Asymptotically tight bounds on the depth of estimated context trees

Álvaro Martín
Universidad de la República
Montevideo, Uruguay
Email: almartin@fing.edu.uy

Gadiel Seroussi
Universidad de la República
Montevideo, Uruguay
Email: gseroussi@ieee.org

Abstract—We study the maximum depth of context tree estimates, i.e., the maximum Markov order attainable by an estimated tree model given an input sequence of length $n$. We consider two classes of estimators: 1) Penalized maximum likelihood (PML) estimators where a context tree $T$ is obtained by minimizing a cost of the form $-\log P_T(x^n) + f(n)|S_T|$, where $P_T(x^n)$ is the ML probability of the input sequence $x^n$ under a context tree $T$, $|S_T|$ is the set of states defined by $T$, and $f(n)$ is an increasing (penalization) function of $n$ (the popular BIC estimator corresponds to $f(n) = \frac{2}{\alpha+1} \log n$, where $\alpha$ is the size of the input alphabet). 2) MDL estimators based on the KT probability assignment. In each case we derive an asymptotic upper bound, $n^{1/2+o(1)}$, and we exhibit explicit input sequences that show that this bound is asymptotically tight up to the term $o(1)$ in the exponent.

I. INTRODUCTION

The range of values for different kinds of Markov order estimates is generally important for the theoretical analysis of the corresponding estimators. For example, the Bayesian Information Criterion (BIC) [1] Markov order estimator, is now known to be strongly consistent even if no restriction is imposed on the set of candidate models [2], although the first consistency results required a known bound on the Markov order [3]. On the other hand, a Minimum Description Length (MDL) [4] Markov order estimator based on the Kricevski-Trofimov (KT) probability assignment [5] is inconsistent if the order estimate is allowed to exceed $C \log n$ for sample sequences of length $n$, where $C$ is a sufficiently large constant [2]. This estimator is strongly consistent if an upper bound $k_n = o(\log n)$ is imposed on the order of candidate models [6].

In addition to the theoretical interest, from a practical computational perspective, the knowledge of an upper bound on the possible outcome of a Markov order estimator can yield significant memory savings in the implementation of such estimators [7].

In this paper we study context tree estimates for tree models [8], in which the conditional probability distribution for the $i$-th symbol of a sequence given all past symbols depends on a finite number of the most recent observations, $x_{i-t}, \ldots, x_{i-1}$. The number $t$ of past symbols that suffice to determine this conditional probability distribution depends on the past symbols themselves and a so-called context tree, whose depth defines the maximum value that $t$ can have, i.e., the Markov order of the model. The consistency of both BIC and KT-based MDL estimators for these models was proved in [9], assuming an upper bound of $o(\log n)$ on the depth of the candidate context trees. The BIC estimator for context trees is proved in [10] to be consistent even if no restriction is imposed on the depth of the candidate models.

In Section II we define tree models and their penalized maximum likelihood (PML) and KT-based MDL estimators (BIC is a special case of PML). In Section III we present our main results, namely, an asymptotic upper bound on the depth of estimated trees for each of these two kinds of estimators, and we exhibit explicit input sequences that show that these bounds are asymptotically tight up to negligible terms in the exponent.

II. CONTEXT TREE ESTIMATES

Let $A$ be a finite alphabet of size $\alpha$, $\alpha \geq 2$. We denote by $A^*$, $A^+$, and $A^n$ the set of finite strings, positive-length strings, and length-$n$ strings over $A$, respectively. The terms string and sequence are used interchangeably. We use the notation $|\cdot|$ for both set cardinality and string length. For a string $u$ we denote by $u^j$ the prefix of length $j$ of $u$, $j \leq |u|$. We let $uv$ denote the concatenation of strings $u$ and $v$, and $\lambda$ denote the empty string.

A context tree $T$ is a rooted full $\alpha$-ary tree, where each of the $\alpha$ edges departing from an internal node is labeled with a different symbol from $A$. We identify a node of $T$ with the concatenation of symbols in the path that ascends from the node to the root. Furthermore, we identify $T$ with the set of its nodes and write, for example, $u \in T$ when the string $u$ is a node of $T$. We refer to the nodes of $T$ as states and, for $x \in A^*$, we say that $x$ selects the state $s$ defined as the deepest node of $T$ such that $s$ is a suffix of $x$.

A tree model, denoted $(T, p_T)$, is comprised of a context tree, $T$, and a model parameter, $p_T$, which defines a conditional probability distribution over $A$ for each state $s$ of $T$, denoted $p_T(\cdot|s)$. This model defines a probability assignment [11] $P_{(T,p_T)}(\cdot)$ given by

$$P_{(T,p_T)}(\cdot) = 1; \quad P_{(T,p_T)}(x) = \prod_{j=1}^{|x|} p_T(x_j|s_{i-1}), \quad x \in A^+, \quad (1)$$

where $s_{i-1}$ denotes the state selected by $x^{i-1}$. For $s = s_{i-1}$ we say that $x_i$ is emitted in state $s$ and that $s$ occurs (at
position $i-1$ in $x$. Notice that each internal node of $T$ occurs at most once in $x$. The internal nodes, which we refer to as transient states [12], play the role of assigning a probability to the first symbols of $x$, similar to the definition of a probability distribution for the initial state in classical Markov chains. The set of permanent states of $T$ [12], denoted $S_T$, is defined as the set of leaves of $T$. The conditional probability distributions associated to permanent states determine the long run statistical properties of the model. Our definitions of permanent and transient states resemble those of recurrent and non-recurrent states in the theory of Markov chains [13]. We notice, however, that the latter two depend on the specific conditional probability distributions associated to the states, which may render a permanent state non-recurrent.

We denote by $n_s$ the number of occurrences of $s$ in $x$, and by $n_s(a)$ the number of times a symbol $x_i = a \in A$ is emitted in state $s$. We omit the dependence on $x$ and $T$ of $n_s(n), n_s$, and other notation; they will be clear from the context.

The empirical probability distribution of symbols conditioned on state $s$ is the distribution $\hat{p}_s$ defined as $\hat{p}_s(a) = n_s(a)/n_s$, $a \in A$. The maximum likelihood (ML) probability of $x$ with respect to a context tree $T$, denoted $\hat{P}_T(x)$, is determined by the tree model $\langle T, \hat{p}_T \rangle$, where $\hat{p}_T(\cdot|s) = \hat{p}_s$ for all states $s$ of $T$, so that

$$-\log \hat{P}_T(x) = \sum_{s \in T, a \in A} n_s(a) \log \frac{n_s(a)}{n_s},$$

(2)

with the convention $n_s(a) \log \frac{n_s(a)}{n_s} = 0$ if $n_s(a) = 0$.

We consider penalized maximum likelihood (PML) estimators of the context tree $T$ from a sample sequence emitted by an unknown model $\langle T, p_T \rangle$. Specifically, for $x \in A^n$, we assign a cost

$$C_T(x) = -\log \hat{P}_T(x) + f(n)|S_T|$$

(3)

to each candidate context tree $T$, where $f(n)$ is a positive and nondecreasing penalization function, with $f(n) \rightarrow \infty$ and $\frac{f(n)}{n} \rightarrow 0$. The estimate $\hat{T}(x)$ of $T$ is obtained by picking a minimum-cost context tree,

$$\hat{T}(x) = \arg\min_T C_T(x).$$

(4)

When $f(n) = \frac{1}{2}(\alpha-1) \log n$, we obtain the BIC estimator [1].

We also consider MDL context tree estimators based on the KT probability assignment. The KT probability assigned to the symbols that occur in state $s$ of a context tree $T$ is

$$KT_s = \prod_{a \in A} \frac{n_s^{(a)} - 1}{n_s^{(a)} - \frac{1}{2}} \left( i + \frac{1}{2} \right),$$

(5)

and the KT probability assigned to $x$ by a context tree $T$ is

$$KT_T(x) = \prod_{u \in T} KT_u(x).$$

(6)

Based on this probability assignment, it is possible to construct a lossless encoding of $x$ of length

$$C_{KT,T} = -\log KT_T + \tau(T),$$

(7)

where $\tau(T)$ is the length of an encoding of $T$ and we ignore integer constraints on code lengths. A context tree MDL estimate based on this encoding is defined as

$$\hat{T}_{KT} = \arg\min_T C_{KT,T}.$$ 

(8)

We assume that there exist positive constants $\xi^-, \xi^+$ such that whenever $T \subset T'$ for context trees $T$ and $T'$, $\tau(\cdot)$ satisfies

$$\xi^- (|T'| - |T|) \leq \tau(T') - \tau(T) \leq \xi^+ (|T'| - |T|).$$

(9)

The natural code [14] used in [12], for example, encodes a context tree $T$ using 1 bit per node and, thus, we have $\xi^- = \xi^+ = 1$ in this case.

### III. MAIN RESULTS

For PML Markov order estimators, an upper bound on the estimated order is derived in [15]. An analogous reasoning yields an upper bound on the number of permanent states in any tree $\hat{T}$, which in turn implies an upper bound of order $O(\frac{n}{\log n})$ on the depth, $d$, of $\hat{T}$. However, this bound is rather loose for general trees $\hat{T}$. The following theorem establishes an asymptotically tight bound.

**Theorem 1.** Let $\hat{d}$ be the depth of a context tree estimated according to (4), and assume $\hat{d} > 2$. If $n$ is sufficiently large we must have

$$\frac{(\hat{d} - 1)^2}{\log \hat{d} - 1} \leq \frac{4n}{f(n)}.$$ 

(10)

The following theorem upper-bounds the depth of a MDL context tree estimate based on the code length function defined in (7), satisfying (9) for some constant $\xi^-$.  

**Theorem 2.** Let $\hat{d}_{KT}$ be the depth of a context tree estimated according to (8) and assume $\hat{d}_{KT} \geq \frac{2(\alpha \xi^- - 1)}{\alpha \xi^-}$. If $n$ is sufficiently large we have

$$\frac{C_{KT}^2}{\log \hat{d}_{KT}} \leq n,$$

(11)

where $C$ is any constant larger than $\left( \frac{\alpha \xi^-}{\alpha \xi^- + 1} \right)^2$.

Notice that, in contrast to plain Markov order estimators where MDL estimates can be much larger than BIC estimates [16], Theorems 1 and 2 imply an upper bound $n^{1/2 + o(1)}$ on both $\hat{d}$ and $\hat{d}_{KT}$. Notice also that the maximum value for the depth of a MDL context tree estimate is much larger than the $o(\log n)$ bound for which consistency is guaranteed [9]; it can be shown that imposing such bound in a source coding application may incur asymptotically significant compression rate degradation for some individual input sequences [16].

The following theorems establish upper bounds on the length of the shortest sequence that estimates a context tree of a given depth $d$. Let $a, b$ be distinct symbols in $A$ and, for
a nonnegative integer \(i\), denote by \(a^i\) the concatenation of \(i\) copies of the symbol \(a\) \((a^0 = \lambda)\). For a multiple \(n \) of \(d + 1\) we let \(\tilde{x}\) be the concatenation of \(\frac{n}{d+1}\) copies of \(a^{d}b\).

**Theorem 3.** Let \(d\) be a positive integer and let \(n\) be the smallest multiple of \(d + 1\) satisfying

\[
\frac{n}{f(n)} \geq \frac{d(d+1)(\alpha - 1)}{\log ed},
\]  

(12)

where \(e\) is the base of the natural logarithm. For the sequence \(\tilde{x} \in A^n\), the depth of \(\hat{T}(\tilde{x})\) is \(d\).

Notice that the bound in (12) asymptotically matches that established in Theorem 1. For the special case of BIC, Theorem 1 implies that \(\hat{d}\) is \(O(n^{1/2})\), and Theorem 3 that \(\hat{d}\) is \(\Omega(n^{1/2})\) for the sequence \(\tilde{x}\). For the KT-based MDL estimator, by Theorem 4 below, the sequence \(\tilde{x}\) yields an estimated context tree whose depth is \(\Omega(n^{1/2})\), while the upper bound in Theorem 2 is \(n^{1/2+o(1)}\).

**Theorem 4.** Let \(d\) be a positive integer and let \(n\) be the smallest multiple of \(d + 1\) satisfying

\[
n > Dd(d + 1),
\]

(13)

where \(D\) is any constant larger than \(\frac{\alpha - 1}{2}\). If \(\hat{d}\) is sufficiently large, for the sequence \(\tilde{x} \in A^n\), the depth of \(\hat{T}_{KT}(\tilde{x})\) is \(d\).

**REFERENCES**


