Context Tree Selection: A Unifying View

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

[a] *aurelien.garivier@telecom-paristech.fr, LTCI, CNRS, Telecom ParisTech*
[b] *florencia@usp.br, Instituto de Matemática e Estatística, Universidade de São Paulo*

Abstract

The present paper investigates non-asymptotic properties of two popular procedures of context tree (or Variable Length Markov Chains) estimation: Rissanen's algorithm Context and the Penalized Maximum Likelihood criterion. First showing how they are related, we prove finite horizon bounds for the probability of over- and under-estimation. Concerning overestimation, no boundedness or loss-of-memory conditions are required: the proof relies on new deviation inequalities for empirical probabilities of independent interest. The underestimation properties rely on loss-of-memory and separation conditions of the process.

These results improve and generalize the bounds obtained in [**?** **?** **?**], refining asymptotic results of [**?** **?**]. Context tree models have been introduced by Rissanen in [**?**] as a parsimonious generalization of Markov models. Since then, they have been widely used in applied probability and statistics.

Keywords: Context Trees, Penalized Maximum Likelihood, Non-asymptotic Model Selection, VLMC, Deviation Inequalities, Martingales

1. Introduction

Context tree models (CTM), first introduced by Jorma Rissanen [**?**] as efficient tools in Information Theory, have since then been successfully studied and used in many fields of probability and statistics, including bio-informatics [**?**], universal coding [**?**], mathematical statistics [**?**] or linguistics [**?**]. Sometimes also called "Variable Length Markov Chain", a context tree process is informally defined as a Markov chain whose memory length may depend on the past symbols. As explained in Section 2, the set of all relevant memory blocks can be represented as a tree, while each tree defines a context tree model.

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 tree has a natural and

*Corresponding author

informative interpretation. In linguistics, for example, [**?**] showed how tree estimation highlights structural discrepancies between Brazilian and European Portuguese.

Of course, practical use of CTM requires the possibility of constructing efficient estimators $\hat{T}$ 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 by Rissanen [**?**], is based on the idea of *tree pruning*. They are somewhat reminiscent of the CART pruning procedures: a measure of discrepancy between a node's children determines whether they have to be removed from the tree. The second family of estimators relies 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 Information Theory, these procedures can be interpreted as applications of the *Minimum Description Length* principle [**?**].

First, only finite trees were considered. In this case, 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 2006, [**?**] proposed to call an estimator $\hat{T}$ *strongly consistent* if for every positive integer K, its truncation $\hat{T}_{|K}$ at level K was equal to $T_0|_K$ eventually almost surely. They showed that, with this definition, PML estimators can be strongly consistent if the penalties are appropriately chosen and if the maximization is restricted to not-too-deep models. [**?**] proved that for finite trees no restriction on the model depth is required in the maximization.

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 thank to 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.

More recently, the problem of deriving *non-asymptotic* bounds for the probability of correct estimation was considered. In [**?**], non-universal inequalities are 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 [**?**], and to PML estimators in [**?**]. Using recent advances in weak dependence theory, all these results strongly rely on mixing hypotheses of the process.

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. 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 [**?** **?** **?**]. To do this, we derive *"self-normalized"* non-asymptotic deviation inequalities, using martingale techniques inspired from proofs of the Law of the Iterated Logarithm [**?** **?**]. These inequalities prove interestersting in completely different fields, as for instance reinforcement learning [**?** **?**]. We also derive bounds on the probability of under-estimation: with high probability, $\hat{T}$ contains all nodes of T_0 whose depth is not too large. However, these results require additional assumptions on the process, namely some loss-of-memory hypotheses and so-called separation conditions which ensure that the process cannot be too well approximated by lower-order models.

The paper is organized as follows. In Section 2 we fix 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. Section 4 contains a discussion on their scope and on possible variants. Finally, the Appendix contains the statement and proof of the self-normalized deviation inequalities.

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, and $A^* = \bigcup_{k\geq 0} A^k$ the set of all finite sequences on alphabet A. The length of the sequence $w \in A^*$ is $|w|$. For $1 \leq i \leq j \leq |w|$, we denote $w_i^j = (w_i, \ldots, w_j) \in A^{j-i+1}$. The empty sequence is represented by the symbol λ and his length is $|\lambda| = 0$.

Given two sequences v and w, we denote by vw the sequence of length $|v| + |w|$ obtained by concatenating the two sequences. In particular, $\lambda w = w\lambda = w$. The concatenation of sequences is also extended to semi-infinite sequences $v = (\ldots, v_{-2}, v_{-1}) = v_{-\infty}^{-1}$.

We say that the sequence s is a *proper suffix* of the sequence w if there exists a sequence u, with $|u| \geq 1$, such that $w = us$. In this case we write $w \succ s$. When $w \succ s$ or $s = w$ we write $w \succeq s$. The longest proper suffix of w is denoted by $\mathrm{suf}(w)$.

Definition 1. A set T of finite or semi-infinite sequences is a *tree* if no sequence $s \in T$ is a suffix of another sequence $w \in T$. It is a *complete tree* if every semi-infinite sequence of A has a suffix in T.

The *height* of the tree T is defined as

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

In the case $h(T) < +\infty$ we say that T is *bounded* and we denote by $|T|$ the cardinality of T. On the other hand, if $h(T) = +\infty$ we say that the tree T is *unbounded*. The elements of T are sometimes

called the *leaves* of T. An *internal node* of T is a proper suffix of a leaf. For any finite sequence w and for any tree T, we define the tree T_w as the set of branches in T which have w as a suffix, that is

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

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

$$T|_K = \{w \in T \colon |w| \leq K\} \cup \{w \colon |w| = K \text{ and } u \succ w, \text{ for some } u \in T\}.$$

Given two trees T_1 and T_2 we say that T_1 is *included* in T_2 (and we denoted it by $T_2 \succeq T_1$) if for any sequence $w \in T_1$ there exists a sequence $u \in T_2$ such that $u \succeq w$; 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

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

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

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

Definition 2. A sequence $w \in A^j$ is a *context* for the process X_t if it satisfies

1. $p(w) > 0$;
2. for any sequence $v \in A^*$ such that $p(v) > 0$ and $v \succeq w$,

$$\mathbb{P}(X_0 = a \,|\, X_{-|v|}^{-1} = v) \;=\; p(a|w), \quad \text{for all } a \in A;$$

3. no suffix of w satisfies 1. and 2.

An *infinite context* is a semi-infinite sequence $w_{-\infty}^{-1}$ such that any of its suffixes w_{-j}^{-1}, $j = 1, 2, \ldots$ is a context.

Definition 2 implies that the set of all contexts (finite or infinite) is a complete tree, also called a *context tree*. For example, i.i.d. processes have a context tree reduced to the set $\{\lambda\}$, and generic Markov chains of order 1 have a context tree equal to A.

Let $d \leq n$ be positive integers. In what follows, we assume that we observe $X_{-d+1}, \ldots, X_0, X_1, \ldots X_n$ distributed from a stationary ergodic process with law $\mathbb{P}$ whose (possibly infinite) context tree is denoted by T_0. We consider a set $\mathscr{T}$ of candidate trees of depth at most d (that may depend on $X_{-d+1}, \ldots X_n$). The goal is to select a tree $T \in \mathscr{T}$ as close as possible from T_0, in some sense that will be formally given below. Several choices are possible, depending on the problem. One may want to choose for $\mathscr{T}$ the set of all complete trees; but in some applications, there are structural restrictions that induce a more restrictive choice of $\mathscr{T}$. For example, in stochastic modelling of natural languages / codified written text, grammatical rules impose constraints on the structure of the possible trees, see [**?**]. Another popular option is to impose a constraint on the number

of occurrences of their contexts in the sample X_1^n. The only hypothese we make on $\mathscr{T}$ is "closed under inclusion", that is :

$$T \in \mathscr{T} \text{ and } T \succeq T' \implies T' \in \mathscr{T} .$$

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 to be observed, similar results would be obtained by using quasi-maximum likelihood estimators [**?**].

We denote by $V_{\mathscr{T}} = \cup_{T \in \mathscr{T}} T$ the set of all candidate contexts. Note that $V_{\mathscr{T}}$ is suffix-closed:

$$w \in V_{\mathscr{T}} \text{ and } w \succeq v \implies v \in V_{\mathscr{T}} .$$

Given a sequence $w \in V_{\mathscr{T}}$ and a 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:

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

Besides, $N_n(w)$ will denote the sum $\sum_{a \in A} N_n(w, a)$. Remark that $N_n(w, a)$ may differ from $N_n(wa)$ by at most 1, if $X_{n-|w|}^n = wa$; this little notational trick will simplify the discussion in the sequel.

Given a tree $T \in \mathscr{T}$, the maximum likelihood of the sequence $X_1, \ldots, X_n$ is given by

$$\hat{\mathbb{P}}_{\mathrm{ML},T}(X_1^n) = \prod_{w \in T} \prod_{a \in A} \hat{p}_n(a|w)^{N_n(w,a)}, \tag{2}$$

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

$$\hat{p}_n(a|w) = \frac{N_n(w, a)}{N_n(w)} \tag{3}$$

if $N_n(w) > 0$ and $\hat{p}_n(a|w) = 1/|A|$ otherwise. For any sequence w 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 shall most often use the *Küllback-Leibler divergence*, defined for two probability measures p and q on A by

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

In the following subsections we introduce the algorithm Context and the penalized maximum likelihood criterion. Both context tree estimation algorithms rely on a procedure CST computing the (compact) *suffix tree* of a sequence X_1^n. This is a tree whose edges are labeled by sequences, and such that each suffix of X_1^n corresponds to exactly one path from the root of the tree to a leaf. Efficient implementations of the CST procedure only require linear time (see [**? ?** , and references therein]).

2.1. Algorithm Context

For all sequences $w \in V_{\mathscr{T}}$ let

$$\Delta_n(w) = \sum_{b\colon bw \in V_{\mathscr{T}}} N_n(bw) D\left(\hat{p}_n(\cdot|bw); \hat{p}_n(\cdot|w)\right).$$

Algorithm Context depends on a non-decreasing sequence of real numbers $(\delta_n)_n$ (convenient choices of $(\delta_n)_n$ will be deduced from Theorem 1 and 2 below). For a sequence X_1^n, let $\mathcal{C}_w(X_1^n)$ be defined for all $w \in V_{\mathscr{T}}$ by the following induction:

$$\mathcal{C}_w(X_1^n) = \begin{cases} 0, & \text{if } N_n(w) \leq 1, \\ \max\{\mathbb{1}\{\Delta_n(w) \geq \delta_n\}, \max_{b\in A} \mathcal{C}_{bw}(X_1^n)\}, & \text{if } N_n(w) > 1. \end{cases}$$

The Context tree estimator $\hat{T}_C$ is the set of all $w \in V_{\mathscr{T}}$ such that $\mathcal{C}_w(X_1^n) = 0$ and $\mathcal{C}_u(X_1^n) = 1$ for all u such that $u \prec w$. It can be computed as the result of $PRUNE(T^c)$, where $T^c = CST(X_1^n)$ and procedure PRUNE is described in Algorithm 1. The algorithm uses a subfunction *ROOT* which returns the root of a tree. Besides, we denote by T_a the subtree of T rooted in the node that shares an edge labelled by a with the root of T.

Algorithm 1 PRUNE(T)

```
if |T| = 1 then
  return T
end if
isLeaf ← true
for all a ∈ A do
  P = PRUNE(T_a)
  if |P| > 1 then
    isLeaf ← false
  end if
  S_a ← P
end for
w ← ROOT(T)
if isLeaf AND Δ_n(w) < δ_n then
  S ← w
end if
return S
```

Remark 1. *We present here the original choice of divergence by Rissanen in $\Delta_n(w)$, but other equivalent possibilities have been proposed in the literature (see for instance [**?**]). For example, $\tilde{\Delta}_n(w) = \max_{a,b} |\hat{p}_n(a|bw) - \hat{p}_n(a|w)|$ or $\max_b D(\hat{p}_n(\cdot|bw); \hat{p}_n(\cdot|w))$.*

2.2. The penalized maximum likelihood criterion

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

$$\hat{T}_{PML}(X_1^n) = \arg\max_{T\in\mathscr{T}} \left\{ \log \hat{\mathbb{P}}_{\mathrm{ML},T}(X_1^n) - \mathrm{pen}(n, |T|) \right\},$$

where $\mathrm{pen}(n,|T|)$ is some penalty function. In the sequel, we assume that the penalty of a model is proportional to its dimension; with this very standard choice, we write $\mathrm{pen}(n,|T|) = |T| f(n)$ for some positive function f.

It may first appear practically impossible to compute $\hat{T}_{PML}(X_1^n)$ [**?**]. Fortunately, Csiszár and Talata showed in their article [**?**] how to adapt the Context Tree Maximizing (CTM) method [**?**] to obtain a simple and efficient algorithm computing $\hat{T}_{PML}(X_1^n)$. For self-containment, we briefly present it here: define recursively, for any $w \in V_{\mathscr{T}}$, the value

$$V_w(X_1^n) = \max\{e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n), \prod_{a\in A:\ aw\in V_{\mathscr{T}}} V_{aw}(X_1^n)\}$$

and the indicator

$$\mathcal{X}_w(X_1^n) = \mathbb{1}\{\prod_{a\in A:\ aw\in V_{\mathscr{T}}} V_{aw}(X_1^n) > e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\}\,.$$

By convention, if $\{a\in A\colon aw\in V_{\mathscr{T}}\} = \emptyset$ or if $N_n(w) = 0$, then $V_w(X_1^n) = e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)$ and $\mathcal{X}_w(X_1^n) = 0$. It is shown in [**?**] that $\hat{T}_{PML}(X_1^n)$ is exactly the set of all words $w \in V_{\mathscr{T}}$ such that $\mathcal{X}_w(X_1^n) = 0$ and $\mathcal{X}_u(X_1^n) = 1$ for all $u \prec w$. Thus, $\hat{T}_{PML}(X_1^n)$ can be computed as $CTM(T^c)$, where $T^c = CST(X_1^n)$ and algorithm CTM is described in Algorithm 2.

Algorithm 2 CTM(T)

```
if |T| = 1 then
  return [T, e^{-f(n)}]
end if
for all a ∈ A do
  [P, V_a] = CTM(T_a)
  S_a ← P
end for
w ← ROOT(T)
if ∏_{a∈A} V_a < e^{-f(n)} P̂_{ML,w}(X_1^n) then
  S ← w
  V ← e^{-f(n)} P̂_{ML,w}(X_1^n)
else
  V ← ∏_{a∈A} V_a
end if
return [S, V]
```

2.3. Results

In this subsection we present the main results of this article. First, we show that there is a close relationship between the estimators $\hat{T}_{PML}(X_1^n)$ and $\hat{T}_C(X_1^n)$.

Proposition 1. *For all sequences X_1^n, if $\delta_n \le f(n)$ then*

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

We now state a new bound on the probability of over-estimation by the Context algorithm.

Theorem 1. *For every positive integer n and for all δ_n it holds that*

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

An immediate consequence of Proposition 1 and Theorem 1 is that the Penalized Maximum Likelihood estimator also avoids over-estimation with overwhelming probability provided that δ_n is chosen appropriately:

Corollary 1. *If $\delta_n \leq f(n)$ then*

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

Remark 2. *Following [**?**], it is sometimes proposed to consider only a limited number $k(n)$ of nodes. A straightforward modification of the proof shows that*

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

and similarly if the Penalized Maximum Likelihood criterion is minimized only over trees containing no more than $k(n)$ nodes fixed a priori, it holds that

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

*In particular, penalties of order $O(\log\log n)$, smaller than the BIC prove to be sufficient to avoid over-estimation if an upper-bound (independent of n) on the depth of the tree is known (as was proved by Finesso for Markov chains, see [**?**] and references therein). More generally, given a function $k(n)$, one can easily choose δ_n so as to ensure consistency.*

The problem of underestimation in context tree models is very different, and requires additional hypotheses on the process $\{X_t \colon t \in \mathbb{Z}\}$. Define the *continuity coefficients* $\{\alpha_k\}_{k\in\mathbb{N}}$ of the process by

$$\alpha_0 := \inf_{x_{-\infty}^{-1}, a\in A} \{\, p(a|x_{-\infty}^{-1}) \,\},$$
$$\alpha_k := \inf_{u\in A^k} \sum_{a\in A} \inf_{x_{-\infty}^{-1}} p(a|x_{-\infty}^{-1}u)\,, \qquad k \geq 1\,.$$

In the sequel, we make the following assumptions.

Assumption 1. *The process $\{X_t \colon t \in \mathbb{Z}\}$ is such that $\alpha_0 > 0$ and*

$$\alpha := \sum_{k\in\mathbb{N}} (1 - \alpha_k) \;<\; +\infty$$

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. As the analysis leaves some choice in the determination of the candidate trees $\mathscr{T}$, we need the following "separation" assumption over $\mathscr{T}$:

Assumption 2. *There exists $\epsilon > 0$ such that for any $u \in T_0|_K$ and any $w \prec u$, there exists $v \in V_{\mathscr{T}}$ satisfying*

$$p(v)\, D\left(p(\cdot|v); p(\cdot|w)\right) > \epsilon\,.$$

Remark 3. *Assumption 2 says that the set of candidate trees $\mathscr{T}$ is rich enough to ensure that if w is not a context of $T_0|_K$, then it can be seen within $V_{\mathscr{T}}$. [?] and [?] give sufficient conditions for Assumption 2 to hold with high probability.*

Theorem 2. *Assume $f(n)$ is such that $f(n)/n \to \infty$ when $n \to \infty$. Then for any $K \in \mathbb{N}$, under Assumptions 1 and 2 there exist constants c_1, c_2 and n_0 such that*

$$\mathbb{P}\big(T_0|_K \preceq \hat{T}_{PML}(X_1^n)\big) \geq 1 - 3e^{\frac{1}{e}}|A|^{2+K}c_1 \exp\left[\frac{-c_2\, n}{64e|A|^3(\alpha+\alpha_0)\log^2\alpha_0}\right].$$

Corollary 2. *For any $K \in \mathbb{N}$, under Assumptions 1 and 2 and if $\delta_n \leq f(n)$, we have*

$$\mathbb{P}\left(T_0|_K \preceq \hat{T}_C(X_1^n)\right) \geq 1 - 3e^{\frac{1}{e}}|A|^{2+K}c_1 \exp\left[\frac{-c_2\, n}{64e|A|^3(\alpha+\alpha_0)\log^2\alpha_0}\right].$$

Remark 4. *Extensions of Theorem 2 can 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 decreases with the sample size. This leads to a rather technical condition, see for instance [?].*

3. Proofs

3.1. Proof of Proposition 1

It is sufficient to prove that $\mathcal{X}_w(X_1^n) \leq \mathcal{C}_w(X_1^n)$ for all $w \in V_{\mathscr{T}}$. Assume that there exists $w \in V_{\mathscr{T}}$ such that $\mathcal{X}_w(X_1^n) = 1$ and $\mathcal{C}_w(X_1^n) = 0$. Note that $\mathcal{C}_w(X_1^n) = 0$ implies $\mathcal{C}_{uw}(X_1^n) = 0$ for all $uw \in V_{\mathscr{T}}$; hence, w can be chosen such that $\mathcal{X}_{bw}(X_1^n) = 0$ for any $bw \in V_{\mathscr{T}}$, $b \in A$.

In this case, we have by definition:

$$\prod_{b\colon bw \in V_{\mathscr{T}}} e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},bw}(X_1^n) = \prod_{b\colon bw \in V_{\mathscr{T}}} V_{bw}(X_1^n) \;>\; e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n).$$

Taking logarithm,

$$\sum_{b\colon bw \in V_{\mathscr{T}}} \sum_{a \in A} N_n(bwa) \log \frac{\hat{p}_n(a|bw)}{\hat{p}_n(a|w)} \;>\; (|\{b\colon bw \in V_{\mathscr{T}}\}| - 1)f(n)$$

and

$$\Delta_n(w) = \sum_{b\colon bw \in V_{\mathscr{T}}} N_n(bw) D\left(\hat{p}_n(\cdot|bw); \hat{p}_n(\cdot|w)\right) > (|\{b\colon bw \in V_{\mathscr{T}}\}| - 1)f(n) \geq \delta_n$$

because as $\mathcal{X}_w(X_1^n) = 1$, we have $|\{b\colon bw \in V_{\mathscr{T}}\}| \geq 2$. This contradicts the fact that $\mathcal{C}_w(X_1^n) = 0$, and concludes the proof.

3.2. Proof of Theorem 1

Let O_n be the event $\left\{\hat{T}_C(X_1^n) \not\preceq T_0\right\}$. Overestimation 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) word u such that $w = us$; thus, O_n can be decomposed as:

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

By definition, for all $b \in A$ it holds that $p(\cdot|w) = p(\cdot|bw)$, and thus:

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

Hence,

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

Using Theorem 3, it follows that

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

where C_n denotes the number of different contexts appearing in X_1^n. But C_n is almost-surely upper-bounded by $n(n-1)/2$, and the result follows.

3.3. Proof of Theorem 2

The proof of the Theorem is analogous to that obtained in [**?**], we include it here for completeness. If U_n denotes the event $\{T_0|_K \not\preceq \hat{T}_{PML}(X_1^n)\}$ then

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

Let $w \prec u \in T_0|_K$ such that $\mathcal{X}_w(X_1^n) = 0$. By Assumption 2 there must exists a tree $T \in \mathscr{T}$ such that

$$\sum_{v \in T_w} p(v) D\left(p(\cdot|v); p(\cdot|w)\right) \;\geq\; \epsilon\,. \tag{4}$$

Then we have

$$\mathbb{P}\left(\,\mathcal{X}_w(X_1^n) = 0\,\right) \;=\; \mathbb{P}\left(\prod_{a \in A\colon\, aw \in V_{\mathscr{T}}} V_{aw}(X_1^n) \;\leq\; e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\right).$$

By definition, for any $aw \in V_{\mathscr{T}}$it can be shown recursively that

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

see for example Lemma 4.4 in [**?**]. Therefore,

$$\begin{aligned}
&\mathbb{P}\Bigg(\prod_{a \in A\colon\, aw \in V_{\mathscr{T}}} V_{aw}(X_1^n) \;\leq\; e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\Bigg) \\
&\qquad\leq\; \mathbb{P}\left(\prod_{u \in T_w} e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},u}(X_1^n) \;\leq\; e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\right) \\
&\qquad=\; \mathbb{P}\left(\sum_{u \in T_w} \log \hat{\mathbb{P}}_{\mathrm{ML},u}(X_1^n) - \log \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n) \leq (|T_w| - 1) f(n)\right),
\end{aligned}$$

by noticing that

$$\prod_{a \in A\colon\, aw \in V_{\mathscr{T}}} \prod_{s \in T_{aw}} e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},s}(X_1^n) = \prod_{u \in T_w} e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},u}(X_1^n)\,.$$

Dividing by n, subtracting $\sum_{u \in T_w} p(u) D\left(p(\cdot|u); p(\cdot|w)\right)$ on both sides and using the inequality in (4) we can write the last probability as

$$\mathbb{P}\left(\sum_{u \in T_w} L_n(u) - L_n(w) \;\leq\; \frac{(|T_w| - 1) f(n)}{n} - \epsilon\right)$$

where for any finite sequence s

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

Then, for all $n \geq n_0$, where n_0 is a sufficiently large integer such that

$$\frac{|T_w| f(n)}{n} \;<\; \frac{\epsilon}{2}$$

we can bound above the last expression by

$$\mathbb{P}\left(\,|L_n(w)| > \frac{\epsilon}{4}\right) \;+\; \sum_{u \in T_w} \mathbb{P}\left(\,|L_n(u)| > \frac{\epsilon}{4|T_w|}\right).$$

Using Corollary A.7(c) in [**?**] we obtain

$$\mathbb{P}\left(\,\mathcal{X}_w(X_1^n) = 0\,\right) \;\leq\; 3e^{\frac{1}{e}} |A|^2 (1 + |T_w|) \exp\left(-\frac{n \min(1, (\epsilon/4|T_w|)^2) \alpha_0^{2(h(T_w)+1)}}{64e|A|^3(\alpha + \alpha_0) \log^2 \alpha_0 \left(h(T_w) + 1\right)}\right),$$

by noticing that $p(w) \geq p(u) \geq \alpha_0^{h(T_w)}$ for any $u \in T_w$. We conclude the proof of Theorem 2 by observing that we only have a finite number of sequences $w \prec u \in T_0|_K$, therefore we obtain

$$\mathbb{P}\left(T_0|_K \preceq \hat{T}_{PML}(X_1^n)\right) \geq 1 - 3e^{\frac{1}{e}}|A|^{2+K}c_1 \exp\left(\frac{-c_2\, n}{64e|A|^3(\alpha+\alpha_0)\log^2\alpha_0}\right),$$

where

$$c_1 = \max_{w\prec u\in T_0|_K}(1+|T_w|) \quad \text{and} \quad c_2 = \min_{w\prec u\in T_0|_K}\left\{\frac{\min(1,(\epsilon/4|T_w|)^2)\alpha_0^{2(h(T_w)+1)}}{h(T_w)+1}\right\}.$$

4. Discussion

Algorithm Context has many variants: in fact, $\Delta_n(w)$ can rely on different (pseudo-) distances rather than on Kullback-Leibler information. Another popular choice (see [**?**]) is

$$\Delta_n(w) = \max_{b,c\in A}\left|\hat{p}_n(c|bw) - \hat{p}_n(c|w)\right|.$$

Similar results can be obtained in this case by combining Theorem 3 and Pinsker's inequality [**?**]:

$$\left(\sup_{B\subset A}\hat{P}_n(A|a_1^k) - P(A|a_1^k)\right)^2 \leq \frac{1}{2}D(\hat{P}_n;P).$$

Then,

$$\begin{aligned}
\mathbb{P}\left[\max_{b\in A}\left|\hat{p}_n(b|a_1^k) - P(b|a_1^k)\right| > \sqrt{\frac{\delta}{N_n}}\right] &\leq \mathbb{P}\left[\sup_{B\subset A}\left(\hat{p}_n(B|a_1^k) - P(B|a_1^k)\right)^2 > \frac{\delta}{N_n}\right] \\
&\leq \mathbb{P}\left[N_n D\left(\hat{p}_n\left(\cdot|a_1^k\right);P\left(\cdot|a_1^k\right)\right) > 2\delta\right] \\
&\leq 2e\left(2\delta\log(n)+|A|\right)\exp\left(-\frac{2\delta}{|A|}\right).
\end{aligned}$$

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 in unavoidable. All what can be expected from the estimator it to highlight evidence of as much dependence structure as possible, while maintaining a limited probability of false discovery. Let us insist once again that no mixing assumption is necessary for Theorem 1. On the other hand, for under-estimation, some conditions like Assumption 1 seem unavoidable.

Appendix 1: Martingale deviation inequalities

This section contains the statement and derivation of two deviation inequalities that are useful to prove the main results of this paper. As we believe they are interesting on their own, we include them in a separate section. The originality is that deviations from expectations are not controlled in L^p norm, but thru the Kullback-Leibler divergence; 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 the 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 [**?**]. As for [**?**], the ingredients of the proofs are mostly inspired by [**?**]. Completely different applications of these inequalities may be found in [**?** **?**].

Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary process of whose (possibly infinite) context tree is T, and let $\mathcal{F}_n$ be the σ-field generated by $(X_j)_{j\leq n}$. For $k\in\mathbb{N}$, $a_1^k\in A^k$, denote $p(b|a_1^k)=\mathbb{P}\left(X_{k+1}=b|X_1^k=a_1^k\right)$. Define

$$\xi_j=\mathbb{1}\{X_{j-k}^{j-1}=a_1^k\},\quad \text{for } j\geq 1,$$

$$\chi_j=\mathbb{1}\{X_{j-k}^{j}=a_1^k b\},\quad \text{for } j\geq 1$$

so that $N_n(a_1^k)=\sum_{j=1}^n\xi_j$ and $N_n(a_1^k,b)=\sum_{j=1}^n\chi_j$.

Denote $\hat{p}_n(b|a_1^k)=S_n(a_1^kb)/N_n(a_1^k)$. From now on, a block a_1^k having a suffix in T_0 and a symbol b are fixed, and there will be no possible confusion in using the shortcuts $p=P(b|a_1^k)$, $N_n=N_n(a_1^k)$, $S_n=N_n(A_1^k,b)$ and $\hat{p}_n=\hat{P}_n(b|a_1^k)=\frac{S_n}{N_n}$. Besides, 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 2.

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

To prove Proposition 2, we introduce a few notations: for every $\lambda>0$, let

$$\phi_p(\lambda)=\log\mathbb{E}\left[\exp\left(\lambda X_1\right)\right]=\log\left(1-p+p\exp\left(\lambda\right)\right).$$

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

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

To proceed, we make use of the so-called 'peeling trick' [**?**]: we divide the interval $\{1,\dots,n\}$ of possible values for N_n into "slices" $\{t_{k-1}+1,\dots,t_k\}$ 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 $t_k = \lfloor (1+\eta)^k \rfloor$. Let D be the first integer such that $t_D \geq n$, that is $D = \left\lceil \frac{\log n}{\log 1+\eta} \right\rceil$. Let $A_k = \{t_{k-1} < N_n \leq t_k\} \cap \{N_n d(\hat{p}_n; p) > \delta\}$. We have:

$$\mathbb{P}\left(N_n d(\hat{p}_n; p) > \delta\right) \leq \mathbb{P}\left(\bigcup_{k=1}^{D} A_k\right) \leq \sum_{k=1}^{D} \mathbb{P}\left(A_k\right). \tag{5}$$

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, $\mathbb{P}\left(A_k\right) = 0$ 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 :
$$\hat{p}_n \geq p \text{ and } d(\hat{p}_n; p) > \frac{\delta}{N_n} \implies \hat{p}_n \geq z.$$

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

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

As $\left(W_t^{\lambda}\right)_{t\geq 0}$ 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:

$$\begin{aligned}
\mathbb{P}\left(\left\{\tilde{t}_{k-1} < N_n \leq t_k\right\} \cap \{\hat{p}_n \geq p\} \cap \{N_n d(\hat{p}_n, p) \geq \delta\}\right) &\leq \mathbb{P}\left(W_n^{\lambda(z)} > \exp\left(\frac{\delta}{1+\eta}\right)\right) \\
&\leq \exp\left(-\frac{\delta}{1+\eta}\right).
\end{aligned} \tag{6}$$

Similarly,

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

so that

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

Finally, by Equation (5),

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

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

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

Remark 5. *The bound of Proposition 2 also holds for $\mathbb{P}\left(N_n d(\hat{p}_n, p) \geq \delta | N_n > 0\right)$: 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 Equation 6, the proof can be rewritten conditionally on $\{N_n > 0\}$; this leads to:

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

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 1. *For any probability distributions P and Q on an alphabet X,*

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

Proof.

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

because

$$\sum_{x\in X}\frac{1-P(x)}{|X|-1} = \sum_{x\in X}\frac{1-Q(x)}{|X|-1} = 1.$$

□

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

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

Theorem 3.

$$\mathbb{P}\left[N_n(a_1^k)D\left(\hat{p}_n\left(\cdot|a_1^k\right);p\left(\cdot|a_1^k\right)\right)>\delta\right]\leq 2e\left(\delta\log(n)+|A|\right)\exp\left(-\frac{\delta}{|A|}\right).$$

Proof. By combining Lemma 1 and Proposition 2, we get

$$\begin{aligned}
\mathbb{P}\left[N_n(a_1^k)D\left(\hat{p}_n\left(\cdot|a_1^k\right);p\left(\cdot|a_1^k\right)\right)>\delta\right] &\leq \mathbb{P}\left[\sum_{b\in A}N_n d\left(\hat{p}_n\left(b|a_1^k\right);p\left(b|a_1^k\right)\right)>\delta\right] \\
&\leq \sum_{b\in A}\mathbb{P}\left[N_n d\left(\hat{p}_n\left(b|a_1^k\right);p\left(b|a_1^k\right)\right)>\frac{\delta}{|A|}\right] \\
&\leq (|A|)2e\left[\frac{\delta}{|A|}\log(n)\right]\exp\left(-\frac{\delta}{|A|}\right) \\
&\leq (|A|)2e\left(\frac{\delta}{|A|}\log(n)+1\right)\exp\left(-\frac{\delta}{|A|}\right) \\
&= 2e\left(\delta\log(n)+|A|\right)\exp\left(-\frac{\delta}{|A|}\right).
\end{aligned}$$

□

Remark 7. *It follows from Remark 5 that the following variant of Theorem 3 holds:*

$$\mathbb{P}\left[N_n(a_1^k)D\left(\hat{p}_n\left(\cdot|a_1^k\right);p\left(\cdot|a_1^k\right)\right)>\delta|N_n>0\right]\leq 2e\left(\delta\log(n)+|A|-1\right)\exp\left(-\frac{\delta}{|A|-1}\right).$$

Acknowledgments

This work was supported by Fapesp (process 2009/09411-8) and USP-COFECUB project. F.L. also thanks CNPq for the support (grant 302162/2009-7).

Context Tree Selection: A Unifying View

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

[a] *aurelien.garivier@telecom-paristech.fr, LTCI, CNRS, Telecom ParisTech*
[b] *florencia@usp.br, Instituto de Matemática e Estatística, Universidade de São Paulo*

Abstract

The present paper investigates non-asymptotic properties of two popular procedures of context tree (or Variable Length Markov Chains) estimation: Rissanen's algorithm Context and the Penalized Maximum Likelihood criterion. First showing how they are related, we prove finite horizon bounds for the probability of over- and under-estimation. Concerning overestimation, no boundedness or loss-of-memory conditions are required: the proof relies on new deviation inequalities for empirical probabilities of independent interest. The underestimation properties rely on loss-of-memory and separation conditions of the process.

These results improve and generalize the bounds obtained in [10, 9, 14], refining asymptotic results of [3, 5]. Context tree models have been introduced by Rissanen in [17] as a parsimonious generalization of Markov models. Since then, they have been widely used in applied probability and statistics.

Keywords: Context Trees, Penalized Maximum Likelihood, Non-asymptotic Model Selection, VLMC, Deviation Inequalities, Martingales

1. Introduction

Context tree models (CTM), first introduced by Jorma Rissanen [17] as efficient tools in Information Theory, have since then been successfully studied and used in many fields of probability and statistics, including bio-informatics [2], universal coding [20], mathematical statistics [3] or linguistics [8]. Sometimes also called "Variable Length Markov Chain", a context tree process is informally defined as a Markov chain whose memory length may depend on the past symbols. As explained in Section 2, the set of all relevant memory blocks can be represented as a tree, while each tree defines a context tree model.

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 tree has a natural and

*Corresponding author

informative interpretation. In linguistics, for example, [8] showed how tree estimation highlights structural discrepancies between Brazilian and European Portuguese.

Of course, practical use of CTM requires the possibility of constructing efficient estimators $\hat{T}$ 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 by Rissanen [17], is based on the idea of *tree pruning*. They are somewhat reminiscent of the CART pruning procedures: a measure of discrepancy between a node's children determines whether they have to be removed from the tree. The second family of estimators relies 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 Information Theory, these procedures can be interpreted as applications of the *Minimum Description Length* principle [1].

First, only finite trees were considered. In this case, 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 2006, [5] proposed to call an estimator $\hat{T}$ *strongly consistent* if for every positive integer K, its truncation $\hat{T}_{|K}$ at level K was equal to $T_0|_K$ eventually almost surely. They showed that, with this definition, PML estimators can be strongly consistent if the penalties are appropriately chosen and if the maximization is restricted to not-too-deep models. [12] proved that for finite trees no restriction on the model depth is required in the maximization.

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 thank to 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.

More recently, the problem of deriving *non-asymptotic* bounds for the probability of correct estimation was considered. In [10], non-universal inequalities are 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 [9], and to PML estimators in [14]. Using recent advances in weak dependence theory, all these results strongly rely on mixing hypotheses of the process.

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. 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 [10, 9, 14]. To do this, we derive *"self-normalized"* non-asymptotic deviation inequalities, using martingale techniques inspired from proofs of the Law of the Iterated Logarithm [16, 4]. These inequalities prove interestersting in completely different fields, as for instance reinforcement learning [13, 6]. We also derive bounds on the probability of under-estimation: with high probability, $\hat{T}$ contains all nodes of T_0 whose depth is not too large. However, these results require additional assumptions on the process, namely some loss-of-memory hypotheses and so-called separation conditions which ensure that the process cannot be too well approximated by lower-order models.

The paper is organized as follows. In Section 2 we fix 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. Section 4 contains a discussion on their scope and on possible variants. Finally, the Appendix contains the statement and proof of the self-normalized deviation inequalities.

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, and $A^* = \bigcup_{k\geq 0} A^k$ the set of all finite sequences on alphabet A. The length of the sequence $w \in A^*$ is $|w|$. For $1 \leq i \leq j \leq |w|$, we denote $w_i^j = (w_i, \ldots, w_j) \in A^{j-i+1}$. The empty sequence is represented by the symbol λ and his length is $|\lambda| = 0$.

Given two sequences v and w, we denote by vw the sequence of length $|v| + |w|$ obtained by concatenating the two sequences. In particular, $\lambda w = w\lambda = w$. The concatenation of sequences is also extended to semi-infinite sequences $v = (\ldots, v_{-2}, v_{-1}) = v_{-\infty}^{-1}$.

We say that the sequence s is a *proper suffix* of the sequence w if there exists a sequence u, with $|u| \geq 1$, such that $w = us$. In this case we write $w \succ s$. When $w \succ s$ or $s = w$ we write $w \succeq s$. The longest proper suffix of w is denoted by $\mathrm{suf}(w)$.

Definition 1. A set T of finite or semi-infinite sequences is a *tree* if no sequence $s \in T$ is a suffix of another sequence $w \in T$. It is a *complete tree* if every semi-infinite sequence of A has a suffix in T.

The *height* of the tree T is defined as

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

In the case $h(T) < +\infty$ we say that T is *bounded* and we denote by $|T|$ the cardinality of T. On the other hand, if $h(T) = +\infty$ we say that the tree T is *unbounded*. The elements of T are sometimes

called the *leaves* of T. An *internal node* of T is a proper suffix of a leaf. For any finite sequence w and for any tree T, we define the tree T_w as the set of branches in T which have w as a suffix, that is

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

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

$$T|_K = \{w \in T \colon |w| \leq K\} \cup \{w \colon |w| = K \text{ and } u \succ w, \text{ for some } u \in T\}.$$

Given two trees T_1 and T_2 we say that T_1 is *included* in T_2 (and we denoted it by $T_2 \succeq T_1$) if for any sequence $w \in T_1$ there exists a sequence $u \in T_2$ such that $u \succeq w$; 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

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

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

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

Definition 2. A sequence $w \in A^j$ is a *context* for the process X_t if it satisfies

1. $p(w) > 0$;
2. for any sequence $v \in A^*$ such that $p(v) > 0$ and $v \succeq w$,

$$\mathbb{P}(X_0 = a \,|\, X_{-|v|}^{-1} = v) \;=\; p(a|w), \quad \text{for all } a \in A;$$

3. no suffix of w satisfies 1. and 2.

An *infinite context* is a semi-infinite sequence $w_{-\infty}^{-1}$ such that any of its suffixes w_{-j}^{-1}, $j = 1, 2, \ldots$ is a context.

Definition 2 implies that the set of all contexts (finite or infinite) is a complete tree, also called a *context tree*. For example, i.i.d. processes have a context tree reduced to the set $\{\lambda\}$, and generic Markov chains of order 1 have a context tree equal to A.

Let $d \leq n$ be positive integers. In what follows, we assume that we observe $X_{-d+1}, \ldots, X_0, X_1, \ldots X_n$ distributed from a stationary ergodic process with law $\mathbb{P}$ whose (possibly infinite) context tree is denoted by T_0. We consider a set $\mathscr{T}$ of candidate trees of depth at most d (that may depend on $X_{-d+1}, \ldots X_n$). The goal is to select a tree $T \in \mathscr{T}$ as close as possible from T_0, in some sense that will be formally given below. Several choices are possible, depending on the problem. One may want to choose for $\mathscr{T}$ the set of all complete trees; but in some applications, there are structural restrictions that induce a more restrictive choice of $\mathscr{T}$. For example, in stochastic modelling of natural languages / codified written text, grammatical rules impose constraints on the structure of the possible trees, see [8]. Another popular option is to impose a constraint on the number of

occurrences of their contexts in the sample X_1^n. The only hypothese we make on $\mathscr{T}$ is "closed under inclusion", that is :

$$T \in \mathscr{T} \text{ and } T \succeq T' \implies T' \in \mathscr{T} .$$

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 to be observed, similar results would be obtained by using quasi-maximum likelihood estimators [11].

We denote by $V_{\mathscr{T}} = \cup_{T \in \mathscr{T}} T$ the set of all candidate contexts. Note that $V_{\mathscr{T}}$ is suffix-closed:

$$w \in V_{\mathscr{T}} \text{ and } w \succeq v \implies v \in V_{\mathscr{T}} .$$

Given a sequence $w \in V_{\mathscr{T}}$ and a 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:

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

Besides, $N_n(w)$ will denote the sum $\sum_{a \in A} N_n(w, a)$. Remark that $N_n(w, a)$ may differ from $N_n(wa)$ by at most 1, if $X_{n-|w|}^n = wa$; this little notational trick will simplify the discussion in the sequel.

Given a tree $T \in \mathscr{T}$, the maximum likelihood of the sequence $X_1, \ldots, X_n$ is given by

$$\hat{\mathbb{P}}_{\mathrm{ML},T}(X_1^n) = \prod_{w \in T} \prod_{a \in A} \hat{p}_n(a|w)^{N_n(w,a)}, \tag{2}$$

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

$$\hat{p}_n(a|w) = \frac{N_n(w, a)}{N_n(w)} \tag{3}$$

if $N_n(w) > 0$ and $\hat{p}_n(a|w) = 1/|A|$ otherwise. For any sequence w 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 shall most often use the *Küllback-Leibler divergence*, defined for two probability measures p and q on A by

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

In the following subsections we introduce the algorithm Context and the penalized maximum likelihood criterion. Both context tree estimation algorithms rely on a procedure CST computing the (compact) *suffix tree* of a sequence X_1^n. This is a tree whose edges are labeled by sequences, and such that each suffix of X_1^n corresponds to exactly one path from the root of the tree to a leaf. Efficient implementations of the CST procedure only require linear time (see [12, 19, and references therein]).

2.1. Algorithm Context

For all sequences $w \in V_{\mathscr{T}}$ let

$$\Delta_n(w) = \sum_{b\colon bw \in V_{\mathscr{T}}} N_n(bw) D\left(\hat{p}_n(\cdot|bw); \hat{p}_n(\cdot|w)\right).$$

Algorithm Context depends on a non-decreasing sequence of real numbers $(\delta_n)_n$ (convenient choices of $(\delta_n)_n$ will be deduced from Theorem 1 and 2 below). For a sequence X_1^n, let $\mathcal{C}_w(X_1^n)$ be defined for all $w \in V_{\mathscr{T}}$ by the following induction:

$$\mathcal{C}_w(X_1^n) = \begin{cases} 0, & \text{if } N_n(w) \leq 1, \\ \max\{\mathbb{1}\{\Delta_n(w) \geq \delta_n\}, \max_{b\in A} \mathcal{C}_{bw}(X_1^n)\}, & \text{if } N_n(w) > 1. \end{cases}$$

The Context tree estimator $\hat{T}_C$ is the set of all $w \in V_{\mathscr{T}}$ such that $\mathcal{C}_w(X_1^n) = 0$ and $\mathcal{C}_u(X_1^n) = 1$ for all u such that $u \prec w$. It can be computed as the result of $PRUNE(T^c)$, where $T^c = CST(X_1^n)$ and procedure PRUNE is described in Algorithm 1. The algorithm uses a subfunction *ROOT* which returns the root of a tree. Besides, we denote by T_a the subtree of T rooted in the node that shares an edge labelled by a with the root of T.

Algorithm 1 PRUNE(T)

```
if |T| = 1 then
  return T
end if
isLeaf ← true
for all a ∈ A do
  P = PRUNE(T_a)
  if |P| > 1 then
    isLeaf ← false
  end if
  S_a ← P
end for
w ← ROOT(T)
if isLeaf AND Δ_n(w) < δ_n then
  S ← w
end if
return S
```

Remark 1. *We present here the original choice of divergence by Rissanen in $\Delta_n(w)$, but other equivalent possibilities have been proposed in the literature (see for instance [10]). For example, $\tilde{\Delta}_n(w) = \max_{a,b} |\hat{p}_n(a|bw) - \hat{p}_n(a|w)|$ or $\max_b D(\hat{p}_n(\cdot|bw); \hat{p}_n(\cdot|w))$.*

2.2. The penalized maximum likelihood criterion

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

$$\hat{T}_{PML}(X_1^n) = \arg\max_{T\in\mathscr{T}} \left\{ \log \hat{\mathbb{P}}_{\mathrm{ML},T}(X_1^n) - \mathrm{pen}(n, |T|) \right\},$$

where $\mathrm{pen}(n,|T|)$ is some penalty function. In the sequel, we assume that the penalty of a model is proportional to its dimension; with this very standard choice, we write $\mathrm{pen}(n,|T|) = |T| f(n)$ for some positive function f.

It may first appear practically impossible to compute $\hat{T}_{PML}(X_1^n)$ [3]. Fortunately, Csiszár and Talata showed in their article [5] how to adapt the Context Tree Maximizing (CTM) method [20] to obtain a simple and efficient algorithm computing $\hat{T}_{PML}(X_1^n)$. For self-containment, we briefly present it here: define recursively, for any $w \in V_{\mathscr{T}}$, the value

$$V_w(X_1^n) = \max\{e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n), \prod_{a\in A:\ aw\in V_{\mathscr{T}}} V_{aw}(X_1^n)\}$$

and the indicator

$$\mathcal{X}_w(X_1^n) = \mathbb{1}\{\prod_{a\in A:\ aw\in V_{\mathscr{T}}} V_{aw}(X_1^n) > e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\}\,.$$

By convention, if $\{a\in A\colon aw\in V_{\mathscr{T}}\} = \emptyset$ or if $N_n(w) = 0$, then $V_w(X_1^n) = e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)$ and $\mathcal{X}_w(X_1^n) = 0$. It is shown in [5] that $\hat{T}_{PML}(X_1^n)$ is exactly the set of all words $w \in V_{\mathscr{T}}$ such that $\mathcal{X}_w(X_1^n) = 0$ and $\mathcal{X}_u(X_1^n) = 1$ for all $u \prec w$. Thus, $\hat{T}_{PML}(X_1^n)$ can be computed as $CTM(T^c)$, where $T^c = CST(X_1^n)$ and algorithm CTM is described in Algorithm 2.

Algorithm 2 CTM(T)

```
if |T| = 1 then
  return [T, e^{-f(n)}]
end if
for all a ∈ A do
  [P, V_a] = CTM(T_a)
  S_a ← P
end for
w ← ROOT(T)
if ∏_{a∈A} V_a < e^{-f(n)} P̂_{ML,w}(X_1^n) then
  S ← w
  V ← e^{-f(n)} P̂_{ML,w}(X_1^n)
else
  V ← ∏_{a∈A} V_a
end if
return [S, V]
```

2.3. Results

In this subsection we present the main results of this article. First, we show that there is a close relationship between the estimators $\hat{T}_{PML}(X_1^n)$ and $\hat{T}_C(X_1^n)$.

Proposition 1. *For all sequences X_1^n, if $\delta_n \le f(n)$ then*

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

We now state a new bound on the probability of over-estimation by the Context algorithm.

Theorem 1. *For every positive integer n and for all δ_n it holds that*

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

An immediate consequence of Proposition 1 and Theorem 1 is that the Penalized Maximum Likelihood estimator also avoids over-estimation with overwhelming probability provided that δ_n is chosen appropriately:

Corollary 1. *If $\delta_n \leq f(n)$ then*

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

Remark 2. *Following [17], it is sometimes proposed to consider only a limited number $k(n)$ of nodes. A straightforward modification of the proof shows that*

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

and similarly if the Penalized Maximum Likelihood criterion is minimized only over trees containing no more than $k(n)$ nodes fixed a priori, it holds that

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

In particular, penalties of order $O(\log\log n)$, smaller than the BIC prove to be sufficient to avoid over-estimation if an upper-bound (independent of n) on the depth of the tree is known (as was proved by Finesso for Markov chains, see [7] and references therein). More generally, given a function $k(n)$, one can easily choose δ_n so as to ensure consistency.

The problem of underestimation in context tree models is very different, and requires additional hypotheses on the process $\{X_t \colon t \in \mathbb{Z}\}$. Define the *continuity coefficients* $\{\alpha_k\}_{k\in\mathbb{N}}$ of the process by

$$\alpha_0 := \inf_{x_{-\infty}^{-1}, a\in A} \{\, p(a|x_{-\infty}^{-1}) \,\},$$
$$\alpha_k := \inf_{u\in A^k} \sum_{a\in A} \inf_{x_{-\infty}^{-1}} p(a|x_{-\infty}^{-1}u)\,, \qquad k \geq 1\,.$$

In the sequel, we make the following assumptions.

Assumption 1. *The process $\{X_t \colon t \in \mathbb{Z}\}$ is such that $\alpha_0 > 0$ and*

$$\alpha := \sum_{k\in\mathbb{N}} (1 - \alpha_k) \;<\; +\infty$$

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. As the analysis leaves some choice in the determination of the candidate trees $\mathscr{T}$, we need the following "separation" assumption over $\mathscr{T}$:

Assumption 2. *There exists $\epsilon > 0$ such that for any $u \in T_0|_K$ and any $w \prec u$, there exists $v \in V_{\mathscr{T}}$ satisfying*

$$p(v)\, D\left(p(\cdot|v); p(\cdot|w)\right) > \epsilon\,.$$

Remark 3. *Assumption 2 says that the set of candidate trees $\mathscr{T}$ is rich enough to ensure that if w is not a context of $T_0|_K$, then it can be seen within $V_{\mathscr{T}}$. [5] and [14] give sufficient conditions for Assumption 2 to hold with high probability.*

Theorem 2. *Assume $f(n)$ is such that $f(n)/n \to \infty$ when $n \to \infty$. Then for any $K \in \mathbb{N}$, under Assumptions 1 and 2 there exist constants c_1, c_2 and n_0 such that*

$$\mathbb{P}\big(T_0|_K \preceq \hat{T}_{PML}(X_1^n)\big) \geq 1 - 3e^{\frac{1}{e}}|A|^{2+K}c_1 \exp\left[\frac{-c_2\, n}{64e|A|^3(\alpha+\alpha_0)\log^2\alpha_0}\right].$$

Corollary 2. *For any $K \in \mathbb{N}$, under Assumptions 1 and 2 and if $\delta_n \leq f(n)$, we have*

$$\mathbb{P}\left(T_0|_K \preceq \hat{T}_C(X_1^n)\right) \geq 1 - 3e^{\frac{1}{e}}|A|^{2+K}c_1 \exp\left[\frac{-c_2\, n}{64e|A|^3(\alpha+\alpha_0)\log^2\alpha_0}\right].$$

Remark 4. *Extensions of Theorem 2 can 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 decreases with the sample size. This leads to a rather technical condition, see for instance [18].*

3. Proofs

3.1. Proof of Proposition 1

It is sufficient to prove that $\mathcal{X}_w(X_1^n) \leq \mathcal{C}_w(X_1^n)$ for all $w \in V_{\mathscr{T}}$. Assume that there exists $w \in V_{\mathscr{T}}$ such that $\mathcal{X}_w(X_1^n) = 1$ and $\mathcal{C}_w(X_1^n) = 0$. Note that $\mathcal{C}_w(X_1^n) = 0$ implies $\mathcal{C}_{uw}(X_1^n) = 0$ for all $uw \in V_{\mathscr{T}}$; hence, w can be chosen such that $\mathcal{X}_{bw}(X_1^n) = 0$ for any $bw \in V_{\mathscr{T}}$, $b \in A$.

In this case, we have by definition:

$$\prod_{b\colon bw \in V_{\mathscr{T}}} e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},bw}(X_1^n) = \prod_{b\colon bw \in V_{\mathscr{T}}} V_{bw}(X_1^n) \;>\; e^{-f(n)}\hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n).$$

Taking logarithm,

$$\sum_{b\colon bw \in V_{\mathscr{T}}} \sum_{a \in A} N_n(bwa) \log \frac{\hat{p}_n(a|bw)}{\hat{p}_n(a|w)} \;>\; (|\{b\colon bw \in V_{\mathscr{T}}\}| - 1)f(n)$$

and

$$\Delta_n(w) = \sum_{b\colon bw \in V_{\mathscr{T}}} N_n(bw) D\left(\hat{p}_n(\cdot|bw); \hat{p}_n(\cdot|w)\right) > (|\{b\colon bw \in V_{\mathscr{T}}\}| - 1)f(n) \geq \delta_n$$

because as $\mathcal{X}_w(X_1^n) = 1$, we have $|\{b\colon bw \in V_{\mathscr{T}}\}| \geq 2$. This contradicts the fact that $\mathcal{C}_w(X_1^n) = 0$, and concludes the proof.

3.2. Proof of Theorem 1

Let O_n be the event $\left\{\hat{T}_C(X_1^n) \not\preceq T_0\right\}$. Overestimation 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) word u such that $w = us$; thus, O_n can be decomposed as:

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

By definition, for all $b \in A$ it holds that $p(\cdot|w) = p(\cdot|bw)$, and thus:

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

Hence,

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

Using Theorem 3, it follows that

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

where C_n denotes the number of different contexts appearing in X_1^n. But C_n is almost-surely upper-bounded by $n(n-1)/2$, and the result follows.

3.3. Proof of Theorem 2

The proof of the Theorem is analogous to that obtained in [14], we include it here for completeness. If U_n denotes the event $\{T_0|_K \not\preceq \hat{T}_{PML}(X_1^n)\}$ then

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

Let $w \prec u \in T_0|_K$ such that $\mathcal{X}_w(X_1^n) = 0$. By Assumption 2 there must exists a tree $T \in \mathscr{T}$ such that

$$\sum_{v \in T_w} p(v) D\left(p(\cdot|v); p(\cdot|w)\right) \;\geq\; \epsilon\,. \tag{4}$$

Then we have

$$\mathbb{P}\left(\,\mathcal{X}_w(X_1^n) = 0\,\right) \;=\; \mathbb{P}\left(\prod_{a \in A\colon\, aw \in V_{\mathscr{T}}} V_{aw}(X_1^n) \;\leq\; e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\right).$$

By definition, for any $aw \in V_{\mathscr{T}}$it can be shown recursively that

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

see for example Lemma 4.4 in [5]. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\prod_{a \in A\colon\, aw \in V_{\mathscr{T}}} V_{aw}(X_1^n) \;\leq\; e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\right) \\
&\qquad\leq\; \mathbb{P}\left(\prod_{u \in T_w} e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},u}(X_1^n) \;\leq\; e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n)\right) \\
&\qquad=\; \mathbb{P}\left(\sum_{u \in T_w} \log \hat{\mathbb{P}}_{\mathrm{ML},u}(X_1^n) - \log \hat{\mathbb{P}}_{\mathrm{ML},w}(X_1^n) \leq (|T_w| - 1) f(n)\right),
\end{aligned}$$

by noticing that

$$\prod_{a \in A\colon\, aw \in V_{\mathscr{T}}} \prod_{s \in T_{aw}} e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},s}(X_1^n) = \prod_{u \in T_w} e^{-f(n)} \hat{\mathbb{P}}_{\mathrm{ML},u}(X_1^n)\,.$$

Dividing by n, subtracting $\sum_{u \in T_w} p(u) D\left(p(\cdot|u); p(\cdot|w)\right)$ on both sides and using the inequality in (4) we can write the last probability as

$$\mathbb{P}\left(\sum_{u \in T_w} L_n(u) - L_n(w) \;\leq\; \frac{(|T_w| - 1) f(n)}{n} - \epsilon\right)$$

where for any finite sequence s

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

Then, for all $n \geq n_0$, where n_0 is a sufficiently large integer such that

$$\frac{|T_w| f(n)}{n} \;<\; \frac{\epsilon}{2}$$

we can bound above the last expression by

$$\mathbb{P}\left(\,|L_n(w)| > \frac{\epsilon}{4}\right) \;+\; \sum_{u \in T_w} \mathbb{P}\left(\,|L_n(u)| > \frac{\epsilon}{4|T_w|}\right).$$

Using Corollary A.7(c) in [14] we obtain

$$\mathbb{P}\left(\,\mathcal{X}_w(X_1^n) = 0\,\right) \;\leq\; 3e^{\frac{1}{e}} |A|^2 (1 + |T_w|) \exp\left(-\frac{n \min(1, (\epsilon/4|T_w|)^2) \alpha_0^{2(h(T_w)+1)}}{64e|A|^3(\alpha + \alpha_0) \log^2 \alpha_0 \left(h(T_w) + 1\right)}\right),$$

by noticing that $p(w) \geq p(u) \geq \alpha_0^{h(T_w)}$ for any $u \in T_w$. We conclude the proof of Theorem 2 by observing that we only have a finite number of sequences $w \prec u \in T_0|_K$, therefore we obtain

$$\mathbb{P}\left(T_0|_K \preceq \hat{T}_{PML}(X_1^n)\right) \geq 1 - 3e^{\frac{1}{e}}|A|^{2+K}c_1 \exp\left(\frac{-c_2\, n}{64e|A|^3(\alpha+\alpha_0)\log^2\alpha_0}\right),$$

where

$$c_1 = \max_{w\prec u\in T_0|_K}(1+|T_w|) \quad \text{and} \quad c_2 = \min_{w\prec u\in T_0|_K}\left\{\frac{\min(1,(\epsilon/4|T_w|)^2)\alpha_0^{2(h(T_w)+1)}}{h(T_w)+1}\right\}.$$

4. Discussion

Algorithm Context has many variants: in fact, $\Delta_n(w)$ can rely on different (pseudo-) distances rather than on Kullback-Leibler information. Another popular choice (see [10]) is

$$\Delta_n(w) = \max_{b,c\in A}\left|\hat{p}_n(c|bw) - \hat{p}_n(c|w)\right|.$$

Similar results can be obtained in this case by combining Theorem 3 and Pinsker's inequality [15]:

$$\left(\sup_{B\subset A}\hat{P}_n(A|a_1^k) - P(A|a_1^k)\right)^2 \leq \frac{1}{2}D(\hat{P}_n;P).$$

Then,

$$\begin{aligned}
\mathbb{P}\left[\max_{b\in A}\left|\hat{p}_n(b|a_1^k) - P(b|a_1^k)\right| > \sqrt{\frac{\delta}{N_n}}\right] &\leq \mathbb{P}\left[\sup_{B\subset A}\left(\hat{p}_n(B|a_1^k) - P(B|a_1^k)\right)^2 > \frac{\delta}{N_n}\right] \\
&\leq \mathbb{P}\left[N_n D\left(\hat{p}_n\left(\cdot|a_1^k\right);P\left(\cdot|a_1^k\right)\right) > 2\delta\right] \\
&\leq 2e\left(2\delta\log(n)+|A|\right)\exp\left(-\frac{2\delta}{|A|}\right).
\end{aligned}$$

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 in unavoidable. All what can be expected from the estimator it to highlight evidence of as much dependence structure as possible, while maintaining a limited probability of false discovery. Let us insist once again that no mixing assumption is necessary for Theorem 1. On the other hand, for under-estimation, some conditions like Assumption 1 seem unavoidable.

Appendix 1: Martingale deviation inequalities

This section contains the statement and derivation of two deviation inequalities that are useful to prove the main results of this paper. As we believe they are interesting on their own, we include them in a separate section. The originality is that deviations from expectations are not controlled in L^p norm, but thru the Kullback-Leibler divergence; 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 the 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 [13]. As for [4], the ingredients of the proofs are mostly inspired by [16]. Completely different applications of these inequalities may be found in [13, 6].

Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary process of whose (possibly infinite) context tree is T, and let $\mathcal{F}_n$ be the σ-field generated by $(X_j)_{j\leq n}$. For $k\in\mathbb{N}$, $a_1^k\in A^k$, denote $p(b|a_1^k)=\mathbb{P}\left(X_{k+1}=b|X_1^k=a_1^k\right)$. Define

$$\xi_j=\mathbb{1}\{X_{j-k}^{j-1}=a_1^k\},\quad \text{for } j\geq 1,$$

$$\chi_j=\mathbb{1}\{X_{j-k}^{j}=a_1^k b\},\quad \text{for } j\geq 1$$

so that $N_n(a_1^k)=\sum_{j=1}^n\xi_j$ and $N_n(a_1^k,b)=\sum_{j=1}^n\chi_j$.

Denote $\hat{p}_n(b|a_1^k)=S_n(a_1^kb)/N_n(a_1^k)$. From now on, a block a_1^k having a suffix in T_0 and a symbol b are fixed, and there will be no possible confusion in using the shortcuts $p=P(b|a_1^k)$, $N_n=N_n(a_1^k)$, $S_n=N_n(A_1^k,b)$ and $\hat{p}_n=\hat{P}_n(b|a_1^k)=\frac{S_n}{N_n}$. Besides, 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 2.

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

To prove Proposition 2, we introduce a few notations: for every $\lambda>0$, let

$$\phi_p(\lambda)=\log\mathbb{E}\left[\exp\left(\lambda X_1\right)\right]=\log\left(1-p+p\exp\left(\lambda\right)\right).$$

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

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

To proceed, we make use of the so-called 'peeling trick' [15]: we divide the interval $\{1,\dots,n\}$ of possible values for N_n into "slices" $\{t_{k-1}+1,\dots,t_k\}$ 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 $t_k = \lfloor (1+\eta)^k \rfloor$. Let D be the first integer such that $t_D \geq n$, that is $D = \left\lceil \frac{\log n}{\log 1+\eta} \right\rceil$. Let $A_k = \{t_{k-1} < N_n \leq t_k\} \cap \{N_n d(\hat{p}_n; p) > \delta\}$. We have:

$$\mathbb{P}\left(N_n d\left(\hat{p}_n; p\right) > \delta\right) \leq \mathbb{P}\left(\bigcup_{k=1}^{D} A_k\right) \leq \sum_{k=1}^{D} \mathbb{P}\left(A_k\right). \tag{5}$$

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, $\mathbb{P}\left(A_k\right) = 0$ 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 :
$$\hat{p}_n \geq p \text{ and } d(\hat{p}_n;p) > \frac{\delta}{N_n} \implies \hat{p}_n \geq z.$$

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

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

As $\left(W_t^{\lambda}\right)_{t\geq 0}$ 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:

$$\begin{aligned}
\mathbb{P}\left(\left\{\tilde{t}_{k-1} < N_n \leq t_k\right\} \cap \{\hat{p}_n \geq p\} \cap \{N_n d(\hat{p}_n, p) \geq \delta\}\right) &\leq \mathbb{P}\left(W_n^{\lambda(z)} > \exp\left(\frac{\delta}{1+\eta}\right)\right) \\
&\leq \exp\left(-\frac{\delta}{1+\eta}\right).
\end{aligned} \tag{6}$$

Similarly,

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

so that

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

Finally, by Equation (5),

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

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

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

Remark 5. *The bound of Proposition 2 also holds for $\mathbb{P}\left(N_n d(\hat{p}_n, p) \geq \delta | N_n > 0\right)$: 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 Equation 6, the proof can be rewritten conditionally on $\{N_n > 0\}$; this leads to:

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

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 1. *For any probability distributions P and Q on an alphabet X,*

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

Proof.

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

because

$$\sum_{x\in X}\frac{1-P(x)}{|X|-1} = \sum_{x\in X}\frac{1-Q(x)}{|X|-1} = 1.$$

□

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

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

Theorem 3.

$$\mathbb{P}\left[N_n(a_1^k)D\left(\hat{p}_n\left(\cdot|a_1^k\right);p\left(\cdot|a_1^k\right)\right)>\delta\right]\leq 2e\left(\delta\log(n)+|A|\right)\exp\left(-\frac{\delta}{|A|}\right).$$

Proof. By combining Lemma 1 and Proposition 2, we get

$$\begin{aligned}
\mathbb{P}\left[N_n(a_1^k)D\left(\hat{p}_n\left(\cdot|a_1^k\right);p\left(\cdot|a_1^k\right)\right)>\delta\right] &\leq \mathbb{P}\left[\sum_{b\in A}N_n d\left(\hat{p}_n\left(b|a_1^k\right);p\left(b|a_1^k\right)\right)>\delta\right]\\
&\leq \sum_{b\in A}\mathbb{P}\left[N_n d\left(\hat{p}_n\left(b|a_1^k\right);p\left(b|a_1^k\right)\right)>\frac{\delta}{|A|}\right]\\
&\leq (|A|)2e\left[\frac{\delta}{|A|}\log(n)\right]\exp\left(-\frac{\delta}{|A|}\right)\\
&\leq (|A|)2e\left(\frac{\delta}{|A|}\log(n)+1\right)\exp\left(-\frac{\delta}{|A|}\right)\\
&= 2e\left(\delta\log(n)+|A|\right)\exp\left(-\frac{\delta}{|A|}\right).
\end{aligned}$$

□

Remark 7. *It follows from Remark 5 that the following variant of Theorem 3 holds:*

$$\mathbb{P}\left[N_n(a_1^k)D\left(\hat{p}_n\left(\cdot|a_1^k\right);p\left(\cdot|a_1^k\right)\right)>\delta|N_n>0\right]\leq 2e\left(\delta\log(n)+|A|-1\right)\exp\left(-\frac{\delta}{|A|-1}\right).$$

Acknowledgments

This work was supported by Fapesp (process 2009/09411-8) and USP-COFECUB project. F.L. also thanks CNPq for the support (grant 302162/2009-7).

References

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

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

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

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

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

[6] Sarah Filippi, Olivier Cappé, and Aurelien Garivier. Optimism in reinforcement learning based on kullback-leibler divergence. *arXiv:1004.5229v2*, abs/1004.5229, 2010.

[7] L. Finesso, C-C. Liu, and P. Narayan. The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory*, 42(5):1488–1497, 1996.

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

[9] A. Galves and F. Leonardi. *Exponential inequalities for empirical unbounded context trees*, volume 60 of *Progress in Probability*, pages 257–270. Birkhauser, 2008.

[10] A. Galves, V. Maume-Deschamps, and B. Schmitt. Exponential inequalities for VLMC empirical trees. *ESAIM Probab. Stat*, 12:43–45, 2008.

[11] Antonio Galves, Aurélien Garivier, and Elisabeth Gassiat. Data selection of context trees and classification. Technical Report, 2010.

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

[13] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems, 2008.

[14] F. Leonardi. Some upper bounds for the rate of convergence of penalized likelihood context tree estimators. *Brazilian Journal of Probability and Statistics (to appear)*, 24(2):321–336, 2010.

[15] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. 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.

[16] J. Neveu. *Martingales temps discret*. Masson, 1972.

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

[18] Z. Talata and T. Duncan. Unrestricted BIC context tree estimation for not necessarily finite memory processes. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 724–728, 28 2009-July 3 2009.

[19] E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260, 1995.

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