BIC Context Tree Estimation for Stationary Ergodic Processes

Zsolt Talata, Member, IEEE, and Tyrone E. Duncan, Fellow, IEEE

Abstract—Context trees of arbitrary stationary ergodic processes with finite alphabets are considered. Such a process is not necessarily a Markov chain, so the context tree may be of infinite depth. Calculated from a sample of size \( n \), the Bayesian information criterion (BIC) is shown to provide a strongly consistent estimator of the context tree of the process, via minimization over hypothetical context trees, without any restriction on the hypothetical context trees. Strong consistency means that the estimated context tree recovers the true one up to a level \( K \), eventually almost surely as \( n \) tends to infinity. Under some conditions on the process, it is shown that the recovery level \( K \) can grow with \( n \) at a specific rate determined by the distribution of the process; thus, the BIC estimator can recover the true context tree to larger and larger depths. The results include for the special case of \( K \) being an arbitrary constant that the strong consistency is satisfied without any assumption on the stationary ergodic process, which itself improves the existing results, where either the true context tree was assumed to be of finite depth or the depth of the hypothetical context trees was bounded by \( o(\log n) \).

Index Terms—Bayesian information criterion (BIC), consistent estimation, context tree, ergodic processes, infinite memory, model selection, suffix tree.

I. INTRODUCTION

LET \( \{X_i, -\infty < i < +\infty\} \) be a stationary ergodic stochastic process with finite alphabet \( A \). \( |A| \) denotes the cardinality of the set \( A \), and \( \emptyset \) denotes the empty set. A string \( s = a_1a_2\ldots a_m \) (\( a_i \in A \)) is denoted also by \( a_1^m \). If \( k > m \), then \( a_k^m \) is the empty string. Otherwise, the length of the string \( a_k^m \) is \( l(a_k^m) = m - k + 1 \). For a string \( s \in A^k \), let

\[
Q(s) = \text{Prob}\{X_{m+k-1} = s\}
\]

and

\[
Q(a|s) = \begin{cases} \text{Prob}\{X_0 = a | X_{-k} = s\}, & \text{if } Q(s) > 0 \\ 0, & \text{if } Q(s) = 0. \end{cases}
\]

A process as above is referred to as process \( Q \).

The process \( Q \) is of finite memory if the conditional probabilities \( \text{Prob}\{X_0 = a | X_{-\infty} = x_{-k}\} \) depend on at most \( k \)-length endings \( x_{-k}^m \) of the pasts \( x_{-k}^m \), for some fixed finite number \( k \). In this case, the process is a Markov chain of order \( k \), and it is described by the collection of the transition probabilities \( Q(a_{k+1}|a_1^k), a_{k+1}^k \in A^{k+1}; |A| - 1 \) of these and \( |A| \) are free parameters of the Markov chain model.

The context tree model, introduced by Rissanen [12], is a refinement of the Markov chain concept. This model uses only those endings of the infinite past which the conditional probabilities of the symbols actually depend on. The length of these relevant endings can vary with the infinite past, and it can be much shorter than its upper bound, which is the Markov chain order. This is why these models are sometimes called variable-length Markov chains [2]. The relevant endings are called contexts for the process, and the model is also called a tree source [17], [18], [20]. The contexts for the process form a tree graph. The context tree of the process is the minimal tree admitting the tree source representation of the process. The depth of the context tree equals the order of the corresponding Markov chain, but the number of parameters of the model, the number of transition probabilities determined by the contexts, can be much smaller than that of the corresponding Markov chain.

If the stationary ergodic process is not of finite memory, the approach of the context tree model is still valid. In this case, the context tree has infinite depth. In this paper, arbitrary stationary ergodic processes are considered, that is, it is not required to know whether the process is of finite memory or not.

The concatenation of the strings \( u \) and \( v \) is denoted by \( uv \). A string \( v \) is called a suffix of a string \( s \) if there exists a string \( u \) such that \( s = uv \), and it is denoted by \( s \supseteq v \). Let \( s \succ v \) for \( s \neq v \), that is, when \( v \) is a proper suffix of \( s \). A suffix of a semi-infinite sequence \( a_{-\infty} = \cdots a_{-k} \cdots a_{-1} \) is defined similarly.

A context tree \( T \) is a set of strings, and perhaps also of semi-infinite sequences, such that each semi-infinite sequence \( a_{-\infty} \) has a unique suffix \( s \) in \( T \).

A context tree \( T \) can be represented by a tree graph. Each edge of the tree graph is labeled by a symbol from \( A \). The root is identified with the empty string. The nodes of the tree graph are identified with the finite suffixes of all \( u \in T \) such that concatenating the symbols along the path from a node to the root gives the string at the node. Thus, the strings \( s \in T \) are identified with the leaves of the tree graph, and the semi-infinite sequences \( a_{-\infty} \in T \) are infinite paths to the root, see Fig. 1. The children of a node \( s \) are the strings \( as, a \in A \); the node \( s \) is the parent of the children.

Any context tree \( T \) has the properties that no \( s_1 \in T \) is a suffix of any other \( s_2 \in T \), and each node except the leaves has
exactly $|A|$ children. The latter property is often referred to as the fact that the context tree is a complete tree.

**Definition 1.1:** A string $s \in A^k$ is a context for a stationary ergodic process $Q$ if $Q(s) > 0$ and

\[
\text{Prob}\{X_0 = a | X^\infty_0 = x^\infty_0\} = Q(\alpha|s), \quad \text{for all } \alpha \in A
\]

whenever $s$ is a suffix of the semi-infinite sequence $x^\infty_0$, and no proper suffix of $s$ has this property. The context tree of the process $Q$ is $T_0$ if $T_0$ is a context tree, contains all contexts for the process $Q$, and none of its elements can be replaced by their proper suffixes without violating either of these properties.

Definition 1.1 implies that the context tree of a process always exists for any stationary ergodic process. Clearly, $s_0 \in T_0$ is a context for the process $Q$ if $Q(s_0) > 0$.

For a context tree $T$, let $d(T)$ denote its depth: $d(T) = \max\{l(s), s \in T\}$, and let $T|_K$ denote the context tree $T$ truncated at level $K \in \mathbb{N}$:

\[
T|_K = \{s' : s' \in T \text{ with } l(s') \leq K \text{ or } s' \text{ is a } K\text{-length suffix of some } s \in T\).
\]

If the context tree of a process has depth $d(T_0) = h_0 < \infty$, the process is a Markov chain of order $h_0$. The context tree of an independent and identically distributed (i.i.d.) process consists of the root only.

Statistical estimation of the context tree $T_0$ from a sample is addressed. The sample $x^\infty_0$ is a realization of $X^\infty_0$. For this task, variants of the context algorithm are widely used [2]. The context algorithm was introduced by Rissanen [12], and it is convenient from computational aspects. Nevertheless, the approach of information criteria is a successful estimation method for several models [16]. In particular, the Bayesian information criterion (BIC) and Rissanen’s minimum description length (MDL) principle [13] with either Krichevsky–Trofimov (KT) code length [10] or normalized maximum likelihood (NML) code length [15] are often used. An information criterion assigns a score to each hypothetical model based on the sample, and the estimator is that model whose score is minimal. For context tree estimation this method was believed to be “computationally infeasible” [7] because of the large number of the hypothetical context trees. However, the context tree maximizing (CTM) algorithm [19], [21] computes the KT context tree estimator with practical computational complexity (see Section IV for a precise formulation), avoiding calculating the value of the information criterion for each hypothetical context tree. Moreover, in [6] it is shown that a modification of the CTM algorithm admits computing the BIC context tree estimator with similar computational complexity.

In the literature, strong consistency of the estimators using the information criteria approach has been proved for various models mostly under the assumption that the number of hypothetical models is finite, see [16] for a review. For Markov chains, it is shown in [5] that the BIC order estimator is strongly consistent without any bound on the hypothetical orders. In [5], a counterexample to the strong consistency of the KT and NML order estimators is also shown if there is no bound on the hypothetical orders, or there is a bound equal to a sufficiently large constant times $\log n$, assuming a bound $O(\log n)$, their strong consistency is proved in [4]. This indicates that the common belief of considering BIC an asymptotic approximation of MDL can be justified only when the number of hypothetical models is finite [5]. Moreover, it implies that the KT or NML context tree estimator cannot be strongly consistent without any bound on the depth of the hypothetical context trees.

**Strongly consistent estimation** of the context tree $T_0$ is considered. If $d(T_0) < \infty$, it conventionally means that the estimated context tree equals $T_0$ eventually almost surely as $n \to \infty$. Here and in the sequel, “eventually almost surely” means that with probability 1 there exists a threshold $n_0$ (depending on the finite realization $x^\infty_0$) such that the claim holds for all $n \geq n_0$.

If $d(T_0) = \infty$, the above notion of strongly consistent estimation needs to be generalized because the entire $T_0$ cannot be recovered from any sample of size $n$. In [6], an estimator of $T_0$ is considered strongly consistent if the estimated context tree truncated at any fixed level $K \in \mathbb{N}$ equals $T_0|_K$ eventually almost surely as $n \to \infty$, see (1). Clearly, for $d(T_0) < \infty$ this definition is equivalent to the conventional one, choosing $K$ greater than $d(T_0)$.

In this paper, the notion of strongly consistent estimation of $T_0$ is further generalized as the level $K \in \mathbb{N}$ is considered being a nondecreasing function of the sample size $n$. If $K = (K(n))$ increases with $n$, a strongly consistent estimator recovers $T_0|_K$ to larger and larger depths as $n \to \infty$. The authors are not aware of prior works investigating increasing recovery level $K$.

**Definition 1.2:** A process $Q$ satisfies the mixing condition if there exists $0 < \psi < 1$ such that for any $k$

\[
\text{Prob}\left\{X^{2(k+1)}_{(s)+k+1} = s | X^k_{(s)+k} = s\right\} - Q(s) \leq \psi^k Q(s)
\]

for each $s \leq s_0$, where $s_0 \in T_0$. (2)

The mixing condition (2) is usually called $\psi$-mixing [14], [4], and is satisfied, for example, by the irreducible and aperiodic Markov chains of any order. Denote

\[
q_{\text{min}} = \inf\{Q(\alpha|s_0) : Q(\alpha|s_0) \neq 0, s_0 \in T_0, \alpha \in A\}.
\]

The main result of this paper is that under the condition $q_{\text{min}} > 0$ and the mixing condition on the stationary ergodic process, the BIC context tree estimator, without any restriction on the hypothetical context trees, is strongly consistent in the above sense for a sequence $K = K(n)$, $n \in \mathbb{N}$, where the increasing rate of $K(n)$ depends on the distribution of the process via a specified quantity but it is at most $O(\log n)$.

Another result of this paper is that for arbitrary stationary ergodic processes the BIC context tree estimator, without any restriction on the hypothetical context trees, is strongly consistent...
in the generalized sense, that is, it recovers the true context tree up to any fixed level, eventually almost surely as \( n \to \infty \). Note that in [6] the consistency is proved in the same sense for arbitrary stationary ergodic processes but a bound of \( \log n \) on the depth of the hypothetical context trees is assumed. In [8], there is no restriction on the hypothetical context trees but the process is assumed to be of finite memory and the consistency is proved in the conventional sense.

In the next section, the consistency results mentioned above are formulated in two theorems, which are proved in Section III. Section IV contains a discussion of the results.

II. CONSISTENCY RESULTS

The BIC has the usual form of the negative logarithm of the maximum likelihood plus one half of the number of free parameters in the model times the logarithm of the sample size.

A process \( Q \) with context tree \( T_0 \) is described by the transition probabilities \( Q(a|s_0), a \in A, s_0 \in T_0 \). Strictly speaking, \( (|A| - 1)|T_0| \) of these parameters are free only if \( Q(s_0) > 0 \) for all \( s_0 \in T_0 \) with \( l(s_0) < \infty \). In [6], all \( s_0 \in T_0 \) with \( l(s_0) < \infty \) and \( Q(s_0) = 0 \) are left out from the context tree model, allowing the context trees not to be complete. With that definition, however, \( (|A| - 1)|T_0| \) may still be larger than the number of free parameters because of the constraints that the probabilities of certain strings must be zero.

The maximum likelihood can be written in an explicit form using the following counts. Let \( N_n(s) \) denote the number of occurrences of the string \( s \in A^l(s) \) in the sample \( x^n_1 \)

\[
N_n(s) = \left\lfloor \frac{n}{l(s)} \right\rfloor = \left\lfloor \frac{n}{l(s)} \right\rfloor + \left\lfloor \frac{n}{l(s) + 1} \right\rfloor.
\] (4)

Given a sample \( x^n_1 \), a reasonable context tree is any context tree \( T \) such that the parent \( u \) of each leaf \( s = au \in T \) has a child \( bu \) (\( b \in A \)) with \( N_n(bu) \geq 1 \). The family of all reasonable context trees is denoted by \( R(x^n_1) \). Clearly, for any \( T \in R(x^n_1) \), we have \( d(T) \leq n - 1 \)

\[
\sum_{a \in A} N_n(sa) = N_n-1(s) \leq n - 1 - l(s)
\]
and

\[
n - 1 - d(T) \leq \sum_{s \in T} N_n-1(s) \leq n - 1.
\]

Regarding a \( T \in R(x^n_1) \) as the context tree of a hypothetical process \( Q' \), the probability of the sample \( x^n_1 \) can be written as

\[
Q'(x^n_1) = \prod_{i=1}^{n} Q'(x_i | x_1^{i-1}) = \prod_{s \in x_1^{i-1} \text{ is for some } s \in T} Q'(x_i | x_1^{i-1}) \times \prod_{s \in T, a \in A} Q'(a|s)^{N_n(sa)} .
\] (5)

For a hypothetical context tree \( T \in R(x^n_1) \), define the maximum likelihood \( ML_T(x^n_1)T \) as the maximum in \( Q'(a|s) \) of the second factor above, that is

\[
ML_T(x^n_1)T = \prod_{s \in T} \prod_{N_n-1(s) \geq 1} \frac{N_n(sa)}{N_n-1(s)} .
\]

This can be written as

\[
ML_T(x^n_1)T = \prod_{s \in \hat{T}} \hat{ML}_\hat{s}(x^n_1).
\] (6)

where

\[
\hat{ML}_\hat{s}(x^n_1) = \begin{cases} \prod_{a \in A} \left[ \frac{N_n(sa)}{N_n-1(s)} \right]^{N_n(sa)} , & \text{if } N_n-1(s) \geq 1, \\ 1, & \text{if } N_n-1(s) = 0. \\ \end{cases}
\] (7)

By definition, for any string \( s \)

\[
\prod_{a \in A} Q(a|s)^{N_n(sa)} \leq \hat{ML}_\hat{s}(x^n_1).
\] (8)

Definition 2.1: Given a sample \( x^n_1 \), the BIC for a reasonable context tree \( T \) is

\[
BIC_{CT}(T) = -\log ML_T(x^n_1)T + \frac{|A| - 1}{2} \log n.
\]

In this paper, logarithms are to the base \( e \).

Remark 2.2: Using (6), the BIC can be written as

\[
BIC_{CT}(T) = \sum_{s \in T} \left( -\log \hat{ML}_\hat{s}(x^n_1) + \frac{|A| - 1}{2} \log n \right). 
\] (9)

This additive property makes the CTM-like implementation of BIC possible [6]. Note that the KT criterion can be written in a similar form but the NML cannot. This is why the NML estimator appears unsuitable for CTM implementation.

Given a string \( u \in A^l(u) \), a \( T \) is said to grow from \( u \) if \( T \) is a set of strings, and perhaps also of semi-infinite sequences, such that each semi-infinite sequence \( a^{\geq l} \geq u \) has a unique suffix \( s \geq u \in \mathcal{T} \), and every \( s \in T \) satisfies \( s \geq u \). Given a context tree \( T' \) and \( U \subseteq T' \), a \( T \) is said to grow from \( U \) if it grows from some \( u \in U \). The set of all \( T \) growing from \( u \) and \( U \) is denoted by \( \mathcal{G}(u) \) and \( \mathcal{G}(U) \), respectively. Obviously \( \mathcal{G}(u) \subseteq \mathcal{G}(U) \) if \( u \in U \). The depth of a \( T \in \mathcal{G}(U) \) is defined similarly to the depth of a context tree.

The recovery level \( K \) of the BIC context tree estimator is investigated to be increasing with the sample size \( n \). The growing rate of the sequence \( K(n) \in \mathbb{N}, n \in \mathbb{N} \), depends on the distribution of the process via the following quantity.
**Definition 2.3:** For a proper suffix $s$ of some $s_0 \in \mathcal{T}_0$, and a sufficiently large $D$ with $l(s) < D < \infty$, let $\Delta h_{\max}^R(s)$ be the maximum in $\mathcal{T}$ of

$$\sum_{a \in A} \frac{1}{|T|} \left( \sum_{ua \in T} Q(ua) \log Q(a|u) - Q(sa) \log Q(a|s) \right)$$

where the maximum is taken for all $T \in \mathcal{G}(s)$ with $d(T) \leq D$. The maximizer $T \in \mathcal{G}(s)$ with $d(T) \leq D$ will be denoted by $T_{\max}^D(s)$.

**Remark 2.4:** The maximum in Definition 2.3 is certainly attained, as the number of $T \in \mathcal{G}(s)$ with $d(T) \leq D$ is finite, but the maximizer is not necessarily unique; in that case, any maximizer can be taken as $T_{\max}^D(s)$.

**Remark 2.5:** Definition 1.1 implies that each $u \in T_{\max}^D(s)$ is a suffix of some $s_0 \in \mathcal{T}_0$.

**Remark 2.6:** Using Jensen’s inequality, for any $s$ with $Q(s) > 0$, $T \in \mathcal{G}(s)$, and $a \in A$

$$\sum_{ua \in T} Q(ua) \log Q(a|u) = Q(s) \sum_{ua \in T, Q(u) > 0} Q(u) \left( \frac{Q(ua)}{Q(s)} \log \frac{Q(ua)}{Q(u)} \right) \geq Q(sa) \log \frac{Q(sa)}{Q(s)}$$

with strict inequality if $Q(a|s) \neq Q(a|u)$ for some $u \in T$. Hence, for any $s < s_0, s_0 \in \mathcal{T}_0$, there exist $T \in \mathcal{G}(s)$ and a constant $D < \infty$ such that $\Delta h_{\max}^D(s) \geq \nu$ for some $\nu > 0$.

The main result of the paper is the following.

**Theorem 2.7:** For any stationary ergodic process satisfying the mixing condition (2) and with $\gamma_{\text{ref}} > 0$, there exists $\alpha > 0$ such that for any nondecreasing sequence $K(n) \in \mathbb{N}$, $n \in \mathbb{N}$, satisfying $K(n) \leq \alpha \log n$

and

$$\min \left\{ \Delta h_{\max}^n(s) : s < s_0 \text{ for some } s_0 \in \mathcal{T}_0 \right\} l(s) = K(n) - 1 \geq \frac{\log(1+\varepsilon)n}{\sqrt{n}}$$

(10)

eventually as $n \rightarrow \infty$ with any $\varepsilon > 0$, the BIC estimator

$$\hat{f}_{\text{BIC}}(x^n) = \arg \min_{T \in \mathcal{R}(x^n)} \text{BIC}_{x^n}(T)$$

satisfies

$$\hat{f}_{\text{BIC}}(x^n) |_{K(n)} = T_0 |_{K(n)}$$

eventually almost surely as $n \rightarrow \infty$.

**Proof:** See Section III.

**Remark 2.8:** In both theorems of this section, the indicated minimum is certainly attained, as the number of reasonable context trees is finite, but the minimizer is not necessarily unique; in that case, any minimizer can be taken as $\arg \min$.

**Remark 2.9:** Theorem 2.7 would still hold if in (10) the lower bound $(\log(1+\varepsilon)n)/\sqrt{n}$ was replaced by a sufficiently large constant times $\log n)/\sqrt{n}$, where the constant depends on the distribution of the process.

**Remark 2.10:** By Remark 2.6, there always exists a sequence $K(n)$ which satisfies the conditions of Theorem 2.7. In particular, $K(n) = \lceil \alpha \log n \rceil$ can be taken if

$$\min \left\{ \Delta h_{\max}^n(s) : s < s_0 \text{ for some } s_0 \in \mathcal{T}_0, l(s) = l \right\} \geq \exp(-\xi l/\alpha)$$

where $0 < \xi < 1/2$ is a constant.

For stationary ergodic processes, the BIC Markov order estimator is strongly consistent without any restriction on the hypotheoretical orders, that is, almost surely as $n \rightarrow \infty$ it eventually equals the true order if the process is a Markov chain [5], and diverges if the process is not of finite memory [6]. The next theorem is the analog of this result for context tree models, as it claims that for arbitrary stationary ergodic processes the BIC context tree estimator is strongly consistent without any restriction on the hypotheoretical context trees. In [8], for Markov chains, that is for processes with $d(T_0) < \infty$, the BIC context tree estimator is shown to be strongly consistent without any restriction on the hypotheoretical context trees. Without any assumption on the stationary ergodic process, in [16] the BIC context tree estimator is shown to be strongly consistent with a bound $o(\log n)$ on the depth of the hypotheoretical context trees.

**Theorem 2.11:** For any stationary ergodic process, the BIC estimator

$$\hat{f}_{\text{BIC}}(x^n) = \arg \min_{T \in \mathcal{R}(x^n)} \text{BIC}_{x^n}(T)$$

satisfies for any constant $K \in \mathbb{N}$

$$\hat{f}_{\text{BIC}}(x^n) |_{K} = T_0 |_{K}$$

eventually almost surely as $n \rightarrow \infty$.

**Proof:** See Section III.

**III. CONSISTENCY PROOFS**

Regarding a context tree $T \in \mathcal{R}(x^n)$ as a hypotheoretical context tree of the process, the hypotheoretical contexts $s \in T$ can be grouped according to their relation to the contexts $s_0 \in \mathcal{T}_0$ up to the level $K \in \mathbb{N}$. Denote $T_{UE,K}$ the set of those strings in $T_0$ that underestimate strings in $T_0$ below the level $K$

$$T_{UE,K} = \{ s \in T : l(s) < K, s \prec s_0 \text{ for some } s_0 \in \mathcal{T}_0 \}$$

(11)

and let $T_{\max}^D(T_{UE,K})$ be the following set of strings underestimated by $T_{UE,K}$

$$T_{\max}^D(T_{UE,K}) = \bigcup_{s \in T_{UE,K}} T_{\max}^D(s)$$

(12)

see Definition 2.3. Note that $T_{\max}^D(T_{UE,K}) \subset T_0$ does not necessarily hold, although for each $s \in T_{\max}^D(T_{UE,K})$ there exists a $s_0 \in \mathcal{T}_0$ such that $s \prec s_0$, see Remark 2.5. Similarly, denote
\[ T^{O_E,K} \] the set of those strings in \( T \) that overestimate strings in \( T_0 \) below the level \( K \)

\[ T^{O_E,K} = \{ s \in T : s > s_0 \text{ for some } s_0 \in T_0 \text{ with } l(s_0) < K \} \quad (13) \]

and let \( T_0(T^{O_E,K}) \) be the set of strings in \( T_0 \) overestimated by \( T^{O_E,K} \)

\[ T_0(T^{O_E,K}) = \{ s_0 \in T_0 : l(s_0) < K \text{ for some } s \in T \} \quad (14) \]

Let

\[ T^{K,D} = (T \setminus T^{UE,K} \cup T^D_0(T^{UE,K})) \setminus T^{O_E,K} \cup T_0(T^{O_E,K}) \]

To prove the claims of Theorems 2.7 and 2.11, it suffices to show that for the specified \( K \) and \( D \)

\[ T^{K,D} \in \mathcal{R}(x^n_T) \text{ and } \mathcal{B}(x^n_T) > \mathcal{B}(x^n_T) \]

simultaneously for all considered context trees \( T \) with \( T|_K \neq T_0|_K \), eventually almost surely as \( n \to \infty \). Since \( T|_K \neq T_0|_K \) implies that either \( T^{UE,K} \neq \emptyset \) or \( T^{O_E,K} \neq \emptyset \), the claims of Theorems 2.7 and 2.11 follow from the sequence of Lemmas 3.1, 3.2 and 3.3 below. The proofs of the lemmas follow them in the same order.

**Lemma 3.1:** For any stationary ergodic process satisfying the mixing condition (2) and with \( q_{\inf} > 0 \), there exists \( \eta' > 0 \) such that for \( D(n) = \eta' \log n \) and any nondecreasing sequence \( K(n) \in \mathbb{N}, n \in \mathbb{N}, \) satisfying \( K(n) \leq D(n) \) and

\[ \min \{ \Delta h_{\max}(s) : l(s) = K(n) - 1 \}
\]

\[ s < s_0 \text{ for some } s_0 \in T_0 \geq \frac{\log^{1+\epsilon} n}{\sqrt{n}} \quad (15) \]

eventually as \( n \to \infty \) with any \( \epsilon > 0 \), it follows that

\[ \mathcal{B}(x^n_T) > \mathcal{B}(x^n_T) \left( (T \setminus T^{UE,K(n)} \cup T^D_0(T^{UE,K(n)})) \right) \quad (16) \]

and

\[ (T \setminus T^{UE,K(n)} \cup T^D_0(T^{UE,K(n)})) \in \mathcal{R}(x^n_T) \quad (17) \]

simultaneously for all \( T \in \mathcal{R}(x^n_T) \) with \( T^{UE,K(n)} \neq \emptyset \), eventually almost surely as \( n \to \infty \).

For any stationary ergodic process, (16) and (17) hold for an arbitrary constant \( K(n) = K \) and a constant \( D(n) = D \) depending on \( K \).

**Lemma 3.2:** For any stationary ergodic process with \( q_{\inf} > 0 \), there exists \( \eta > 0 \) such that for any sequence \( K(n) \in \mathbb{N}, n \in \mathbb{N}, \) with \( K(n) \leq \eta \log n \), it follows that

\[ \mathcal{B}(x^n_T) > \mathcal{B}(x^n_T) \left( (T \setminus T^{O_E,K(n)} \cup T_0(T^{O_E,K(n)}) \right) \quad (18) \]

and

\[ (T \setminus T^{O_E,K(n)} \cup T_0(T^{O_E,K(n)}) \in \mathcal{R}(x^n_T) \quad (19) \]

simultaneously for all \( T \in \mathcal{R}(x^n_T) \) with \( T^{O_E,K(n)} \neq \emptyset \) and \( d(T^{O_E,K(n)}) \leq \eta \log n \), eventually almost surely as \( n \to \infty \).

**Lemma 3.3:** For any stationary ergodic process, for any \( \eta > \zeta > 0 \) and any sequence \( K(n) \in \mathbb{N}, n \in \mathbb{N}, \) with \( K(n) \leq \zeta \log n \), it follows that

\[ T^{O_E,K(n)} \notin \mathcal{B}(x^n_T) \]

simultaneously for all \( T \in \mathcal{R}(x^n_T) \) with \( d(T^{O_E,K(n)}) \leq \eta \log n \), eventually almost surely as \( n \to \infty \).

**Proof of Lemma 3.1:** Using (9), (16) is equivalent to

\[ \sum_{s \in T^{UE,K(n)}} \log \mathcal{M}_{\text{es}}(x^n_T) \]

\[ - \sum_{u \in T_0^D(T^{UE,K(n)})} \log \mathcal{M}_{\text{es}}(x^n_T) \]

\[ \leq \left( \frac{1}{2} - 1 \right) 
\]

\[ \times \left( |T^{UE,K(n)}| - |T_0^D(T^{UE,K(n)})| \right) \log n \]

which, by (7), follows from

\[ \sum_{u \in T_0^D(T^{UE,K(n)})} \frac{N_n(ua)}{n} \log \frac{N_n(ua)}{N_{n-1}(u)} 
\]

\[ - \sum_{s \in T^{UE,K(n)}, a \in A} \frac{N_n(sa)}{n} \log \frac{N_n(sa)}{N_{n-1}(s)} \]

\[ > \left( \frac{1}{2} - 1 \right) 
\]

\[ \times |T^{UE,K(n)}| \log n \]

(20)

simultaneously for all \( T \in \mathcal{R}(x^n_T) \) with \( T^{UE,K(n)} \neq \emptyset \), eventually almost surely as \( n \to \infty \). Therefore, it suffices to show (20) as it implies the assertion (16) of the lemma.

Assume the mixing condition (2) and that \( q_{\inf} > 0 \). Let \( \eta > 0 \) be arbitrary first, and let \( K(n) \) and \( D(n) \) be as in the claim of the lemma being proven. Using (8), the left side of (20) can be lower bounded by

\[ \sum_{u \in T_0^D(T^{UE,K(n)})} \frac{N_n(ua)}{n} \log \frac{N_n(ua)}{N_{n-1}(ua)} - \sum_{s \in T^{UE,K(n)}, a \in A} \frac{N_n(sa)}{n} \log \frac{N_n(sa)}{N_{n-1}(sa)} \]

that can be written as

\[ \sum_{u \in T_0^D(T^{UE,K(n)})} \frac{N_n(ua) - Q(ua)}{n} \log Q(ua) 
\]

\[ - \sum_{s \in T^{UE,K(n)}, a \in A} \frac{N_n(sa) - Q(sa)}{n} \log Q(sa) 
\]

\[ - \sum_{s \in T^{UE,K(n)}, a \in A} \frac{N_n(sa)}{n} \log \frac{N_n(sa)}{N_{n-1}(sa)/Q(a)} 
\]

\[ + \sum_{u \in T_0^D(T^{UE,K(n)})} Q(ua) \log Q(ua) 
\]

\[ - \sum_{s \in T^{UE,K(n)}, a \in A} Q(sa) \log Q(sa) \]

(21)
Theorem 1.b and the remark after that in [4] imply that for any stationary ergodic process satisfying the mixing condition (2) and with \(q_{\inf} > 0\) there exist \(\gamma > 0\) and \(\eta' > 0\) such that

\[
\left| \frac{N_n(s)}{n} - Q(s) \right| \leq \frac{\gamma \sqrt{Q(s) \log n}}{\sqrt{n}}
\]

simultaneously for all strings \(s \leq s_0\), where \(s_0 \in \mathcal{T}_0\), with \(q(s) \leq \eta\log n\), eventually almost surely as \(n \to \infty\). Note that in [4] the bound in the mixing condition (2) is required to hold for all strings \(s\), not only for \(s \leq s_0\), where \(s_0 \in \mathcal{T}_0\), and that implies (22) simultaneously for all strings \(s\) with \(q(s) \leq \eta' \log n\), eventually almost surely as \(n \to \infty\). Let \(\eta' > 0\) be chosen according to the above. Then, since all \(u \in \mathcal{T}_0' \log n (\mathcal{T}_{\text{UE}}, K(n))\) and \(u \in \mathcal{T}_{\text{UE}}, K(n)\) satisfy \(u \leq s_0\) for some \(s_0 \in \mathcal{T}_0\), and \(l(u) \leq \eta' \log n\), see (12), (11), and Definition 2.3, it follows from (22) that the first and respectively the second term in (21)

\[
\sum_{u \in \mathcal{T}_0' \log n (\mathcal{T}_{\text{UE}}, K(n)), a \in A} \left( \frac{N_n(ua)}{n} - Q(ua) \right) \log Q(a|u) \\
\geq |A| \left| \mathcal{T}_0' \log n \left( \mathcal{T}_{\text{UE}}, K(n) \right) \right| (\log q_{\inf}) \frac{\gamma \log n}{\sqrt{n}}
\]

and

\[
- \sum_{s \in \mathcal{T}_{\text{UE}}, K(n), a \in A} \left( \frac{N_n(sa)}{n} - Q(sa) \right) \log Q(a|s) \\
\geq |A| \left| \mathcal{T}_{\text{UE}}, K(n) \right| (\log q_{\inf}) \frac{\gamma \log n}{\sqrt{n}}
\]

simultaneously for all \(T \in \mathcal{R}(x^T)\) with \(T_{\text{UE}}, K(n) \neq \emptyset\), eventually almost surely as \(n \to \infty\).

For the third term in (21)

\[
- \sum_{s \in \mathcal{T}_{\text{UE}}, K(n), a \in A} \frac{N_n(sa)}{n} \log Q(a|s) \\
= - \sum_{s \in \mathcal{T}_{\text{UE}}, K(n)} \frac{N_{n-1}(s)}{n} D \left( \frac{N_n(s)}{N_{n-1}(s)} \mid Q(s) \right) \\
\geq (i) \sum_{s \in \mathcal{T}_{\text{UE}}, K(n)} \frac{N_{n-1}(s)}{n} \frac{1}{q_{\inf}} \sum_{a \in A} \left( \frac{N_n(sa)}{N_{n-1}(s)} - Q(a|s) \right)^2 \\
\geq (ii) - \frac{\gamma^2}{q_{\inf}} \left| \mathcal{T}_{\text{UE}}, K(n) \right| \log \frac{\log^2 n}{n}
\]

simultaneously for all \(T \in \mathcal{R}(x^T)\) with \(T_{\text{UE}}, K(n) \neq \emptyset\), eventually almost surely as \(n \to \infty\). Here \(D(\|\|)\) denotes the information divergence, and (i) follows as \(D(P_1 \| P_2) \leq \sum_{a \in A} (P_1(a) - P_2(a))^2 / P_2(a)\) for any two probability distributions on \(A\), see, e.g., Lemma 6.3 in [6]. Moreover, it follows from (22), see Section 3 of [4], that

\[
\left| \frac{N_n(sa)}{N_{n-1}(s)} - Q(a|s) \right| \leq \frac{\gamma \log n}{\sqrt{N_{n-1}(s)}},
\]

simultaneously for all strings \(s \leq s_0\), where \(s_0 \in \mathcal{T}_0\), with \(Q(s) > 0\) and \(l(s) \leq \eta' \log n\), eventually almost surely as \(n \to \infty\). This implies (ii) in (25).

For the fourth and fifth terms in (21)

\[
\sum_{u \in \mathcal{T}_0', \log n (\mathcal{T}_{\text{UE}}, K(n)), a \in A} Q(ua) \log Q(a|u) \\
- \sum_{s \in \mathcal{T}_{\text{UE}}, K(n), a \in A} Q(sa) \log Q(a|s) \\
= (i) \sum_{s \in \mathcal{T}_{\text{UE}}, K(n), a \in A} \left( \sum_{u \in \mathcal{T}_0', \log n (a)} Q(ua) \log Q(a|u) \\
- Q(sa) \log Q(a|s) \right) \\
= (ii) \sum_{s \in \mathcal{T}_{\text{UE}}, K(n)} \left| \mathcal{T}_0', \log n (a) \right| \log \frac{\log^2 n}{n} \\
\geq (iii) \sum_{s \in \mathcal{T}_{\text{UE}}, K(n)} \left| \mathcal{T}_0', \log n (a) \right| \frac{\log^{1+\epsilon} n}{\sqrt{n}}
\]

eventually as \(n \to \infty\), where (i) and (ii) follow from the definition (12) of \(T_0^D(\mathcal{T}_{\text{UE}}, K(n))\), and (iii) from Definition 2.3. If \(l(s) = K(n) - 1\), (iii) is provided by the condition (15). If \(l(s) < K(n) - 1\), (iii) follows from (15) too, because \(K(n)\) is nondecreasing, \(\eta' \log n\) is increasing, and the right of (15) is decreasing in \(n\).

Combining the bounds (23), (24), (25), and (26), the expression (21) can be lower bounded by

\[
|A| \left( \frac{\mathcal{T}_0', \log n \left( \mathcal{T}_{\text{UE}}, K(n) \right) }{\log q_{\inf}} \right) \frac{\gamma \log n}{\sqrt{n}} \\
+ \left| \mathcal{T}_{\text{UE}}, K(n) \right| \frac{\log q_{\inf}}{\sqrt{n}} \\
+ \frac{\gamma^2}{q_{\inf}} \left| \mathcal{T}_{\text{UE}}, K(n) \right| \frac{\log^2 n}{n} \\
- \frac{\gamma^2}{q_{\inf}} \left| \mathcal{T}_{\text{UE}}, K(n) \right| \log \frac{\log^2 n}{n} \\
\times \left( \log^2 n - 2\gamma |A| \log q_{\inf} \right) - \frac{|A|^2 \log^2 n}{q_{\inf}} \frac{\log n}{\sqrt{n}}
\]

simultaneously for all \(T \in \mathcal{R}(x^T)\) with \(T_{\text{UE}}, K(n) \neq \emptyset\), eventually almost surely as \(n \to \infty\). Due to the assumption \(q_{\inf} > 0\) of the lemma, (27) yields (20), and the assertion (16) follows. The second assertion (17) of the lemma is implied by (22) because \(d(T_0', \log n (\mathcal{T}_{\text{UE}}, K(n))) \leq \eta' \log n\), see (12) and Definition 2.3.

If \(K(n) = K\) is a constant, then (20) follows immediately without assuming the mixing condition (2) and that \(q_{\inf} > 0\). Indeed, if \(|T_{\text{UE}}, K| < \mathcal{T}_0^D(\mathcal{T}_{\text{UE}}, K) \leq \infty\), the ergodic theorem implies that the left side of (20) converges almost surely to

\[
\sum_{u \in \mathcal{T}_0', \log n (\mathcal{T}_{\text{UE}}, K(n)), a \in A} Q(ua) \log Q(a|u) \\
- \sum_{s \in \mathcal{T}_{\text{UE}}, K(n), a \in A} Q(sa) \log Q(a|s)
\]
By Remark 2.6, there exist constants $D < \infty$ and $\nu > 0$ such that $\Delta t_{\text{max}}^{D}(s) \geq \nu$ simultaneously for all $s < s_0$, where $s_0 \in T_0$, with $I(s) < K$. Then, similarly as in (26), (28) can be lower bounded by $|T_0^{D}(T^{OE,K}(n))| \nu$, and (20) follows. The second assertion (17) of the lemma follows from the ergodic theorem as well in this case.

Proof of Lemma 3.2: Using (9), (18) is equivalent to
\[
\sum_{s \in T^{OE,K}(n)} \log \mathcal{M}_{s}(x_n) - \sum_{s \in T_0(T^{OE,K}(n))} \log \mathcal{M}_{s_0}(x_n)< \frac{|A|-1}{2} \left(\left(\sum_{s \in T^{OE,K}(n)} I(s) - |T_0(T^{OE,K}(n))|\right)\log n\right), \tag{29}
\]
Using (7) and (8), the left of (29) can be upper bounded by
\[
\sum_{s \in T^{OE,K}(n)} N_\alpha(s) \log \frac{N_\alpha(s)}{N_{s-1}(s)} - \sum_{s \in T_0(T^{OE,K}(n))} N_\alpha(s_0) \log Q(a|s_0), \tag{30}
\]
Each $s \in T^{OE,K}(n)$ has a suffix $s_0 \in T_0(T^{OE,K}(n))$ with $I(s_0) < K(n)$, see (13) and (14). This immediately implies the assertion (19) since $T \in \mathcal{R}(x^T_1)$, and that
\[
\sum_{s \in T_0(T^{OE,K}(n))} N_\alpha(s_0) \log Q(a|s_0)\geq \sum_{s \in T_0(T^{OE,K}(n))} \left(\sum_{s \in T^{OE,K}(n)} N_\alpha(s) \log Q(a|s)\right)
+ (\log q_{\text{inf}}) \left(\sum_{s \in T_0(T^{OE,K}(n))} N_\alpha(s_0)\right)
- \sum_{s \in T^{OE,K}(n)} N_\alpha(s_0) \log Q(a|s) \right), \tag{31}
\]
because $Q(a|s) = Q(a|s_0)$ if $s \geq s_0$, see Definition 1.1. Clearly
\[
\sum_{s \in T_0(T^{OE,K}(n))} N_\alpha(s_0) - \sum_{s \in T^{OE,K}(n)} N_\alpha(s) \leq d(T^{OE,K}(n)) \leq \eta \log n \tag{32}
\]
where the last inequality is a condition of the lemma. Combining (31) and (32), (30) can be further bounded by
\[
\sum_{s \in T_0(T^{OE,K}(n))} N_\alpha(s_0) \log \frac{N_\alpha(s_0)}{N_{s-1}(s)} \log Q(a|s) \times D\left(\frac{N_\alpha(s)}{N_{s-1}(s)} \log Q(a|s)\right)
+ \eta \log q_{\text{inf}} \log n \leq \sum_{s \in T_0(T^{OE,K}(n))} \sum_{s \in T^{OE,K}(n)} N_\alpha(s_0) \log \frac{N_\alpha(s_0)}{N_{s-1}(s)} \log Q(a|s) \times D\left(\frac{N_\alpha(s)}{N_{s-1}(s)} \log Q(a|s)\right)
+ \eta \log q_{\text{inf}} \log n, \tag{33}
\]
It is known, see the proof of Lemma 3.2 in [6] and also the proof of Theorem 4 in [4], that for any stationary ergodic process, to any $\delta > 0$ there exists $\eta' > 0$ such that for any $T'$ growing from some $s_0 \in T_0$ with $d(T') \leq \eta' \log n$
\[
\sum_{s \in T'} N_{n-1}(s) \log \left|\left|\frac{N_\alpha(s)}{N_{s-1}(s)} \log Q(a|s)\right|\right| < \delta \frac{|A|}{q_{\text{inf}} |T'| \log n}
\]
simultaneously for all $T'$ and $s_0$ as above, eventually almost surely as $n \to \infty$. Thus, if $\eta \leq \eta'$, the right of (33) can be upper bounded by
\[
\sum_{s \in T_0(T^{OE,K}(n))} \frac{2A}{q_{\text{inf}}} \log \left|\left|\frac{N_\alpha(s)}{N_{s-1}(s)} \log Q(a|s)\right|\right| \log n
\leq \sum_{s \in T_0(T^{OE,K}(n))} \frac{2A}{q_{\text{inf}}}
\times \left(\left|\left|\frac{N_\alpha(s)}{N_{s-1}(s)} \log Q(a|s)\right|\right| - 1\right) \log n
+ \eta \log q_{\text{inf}} \log n
= \left(\frac{2A}{q_{\text{inf}}} + \eta \log q_{\text{inf}}\right) \log n \tag{34}
\]
simultaneously for all $T \in \mathcal{R}(x^T_1)$ with $T^{OE,K}(n) \neq \emptyset$ and $d(T^{OE,K}(n)) \leq \eta \log n$, eventually almost surely as $n \to \infty$. Choosing
\[
\delta = \min \left(\eta', \frac{|A|-1}{8|A|}\right) \quad \text{and} \quad \eta = \min \left(\eta', \frac{|A|-1}{2|A|}\right)
\tag{34}
\]
implies (29) and the assertion (18).

Proof of Lemma 3.3: The proof technique is inspired by the proof of Theorem 1 in [8], which itself is inspired by the proof of Proposition 2 in [5]. A similar technique is applied in the proof of Barron’s lemma on competitive optimality of the Shannon code [1] also.

Fix $\eta > 0$. It is shown that
\[
Q\left(\left\{x^T_1 : T^{OE,K}(n) \subseteq \hat{G}_{\text{BKC}}(x^T_1) \right\} \right. \quad \text{for some} \quad T \in \mathcal{R}(x^T_1)
\quad \text{with} \quad d\left(T^{OE,K}(n) \right) \geq \eta \log n \right) \leq n^{-2} \tag{35}
\]
for sufficiently large $n$, that yields the assertion of the lemma by the Borel-Cantelli lemma.

Let $T^{K(n)}_0 = \{s_0 \in T_0 : I(s_0) < K(n)\}$. For $T' \in G(s_0) \subseteq G(T^{K(n)}_0)$, denote $r(T') = s_0$. Recalling (13) and (9), (35) is equivalent to
\[
Q\left(\left\{x^T_1 : T' \subseteq \hat{G}_{\text{BKC}}(x^T_1) \right\} \quad \text{for some}\quad T' \in G(T^{K(n)}_0)
\quad \text{with} \quad \eta \log n \leq d(T') \leq n \right) \leq n^{-2} \tag{36}
\]
for sufficiently large $n$. Defining $B_{n,T'} = \{x^T_1 : T' \subseteq \hat{G}_{\text{BKC}}(x^T_1)\}$, (36) is further equivalent to
\[
Q\left(\bigcup_{T' \in G(T^{K(n)}_0) \cap \eta \log n \leq d(T') \leq n} B_{n,T'} \right) \leq n^{-2} \tag{37}
\]
for sufficiently large $n$.

First, by (5) for any $T' \in G(T_0, K(n))$ there is the equality

\[
Q(B_{n,T'}) = \sum_{x_1^n \in B_{n,T'}} Q(x_1^n) = \sum_{x_1^n \in B_{n,T'}} \prod_{i=1}^{n} Q(x_i | x_1^{i-1}) \prod_{a \in A} Q(a|s_0)^{N_n(a)}
\]

Using (8), for the last term in (38) the following bound is obtained

\[
\prod_{a \in A} Q(a|r(T'))^{N_n(r(T')a)} \leq \overline{M}_{r(T')}(x_1^n), \quad (39)
\]

Since (9) implies

\[
- \sum_{s \in T'} \log \overline{M}_{r(T')}(x_1^n) + \frac{|A| - 1}{2} |T'| \log n < \log \overline{M}_{r(T')}(x_1^n) + \frac{|A| - 1}{2} \log n
\]

which is equivalent to

\[
\overline{M}_{r(T')}(x_1^n) < n^{\frac{|A| - 1}{2} |T'|} \prod_{s \in T'} \overline{M}_{r(T')}(x_1^n), \quad x_1^n \in B_{n,T'}
\]

The bound (39) can be written as

\[
\prod_{a \in A} Q(a|r(T'))^{N_n(r(T')a)} < n^{\frac{|A| - 1}{2} |T'|} \prod_{s \in T'} \overline{M}_{r(T')}(x_1^n), \quad x_1^n \in B_{n,T'}, \quad (40)
\]

The proof relies upon the comparison of the maximum likelihood to the KT distribution. Denote the equation shown at the bottom of the page if $N_{n-1}(s) \geq 1$, and $\hat{P}_{KT,s}(x_1^n) = 1$ if $N_{n-1}(s) = 0$. Consider the standard bound (see, e.g., [6, eq. (17)])

\[
\log \hat{P}_{KT,s}(x_1^n) - \log \overline{M}_{s}(x_1^n) + \frac{|A| - 1}{2} \log N_{n-1}(s) < C
\]

for any string $s$ with $N_{n-1}(s) \geq 1$, where $C$ is a constant depending only on the alphabet size $|A|$. This implies

\[
\log \overline{M}_{s}(x_1^n) \leq \log \hat{P}_{KT,s}(x_1^n) + \frac{|A| - 1}{2} \log N_{n-1}(s) + C
\]

which is equivalent to

\[
\overline{M}_{s}(x_1^n) \leq \hat{P}_{KT,s}(x_1^n) N_{n-1}(s)^{\frac{|A| - 1}{2}} e^C
\]

Using $(\prod_{i=1}^{l} c_i)^{1/l} \leq (1/l) \sum_{i=1}^{l} c_i$, $c_i > 0$, we get

\[
\prod_{a \in A} Q(a|r(T'))^{N_n(r(T')a)} < n^{\frac{|A| - 1}{2} |T'|} \prod_{s \in T'} \hat{P}_{KT,s}(x_1^n)
\]

Hence, the bound (40) can be written as

\[
\prod_{a \in A} Q(a|r(T'))^{N_n(r(T')a)} < n^{\frac{|A| - 1}{2} |T'|} \prod_{s \in T'} \hat{P}_{KT,s}(x_1^n)
\]

Applying (41) to (38), it follows that

\[
Q(B_{n,T'}) \leq n^{\frac{|A| - 1}{2} |T'|} \prod_{s \in T'} Q(x_i | x_1^{i-1}) \prod_{x_1^n \in B_{n,T'}} \prod_{a \in A} Q(a|s_0)^{N_n(a)} \prod_{s \in T'} \hat{P}_{KT,s}(x_1^n)
\]

Although $(T_0 \setminus \{r(T')\}) \cup T'$ is a context tree, the function

\[
\prod_{x_1^n \in B_{n,T'}} \prod_{i=1}^{l} Q(x_i | x_1^{i-1}) \prod_{a \in A} Q(a|s_0)^{N_n(a)} \prod_{s \in T'} \hat{P}_{KT,s}(x_1^n)
\]

can be written as

\[
\prod_{a \in A} Q(a|r(T'))^{N_n(r(T')a)} < n^{\frac{|A| - 1}{2} |T'|} \prod_{s \in T'} \hat{P}_{KT,s}(x_1^n)
\]

which implies

\[
\hat{P}_{KT,s}(x_1^n) = \prod_{a \in A} Q(a|s_0)^{N_n(a)} \prod_{s \in T'} \hat{P}_{KT,s}(x_1^n)
\]

\[
\prod_{a \in A} Q(a|r(T'))^{N_n(r(T')a)} < n^{\frac{|A| - 1}{2} |T'|} \prod_{s \in T'} \hat{P}_{KT,s}(x_1^n)
\]
may not define a probability mass function on \( \{x^n_0 \in A^n\} \) because \( \sum_{s \in T} N_{n-1}(s) \) may be less than \( N_{n-1}(T') \). But considering \( N_{n-1}(T') = \sum_{s \in T} N_{n-1}(s) \leq d(T') \), it holds that
\[
\sum_{x^n_0 \in B_{n,T'}} \prod_{s \in T'} Q(x_i | x^n_1) \prod_{s \in T \setminus T'} Q(d|s)N_{n-1}(s) \leq A|d(T')|.
\]

Hence, the bound (42) can be written as
\[
Q(B_{n,T'}) \leq n^{(\frac{d(T')}{2})} e^{d(T')} |d(T')|^{-\frac{1}{2}} |d(T')| A|d(T')|.
\]  
(43)

From \( T' \in G(\tau(T')) \), it is clear that
\[
|T'| \geq |A|d(T') - \|\tau(T')\| \geq |A|d(T') - K(n).
\]  
(44)

Using (44), the bound (43) can be written as
\[
Q(B_{n,T'}) \leq |A|K(n)n^{(\frac{d(T')}{2})} e^{d(T')} |d(T')|^{-\frac{1}{2}} |d(T')| |T'|.
\]  
(45)

Now, by (45), the left of (37) can be bounded as
\[
Q \left( \bigcup_{T' \in G(\tau(T^n_0(n))) \cap \mathbb{N} \log n \leq d(T') \leq n} B_{n,T'} \right)
\leq \sum_{T' \in G(\tau(T^n_0(K(n))) \cap \mathbb{N} \log n \leq d(T') \leq n} Q(B_{n,T'})
\leq |A|^{K(n)}n^{(\frac{d(T')}{2})} e^{d(T')} |d(T')|^{-\frac{1}{2}} |d(T')| |T'|.
\]  
(46)

In the above sum, the terms will be grouped according to the cardinality of \( T' \). Using (44) again, \( d(T') \geq \eta \log n \) implies \( |T'| \geq |A|(\eta \log n - K(n)) \). Note that it would be sufficient to consider only those \( T' \in G(\tau(T^n_0(K(n))) \) for which \( T' \subseteq T^{O.E.K(n)} \) for some \( T \in \mathcal{R}(x^n_0) \), \( x^n_0 \in A^n \). Such \( T' \) satisfy that \( |T'| \leq |A|^2 \). Using this restriction on the cardinality of \( T' \), however, would not improve the results. Hence, (46) is continued by
\[
\leq |A|^{K(n)}n^{(\frac{d(T')}{2})} e^{d(T')} |d(T')|^{-\frac{1}{2}} |d(T')| |T'|
\leq |A|^{K(n)}n^{(\frac{d(T')}{2})} e^{d(T')} |d(T')|^{-\frac{1}{2}} |d(T')| m = |A|^{K(n)}n^{(\frac{d(T')}{2})} e^{d(T')} |d(T')|^{-\frac{1}{2}} |d(T')| m.
\]  
(47)

It is known, see, e.g., [8, Lemma 2], that the number of context trees with \( m \) leaves is upper bounded by \( 16^m \). Since \( \tau(T') \in T^n_0(K(n)) \), this gives
\[
\left| \left\{ T' : T' \in G(\tau(T^n_0(K(n))) \cap |T'| = m \right\} \right| \leq |T^n_0(K(n))| m = |A|^{K(n)} 16^m.
\]

Hence, (47) can be continued by
\[
\leq |A|^{2K(n)}n^{(\frac{d(T')}{2})} \sum_{m \geq |A| \log n - K(n)} e^{m(C+\log 16+|A|^{-1} \log |A|)} \times m \leq |A|^{2K(n)}n^{(\frac{d(T')}{2})} \sum_{m \geq |A| \log n - K(n)} e^{m(C+\log 16+|A|^{-1} \log |A|)} \times m
\]  
(48)

A straightforward calculation [8] shows that
\[
\sum_{m \geq |A| \log n - K(n)} e^{m(C+\frac{d(T')}{2}) \log m} \leq e^{-m_0} (C' + \frac{d(T')}{2} \log m_0)
\]
for any constant \( C' > 0 \), if \( m_0 \) is sufficiently large. Applying this with \( C' = C + \log 16 + |A|^{-1} \log |A| \), (48) can be continued by
\[
\leq |A|^{2K(n)}n^{(\frac{d(T')}{2})} \exp \left\{ |A|(\eta \log n - K(n)) - 1 \right\}
\times (C' - \frac{d(T')}{2} \log (|A|(\eta \log n - K(n)) - 1)) \}
\]  
(49)

if \( n \) is sufficiently large. With \( 0 < \zeta < \eta \) and \( K(n) \leq \zeta \log n \), (49) yields (37) and the proof is complete.

**IV. Discussion**

Arbitrary stationary ergodic processes were considered, whose context trees may have infinite depth. It was shown that the BIC context tree estimator, without any restriction on the hypothetical context trees, recovers the true context tree up to any fixed level, eventually almost surely as \( n \to \infty \). More than that, it was proved that under some conditions on the process the above level can increase with \( n \) at a specific rate determined by the distribution of the process; thus, the BIC estimator can recover the true context tree to larger and larger depths. Although the proved growing rate of the recovery level appears to be the best achievable, it remains open if it can be achieved under weaker conditions on the process than those assumed. Other information criteria, such as the popular KT or NML versions of MDL, were not considered as they are known not to be consistent without any bound on the hypothetical context trees [5].

In this paper, as often in the literature [20], [21], [3], [8], the context trees were required to be complete, that is, each non-leaf node of the context tree has as many children as the alphabet size. In [6], this property is not assumed and the strings with finite length and zero probability are left out from the context tree of the process. Additionally, in the generalized context tree models [11] the nonbranching paths in the not necessarily complete context tree graphs are represented by single edges labeled by strings rather than single symbols. In [22], [3] and
[8], an indeterminate symbol ε is added to the alphabet. This extra symbol is used to fill in the unknown infinite past of a finite-length observation of the process, generating sufficient past for the first symbols in the observation. Here, this technique was avoided because it increases the penalty term of BIC. However, it remains open whether the results here still hold if the context trees are not assumed to be complete, as it also affects the penalty term of BIC.

The number of occurrences of strings was defined to be the actual one. In [6], the index \( i \) in the definition (4) runs from \( D(n) + 1 \), where \( D(n) \) is the upper bound on the depth of the hypothetical context trees, to guarantee that the sum of the counts equals \( n - D(n) \) uniformly for all hypothetical context trees. Since no such upper bound was assumed in this paper, the counts could not be defined this way. Another way of defining the counts is getting the index \( i \) run from \( d(T) + 1 \), where \( T \) is a hypothetical context tree, to guarantee that the sum of the counts equals \( n - d(T) \) for the hypothetical context tree \( T \). Then the number of occurrences of a certain hypothetical context would vary with which hypothetical context tree is being considered. This kind of definition would cause difficulties in the practical computation of the estimator.

Algorithms to compute the BIC context tree estimator were not considered in this paper. Calculating the BIC estimator via computing the value of BIC for each hypothetical context tree would not be feasible in practice, since the number of the hypothetical context trees based on a sample is very large. In [6], a CTM-like algorithm is presented to compute the BIC estimator, with a bound \( O(\log n) \) on the depth of the hypothetical context trees, from the sample of size \( n \) in linear time \( O(n) \). Without such a bound, the computational complexity of that algorithm would be of higher order. In [8], the idea of [6] is adapted to efficient tree representation algorithms [9], and a method is shown to compute the BIC estimator, without any restriction on the hypothetical context trees, in linear time. In [6], the on-line version of the algorithm is proved to compute the estimator on-line, that is, simultaneously for all subsamples of sizes \( i \leq n \), in time \( O(n\log n) \). In [8], the on-line version of the algorithm is not considered, although it is clear that a quadratic computational complexity \( O(n^2) \) is achievable based on the off-line algorithm.

ACKNOWLEDGMENT

The authors acknowledge the Associate Editor and the reviewers’ effort to help revise the paper.

REFERENCES


Zsolt Talata (M’06) received the M.S. degree in electrical engineering and the Ph.D. degree in applied mathematics from Budapest University of Technology and Economics, Hungary, in 2000 and 2005, respectively. He received a Ph.D. scholarship from the Rényi Institute of Mathematics, the Hungarian Academy of Sciences at numerous other institutions.

Dr. Duncan is a member of the editorial boards of Communications on Stochastic Analysis and Risk and Decision Analysis and was on the editorial board of the SIAM Journal on Control and Optimization (1994–2007) as an Associate Editor and a Corresponding Editor. He is a member of AMS, MAA, and SIAM. He is an IFAC Fellow.