Available online at www.sciencedirect.com
SciVerse ScienceDirect
Stochastic Processes and their Applications 121 (2011) 2488–2506
stochastic processes and their applications
www.elsevier.com/locate/spa

Context tree selection: A unifying view

A. Garivier[a,*], F. Leonardi[b]

[a] *LTCI, CNRS, Telecom ParisTech, 75634 Paris Cedex 13, France*
[b] *Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil*

Received 10 November 2010; received in revised form 27 June 2011; accepted 30 June 2011
Available online 14 July 2011

Abstract

Context tree models have been introduced by Rissanen in [25] as a parsimonious generalization of Markov models. Since then, they have been widely used in applied probability and statistics. The present paper investigates non-asymptotic properties of two popular procedures of context tree estimation: Rissanen's algorithm Context and penalized maximum likelihood. First showing how they are related, we prove finite horizon bounds for the probability of over- and under-estimation. Concerning over-estimation, no boundedness or loss-of-memory conditions are required: the proof relies on new deviation inequalities for empirical probabilities of independent interest. The under-estimation properties rely on classical hypotheses for processes of infinite memory. These results improve on and generalize the bounds obtained in Duarte et al. (2006) [12], Galves et al. (2008) [18], Galves and Leonardi (2008) [17], Leonardi (2010) [22], refining asymptotic results of Bühlmann and Wyner (1999) [4] and Csiszár and Talata (2006) [9].

Keywords: Algorithm Context; Penalized maximum likelihood; Model selection; Variable length Markov chains; Bayesian information criterion; Deviation inequalities

1. Introduction

Context tree models (CTM), first introduced by Jorma Rissanen in [25] as efficient tools in Information Theory, have been successfully studied and used since then in many fields of Probability and Statistics, including Bioinformatics [2,5], Universal Coding [27], Mathematical

* Corresponding author.
E-mail addresses: aurelien.garivier@telecom-paristech.fr, garivier@telecom-paristech.fr (A. Garivier), florencia@usp.br (F. Leonardi).

0304-4149/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.spa.2011.06.012

Statistics [4] or Linguistics [15]. Sometimes also called Variable Length Markov Chain (VLMC), a context tree process is informally defined as a Markov chain whose memory length depends on past symbols. This property makes it possible to represent the set of memory sequences as a tree, called the *context tree* of the process.

A remarkable tradeoff between expressivity and simplicity explains this success: no more difficult to handle than Markov chains, they appear to be much more flexible and parsimonious, including memory only where necessary. Not only do they provide more efficient models for fitting the data: it appears also that, in many applications, the shape of the context tree has a natural and informative interpretation. In Bioinformatics, the contexts trees of a sample have been useful to test the relevance of protein families databases [5] and in Linguistics, tree estimation highlights structural discrepancies between Brazilian and European Portuguese [15].

Of course, practical use of CTM requires the possibility of constructing efficient estimators of the model T_0 generating the data. It could be feared that, as a counterpart of the model multiplicity, increased difficulty would be encountered in model selection. Actually, this is not the case, and soon several procedures have been proposed and proved to be consistent. Roughly speaking, two families of context tree estimators are available. The first family, derived from the so-called algorithm *Context* introduced by Rissanen in [25], is based on the idea of *tree pruning*. They are somewhat reminiscent of the CART [3] pruning procedures: a measure of discrepancy between a node's children determines whether they have to be removed from the tree or not. The second family of estimators are based on a classical approach of mathematical statistics: *Penalized Maximum Likelihood* (PML). For each possible model, a criterion is computed which balances the quality of fit and the complexity of the model. In the framework of Information Theory, these procedures can be interpreted as derivations of the *Minimum Description Length* principle [1].

In the case of bounded memory processes, the problem of consistent estimation is clear: an estimator $\hat{T}$ is strongly consistent if it is equal to T_0 eventually almost surely as the sample size grows to infinity. As soon as 1983, Rissanen proved consistency results for the algorithm Context in this case. But later, the possibility of handling infinite memory processes was also addressed. In [9], an estimator $\hat{T}$ is called *strongly consistent* if for every positive integer K, its truncation $\hat{T}_{|K}$ at level K is equal to the truncation $T_0|_K$ of T_0 eventually almost surely. With this definition, PML estimators are shown to be strongly consistent if the penalties are appropriately chosen and if the maximization is restricted to a proper set of models. This last restriction was proven to be unnecessary in the finite memory case [19].

More recently, the problem of deriving *non-asymptotic* bounds for the probability of incorrect estimation was considered. In [18], non-universal inequalities were derived for a version of the algorithm Context in the case of finite context trees. These results were generalized to the case of infinite trees in [17], and to PML estimators in [22]. Using recent advances in weak dependence theory, all these results strongly rely on mixing hypotheses of the process.

For all these results, a distinction has to be made between two potential errors: under- and over-estimation. A context of T_0 is said to be *under-estimated* if one of its proper suffixes appears in the estimated tree $\hat{T}$, whereas it is called *over-estimated* if it appears as an internal node of $\hat{T}$. Over- and under-estimation appear to be of different natures: while under-estimation is eventually avoided by the existence of a strictly positive distance between a process and all processes with strictly smaller context trees, controlling over-estimation requires bounds on the fluctuations of empirical processes.

In this article, we present a unified analysis of the two families of context tree estimators. We contribute to a completely non-asymptotic analysis: we show that for appropriate parameters

and measure of discrepancy, the PML estimator is always smaller than the estimator given by the algorithm Context. To our knowledge, this is the first result comparing this two context tree selection methods.

Without restrictions on the (possibly infinite) context tree T_0, we prove that both methods provide estimators that are with high probability sub-trees of T_0 (i.e., a node that is not in T_0 does not appear in $\hat{T}$). These bounds are more precise and do not require the conditions assumed in [18,17,22]. For this purpose, we derive "*self-normalized*" non-asymptotic deviation inequalities, using martingale techniques inspired from proofs of the Law of the Iterated Logarithm [24,8]. These inequalities prove interesting in other fields, as for instance in reinforcement learning [21,14]. On the other hand, we derive upper bounds on the probability of under-estimation by assuming classical mixing conditions on the process generating the sample: with high probability, $\hat{T}$ contains every node of T_0 at moderate height. This result is based on exponential inequalities derived for a wider class of processes than in [18,17,22].

Our upper bounds on the probability of over- and under-estimation imply strong consistency of the PML estimators for a larger class of penalizing functions than in [22]. Similarly, in the case of the algorithm Context the strong consistency can also be derived for suitable threshold parameters, generalizing the convergence in probability for this estimator obtained previously in [12].

The paper is organized as follows. In Section 2 we set notation and definitions, we describe in detail the algorithms and we state our main results. The proof of these results is given in Section 3. In Section 4 we briefly discuss our results. Appendix A contains the statement and proof of the self-normalized deviation inequalities and Appendix B is devoted to the presentation of exponential inequalities for weak dependent processes.

2. Notations and results

In what follows, A is a finite alphabet; its size is denoted by $|A|$. A^j denotes the set of all sequences of length j over A, in particular A^0 has only one element, the empty sequence. We denote by $A^* = \bigcup_{k\geq 0} A^k$ the set of all finite sequences on alphabet A and A^∞ will denote the set of all semi-infinite sequences $v = (\ldots, v_{-2}, v_{-1})$ of symbols in A. The length of the sequence $w \in A^*$ is $|w|$. For $1 \leq i \leq j \leq |w|$, we denote $w_i^j = (w_i, \ldots, w_j) \in A^{j-i+1}$ and $v_{-\infty}^{-1}$ denotes the semi-infinite sequence $(\ldots, v_{-2}, v_{-1}) \in A^\infty$. Given $v \in A^* \cup A^\infty$ and $w \in A^*$, we denote by vw the sequence obtained by concatenating the two sequences v and w. We say that the sequence $s \in A^*$ is a *suffix* of the sequence $w \in A^* \cup A^\infty$ if there exists a sequence $u \in A^* \cup A^\infty$ such that $w = us$. In this case we write $w \succeq s$ or $s \preceq w$. When $|u| \geq 1$ we say that s is a *proper* suffix of w and we write $w \succ s$ or $s \prec w$.

A set $T \subset A^* \cup A^\infty$ is a *tree* if no sequence $s \in T$ is a proper suffix of another sequence $w \in T$. The *height* of the tree T is defined as

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

If $h(T) < +\infty$ we say that T is *bounded* and we denote by $|T|$ the cardinality of T. If $h(T) = +\infty$ we say that T is *unbounded*. The elements of T are also called the *leaves* of T. An *internal node* of T is a proper suffix of a leaf. For any sequence $w \in A^* \cup A^\infty$ and for any tree T, we define the tree T_w as the set of leaves in T which have w as a suffix, that is

$$T_w = \{u \in T : u \succeq w\}.$$

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

$$T|_K = \{w \in T : |w| \leq K\} \cup \{w \in A^K : w \prec u \text{ for some } u \in T\}.$$

Given two trees T_1 and T_2 we say that T_1 is *included* in T_2 (denoted by $T_1 \preceq T_2$ or $T_2 \succeq T_1$) if for any sequence $w \in T_1$ there exists a sequence $u \in T_2$ such that $w \preceq u$; in other words, all leaves of T_1 are either leaves or internal nodes of T_2.

Consider a stationary ergodic stochastic process $\{X_t : t \in \mathbb{Z}\}$ over A. Given a sequence $w \in A^*$ we denote by

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

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

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

Definition 2.1. A sequence $w \in A^*$ is a *finite context* for the process $\{X_t : t \in \mathbb{Z}\}$ if it satisfies

1. $p(w) > 0$;
2. for any sequence $v \in A^*$ such that $p(v) > 0$ and $v \succeq w$,
$$\mathbb{P}(X_0 = a \mid X_{-|v|}^{-1} = v) = p(a|w), \quad \text{for all } a \in A;$$
3. no proper suffix of w satisfies 1 and 2.

An *infinite context* is a semi-infinite sequence $w_{-\infty}^{-1} \in A^\infty$ such that any of its finite suffixes w_{-j}^{-1}, $j = 1, 2, \ldots$ is a context. In what follows the term *context* will refer to a finite or infinite context.

It can be seen that the set of all contexts of the process $\{X_t : t \in \mathbb{Z}\}$ is a tree. This is called the *context tree* of the process. For example, the context tree of an i.i.d. process is A^0 and the context tree of a generic Markov chain of order 1 is $A^1 = A$. In what follows, we will denote by T_0 the context tree of the process $\{X_t : t \in \mathbb{Z}\}$.

Let $d \leq n$ be positive integers. Let $X_{-d+1}, \ldots, X_0, X_1, \ldots, X_n$ be a sequence distributed according to $\mathbb{P}$. For any sequence $w \in A^*$ and any symbol $a \in A$ we denote by $N_n(w, a)$ the number of occurrences of symbol a in X_1^n that are preceded by an occurrence of w, that is:

$$N_n(w, a) = \sum_{t=1}^{n} \mathbb{1}\{X_{t-|w|}^{t-1} = w, X_t = a\}. \tag{2.2}$$

The sum $\sum_{a \in A} N_n(w, a)$ is denoted by $N_n(w)$.

We will denote by $\mathcal{V}_n$ the set of all sequences $w \in A^*$ that appear at least once in the sample, that is

$$\mathcal{V}_n = \{w \in A^* : N_n(w) \geq 1\}.$$

Definition 2.3. We will say that a tree $T \subset \mathcal{V}_n$ is *acceptable* if it satisfies the following conditions:

1. $h(T) \leq d$; and
2. every sequence $w \in A^*$ such that $N_n(w) \geq 1$ belongs to T or has a proper suffix that belongs to T.

Then, our set of candidate trees, denoted by $\mathscr{T}_n$, will be the set of all acceptable trees. Our goal is to select a tree $T \in \mathscr{T}_n$ as close as possible to T_0, in some sense that will be formally given below. Note that d may depend on n, so that the set of candidate trees is allowed to grow with the sample size. The symbols $X_{-d+1}, \ldots, X_0$ are only observed to ensure that, for every candidate tree T, the context of X_i in T is well defined, for every $i = 1, \ldots, n$. Alternatively, if $X_{-d+1}, \ldots, X_0$ were not assumed observed, similar results would be obtained by using quasi-maximum likelihood estimators [16]. Given a tree $T \subset \cup_{j=1}^{d} A^j$, the maximum likelihood of the sequence $X_1, \ldots, X_n$ is given by

$$\hat{\mathbb{P}}_{\mathrm{ML},T}(X_1^n) = \prod_{w \in T} \prod_{a \in A} \hat{p}_n(a|w)^{N_n(w,a)}, \tag{2.4}$$

where the empirical probabilities $\hat{p}_n(a|w)$ are

$$\hat{p}_n(a|w) = \frac{N_n(w,a)}{N_n(w)} \tag{2.5}$$

if $N_n(w) > 0$ and $\hat{p}_n(a|w) = 1/|A|$ otherwise. For any sequence $w \in A^*$ we define

$$\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n) = \prod_{a \in A} \hat{p}_n(a|w)^{N_n(w,a)}.$$

Hence, we have

$$\hat{\mathbb{P}}_{\mathrm{ML},T}(X_1^n) = \prod_{w \in T} \hat{\mathbb{P}}_{\mathrm{ML},w}(x_1^n).$$

In order to measure discrepancy between two probability measures over A we use the *Kullback–Leibler divergence*, defined for two probability measures P and Q on A by

$$D(P; Q) = \sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)}$$

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 $P(a) > Q(a) = 0$.

2.1. The algorithm Context

The algorithm Context introduced by Rissanen in [25] computes, for each node of a given tree, a discrepancy measure between the transition probability associated to this context and the corresponding transition probabilities of the nodes obtained by concatenating a single symbol to the context. Beginning with the largest leaves of a candidate tree, if the discrepancy measure is greater than a given threshold, the contexts are maintained in the tree; otherwise, they are pruned. The procedure continues until no more pruning of the tree can be performed.

For all sequences $w \in \mathcal{V}_n$ let

$$\Delta_n(w) = \sum_{b : bw \in \mathcal{V}_n} N_n(bw) D(\hat{p}_n(\cdot|bw); \hat{p}_n(\cdot|w)).$$

Remark 2.6. We use here the original choice of divergence $\Delta_n(w)$ proposed by Rissanen in [25], but other possibilities have been proposed in the literature (see for instance [4,18]).

We will denote the threshold used in algorithm Context on samples of length n by δ_n, where $(\delta_n)_{n\in\mathbb{N}}$ is a sequence of positive real numbers such that $\delta_n \to +\infty$ and $\delta_n/n \to 0$ when $n \to +\infty$. For a sequence X_1^n, let $\mathcal{C}_w(X_1^n) \in \{0,1\}$ be an indicator function defined for all $w \in \mathcal{V}_n$ by the following induction:

$$\mathcal{C}_w(X_1^n) = \begin{cases} 0, & \text{if } N_n(w) \le 1 \text{ or } |w| \ge d, \\ \max\{\mathbb{1}\{\Delta_n(w) \ge \delta_n\}, \max\limits_{b\in A} \mathcal{C}_{bw}(X_1^n)\}, & \text{if } N_n(w) > 1 \text{ and } |w| < d. \end{cases} \tag{2.7}$$

With these definitions, the context tree estimator $\hat{T}_C(X_1^n)$ is the set given by

$$\hat{T}_C(X_1^n) = \{w \in \mathcal{V}_n : \mathcal{C}_w(X_1^n) = 0 \text{ and } \mathcal{C}_u(X_1^n) = 1 \text{ for all } u \prec w\}. \tag{2.8}$$

2.2. The penalized maximum likelihood criterion

The penalized maximum likelihood criterion for the sequence X_1^n is defined by

$$\hat{T}_{\mathrm{PML}}(X_1^n) = \arg\max_{T\in\mathcal{T}_n} \left\{ \log \hat{\mathbb{P}}_{\mathrm{ML},T}(X_1^n) - |T| f(n) \right\}, \tag{2.9}$$

where $f(n)$ is some positive function such that $f(n) \to +\infty$ and $f(n)/n \to 0$ when $n \to \infty$.

This class of context tree estimators was first considered by Csiszár and Talata in [9], who introduced the Bayesian Information Criterion (BIC) for context trees and proved its consistency. The BIC leads to the choice of the penalty function $f(n) = (|A| - 1)\log(n)/2$. It may first appear practically impossible to compute $\hat{T}_{\mathrm{PML}}(X_1^n)$, because the maximization in (2.9) must be performed over the set of all candidate trees. Fortunately, Csiszár and Talata showed in their article [9] how to adapt the Context Tree Maximizing (CTM) method [27] in order to obtain a simple and efficient algorithm computing $\hat{T}_{\mathrm{PML}}(X_1^n)$. As the representation of the estimator $\hat{T}_{\mathrm{PML}}(X_1^n)$ given by this algorithm is important for the proof of our results, we briefly present it here. Define recursively, for any $w \in \mathcal{V}_n$, with $|w| < d$, the value

$$V_w(X_1^n) = \max\left\{ \mathrm{e}^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n), \prod_{b\in A : bw\in\mathcal{V}_n} V_{bw}(X_1^n) \right\} \tag{2.10}$$

and the indicator

$$\mathcal{X}_w(X_1^n) = \mathbb{1}\left\{ \prod_{b\in A : bw\in\mathcal{V}_n} V_{bw}(X_1^n) > \mathrm{e}^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n) \right\}. \tag{2.11}$$

By convention, if $\{b \in A : bw \in \mathcal{V}_n\} = \emptyset$ or if $|w| = d$ then $V_w(X_1^n) = \mathrm{e}^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)$ and $\mathcal{X}_w(X_1^n) = 0$. As shown in [9], it holds that

$$\hat{T}_{\mathrm{PML}}(X_1^n) = \{w \in \mathcal{V}_n : \mathcal{X}_w(X_1^n) = 0 \text{ and } \mathcal{X}_u(X_1^n) = 1 \text{ for all } u \prec w\}. \tag{2.12}$$

2.3. Results

In this subsection we present the main results of this article. First, we show that the empirical tree given by the algorithm Context is always included in the tree given by the penalized maximum likelihood estimator, if the threshold δ_n is smaller than the penalization function $f(n)$.

Proposition 2.13. *For any $n \geq 1$ and all sequences X_1^n, if $\delta_n \leq f(n)$ then*

$$\hat{T}_{\mathrm{PML}}(X_1^n) \preceq \hat{T}_C(X_1^n).$$

In the sequel we will assume that the cutoff sequence of the algorithm Context equals the penalization term of the penalized maximum likelihood estimator, in order to allow a unified treatment of the two algorithms. That is, we will assume that $\delta_n = f(n)$ for any $n \geq 1$.

We now state a new bound on the probability of over-estimation that does not require any mixing hypotheses on the underlying process.

Theorem 2.14. *For every $n \geq 1$ it holds that*

$$\mathbb{P}\left(\hat{T}(X_1^n) \preceq T_0\right) \geq 1 - e\left(\delta_n \log(n) + |A|^2\right) n^2 \exp\left(-\frac{\delta_n}{|A|^2}\right), \tag{2.15}$$

where $\hat{T}(X_1^n) = \hat{T}_{\mathrm{PML}}(X_1^n)$ or $\hat{T}(X_1^n) = \hat{T}_C(X_1^n)$.

Remark 2.16. Theorem 2.14 is proven without assuming any bound on the height of the hypothetical trees. That is, the result remains valid even if $d = -\infty$. But if the candidate trees have only a limited number of nodes, possibly depending on n (see, e.g., [25,9]), a straightforward modification of the proof shows that

$$\mathbb{P}\left(\hat{T}(X_1^n) \preceq T_0\right) \geq 1 - 2e\left(\delta_n \log(n) + |A|^2\right) k(n) \exp\left(-\frac{\delta_n}{|A|^2}\right),$$

where $k(n)$ is the maximal number of nodes of a candidate tree. In particular, if the height of the trees is smaller than a function $d(n)$ (possibly constant) then $k(n) = |A|^{d(n)}$.

The problem of under-estimation in context tree models is very different, and requires additional hypotheses on the process $\{X_t : t \in \mathbb{Z}\}$. For any $w \in A^*$ with $p(w) > 0$ define the coefficient

$$\beta(w, r) = \max_{u \in A^r} \max_{a \in A} \{|p(a|w) - p(a|uw)|\}.$$

The *continuity rate* of the process $\{X_t : t \in \mathbb{Z}\}$ is the sequence $\{\beta_k\}_{k \in \mathbb{N}}$ where

$$\beta_k = \max_{w \in A^k} \sup_{r \geq 1} \{\beta(w, r)\}.$$

Define also the *non-nullness* coefficient

$$\alpha_0 := \sum_{a \in A} \inf_{w \in T_0} \{p(a|w)\}. \tag{2.17}$$

Our under-estimation error bounds will rely on the following assumption.

Assumption 1. The process $\{X_t : t \in \mathbb{Z}\}$ satisfies the following conditions

1. $\alpha_0 > 0$ (weakly non-nullness) and
2. $\beta := \sum_{k \in \mathbb{N}} \beta_k < +\infty$ (summable continuity rate).

These are classical hypotheses for processes of infinite memory, which are also referred to as *chains of type A*, see for instance [13] and references therein.

To establish upper bounds for the probability of under-estimation we will consider the truncated tree $T_0|_K$, for any given constant $K \in \mathbb{N}$. Note that in the case T_0 is a finite tree, $T_0|_K$ coincides with T_0 for a sufficiently large constant K. The bounds are stated in the following theorem.

Theorem 2.18. *Assume the process $\{X_t : t \in \mathbb{Z}\}$ satisfies Assumption 1. Let $K \in \mathbb{N}$ and let d be such that*

$$\min_{w \prec u \in T_0|_K} \max_{r \le d-|w|} \{\beta(w, r)\} \ge \epsilon > 0. \tag{2.19}$$

Then, there exists $n_0 \in \mathbb{N}$ such that for any $n \ge n_0$ it holds that

$$\mathbb{P}\left(T_0|_K \preceq \hat{T}(X_1^n)|_K\right) \ge 1 - 3e^{\alpha_0/32e^2|A|^2(|A|\beta+2\alpha_0)}|A|^{2+K} \times \exp\left[\frac{-n\epsilon^2\left[p_{\min}^d - \frac{8|A|df(n)}{\epsilon^2 n}\right]^2}{16(d+1)}\right], \tag{2.20}$$

where $\hat{T}(X_1^n) = \hat{T}_{\mathrm{PML}}(X_1^n)$ or $\hat{T}(X_1^n) = \hat{T}_C(X_1^n)$ and $p_{\min} = \min_{a\in A, w\in A^d}\{p(a|w) : p(a|w) > 0\}$.

Remark 2.21. It can be seen that for any $K \in \mathbb{N}$ there is always a value of d such that (2.19) holds. This hypothesis can be avoided by letting d increase with the sample size n and by controlling the upper bounds in (2.20). Extensions of Theorem 2.18 can also be obtained by allowing K to be a function of the sample size n. In this case, the rate at which K increases must be controlled together with the rate at which ϵ and $p_{\min}$ decrease with the sample size. This leads to a rather technical condition, see for instance [26].

Finally, the next theorem states the strong consistency of the estimators $\hat{T}_C(X_1^n)$ and $\hat{T}_{\mathrm{PML}}(X_1^n)$ for appropriate threshold parameters and penalizing functions, respectively.

Theorem 2.22. *Assume the hypotheses of Theorem 2.18 are met. Then for any threshold parameter $(\delta_n)_{n\in\mathbb{N}}$ such that*

$$\sum_{n\in\mathbb{N}} \exp\left(-\frac{\delta_n}{|A|^2} + \log(\delta_n \log(n))\right) < +\infty$$

we have $\hat{T}_C(X_1^n)|_K = T_0|_K$ eventually almost surely as $n \to +\infty$. Similarly, if we choose $f(n) = \delta_n$ we have $\hat{T}_{\mathrm{PML}}(X_1^n)|_K = T_0|_K$ eventually almost surely as $n \to +\infty$.

3. Proofs

3.1. Proof of Proposition 2.13

We must prove that a leaf in $\hat{T}_{\mathrm{PML}}(X_1^n)$ is always a leaf or an internal node in $\hat{T}_C(X_1^n)$. By the characterization of $\hat{T}_C(X_1^n)$ and $\hat{T}_{\mathrm{PML}}(X_1^n)$ given by Eqs. (2.8) and (2.12), respectively, this is equivalent to proving that $\mathcal{X}_w(X_1^n) \le \mathcal{C}_w(X_1^n)$ for all $w \in \mathcal{V}_n$ with $|w| < d$. In fact, assume that $\mathcal{X}_w(X_1^n) = 1$ implies $\mathcal{C}_w(X_1^n) = 1$, and take $w \in \hat{T}_{\mathrm{PML}}(X_1^n)$; then, either $|w| = d$ and $w \in \hat{T}_C(X_1^n)$, or it holds that for all $u \prec w$, $\mathcal{X}_u(X_1^n) = 1$, which implies by assumption that

$\mathcal{C}_u(X_1^n) = 1$. Now, if $\mathcal{C}_w(X_1^n) = 0$, then $w \in \hat{T}_C(X_1^n)$; otherwise, w is a proper suffix of a sequence $v \in T_C(X_1^n)$. In any case, w is a leaf or an internal node of $\hat{T}_C(X_1^n)$.

Assume there exists $w \in \mathcal{V}_n$, $|w| < d$, such that $\mathcal{X}_w(X_1^n) = 1$ and $\mathcal{C}_w(X_1^n) = 0$. Note that by (2.7), $\mathcal{C}_w(X_1^n) = 0$ implies $\mathcal{C}_{uw}(X_1^n) = 0$ for all $uw \in \mathcal{V}_n$, $|uw| \le d$; hence, w can be chosen such that $\mathcal{X}_{bw}(X_1^n) = 0$ for any $bw \in \mathcal{V}_n$, $b \in A$. In this case we have, by the definitions (2.10) and (2.11) that

$$\mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n) < \prod_{b:bw\in\mathcal{V}_n} V_{bw}(X_1^n) \tag{3.1}$$

$$= \prod_{b:bw\in\mathcal{V}_n} \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},bw}(X_1^n). \tag{3.2}$$

The equality in the second line of the last expression follows by the fact that $\mathcal{X}_{bw}(X_1^n) = 0$ for any $bw \in \mathcal{V}_n, b \in A$; therefore we must have $V_{bw}(X_1^n) = \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},bw}(X_1^n)$ for any $bw \in \mathcal{V}_n, b \in A$.

Now, observe that for any $a \in A$, $N_n(w,a) = \sum_{b:bw\in\mathcal{V}_n} N_n(bw,a)$ and $|\{b: bw \in \mathcal{V}_n\}| \ge 2$. If not, $N_n(w,a)$ would be equal to $N_n(cw,a)$ for some $c \in A$ and for all $a \in A$, implying that $\hat{\mathbb{P}}_{\mathrm{ML},cw}(X_1^n) = \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)$; hence

$$\prod_{b:bw\in\mathcal{V}_n} V_{bw}(X_1^n) = V_{cw}(X_1^n) = \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},cw}(X_1^n) = \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)$$

and thus, by definition, $\mathcal{X}_w(X_1^n) = 0$. Using these facts, and taking logarithm on both sides of Inequality (3.1), we obtain

$$\begin{aligned}(|\{b: bw \in \mathcal{V}_n\}| - 1)\, f(n) &< \sum_{b:bw\in\mathcal{V}_n}\sum_{a\in A} N_n(bw,a)\log\frac{\hat{p}_n(a|bw)}{\hat{p}_n(a|w)} \\ &= \sum_{b:bw\in\mathcal{V}_n} N_n(bw)D\left(\hat{p}_n(\cdot|bw);\, \hat{p}_n(\cdot|w)\right) = \Delta_n(w).\end{aligned}$$

Therefore, if $\delta_n \le f(n)$ we have $\delta_n < \Delta_n(w)$ which contradicts the fact that $\mathcal{C}_w(X_1^n) = 0$. This concludes the proof of Proposition 2.13.

3.2. *Proof of Theorem 2.14*

We will prove the result for the case $\hat{T}(X_1^n) = \hat{T}_C(X_1^n)$. The case $\hat{T}(X_1^n) = \hat{T}_{\mathrm{PML}}(X_1^n)$ follows straightforwardly from Proposition 2.13 and equality $f(n) = \delta_n$.

Let O_n be the event $\{\hat{T}_C(X_1^n) \not\preceq T_0\}$. Over-estimation occurs if at least one internal node w of $\hat{T}_C(X_1^n)$ has a (non-necessarily proper) suffix s in T_0; that is, if there exists a (possibly empty) sequence u such that $w = us$. Thus, with a little abuse of notation O_n can be written as

$$O_n = \bigcup_{s\in T_0}\bigcup_{u\in A^*}\{\Delta_n(us) > \delta_n\}.$$

For any sequence $w \in A^*$ we have that $\hat{p}_n(\cdot|w)$ are the maximum likelihood estimators of the transition probabilities $p(\cdot|w)$, therefore we have that

$$\Delta_n(w) = \sum_{b\in A} N_n(bw)D\left(\hat{p}_n(\cdot|bw);\, \hat{p}_n(\cdot|w)\right)$$

$$
\begin{aligned}
&= \sum_{b\in A} N_n(bw) \sum_{a\in A} \left(\hat{p}(a|bw)\log\hat{p}(a|bw) - \hat{p}(a|bw)\log\hat{p}(a|w)\right)\\
&= \left(\sum_{b\in A} N_n(bw) \sum_{a\in A} \hat{p}(a|bw)\log\hat{p}(a|bw)\right) - \sum_{b\in A}\sum_{a\in A} N_n(bw,a)\log\hat{p}(a|w)\\
&= \left(\sum_{b\in A} N_n(bw) \sum_{a\in A} \hat{p}(a|bw)\log\hat{p}(a|bw)\right) - \sum_{a\in A} N_n(w,a)\log\hat{p}(a|w)\\
&\le \left(\sum_{b\in A} N_n(bw) \sum_{a\in A} \hat{p}(a|bw)\log\hat{p}(a|bw)\right) - \sum_{a\in A} N_n(w,a)\log p(a|w)\\
&= \left(\sum_{b\in A} N_n(bw) \sum_{a\in A} \hat{p}(a|bw)\log\hat{p}(a|bw)\right) - \sum_{b\in A}\sum_{a\in A} N_n(bw,a)\log p(a|w)\\
&= \sum_{b\in A} N_n(bw) \sum_{a\in A} \left(\hat{p}(a|bw)\log\hat{p}(a|bw) - \hat{p}(a|bw)\log p(a|w)\right)\\
&= \sum_{b\in A} N_n(bw) D\left(\hat{p}_n(\cdot|bw);\, p(\cdot|w)\right).
\end{aligned}
$$

Hence, as for all $b \in A$ it holds that $p(\cdot|w) = p(\cdot|bw)$ we obtain

$$
\mathbb{P}(\Delta_n(w) > \delta_n) \le \mathbb{P}\left(\sum_{b\in A} N_n(bw) D\left(\hat{p}_n(\cdot|bw);\, p(\cdot|bw)\right) > \delta_n\right).
$$

Using Theorem A.7, stated in Appendix A, it follows that

$$
\begin{aligned}
\mathbb{P}(O_n) &\le \sum_{s\in T_0}\sum_{u\in A^*} \mathbb{P}\left(\Delta_n(us) > \delta_n\right)\\
&\le \sum_{s\in T_0}\sum_{u\in A^*} \mathbb{P}\left(\sum_{b\in A} N_n(bus) D\left(\hat{p}_n(\cdot|bus);\, p(\cdot|bus)\right) > \delta_n\right)\\
&\le \sum_{s\in T_0}\sum_{u\in A^*}\sum_{b\in A} \mathbb{P}\left(N_n(bus) D\left(\hat{p}_n(\cdot|bus);\, p(\cdot|bus)\right) > \frac{\delta_n}{|A|}\,\middle|\, N_n(bus) > 0\right)\\
&\quad \times \mathbb{P}\left(N_n(bus) > 0\right)\\
&\le 2e\left(\delta_n \log n + |A|^2\right)\exp\left(-\frac{\delta_n}{|A|^2}\right)\sum_{s\in T_0}\sum_{u\in A^*}\sum_{b\in A} \mathbb{P}\left(N_n(bus) > 0\right)\\
&\le 2e\left(\delta_n \log n + |A|^2\right)\exp\left(-\frac{\delta_n}{|A|^2}\right)\mathbb{E}[C_n],
\end{aligned}
$$

where C_n denotes the number of different contexts of the symbols in X_1^n. But C_n is always upper bounded by the number $n(n-1)/2$ of (non-necessarily distinct) contexts of X_1^n, and the result follows.

3.3. Proof of Theorem 2.18

In this case we will prove the result for the case $\hat{T}(X_1^n) = \hat{T}_{\mathrm{PML}}(X_1^n)$. The case $\hat{T}(X_1^n) = \hat{T}_C(X_1^n)$ follows again from Proposition 2.13 and the assumption that $\delta_n = f(n)$.

If U_n denotes the event $\{T_0|_K \not\preceq \hat{T}_{\mathrm{PML}}(X_1^n)|_K\}$ then

$$U_n \subset \bigcup_{w \prec u \in T_0|_K} \{\mathcal{X}_w(X_1^n) = 0\}.$$

Let $w \prec u \in T_0|_K$. Then we have

$$\mathbb{P}\left(\mathcal{X}_w(X_1^n) = 0\right) = \mathbb{P}\left(\prod_{a\in A: aw\in\mathcal{V}_n} V_{aw}(X_1^n) \le \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\right). \tag{3.3}$$

By hypothesis, there exists $r \le d - |w|$ and $s \in A^r$ such that

$$\max_{a\in A} |p(a|w) - p(a|sw)| \ge \epsilon.$$

If $s = (s_1 \dots s_r)$, denote by $A_i = A \setminus \{s_i\}$ and let T be the tree given by

$$T = \cup_{i=2}^r \cup_{b\in A_i} \{bs_i^r w\} \cup \{sw\}.$$

By definition, for any $aw \in \mathcal{V}_n$ it can be shown recursively that

$$V_{aw}(X_1^n) = \max_{T'\in\mathscr{T}_n} \prod_{v\in T'_{aw}} \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},v}(X_1^n)$$

see for example Lemma 4.4 in [9]. Therefore,

$$\begin{aligned}&\mathbb{P}\left(\prod_{a\in A: aw\in\mathcal{V}_n} V_{aw}(X_1^n) \le \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\right)\\ &\quad\le \mathbb{P}\left(\prod_{u\in T} \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},u}(X_1^n) \le \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\right)\end{aligned} \tag{3.4}$$

by noticing that

$$\begin{aligned}\prod_{a\in A: aw\in\mathcal{V}_n} \max_{T'\in\mathscr{T}_n} \prod_{v\in T'_{aw}} \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},v}(X_1^n) &\ge \prod_{a\in A: aw\in\mathcal{V}_n} \prod_{v\in T_{aw}} \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},v}(X_1^n)\\ &\ge \prod_{u\in T} \mathrm{e}^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},u}(X_1^n).\end{aligned}$$

Applying logarithm and using that $N_n(w,a) = \sum_{u\in T} N_n(u,a)$ for any $a \in A$ we can write the probability in (3.4) by

$$\begin{aligned}&\mathbb{P}\left(\sum_{u\in T} N_n(u) D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \le (|T|-1) f(n)\right)\\ &\quad\le \mathbb{P}\left(N_n(sw) D(\hat{p}_n(\cdot|sw);\, \hat{p}_n(\cdot|w)) \le (|T|-1) f(n)\right).\end{aligned} \tag{3.5}$$

Define the events $A_n^{s,w}$ and $B_n^{s,w}$ by

$$\begin{aligned}A_n^{s,w} &= \left\{X_1^n : N_n(sw) D(\hat{p}_n(\cdot|sw);\, \hat{p}_n(\cdot|w)) \le (|T|-1) f(n)\right\}\\ B_n^{s,w} &= \left\{X_1^n : D(\hat{p}_n(\cdot|sw);\, \hat{p}_n(\cdot|w)) > \epsilon^2/8\right\}.\end{aligned}$$

Then we can bound above the probability in (3.5) by $\mathbb{P}\left(A_n^{s,w} \cap B_n^{s,w}\right) + \mathbb{P}\left((B_n^{s,w})^c\right)$. To bound the first term note that by Lemma B.11, if n satisfies

$$\frac{f(n)}{n} < \frac{\epsilon^2 p(sw)}{8(|T|-1)}$$

then, using the bound $|T| - 1 \leq |A|r \leq |A|d$ we obtain

$$\begin{aligned}\mathbb{P}\left(A_n^{s,w} \cap B_n^{s,w}\right) &\leq \mathbb{P}\left(N_n(sw) \leq \frac{8(|T|-1)f(n)}{\epsilon^2}\right) \\ &\leq \mathrm{e}^{\alpha_0/8\mathrm{e}^2|A|^2(|A|\beta+2\alpha_0)}|A| \exp\left(\frac{-n\left[p(sw) - \frac{8|A|\mathrm{d}f(n)}{\epsilon^2 n}\right]^2}{|sw|+1}\right).\end{aligned}$$

On the other hand, by Lemma B.13 we have

$$\mathbb{P}\left((B_n^{s,w})^c\right) \leq 2\mathrm{e}^{\alpha_0/32\mathrm{e}^2|A|^2(|A|\beta+2\alpha_0)}(|A|+1)\exp\left[-n\frac{\epsilon^2 p(sw)^2}{16(|sw|+1)}\right].$$

We conclude the proof of Theorem 2.18 by observing that we only have a finite number of sequences $w \prec u \in T_0|_K$, therefore we obtain

$$\mathbb{P}(U_n) \leq 3\mathrm{e}^{\alpha_0/32\mathrm{e}^2|A|^2(|A|\beta+2\alpha_0)}|A|^{2+K} \exp\left(\frac{-n\epsilon^2\left[p_{\min}^d - \frac{8|A|\mathrm{d}f(n)}{\epsilon^2 n}\right]^2}{16(d+1)}\right).$$

3.4. *Proof of Theorem 2.22*

The statement of the theorem follows straightforward from Theorems 2.14 and 2.18 and the Borel–Cantelli Lemma, by noticing that the upper bounds for

$$\mathbb{P}(\hat{T}_C(X_1^n)|_K \neq T_0|_K) \leq \mathbb{P}(\hat{T}_C(X_1^n)|_K \not\preceq T_0|_K) + \mathbb{P}(T_0|_K \not\preceq \hat{T}_C(X_1^n)|_K)$$

are summable in n. The same reasoning applies to $\hat{T}_{\mathrm{PML}}(X_1^n)$ when $f(n) = \delta_n$.

4. Discussion

In this paper we showed a relation between two classical algorithms for context tree selection. We proved that for a proper set of parameters, the Penalized Maximum Likelihood estimator always yields a smaller tree than the tree given by the algorithm Context. This relation between the empirical context trees allows us to derive, in an unified way, non-asymptotic bounds for the probability of over- and under-estimation of the context tree generating the sample. The tree may be unbounded, and our results apply to processes that do not necessarily have a finite memory.

Concerning under-estimation, we assume the process satisfies some conditions that implies exponential inequalities for the empirical probabilities. These inequalities were obtained in [17] under a stronger *non-nullness* assumption; namely, that the transition probabilities were lower bounded by a positive constant. In this paper we show that the results also hold for a larger class of processes. It is conjectured that similar results cannot be obtained without assuming any non-nullness nor mixing condition of the process.

Concerning over-estimation no mixing assumption is necessary for Theorem 2.14 to hold. This improves on and generalizes the results obtained in [17,22]. Our proof is based on deviation inequalities obtained for empirical Kullback–Leibler divergence, instead of L^p norm; it appears that this pseudo-metric is more intrinsic for binomial distributions (and partially also for multinomial distributions), as the binary Kullback–Leibler divergence is the rate function of a Large Deviations Principle. Deriving similar inequalities is also possible for other distributions and thus other pseudo-metrics, or by using upper bounds of the Legendre transform of the distribution, as in [21]. These type of inequalities are interesting on their own and prove useful in various settings: other applications of similar bounds may be found in [21,14,20].

From the point of view of most applications, over- and under-estimation play a different role. In fact, data-generating processes can often not be assumed to have finite memory: the whole dependence structure cannot be recovered from finitely many observations and under-estimation is unavoidable. All what can be expected from the estimator is to highlight evidence of as much dependence structure as possible, while maintaining a limited probability of false discovery.

Our results imply the strong consistency of the algorithm Context for processes of infinite memory, generalizing the convergence in probability of this estimator previously obtained in [12]. Likewise, the strong consistency for the PML estimator is also derived for a larger class of penalizing functions than in [22].

Acknowledgments

This work was supported by MacLinC project (Proc. USP 11.1.9367.1.5), Fapesp (grant 2009/09411-8) and USP-COFECUB (grant 2009.1.820.45.8). FL is partially supported by a CNPq fellowship (grant 302162/2009-7).

Appendix A. Martingale deviation inequalities

This section contains the statement and derivation of two deviation inequalities that are useful to prove the results of this paper. As they are interesting on their own, we include them in a separate section. The ingredients of the proofs are mostly inspired by [24], see also [8].

We briefly recall some notation so as to keep this section self-contained. Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary process whose (possibly infinite) context tree is T_0, and let $\mathcal{F}_n$ be the σ-field generated by $(X_j)_{j\le n}$. For $k \in \mathbb{N}$, $w \in A^k$, denote $p(b|w) = \mathbb{P}\left(X_{k+1} = b | X_1^k = w\right)$. For $j \ge 1$, define

$$\xi_j = \mathbb{1}\{X_{j-k}^{j-1} = w\} \quad \text{and} \quad \chi_j = \mathbb{1}\{X_{j-k}^{j} = wb\},$$

so that $N_n(w) = \sum_{j=1}^n \xi_j$ and $N_n(w,b) = \sum_{j=1}^n \chi_j$. Denote $\hat{p}_n(b|w) = N_n(w,b)/N_n(w)$. The Kullback–Leibler divergence between Bernoulli variables will be denoted by d: for all $p, q \in [0,1]$,

$$d(p;q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}.$$

Proposition A.1. *Let k be a positive integer, let $w \in A^k$ such that there exists $u \in T_0$ with $w \succeq u$, and let $b \in A$. Then for any $\delta > 0$*

$$\mathbb{P}\left[N_n(w) d\left(\hat{p}_n(b|w); p(b|w)\right) > \delta\right] \le 2e\,\lceil \delta\log(n)\rceil \exp(-\delta).$$

Proof. Denote by $p = p(b|w)$, $N_n = N_n(w)$, $S_n = N_n(w, b)$ and $\hat{p}_n = S_n/N_n$. For every $\lambda > 0$, let

$$\phi_p(\lambda) = \log\left(1 - p + p\exp\left(\lambda\right)\right).$$

Let also $W_0^\lambda = 1$ and for $t \geq 1$,

$$W_t^\lambda = \exp(\lambda S_t - N_{t-1}\phi_p(\lambda)).$$

First, note that $\left(W_t^\lambda\right)_{t\geq 0}$ is a martingale relative to $(\mathcal{F}_t)_{t\geq 0}$ with expectation $\mathbb{E}[W_0^\lambda] = 1$. In fact,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\left(S_{t+1} - S_t\right)\right)|\mathcal{F}_t\right] &= \mathbb{E}\left[\exp\left(\lambda\chi_{t+1}\right)|\mathcal{F}_t\right] \\ &= \exp\left(\xi_t\phi_p\left(\lambda\right)\right) \\ &= \exp\left(\left(N_t - N_{t-1}\right)\phi_p\left(\lambda\right)\right)\end{aligned}$$

which can be rewritten as

$$\mathbb{E}\left[\exp\left(\lambda S_{t+1} - N_t\phi_p\left(\lambda\right)\right)|\mathcal{F}_t\right] = \exp\left(\lambda S_t - N_{t-1}\phi_p\left(\lambda\right)\right).$$

To proceed, we make use of the so-called 'peeling trick' [23]: we divide the interval $\{1, \ldots, n\}$ of possible values for N_n into "slices" $\{t_{k-1} + 1, \ldots, t_k\}$ of geometrically increasing size, and treat the slices independently. We may assume that $\delta > 1$, since otherwise the bound is trivial. Take $\eta = 1/(\delta - 1)$, let $t_0 = 0$ and for $k \in \mathbb{N}^*$, let $t_k = \left\lfloor(1+\eta)^k\right\rfloor$. Let m be the first integer such that $t_m \geq n$, that is $m = \left\lceil \frac{\log n}{\log 1+\eta}\right\rceil$. Let $A_k = \{t_{k-1} < N_n \leq t_k\} \cap \left\{N_n d\left(\hat{p}_n; p\right) > \delta\right\}$. We have:

$$\mathbb{P}\left(N_n d\left(\hat{p}_n; p\right) > \delta\right) \leq \mathbb{P}\left(\bigcup_{k=1}^{m} A_k\right) \leq \sum_{k=1}^{m}\mathbb{P}(A_k). \tag{A.2}$$

We upper bound the probability of $A_k \cap \{\hat{p}_n > p\}$, the same arguments can easily be transposed for left deviations. Let s be the smallest integer such that $\delta/(s+1) \leq d(1; p)$; if $N_n \leq s$, then $N_n d(\hat{p}_n, p) \leq s d(\hat{p}_n, p) \leq s d(1, p) < \delta$ and $\mathbb{P}(N_n d(\hat{p}_n, p) \geq \delta, \hat{p}_n > p) = 0$. Thus, $\mathbb{P}(A_k) = 0$ for all k such that $t_k \leq s$.

Take k such that $t_k > s$, and let $\tilde{t}_{k-1} = \max\{t_{k-1}, s\}$. Let $x \in]p, 1]$ be such that $d(x; p) = \delta/N_n$, and let $\lambda(x) = \log(x\,(1-p)) - \log(p\,(1-x))$, so that $d(x; p) = \lambda(x)x - \phi_p(\lambda)$. Let z such that $z \geq p$ and $d(z, p) = \delta/(1+\eta)^k$. Observe that:

- if $N_n > t_{k-1}$, then
$$d(z; p) = \frac{\delta}{(1+\eta)^k} \geq \frac{\delta}{(1+\eta)N_n};$$
- if $N_n \leq t_k$ then, as
$$d(\hat{p}_n; p) > \frac{\delta}{N_n} > \frac{\delta}{(1+\eta)^k} = d(z; p),$$
we have:
$$\hat{p}_n \geq p \quad \text{and} \quad d(\hat{p}_n; p) > \frac{\delta}{N_n} \Longrightarrow \hat{p}_n \geq z.$$

Hence, on the event $\{t_{k-1} < N_n \leq t_k\} \cap \left\{\hat{p}_n \geq p\right\} \cap \left\{d(\hat{p}_n; p) > \frac{\delta}{N_n}\right\}$ it holds that

$$\lambda(z)\hat{p}_n - \phi_p(\lambda(z)) \geq \lambda(z)z - \phi_p(\lambda(z)) = d(z; p) \geq \frac{\delta}{(1+\eta)N_n}.$$

Putting everything together,

$$\begin{aligned}\left\{\tilde{t}_{k-1} < N_n \le t_k\right\} \cap \left\{\hat{p}_n \ge p\right\} \cap \left\{d(\hat{p}_n; p) \ge \frac{\delta}{N_n}\right\} &\subset \left\{\lambda(z)\hat{p}_n - \phi_p(\lambda(z)) \ge \frac{\delta}{N_n(1+\eta)}\right\} \\ &\subset \left\{\lambda(z)S_n - N_n\phi_p(\lambda(z)) \ge \frac{\delta}{1+\eta}\right\} \\ &\subset \left\{W_n^{\lambda(z)} > \exp\left(\frac{\delta}{1+\eta}\right)\right\}.\end{aligned}$$

As $\left(W_t^{\lambda}\right)_{t\ge 0}$ is a martingale, $\mathbb{E}\left[W_n^{\lambda(z)}\right] = \mathbb{E}\left[W_0^{\lambda(z)}\right] = 1$, and the Markov inequality yields:

$$\begin{aligned}\mathbb{P}\left(\left\{\tilde{t}_{k-1} < N_n \le t_k\right\} \cap \left\{\hat{p}_n \ge p\right\} \cap \left\{N_n d(\hat{p}_n, p) \ge \delta\right\}\right) &\le \mathbb{P}\left(W_n^{\lambda(z)} > \exp\left(\frac{\delta}{1+\eta}\right)\right) \\ &\le \exp\left(-\frac{\delta}{1+\eta}\right). \qquad \text{(A.3)}\end{aligned}$$

Similarly,

$$\mathbb{P}\left(\left\{\tilde{t}_{k-1} < N_n \le t_k\right\} \cap \left\{\hat{p}_n \le p\right\} \cap \left\{N_n d(\hat{p}_n, p) \ge \delta\right\}\right) \le \exp\left(-\frac{\delta}{1+\eta}\right),$$

so that

$$\mathbb{P}\left(\left\{\tilde{t}_{k-1} < N_n \le t_k\right\} \cap \left\{N_n d(\hat{p}_n, p) \ge \delta\right\}\right) \le 2\exp\left(-\frac{\delta}{1+\eta}\right).$$

Finally, by Eq. (A.2),

$$\mathbb{P}\left(\bigcup_{k=1}^{m}\left\{\tilde{t}_{k-1} < N_n \le t_k\right\} \cap \left\{N_n d(\hat{p}_n, p) \ge \delta\right\}\right) \le 2m\exp\left(-\frac{\delta}{1+\eta}\right).$$

But as $\eta = 1/(\delta-1)$, $m = \left\lceil \frac{\log n}{\log(1+1/(\delta-1))} \right\rceil$ and as $\log(1+1/(\delta-1)) \ge 1/\delta$, we obtain:

$$\mathbb{P}\left(N_n d(\hat{p}_n, p) \ge \delta\right) \le 2\left\lceil \frac{\log n}{\log\left(1+\frac{1}{\delta-1}\right)} \right\rceil \exp(-\delta+1) \le 2e\,\lceil \delta\log(n)\rceil \exp(-\delta). \quad \square$$

Remark A.4. The bound of Proposition A.1 also holds for $\mathbb{P}\left(N_n d(\hat{p}_n, p) \ge \delta | N_n > 0\right)$: in fact, as

$$\begin{aligned}1 = \mathbb{E}\left[W_n^{\lambda(z)}\right] &= \mathbb{E}\left[W_n^{\lambda(z)} | N_n > 0\right]\mathbb{P}(N_n > 0) + \mathbb{E}\left[W_n^{\lambda(z)} | N_n = 0\right]\mathbb{P}(N_n = 0) \\ &= \mathbb{E}\left[W_n^{\lambda(z)} | N_n > 0\right]\mathbb{P}(N_n > 0) + 1 - \mathbb{P}(N_n > 0),\end{aligned}$$

it follows that $\mathbb{E}\left[W_n^{\lambda(z)} | N_n > 0\right] = 1$ and starting from Eq. (A.3), the proof can be rewritten conditionally on $\{N_n > 0\}$; this leads to:

$$\mathbb{P}\left(N_n d(\hat{p}_n, p) \ge \delta | N_n > 0\right) \le 2e\,\lceil \delta\log(n)\rceil \exp(-\delta).$$

However, in general no such result can be proved for $\mathbb{P}\left(N_n d(\hat{p}_n, p) \geq \delta | N_n > k\right)$ for positive values of k.

To proceed, we need the following lemma:

Lemma A.5. *For any probability distributions P and Q on the finite alphabet A,*

$$D(P; Q) \leq \sum_{x \in A} d\left(P(x); Q(x)\right).$$

Proof.

$$\begin{aligned}\sum_{x \in A} d\left(P(x); Q(x)\right) - D(P; Q) &= \sum_{x \in A} (1 - P(x)) \log \frac{1 - P(x)}{1 - Q(x)} \\ &= (|X| - 1) \sum_{x \in A} \frac{1 - P(x)}{|A| - 1} \log \left(\frac{(1 - P(x))/(|A| - 1)}{(1 - Q(x))/(|A| - 1)} \right) \\ &\geq 0\end{aligned}$$

because the sum in the next-to-last line is the Kullback–Leibler divergence between the probability distributions R and S defined on A by:

$$R(x) = \frac{1 - P(x)}{|A| - 1} \quad \text{and} \quad S(x) = \frac{1 - Q(x)}{|A| - 1}. \quad \square$$

Remark A.6. Obviously, this lemma is suboptimal for $|A| = 2$ by a factor 2. For larger alphabets, it does not appear possible to improve on this bound for all P and Q.

We are now in position to state the deviation result we use in order to upper bound the probability of over-estimation:

Theorem A.7. *Let k be a positive integer and let $w \in A^k$. Then, for any $\delta > 0$*

$$\mathbb{P}\left[N_n(w) D\left(\hat{p}_n\left(\cdot|w\right); p\left(\cdot|w\right)\right) > \delta\right] \leq 2e\left(\delta \log(n) + |A|\right) \exp\left(-\frac{\delta}{|A|}\right).$$

Proof. By combining Lemma A.5 and Proposition A.1, we get

$$\begin{aligned}\mathbb{P}\left[N_n(w) D\left(\hat{p}_n\left(\cdot|w\right); p\left(\cdot|w\right)\right) > \delta\right] &\leq \mathbb{P}\left[\sum_{b \in A} N_n d\left(\hat{p}_n\left(b|w\right); p\left(b|w\right)\right) > \delta\right] \\ &\leq \sum_{b \in A} \mathbb{P}\left[N_n d\left(\hat{p}_n\left(b|w\right); p\left(b|w\right)\right) > \frac{\delta}{|A|}\right] \\ &\leq 2|A|e\left\lceil \frac{\delta}{|A|} \log(n) \right\rceil \exp\left(-\frac{\delta}{|A|}\right) \\ &\leq 2|A|e\left(\frac{\delta}{|A|} \log(n) + 1\right) \exp\left(-\frac{\delta}{|A|}\right) \\ &= 2e\left(\delta \log(n) + |A|\right) \exp\left(-\frac{\delta}{|A|}\right). \quad \square\end{aligned}$$

Remark A.8. It follows from Remark A.4 that the following variant of Theorem A.7 holds:

$$\mathbb{P}\left[N_n(w) D\left(\hat{p}_n\left(\cdot|w\right); p\left(\cdot|w\right)\right) > \delta | N_n(w) > 0\right] \leq 2e\left(\delta \log(n) + |A| - 1\right) \exp\left(-\frac{\delta}{|A|-1}\right).$$

Appendix B. Exponential inequalities for weak dependent processes

In this section we state some results providing exponential inequalities for processes satisfying Assumption 1 and prove two lemmas that are useful in the proof of Theorem 2.18. The first result is a version of Theorem 3.1 in [17] that we state under weaker conditions, given by Assumption 1.

Proposition B.9. *Assume the process $\{X_t : t \in \mathbb{Z}\}$ satisfies Assumption 1. Then for any $w \in A^*$, any $a \in A$ and any $t > 0$ the following inequality holds*

$$\mathbb{P}(|N_n(w,a) - np(wa)| > t) \leq \mathrm{e}^{\alpha_0/8\mathrm{e}^2(|A|\beta+2\alpha_0)} \exp\left(\frac{-t^2}{|wa|n}\right).$$

Proof. Theorem 3.1 in [17] was proven for a process satisfying a stronger non-nullness hypothesis than our Assumption 1, namely that $\inf_{w\in T_0}\{p(a|w)\} > 0$ for any $a \in A$. But the proof of the theorem is based on results obtained in [7] and [11] that also hold for processes satisfying our weaker assumption. Moreover, the upper bound in Theorem 3.1 in [17] depends on the coefficient

$$\alpha := \sum_{k\geq 0}(1-\alpha_k),$$

where for $k \geq 1$

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

But it can be shown that for any $k \geq 1$ we have $1 - \alpha_k \leq |A|\beta_k$, as noted by [10] in their proof of Lemma 3. Therefore $\alpha \leq |A|\beta + \alpha_0$ and Theorem 3.1 in [17] takes the form of Proposition B.9. □

As a consequence of this result we have the following lemma, proven in [22, Corollary A.7].

Lemma B.10. *Assume the process $\{X_t : t \in \mathbb{Z}\}$ satisfies Assumption 1. Then for any $w \in A^*$, any $a \in A$ and any $t > 0$ the following inequality holds*

$$\mathbb{P}\left(|\hat{p}_n(a|w) - p(a|w)| > t\right) \leq \mathrm{e}^{\alpha_0/32\mathrm{e}^2|A|^2(|A|\beta+2\alpha_0)}(|A|+1)\exp\left(\frac{-nt^2p(w)^2}{|w|+1}\right).$$

Now, we prove Lemmas B.11 and B.13. These two results are useful in the proof of Theorem 2.18.

Lemma B.11. *Assume the process $\{X_t : t \in \mathbb{Z}\}$ satisfies Assumption 1. Then for any $w \in A^*$ and any $t > 0$ such that $t < np(w)$ we have*

$$\mathbb{P}(N_n(w) \leq t) \leq \mathrm{e}^{\alpha_0/8\mathrm{e}^2|A|^2(|A|\beta+2\alpha_0)}|A| \exp\left(\frac{-n\left[p(w) - \frac{t}{n}\right]^2}{|w|+1}\right). \tag{B.12}$$

Proof. Using that $N_n(w) = \sum_{a\in A} N_n(w,a)$, $p(w) = \sum_{a\in A} p(wa)$ and $t - np(w) < 0$ we have that

$$\begin{aligned}\mathbb{P}(N_n(w) \le t) &= \mathbb{P}\left(\sum_{a\in A}[N_n(w,a) - np(wa)] \le t - np(w)\right)\\ &\le \sum_{a\in A}\mathbb{P}\left(|N_n(w,a) - np(wa)| \ge \frac{np(w)-t}{|A|}\right).\end{aligned}$$

Using Proposition B.9 we can bound above the right hand side of the last inequality by

$$\mathrm{e}^{\alpha_0/8\mathrm{e}^2(|A|\beta+2\alpha_0)}|A|\exp\left[-\frac{[np(w)-t]^2}{|A|^2(|w|+1)n}\right].$$

This implies the bound in (B.12). □

Lemma B.13. *Assume the process $\{X_t : t \in \mathbb{Z}\}$ satisfies Assumption 1. Let $u, w \in A^*$ and $b \in A$ such that $p(b|u) - p(b|w) > 0$. Then, for any $t < [p(b|u) - p(b|w)]^2/8$ we have that*

$$\begin{aligned}\mathbb{P}\left(D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \le t\right) \le\ & 2\mathrm{e}^{\alpha_0/32\mathrm{e}^2|A|^2(|A|\beta+2\alpha_0)}(|A|+1)\\ &\times \exp\left[-n\frac{t}{2}\min\left(\frac{p(w)^2}{|w|+1}, \frac{p(u)^2}{|u|+1}\right)\right].\end{aligned}$$

Proof. By Pinsker's inequality (see, e.g., [6, Section A.2] for a proof) we have that

$$\begin{aligned}D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) &\ge \frac{1}{2}\left[\sum_{a\in A}|\hat{p}_n(a|u) - \hat{p}_n(a|w)|\right]^2\\ &\ge \frac{1}{2}\left(\hat{p}_n(b|u) - \hat{p}_n(b|w)\right)^2. \qquad \text{(B.14)}\end{aligned}$$

Now, set $\nu = \frac{1}{8}[p(b|u) - p(b|w)]^2$ and define the events

$$C_{n,\nu}^{b,w,u} = \{X_1^n : |\hat{p}_n(b|u) - p(b|u)| \le \sqrt{\nu/2}\} \cap \{X_1^n : |\hat{p}_n(b|w) - p(b|w)| \le \sqrt{\nu/2}\}. \qquad \text{(B.15)}$$

Then, if $t < \nu$ we have that the event

$$\{X_1^n : D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \le t\} \cap C_{n,\nu}^{b,w,u} = \emptyset.$$

To see this note that by (B.14), if (B.15) holds then

$$D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \ge \frac{1}{2}\left[(p(b|u) - \sqrt{\nu/2}) - (p(b|w) + \sqrt{\nu/2})\right]^2 = \nu > t.$$

Therefore, using the bounds in Lemma B.10 we obtain for any $t < \nu$ that

$$\begin{aligned}\mathbb{P}\left(D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \le t\right) &\le \mathbb{P}\left(|\hat{p}_n(b|u) - p(b|u)| \ge \sqrt{\nu/2}\right)\\ &\quad + \mathbb{P}\left(|\hat{p}_n(b|w) - p(b|w)| \ge \sqrt{\nu/2}\right)\\ &\le 2\mathrm{e}^{\alpha_0/32\mathrm{e}^2|A|^2(|A|\beta+2\alpha_0)}(|A|+1)\exp\left[-n\frac{\nu}{2}\min\left(\frac{p(w)^2}{|w|+1}, \frac{p(u)^2}{|u|+1}\right)\right].\end{aligned}$$

□

References

[1] A. Barron, J. Rissanen, B. Yu, The minimum description length principle in coding and modeling, IEEE Trans. Inform. Theory 44 (6) (1998) 2743–2760. Information theory: 1948–1998.

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

[3] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, in: Wadsworth Statistics/Probability Series, Wadsworth Advanced Books and Software, Belmont, CA, 1984.

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

[5] J.R. Busch, P.A. Ferrari, A.G. Flesia, R. Fraiman, S.P. Grynberg, F. Leonardi, Testing statistical hypothesis on random trees and applications to the protein classification problem, Ann. Appl. Stat. 3 (2) (2009).

[6] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, Cambridge, 2006.

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

[8] 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.

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

[10] I. Csiszár, Z. Talata, On rate of convergence of statistical estimation of stationary ergodic processes, IEEE Trans. Inform. Theory 56 (8) (2010) 3637–3641.

[11] J. Dedecker, C. Prieur, New dependence coefficients. Examples and applications to statistics, Probab. Theory Related Fields 132 (2005) 203–236.

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

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

[14] S. Filippi, O. Cappé, A. Garivier, Optimism in reinforcement learning based on Kullback–Leibler divergence, in: 48th Annual Allerton Conference on Communication, Control, and Computing, 2010.

[15] A. Galves, C. Galves, J. Garcia, N.L. Garcia, F. Leonardi, Context tree selection and linguistic rhythm retrieval from written texts, 2010, pp. 1–25. ArXiv:0902.3619.

[16] A. Galves, A. Garivier, E. Gassiat, Data selection of context trees and classification, Technical Report, 2010.

[17] A. Galves, F. Leonardi, Exponential Inequalities for Empirical Unbounded Context Trees, in: Progress in Probability, vol. 60, Birkhauser, 2008, pp. 257–270.

[18] A. Galves, V. Maume-Deschamps, B. Schmitt, Exponential inequalities for VLMC empirical trees, ESAIM Probab. Stat. 12 (2008) 43–45.

[19] A. Garivier, Consistency of the unlimited BIC context tree estimator, IEEE Trans. Inform. Theory 52 (10) (2006) 4630–4635.

[20] A. Garivier, O. Cappé, The KL-UCB algorithm for bounded stochastic bandits and beyond, in: 23rd Conf. Learning Theory, COLT, Budapest, Hungary, 2011.

[21] A. Garivier, E. Moulines, On upper-confidence bound policies for non-stationary bandit problems, 2008. arxiv.org:0805.3415.

[22] F. Leonardi, Some upper bounds for the rate of convergence of penalized likelihood context tree estimators, Braz. J. Probab. Stat. 24 (2) (2010) 321–336.

[23] P. Massart, Concentration Inequalities and Model Selection, in: Lecture Notes in Mathematics, vol. 1896, Springer, Berlin, 2007, Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[24] J. Neveu, Martingales à Temps Discret, Masson, 1972.

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

[26] Z. Talata, T. Duncan, Unrestricted BIC context tree estimation for not necessarily finite memory processes, in: Information Theory, 2009. ISIT 2009. IEEE International Symposium on, 28 2009-July 3 2009, pp. 724–728.

[27] F.M.J. Willems, Y.M. Shtarkov, T.J. Tjalkens, The context-tree weighting method: basic properties, IEEE Trans. Inf. Theory 41 (3) (1995) 653–664.