hand side of (3.4) by

$$\begin{aligned}
&\mathbb{P}\Big[\sum_{a\in A} N_n(w,a)\log p(a|w) - \sum_{u\in\mathcal{T}_w(x_1^n)} \log \hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n) < \big(1 - |\mathcal{T}_w(x_1^n)|\big) f(n)\Big] \\
&\quad = \mathbb{P}\Big[-\sum_{a\in A}\sum_{u\in\mathcal{T}_w(x_1^n)} N_n(u,a)\log\frac{\hat{p}_n(a|u)}{p(a|u)} < \big(1 - |\mathcal{T}_w(x_1^n)|\big) f(n)\Big].
\end{aligned}$$

This equality follows by substituting $N_n(w,a)$ by $\sum_{u\in\mathcal{T}_w(x_1^n)} N_n(u,a)$ and the fact that $p(a|u) = p(a|w)$ for all $u \in \mathcal{T}_w(x_1^n)$, remembering that $w \in \mathcal{T}_0$. Observe that

$$\begin{aligned}
\sum_{a\in A}\sum_{u\in\mathcal{T}_w(x_1^n)} N_n(u,a)\log\frac{\hat{p}_n(a|u)}{p(a|u)} &= \sum_{u\in\mathcal{T}_w(x_1^n)} N_n(u)\sum_{a\in A}\hat{p}_n(a|u)\log\frac{\hat{p}_n(a|u)}{p(a|u)} \\
&= \sum_{u\in\mathcal{T}_w(x_1^n)} N_n(u) D\big(\hat{p}_n(\cdot|u) \parallel p(\cdot|u)\big).
\end{aligned}$$

Then, using Lemma A.1 and dividing by $n-d$ we have that

$$\begin{aligned}
&\mathbb{P}\Big[-\sum_{u\in\mathcal{T}_w(x_1^n)} N_n(u) D\big(\hat{p}_n(\cdot|u) \parallel p(\cdot|u)\big) < \big(1 - |\mathcal{T}_w(x_1^n)|\big) f(n)\Big] \\
&\quad \leq \mathbb{P}\Big[-\sum_{u\in\mathcal{T}_w(x_1^n)} \frac{N_n(u)}{n-d}\sum_{a\in A}\frac{[\hat{p}_n(a|u) - p(a|u)]^2}{p(a|u)} < \frac{(1 - |\mathcal{T}_w(x_1^n)|) f(n)}{n-d}\Big].
\end{aligned}$$

As $\mathcal{X}_w(x_1^n) = 1$ it follows that $|\mathcal{T}_w(x_1^n)| \geq 2$. On the other hand, $1 \leq N_n(u) \leq n-d$ for any $u \in \mathcal{T}_w(x_1^n)$. Therefore, we can bound above the right-hand side of the last expression by

$$\sum_{u\in\mathcal{T}_w(x_1^n)}\sum_{a\in A}\mathbb{P}\Big[\big|\hat{p}_n(a|u) - p(a|u)\big| > \sqrt{\frac{f(n)p(a|u)}{2|A|(n-d)}}, N_n(u) \geq 1\Big].$$

Hence, using part (b) of Corollary A.7 we can bound above this expression by

$$e^{1/e}(|A|+1)|A|^{d+1}\exp\left[-\frac{f(n)\alpha_0^{2(d+1)}}{64e(\alpha+\alpha_0)|A|^3(d+1)}\right].$$

This finishes the proof of Theorem 2.7, by taking

$$c_3 = 2e^{1/e}|A|^2 \quad \text{and} \quad c_4 = \frac{\alpha_0^2}{64e(\alpha+\alpha_0)|A|^3}.$$

Proof of Corollary 2.8. It follows from the Borel–Cantelli lemma and Theorem 2.7, by noting that

$$\mathbb{P}[\hat{\mathcal{T}}(x_1^n)|_K \neq \mathcal{T}_0|_K] \leq \mathbb{P}[U_n^K] + \mathbb{P}[O_n^K]$$

and the right-hand side is summable in n when condition (2.3) is satisfied. □

4 Final remarks

The present paper presents upper bounds for the rate of convergence of penalized likelihood context tree estimators. We obtain an exponential bound for the underestimation event and an under-exponential bound in the case of the overestimation event. These results generalizes the previous work by Dorea and Zhao (2006), who obtained similar bounds in the case of the estimation of the order of a Markov chain, using also penalized likelihood criteria. One question that still remains open is if these bounds are optimal, as in the case of an estimator introduced in Finesso, Liu and Narayan (1996) for the estimation of the order of a Markov chain. They prove that in the case of their estimator, the constant appearing in the underestimation bound is optimal, and that the overestimation bound cannot be exponential if the estimator is universal, as in our case. The answer to these questions are important subjects for future work in this area.

Appendix

A.1 The context tree maximizing principle

The following definitions and results were taken from Csiszár and Talata (2006) and were included for completeness. Definitions A.2 and A.3 and Lemmas A.4 and A.5 were originally proven for the usual penalizing term $f(n) = \frac{|A|-1}{2}\log n$, but can be adapted in a straightforward way to our setting.

Given two probability distributions p and q over A, the *Kullback–Leibler divergence* is defined by

$$D(p \parallel q) = \sum_{a\in A} p(a)\log\frac{p(a)}{q(a)},$$

where, by convention, $p(a)\log\frac{p(a)}{q(a)}=0$ if $p(a)=0$ and $p(a)\log\frac{p(a)}{q(a)}=+\infty$ if $p(a)>q(a)=0$. Using Jensen's inequality it can be seen that $D(p\parallel q)\geq 0$ for all p and q and $D(p\parallel q)=0$ if and only if $p(a)=q(a)$ for all $a\in A$.

Lemma A.1. *If p and q are two probability distributions over A then*

$$D(p\parallel q)\leq\sum_{a\in A}\frac{[p(a)-q(a)]^2}{q(a)}.$$

Proof. See Csiszár and Talata (2006), Lemma 6.3. □

Assume we have a sample $x_1,\dots,x_n$ of the process X_t. Consider the full tree A^d, and let S^d denote the set of all sequences of length at most d, that is $S^d=\bigcup_{j=0}^d A^j$.

Definition A.2. Given a sequence $w\in S^d$ with $N_n(w)\geq 1$, we define recursively, starting from the sequences of the full tree A^d, the value

$$V_w(x_1^n)=\begin{cases}\max\left\{e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n),\prod_{a\in A}V_{aw}(x_1^n)\right\}, & \text{if } 0\leq\ell(w)<d,\\ e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n), & \text{if } \ell(w)=d\end{cases}$$

and the indicator

$$\mathcal{X}_w(x_1^n)=\begin{cases}1, & \text{if } 0\leq\ell(w)<d \quad\text{and}\quad \prod_{a\in A}V_{aw}(x_1^n)>e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n),\\ 0, & \text{if } 0\leq\ell(w)<d \quad\text{and}\quad \prod_{a\in A}V_{aw}(x_1^n)\leq e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n),\\ 0, & \text{if } \ell(w)=d.\end{cases}$$

Definition A.3. Given $w\in S^d$ with $N_n(w)\geq 1$, the maximizing tree assigned to the sequence w is the tree

$$\mathcal{T}_w(x_1^n)=\{u\in S^d:\mathcal{X}_u(x_1^n)=0,\mathcal{X}_v(x_1^n)=1 \text{ for all } w\preceq v\prec u\}$$

if $\mathcal{X}_w(x_1^n)=1$ and $\mathcal{T}_w(x_1^n)=\{w\}$ if $\mathcal{X}_w(x_1^n)=0$.

Lemma A.4. *For any $w\in S^d$ with $N_n(w)\geq 1$,*

$$V_w(x_1^n)=\max_{\mathcal{T}\in\mathcal{F}_w^d(x_1^n)}\prod_{u\in\mathcal{T}}e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n)=\prod_{u\in\mathcal{T}_w(x_1^n)}e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n).$$

Moreover, if $\mathcal{X}_w(x_1^n)=1$ then

$$\prod_{a\in A}V_{aw}(x_1^n)=\prod_{u\in\mathcal{T}_w(x_1^n)}e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},u}(x_1^n).$$

Proof. The first two equalities follows directly from Csiszár and Talata (2006), Lemma 4.4. To prove the last equality it is sufficient to observe that, by Definition A.3, if $\mathcal{X}_w(x_1^n) = 1$ we have

$$\mathcal{T}_w(x_1^n) = \bigcup_{a \in A} \mathcal{T}_{aw}(x_1^n)$$

and this union is disjoint. □

Lemma A.5. *The context tree estimator* $\hat{\mathcal{T}}(x_1^n)$ *in* (2.5) *equals the maximizing tree assigned to the empty string* λ, *that is*,

$$\hat{\mathcal{T}}(x_1^n) = \mathcal{T}_\lambda(x_1^n).$$

Proof. See Csiszár and Talata (2006), Proposition 4.3. □

From this result it follows that in order to obtain the tree maximizing the penalized maximum likelihood criteria it is sufficient to assign to each sequence $w \in S^d$, with $N_n(w) \geq 1$, the indicator $\mathcal{X}_w(x_1^n)$ and then to get the maximizing tree $\mathcal{T}_\lambda(x_1^n)$. The computational cost of this algorithm is linear in n if $d(n) = o(\log n)$, as proven by Csiszár and Talata (2006).

A.2 Exponential inequalities for empirical probabilities

The following result was proven in Galves and Leonardi (2008); we omit its proof here.

Theorem A.6. *Assume the process* X_t *satisfies Assumption* 1, *then for any finite sequence* w, *any symbol* $a \in A$ *and any* $t > 0$ *the following inequality holds*:

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

where

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

As a consequence of Theorem A.6 we obtain the following corollary.

Corollary A.7. *For any finite sequence* w *and any* $t > 0$ *the following inequalities hold*:

(a) *For any* $n > \frac{t}{p(w)} + d$,

$$\mathbb{P}\big(N_n(w) \leq t\big) \leq e^{1/e}|A| \exp\left[-\frac{(n-d)(p(w) - t/(n-d))^2 C}{|A|^2(\ell(w)+1)}\right];$$

(b) *For any* $a \in A$,

$$\mathbb{P}\big(|\hat{p}_n(a|w) - p(a|w)| > t, N_n(w) \geq 1\big) \leq e^{1/e}(|A|+1)\exp\Big[-\frac{(n-d)t^2 p(w)^2 C}{4|A|^2(\ell(w)+1)}\Big];$$

(c) *Let* $L_n(w) = \sum_{a\in A} p(wa)\log p(a|w) - \frac{N_n(w,a)}{n-d}\log \hat{p}_n(a|w)$. *Then,*

$$\mathbb{P}[|L_n(w)| > t, N_n(w) \geq 1] \leq 3e^{1/e}|A|^2 \exp\Big[-\frac{(n-d)\min(t,t^2)p(w)^2\alpha_0 C}{8|A|^3\log^2\alpha_0(\ell(w)+1)}\Big].$$

Proof. To prove (a) observe that, if $n > \frac{t}{p(w)} + d$, we have

$$\begin{aligned}\mathbb{P}\big(N_n(w) \leq t\big) &= \mathbb{P}\Big(\sum_{a\in A} N_n(w,a) - (n-d)p(wa) \leq t - (n-d)p(w)\Big)\\ &\leq \mathbb{P}\Big(|A|\max_{a\in A}|N_n(w,a) - (n-d)p(wa)| \geq (n-d)p(w) - t\Big)\\ &\leq \sum_{a\in A}\mathbb{P}\Big(|N_n(w,a) - (n-d)p(wa)| \geq \frac{(n-d)p(w)-t}{|A|}\Big).\end{aligned}$$

Using Theorem A.6 we can bound above the last expression by

$$e^{1/e}|A|\exp\Big[-\frac{(n-d)[p(w)-t/(n-d)]^2 C}{|A|^2(\ell(w)+1)}\Big].$$

This finishes the proof of (a). To prove (b) observe that

$$p(a|w) = \frac{(n-d)p(wa)}{(n-d)p(w)}.$$

Then, summing and substracting the term $\frac{N_n(w,a)}{(n-d)p(w)}$ we obtain

$$\begin{aligned}\Big|\frac{N_n(w,a)}{N_n(w)} - \frac{(n-d)p(wa)}{(n-d)p(w)}\Big| &\leq \frac{N_n(w,a)}{N_n(w)(n-d)p(w)}|(n-d)p(w) - N_n(w)|\\ &\quad + \frac{1}{(n-d)p(w)}|N_n(w,a) - (n-d)p(wa)|.\end{aligned}$$

Therefore we have

$$\begin{aligned}&\mathbb{P}\big(|\hat{p}_n(a|w) - p(a|w)| > t, N_n(w) \geq 1\big)\\ &\quad\leq \mathbb{P}\Big(|(n-d)p(w) - N_n(w)| > \frac{t(n-d)p(w)}{2}\Big)\\ &\qquad + \mathbb{P}\Big(|N_n(w,a) - (n-d)p(wa)| > \frac{t(n-d)p(w)}{2}\Big).\end{aligned}$$

We can write $N_n(w) = \sum_{b\in A} N_n(w,b)$ and $p(w) = \sum_{b\in A} p(wb)$, then the right-hand side of the last inequality can be bounded above by the sum

$$\sum_{b\in A} \mathbb{P}\Big(|N_n(w,b) - (n-d)p(wb)| > \frac{t(n-d)p(w)}{2|A|}\Big) + \mathbb{P}\Big(|N_n(w,a) - (n-d)p(wa)| > \frac{t(n-d)p(w)}{2}\Big).$$

Using Theorem A.6 we can bound above this expression by

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

Finally, to prove (c) we have that

$$\begin{aligned}\mathbb{P}[|L_n(w)| > t, N_n(w) \geq 1] &\leq \mathbb{P}\Big[\Big|\sum_{a\in A} \log p(a|w)\Big(p(wa) - \frac{N_n(w,a)}{n-d}\Big)\Big| > \frac{t}{2}\Big] \\ &\quad + \mathbb{P}\Big[\Big|\sum_{a\in A} \frac{N_n(w,a)}{n-d} \log \frac{\hat{p}_n(a|w)}{p(a|w)}\Big| > \frac{t}{2}, N_n(w) \geq 1\Big].\end{aligned}$$

Observe that

$$\begin{aligned}&\mathbb{P}\Big[\Big|\sum_{a\in A} \log p(a|w)\Big(p(wa) - \frac{N_n(w,a)}{n-d}\Big)\Big| > \frac{t}{2}\Big] \\ &\quad \leq \sum_{a\in A} \mathbb{P}\Big[|N_n(w,a) - (n-d)p(wa)| > \frac{(n-d)t}{2|\log p(a|w)||A|}\Big]\end{aligned}$$

and this last expression can be bounded above using Theorem A.6 by

$$e^{1/e}|A|\exp\Big[\frac{-(n-d)t^2 C}{4|A|^2 \log^2 \alpha_0(\ell(w)+1)}\Big]. \tag{A.1}$$

On the other hand, using the definition of the *Kullback–Leibler divergence* and Lemma A.1 we obtain

$$\begin{aligned}&\mathbb{P}\Big[\Big|\sum_{a\in A} \frac{N_n(w,a)}{n-d} \log \frac{\hat{p}_n(a|w)}{p(a|w)}\Big| > \frac{t}{2}, N_n(w) \geq 1\Big] \\ &\quad \leq \mathbb{P}\Big[D\big(\hat{p}_n(\cdot|w) \parallel p(\cdot|w)\big) > \frac{t}{2}, N_n(w) \geq 1\Big] \\ &\quad \leq \sum_{a\in A} \mathbb{P}\Big[\big|p(a|w) - \hat{p}_n(a|w)\big| > \sqrt{\frac{tp(a|w)}{2|A|}}, N_n(w) \geq 1\Big]\end{aligned}$$

and the last expression can be bounded above using part (b) of this corollary by the expression

$$2e^{1/e}|A|^2 \exp\left[-\frac{(n-d)tp(w)^2\alpha_0 C}{8|A|^3(\ell(w)+1)}\right]. \tag{A.2}$$

Summing (A.1) and (A.2) we obtain the bound in part (c) and we conclude the proof of Corollary A.7. □

Acknowledgments

The author is thankful to Antonio Galves, Aurélien Garivier and Eric Moulines for interesting discussions about the subject. She also thanks the anonymous referees for useful comments to improve the presentation of this paper. This work is part of FAPESP's project *Consistent estimation of stochastic processes with variable length memory* (process 2009/09411-8) and CNPq's project *Stochastic systems with variable length interactions* (Edital Universal/CNPq-MCT 476501/2009-1).

References

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

Bühlmann, P. and Wyner, A. J. (1999). Variable length Markov chains. *The Annals of Statistics* **27** 480–513. MR1714720

Comets, F., Fernández, R. and Ferrari, P. (2002). Processes with long memory: Regenerative construction and perfect simulation. *The Annals of Applied Probability* **12** 921–943. MR1925446

Csiszár, I. and Talata, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Transactions on Information Theory* **52** 1007–1016. MR2238067

Dorea, C. and Zhao, L. (2006). Exponential bounds for the probability of wrong determination of the order of a Markov chain by using the edc criterion. *Journal of Statistical Planning and Inference* **136** 3689–3697. MR2284671

Duarte, D., Galves, A. and Garcia, N. (2006). Markov approximation and consistent estimation of unbounded probabilistic suffix trees. *Bulletin of the Brazilian Mathematical Society* **37** 581–592. MR2284889

Ferrari, F. and Wyner, A. (2003). Estimation of general stationary processes by variable length Markov chains. *Scandinavian Journal of Statistics* **30** 459–480. MR2002222

Finesso, L., Liu, C.-C. and Narayan, P. (1996). The optimal error exponent for Markov order estimation. *IEEE Transactions on Information Theory* **42** 1488–1497. MR1426225

Galves, A. and Leonardi, F. (2008). Exponential inequalities for empirical unbounded context trees. In *Progress in Probability* **60** 257–269. Birkhauser, Basel. MR2477385

Galves, A., Maume-Deschamps, V. and Schmitt, B. (2008). Exponential inequalities for VLMC empirical trees. *ESAIM: Probability and Statistics* **12** 43–45. MR2374639

Garivier, A. (2006). Consistency of the unlimited BIC context tree estimator. *IEEE Transactions on Information Theory* **52** 4630–4635. MR2300844

Leonardi, F. (2006). A generalization of the PST algorithm: Modeling the sparse nature of protein sequences. *Bioinformatics* **22** 1302–1307.

Rissanen, J. (1983). A universal data compression system. *IEEE Transactions on Information Theory* **29** 656–664. MR0730903

Willems, F. (1998). The context-tree weighting method: Extensions. *IEEE Transactions on Information Theory* **44** 792–798. MR1607711

Willems, F., Shtarkov, Y. and Tjalkens, T. (1995). The context-tree weighting method: Basic properties. *IEEE Transactions on Information Theory* **41** 653–664.

Instituto de Matemática e Estatística
Universidade de São Paulo
Rua do Matão 1010—Cidade Universitária
CEP 05508-090, São Paulo, SP
Brazil
E-mail: florencia@usp.br