The Context-Tree Weighting Method: Extensions

Frans M. J. Willems, Member, IEEE

Abstract—First we modify the basic (binary) context-tree weighting method such that the past symbols \(x_{1−D}, x_{2−D}, \ldots, x_0\) are not needed by the encoder and the decoder. Then we describe how to make the context-tree depth \(D\) infinite, which results in optimal redundancy bounds for all tree sources, while the number of records in the context tree is not larger than \(2T−1\). Here \(T\) is the length of the source sequence. For this extended context-tree weighting algorithm we show that with probability one the compression ratio is not larger than the source entropy for source sequence length \(T → ∞\) for stationary and ergodic sources.

Index Terms—Binary stationary and ergodic sources, cumulative redundancy bounds, modeling procedure, sequential data compression, tree sources, universal source coding.

I. INTRODUCTION

The context-tree weighting method, first presented in [7], appears to be an efficient implementation for weighting (mixing) the coding distributions (universal over the parameters) corresponding to all tree models in class \(C_D\), i.e., the set of all tree models \(S\) whose maximum depth does not exceed \(D\). A tree model is determined by a proper and complete set \(S\) of suffixes. Together these suffixes form a tree that is grown in negative \(t\) (i.e., time) direction. Each semi-infinite sequence \(\cdots \rightarrow x_i \rightarrow x_{i+1} \rightarrow x_{i+2} \rightarrow \cdots\) has a unique suffix in \(S\), i.e., it passes through a unique leaf in the corresponding model tree. We restrict ourselves here to binary sources. Then this suffix (leaf) determines the probability that \(x_1\), i.e., the next symbol generated by the (binary) source, is a 1.

The analysis of the context-tree weighting method turns out to be very straightforward (see [8]). It shows that the performance is as good as we can possibly hope, not only asymptotically but also for finite sequence lengths. Here we will propose two extensions to the basic context-tree weighting method and derive an interesting consequence of these extensions. We will use the notation of [8]. Codewords are assumed to be binary, logarithms have base 2, and information quantities are expressed in bits.

II. CODING WITHOUT KNOWLEDGE OF PAST SYMBOLS \(x_{1−D}, x_{2−D}, \ldots, x_0\) AND \(x_0\)

A. A Simple Adaptation of the Basic Context-Tree Weighting Method

In its basic form (described in [8]), where we assumed that the actual tree model \(S \in C_D\), the context-tree weighting method needs, for processing the symbol \(x_t\), for \(t = 1, 2, \ldots, T\), the context \(x_t[−1] = (x_{1−D} x_{2−D} \cdots x_{1−1})\) for this symbol. Here \(D\) is the depth of the context tree. This implies that for processing \(x_1\) the encoder and decoder must have access to \(x_{1−D} x_{2−D} \cdots x_0\) and for \(x_2\) they need \(x_{2−D} x_{3−D} \cdots x_1\), etc. This is an unpleasant fact. The past symbols \(x_{1−D}, x_{2−D}, \ldots, x_0\) may not be available at all to the encoder and the decoder. A straightforward way to circumvent this problem is to start processing only after a full context is available to both, i.e., to start processing with \(x_{T+1}\). The encoder then sends the first context \(x_1 x_2 \cdots x_{T+1}\) to the decoder in an uncoded way. This requires \(D\) binary-code digits.

To study coding methods that do not assume availability of the past symbols \(x_{1−D}, x_{2−D}, \cdots, x_0\) we first consider the case where the encoder and the decoder do know the tree model \(S\) of the source.

B. Known Model

In the situation where the encoder and the decoder already know the tree model \(S\), the source symbol \(x_t\) can be transmitted in an uncoded way, if its available context \(x_1 x_2 \cdots x_{T−1}\) does not have a suffix\(^1\) in \(S\). If, on the other hand, the available context of a symbol does have a suffix in \(S\) we use the Krichevsky–Trofimov estimator [2] that corresponds to this suffix. This results in the coding distribution

\[
P'_s(x_t | S) \geq 2^{-s(x_t)} \prod_{s \in S} P_r(a'_s(x_t), b'_s(x_t)),
\]

for all \(x_t \in \{0, 1\}, t = 0, 1, \ldots, T\). (1)

Here \(a'_s(x_t)\) denotes the number of zeros that occurred in \(x_t\) at instants \(\tau = 1, t\) such that \(s\) is a suffix of \(x_{T−1}\), and \(b'_s(x_t)\) denotes the number of ones in \(x_t\) at instants \(\tau = 1, t\), for which \(s\) is a suffix of \(x_{T−1}\), for \(s \in T_D\). Furthermore, the estimator \(P_r(\cdot, \cdot)\) is defined as

\[
P_r(a, b) \equiv \frac{1}{2} \cdot \frac{a}{b} \cdot \frac{b}{a+1} \cdot \cdots \cdot \frac{b}{a+b},
\]

for \(a > 0\) and \(b > 0\), etc. (2)

For the number of uncoded symbols \(\Delta_S(x_t)\), i.e., the number of symbols \(x_t\) in \(x_{T−1}\) for which the available context \(x_{T−1}\) does not have a suffix in \(S\), we can write

\[
\Delta_S(x_t) = t - \sum_{s \in S} (a'_s(x_t) + b'_s(x_t)).
\]

(3)

Note that (1) is a sequentially available coding distribution. Observe that the number of uncoded symbols \(\Delta_S(x_t)\) in the source sequence \(x_{T−1}\) depends on both the suffix set \(S\) and the source sequence \(x_{T−1}\).

We always have that

\[
\Delta_S(x_{T−1}) \leq \Delta_S^{\text{max}} \equiv \max_{s \in S} l(s)
\]

where \(l(s)\) is the length of \(s\). The number of uncoded symbols is also called the number of missing contexts.

Let

\[
\gamma(z) \equiv \begin{cases} \frac{z}{2} \log z + 1, & \text{for } 0 \leq z < 1 \\ \frac{z}{2} \log z + 1, & \text{for } z \geq 1 \end{cases}
\]

(4)

thus \(\gamma(\cdot)\) is the “smallest” convex-\(\gamma\) continuation of \(\frac{1}{2} \log z + 1\) for \(0 \leq z < 1\) satisfying \(\gamma(0) = 0\). For the individual (cumulative) redundancy resulting from this coding distribution for the source sequence \(x_{T−1}\) with respect to the tree source with model \(S\) and parameter vector \(\Theta_S\), we now obtain for all past source sequences

\[^1\text{Also if } x_1 x_2 \cdots x_{T−1} \in S \text{ we say that } x_1 x_2 \cdots x_{T−1} \text{ has a suffix in } S.\]
The first inequality follows from the fact that \( a_s \geq a'_s = a'_s(x'_T) \) and \( b_s \geq b'_s = b'_s(x'_T) \), where \( a_s = a_s(x'_T|x_{-D}^0) \), respectively \( b_s = b_s(x'_T|x_{-D}^0) \), is the number of times that \( x_r = 0 \), respectively \( x_r = 1 \), in \( x'_T \) for \( 1 \leq \tau \leq T \) such that \( x^\tau_{-D} = s \) for \( s \in \mathcal{T}_D \).

The second inequality is obtained as in [8, eq. (23)], where the parameter redundancy is bounded. Observe, however, that here

\[
\sum_{s \in \mathcal{S}} (a'_s + b'_s) = T - \Delta_S(x'_T).
\]

In the case of a known tree model, if the past symbols \( x_{-D}, x_{-D+1}, \ldots \) are available to the encoder and the decoder, the (parameter) redundancy can be upper-bounded (see [8, eq. (23)]) by \( |\mathcal{S}| \gamma (T/|\mathcal{S}|) \). Therefore, regarding these upper bounds, we may conclude that noting the past symbols costs at most \( \Delta_S(x'_T) \) bits, if the encoder and the decoder do know the model \( \mathcal{S} \). We say that the starting redundancy is upper-bounded by \( \Delta_S(x'_T) \). Note that \( \Delta_S(x'_T) \leq \Delta_S^{\max} \leq D \) since we assume that \( \mathcal{S} \in \mathcal{C}_D \). Therefore, this method never performs worse than the (straightforward) method that was described in the previous subsection.

### C. Unknown Model

Again suppose that the model \( \mathcal{S} \in \mathcal{C}_D \). In this subsection we will show that, also in this case where the encoder and decoder do not know the model \( \mathcal{S} \), we lose not more than \( \Delta_S(x'_T) \) bits, if they do not have access to the past symbols \( x_{-D}, x_{-D+1}, \ldots, x_0 \). In other words, the starting redundancy is again not more than \( \Delta_S(x'_T) \).

We demonstrate this by modifying the basic context-tree weighting method.

The starting point of this modification is that the encoder and decoder assign to all the unknown past symbols \( x_{-D}, x_{-D+1}, x_{-D+2}, \ldots \) and \( x_0 \), the value \( \varepsilon \). The value \( \varepsilon \) is the indeterminate symbol value. Because of the alphabet extension we must change the binary context tree \( \mathcal{T}_D \) into a ternary context tree \( \mathcal{T}_D \). A node in this tree corresponds to a string of symbols from the alphabet \( \{0, \varepsilon, 1\} \). To each node \( s \) in this ternary context tree, there corresponds a count \( \tilde{a}_s(x'_T|\varepsilon^D) \) that denotes the number of zeros that occur in \( x'_T \) at instants \( 1 \leq \tau \leq t \), such that \( s \) is a suffix of \( \varepsilon^D x^\tau_{-D} \), and a count \( \tilde{b}_s(x'_T|\varepsilon^D) \) that stands for the number of ones in \( x^\tau_{-D} \) at instants \( 1 \leq \tau \leq t \), where \( s \) is a suffix of \( \varepsilon^D x^\tau_{-D} \), for all \( s \in \mathcal{T}_D \).

The definition of the weighted distribution is now slightly different from the basic one given in [8].

\[
\log \frac{P_s(x'_T|x_{-D}, \mathcal{S}, \Theta_{\mathcal{S}})}{P_{11}(x'_T|\mathcal{S})} = \log \frac{\prod_{s \in \mathcal{S}} (1 - \theta_s) \theta_s b_s}{2^{-\Delta_S(x'_T)} \prod_{s \in \mathcal{S}} P_s(a'_s, b'_s)} \leq \log \frac{\prod_{s \in \mathcal{S}} (1 - \theta_s) \theta_s b_s}{2^{-\Delta_S(x'_T)} \prod_{s \in \mathcal{S}} P_s(a'_s, b'_s)} \leq \sum_{s \in \mathcal{S}} \log \frac{(1 - \theta_s) \theta_s b_s}{P_s(a'_s, b'_s)} + \Delta_S(x'_T) \leq |\mathcal{S}| \gamma (T/|\mathcal{S}|) + \Delta_S(x'_T).
\]

(5)

\[\text{Definition 1:} \quad \text{To each node} \ s \in \mathcal{T}_D \text{ we assign a weighted probability which is defined as}\]

\[
\tilde{P}_w^{\varepsilon} = \begin{cases} \tilde{P}_w(a_s, b_s) + \frac{1}{3} \tilde{P}_w^{\varepsilon}(x'_T|\varepsilon^D), & \text{for} \ 0 \leq l(s) < D \\ \tilde{P}_w(a_s, b_s), & \text{for} \ l(s) = D. \end{cases}
\]

(6)

The corresponding coding distribution is defined as

\[
\tilde{P}_c(x'_T|\varepsilon^D) = \tilde{P}_w^{\varepsilon}(x'_T|\varepsilon^D), \quad \text{for all} \ x'_T \in \{0,1\}^T, \ t = 0, 1, \ldots, T.
\]

(7)

Just like before, we can prove that this coding distribution satisfies [8, eq. (6)], i.e., that it is a probability distribution. Moreover it is sequentially updatable. And again the computational and storage complexity needed to update this distribution is not larger than linear in \( T \). Before we continue with deriving an upper bound on the redundancy of this modified context-tree weighting method, we give an example.

**Example 1:** Suppose that a binary source generated the sequence \( x'_T = 0110100 \). For \( D = 2 \) we have plotted the context tree \( \mathcal{T}_D \) in Fig. 1. Node \( s \in \mathcal{T}_D \) which is equal to the Krichevsky–Trofimov estimate \( P_s(a_s, b_s) \), and the weighted probability \( \tilde{P}_w^{\varepsilon}(x'_T|\varepsilon^D) \). Nodes at depth \( D = 2 \) are listed only with their weighted probability \( \tilde{P}_w^{\varepsilon}(x'_T|\varepsilon^D) \) which is equal to the Krichevsky–Trofimov estimate \( P_s(a_s, b_s) \). There the coding probability \( P_s(x'_T) = \tilde{P}_w^{\varepsilon}(x'_T|\varepsilon^D) \) for the sequence 0110100 turns out to be 9/4096.

Note that for a context path to have nonzero counts it is necessary to contain no \( \varepsilon \)'s at all, or to have the structure \( \varepsilon^D x_1 x_2 \cdots \varepsilon^D x_{D-d} \) for \( d = 1, D \). In the latter case, the sum of the counts cannot exceed 1. There are only \( D \) such paths. In the figure, we see a path \( \varepsilon \varepsilon \) (leaving the root node), and a path \( \varepsilon \varepsilon \varepsilon \).

We are now ready to state a theorem that upper-bounds the redundancy of this new weighting method. First we repeat [8, Definition 2]. The cost \( \Gamma_{\mathcal{D}}(\mathcal{S}) \) of a model \( \mathcal{S} \) with respect to class \( \mathcal{C}_D \) is defined as

\[
\Gamma_{\mathcal{D}}(\mathcal{S}) \triangleq |\mathcal{S}| - 1 + \left| \{s \in \mathcal{S} : l(s) \neq D \} \right|
\]

(8)

where it is assumed that \( \mathcal{S} \in \mathcal{C}_D \).
Theorem 1: For any tree source with unknown model $S \in \mathcal{C}_D$ and unknown parameter vector $\theta_S$, the individual redundancies with respect to $(S, \theta_S)$ are upper-bounded by

$$r(x_i^T | x_{i-1}^d, S, \theta_S) \leq L(x_i^T) - \frac{1}{P_n(x_i^T | x_{i-1}^d, S, \theta_S)}$$

$$< \Gamma_D(S) + |S| \gamma \left( \frac{T - \Delta_S(x_i^T)}{|S|} \right) + \Delta_S(x_i^T) + 2 \quad (9)$$

for all $x_i^T \in \{0, 1\}^T$, for all sequences of past symbols $x_{i-1}^d$, if we use the coding distribution specified in (7).

Proof: Consider a sequence $x_i^T \in \{0, 1\}^T$. As in [8, Proof of Theorem 2], we split the individual redundancy into three terms

$$r(x_i^T | x_{i-1}^d, S, \theta_S) = L(x_i^T) - \log \frac{P_n(x_i^T | x_{i-1}^d, S, \theta_S)}{P_n(x_{i-1}^d | S)} + \log \frac{P_n(x_{i-1}^d | S)}{P_n(x_{i-1}^d | S)}$$

$$+ \log \frac{P_n(x_i^T | x_{i-1}^d, S, \theta_S)}{P_n(x_i^T | x_{i-1}^d, S, \theta_S)} + \left( L(x_i^T) - \log \frac{1}{P_n(x_i^T)} \right) \quad (10)$$

The last term in (10), the coding redundancy term, is upper-bounded by 2. For the middle term, the parameter plus starting redundancy term, we use the bound given by (5)

$$\log \frac{P_n(x_i^T | x_{i-1}^d, S, \theta_S)}{P_n(x_{i-1}^d | S)} \leq |S| \gamma \left( \frac{T - \Delta_S(x_i^T)}{|S|} \right) + \Delta_S(x_i^T) \quad (11)$$

What remains to be investigated is the first term, the model redundancy term. We can lower-bound $P_n(x_i^T | S)$ in terms similar to the terms that form $P_n(x_i^T | S)$ in (1), as we shall soon see.

Consider the tree model $S$ which we know to be contained in the context tree $T_D$. Leaves of $S$ are nodes $s$ in the context tree $T_D$ for which $s \in S$ and internal nodes $s'$ of $S$ are nodes in $T_D$ that are a suffix of some string $s \in S$, $s' \neq s$.

First observe that for nodes $s \in S$ we have that

$$\tilde{a}_s(x_i^T | D) = \tilde{a}_s(x_i^T),$$

and

$$\tilde{b}_s(x_i^T | D) = \tilde{b}_s(x_i^T). \quad (12)$$

Next note that for $s \in T_D$

$$\tilde{P}_n(x_i^T | D) \geq \begin{cases} \frac{1}{2} P_n(a'_s(x_i^T), b'_s(x_i^T)), & \text{if } s \text{ is a leaf of } S \text{ with } l(s) \neq D \\ P_n(a'_s(x_i^T), b'_s(x_i^T)), & \text{if } s \text{ is a leaf of } S \text{ with } l(s) = D \\ \frac{1}{2} \tilde{P}_n(x_i^T | D) \tilde{P}_n(x_i^T), & \text{if } s \text{ is an internal node of } S \text{, and } x_i^{t(s)} \neq s \\ \frac{1}{2} \tilde{P}_n(x_i^T | D) \tilde{P}_n(x_i^T), & \text{if } s \text{ is an internal node of } S \text{, and } x_i^{t(s)} = s. \end{cases} \quad (13)$$

Combining these inequalities, starting in the leaves of $S$ and working towards the root of the context tree we see that we lose 1 bit in each internal node $s$ of $S$ and leaf $s \in S$ not at depth $D$, which adds up to $\Gamma_D(S)$ in total. In addition to this we get an increase of 1 bit in internal nodes of $S$ that occur as a prefix of the source sequence $x_i^T$. Such a one-bit increase corresponds to a missing context or, in other words, to an uncoded symbol. These additional costs add up to $\Delta_S(x_i^T)$.

We get as lower bound for the coding probability

$$\tilde{P}_n(x_i^T) = \tilde{P}_n(x_i^T | D) \geq 2^{-T_\Delta(S)} - \Delta_S(x_i^T)$$

$$\prod_{s \in S} P_n(a'_s(x_i^T), b'_s(x_i^T)). \quad (14)$$

Combining this with (1), as in [8, eq. (25)], this results in the following upper bound for the model redundancy:

$$\log \frac{P_n(x_i^T | S)}{P_n(x_i^T)} \leq \Gamma_D(S). \quad (15)$$

Substitution of (11), (15), and the 2 bits for the coding redundancy in (10) finally leads to the theorem.

Theorem 1 states that also in the case where the past symbols are not available to the encoder and the decoder, the loss of not knowing the model $S$ is bounded by $\Gamma_D(S)$ bits. This is completely identical to the basic context-tree weighting result. In both the known and the unknown model situation, the loss of not having access to the past symbols $x_{i-1}^d, x_{i-2}^d, \cdots, x_0$, i.e., the starting redundancy, is never more than $\Delta_S(x_i^T)$.

Although the context tree $T_D$ is, in principle, ternary, it is possible to show that $\tilde{P}_n(x_i^T)$ is a weighting over coding distributions $P_n(x_i^T | S)$ for all binary-tree models $S \in \mathcal{C}_D$. This is shown in the Appendix.

Although the coding distribution (6) suggests the use of a ternary context tree, this is not necessary at all. Since only binary contexts and contexts of the form $e_i^d, e_i^{d-1}$ for $d \geq 1$, $D$ can actually occur, it suffices to implement a binary context tree in which each node contains a Boolean variable that indicates whether or not this node has a tail. A node $s \in T_D$ has a tail, if $x_i^{t(s)} = s$, and $l(s) < D$. The tail corresponds to the missing part of the context. If a node $s$ has a tail, the product of the weighted probabilities of the children $0$'s and $1$'s of this node should be multiplied by $1/2$.

III. INFINITE-DEPTH CONTEXT-TREE WEIGHTING

In the previous section, we have dealt with a first unpleasant property of the basic context-tree weighting method, the fact that the past symbols were needed by the encoder and the decoder. A second shortcoming of the basic method is that the depth $D$ of the context tree $T_D$ is assumed to be finite. Only for models that fit into this finite-depth context tree, the weighting method achieves desirable redundancy bounds. The second result in this correspondence concerns a generalization of the basic context-tree weighting method to the situation where the context-tree depth is not bounded. In this case, we can still achieve a storage complexity (number of stored records) which does not increase faster than linear in the sequence length $T$.

The first observation that leads to this result is that after having processed the source sequence $x_i^T$, we have seen $t$ semi-infinite contexts $\cdots e_{i-\tau}x_{\tau} \cdots$ for each $\tau = 1, t$ if we assume an infinite-depth context tree. It is important to note that all these contexts differ from each other.

In Fig. 2, we have depicted all these contexts up to the last $e$-edge for $x_i^T = 110100$. We can observe the contexts $\cdots e\cdot x_{i-1}$, one for each $\tau = 1, t$ if we assume an infinite-depth context tree. It is important to note that all these contexts differ from each other.

Note that as in the previous section we assume that $x_{i-1} = 110100$, and $x_{i-1} = 110100$ for $x_0$. Note that in the previous section we assume that $x_{i-1} = 110100$. Instead of labeling the edges that connect the nodes with values from the alphabet $\{0, 1\}$ we label them with the time-index of the symbol in the last context that went through this edge. For example, the last context $\cdots e_{i-10}x_{i-10}$ was formed by symbols $\cdots x_{i-11}x_{i-10}x_{i-7}e_5$, and therefore the indices $\cdots e_{i-10}x_{i-10}$ are found along the path representing this last context. The context corresponding to symbol $x_{i-1}$ was $\cdots e_{i-10}$. Therefore, after having processed symbol $x_{i-10}$, the edges corresponding to context $\cdots e_{i-11}$ were labeled with $\cdots e_{i-10}$. While processing the symbol $x_{i-10}$ with context $e_{i-10}$, the last two labels on this path were updated again and changed from 23 into 45. Note that a straightforward
implementation of the structure that we have just described would yield a number of nodes that grows quadratically in $T$.

However, a second observation is that all unique nodes that correspond to the same context are equivalent and can be replaced by a single record (see Fig. 3). A node $s$ is said to be unique if $s$ occurs only once as a context in $\cdots \varepsilon x_1 x_2 \cdots x_1$. For example, for source sequence $110100$ the nodes $\cdots , \varepsilon 11010, 11010, 1010, 010,$ are all unique and correspond to the same context (i.e., $11010$). They all can be replaced by a single record (we call this a leaf-record). We can label this record with the node closest to the root among the equivalent nodes. Furthermore, this record should contain a pointer to the position in the source sequence where the segment corresponding to the equivalent nodes occurs. This segment is formed by the unique edges in the context. The pointer contains the index of the most recent (closest to the root) edge in the corresponding source segment. This segment is formed by the edges through which only all contexts in the mentioned set pass. In the figure, the segment corresponding to contexts $\cdots \varepsilon 110$ and $\cdots \varepsilon 11010$ is formed by the edges 1 and 0 labeled with $x_4$ and $x_5$. The most recent edge in the segment is $x_5$. The pointer is therefore 5. Also the length of the corresponding segment should be stored now (two edges, $x_4$ and $x_5$ in the figure, the length is therefore 2).

In addition to the pointer to the most recent edge (symbol) of the corresponding source segment, the length of that segment, and the $a$ and $b$ counts, an internal record contains pointers to its 0-, $\varepsilon$-, and 1-child records. Two of these pointers are non-nil so an internal node has two or three children. Updating the tree with a new context always creates a new leaf record. Therefore, after having processed the entire sequence $x_1^n$, the total number of produced leaf records will be $T$ while the number of internal records is at most $T - 1$. This results in a storage complexity ($< 2T - 1$ records) which grows not faster than linear in the source sequence length $T$. Note that also the sequence itself should be stored but this is also linear in $T$. It should be mentioned here that the described implementation of the context tree strongly relates to the DAWG concept proposed by Blumer et al. [1].
Apart from maintaining the context-tree structure, the counts \( \tilde{a}_s \) and \( \tilde{b}_s \) (note that now \( s \in \mathcal{T}_w \)), the estimated probabilities \( \tilde{P}_w(\tilde{a}_s, \tilde{b}_s) \), and the weighted probabilities \( \tilde{P}_w, \infty \) should be updated. In accordance to (6), taking \( D = \infty \), we can now define the weighted distribution and the resulting coding distribution.

**Definition 2:** To each node \( s \in \mathcal{T}_w, \infty \), assign a weighted probability which is defined as

\[
\tilde{P}_w, \infty (s) = \left\{ \begin{array}{ll}
\frac{1}{2} P_a(\tilde{a}_s, \tilde{b}_s) + \frac{1}{2} P_\infty(s) & \text{for nonunique } s \\
\frac{1}{2} \gamma & \text{for unique } s.
\end{array} \right.
\]

(16)

The corresponding coding distribution is defined as

\[
\tilde{P}_w(x_1^T) \triangleq \tilde{P}_w, \infty (s) \mathbb{1}(x_1^T),
\]

for all \( s \in \{0, 1\}^T \), \( t = 0, 1, \ldots, T \).

(17)

First note that equivalent nodes all have the same counts and therefore the same estimated probability. Note also that since the estimated probabilities of the equivalent nodes in a leaf record would all be equal to \( 1/2 \), the weighted probability of all these nodes is also \( 1/2 \). In an internal record the situation is slightly more complicated. Although the estimated probabilities are equal for all equivalent nodes, this does not hold for the weighted probabilities of these nodes. They are, however, easy to calculate and only the weighted probability corresponding to the node closest to the root is actually needed (by its parent). For example, for the equivalent nodes 10 and 0 in Fig. 3, the weighted probabilities are

\[
\tilde{P}_w, \infty (s) = \frac{1}{2} P_\infty(s) + \frac{1}{2} \tilde{P}_{10, w, \infty},
\]

and

\[
\tilde{P}_w, \infty (s) = \frac{1}{2} P_\infty(s) + \frac{1}{2} \tilde{P}_{0, w, \infty},
\]

respectively. Here

\[
P_\infty(s) = P_\infty(\tilde{a}_0, \tilde{b}_0) = P_\infty(\tilde{a}_0, \tilde{b}_0).
\]

Only \( \tilde{P}_w, \infty \) is really needed (by the root record labeled \( \lambda \)).

The individual redundancy of this infinite-depth context-tree weighting method is as expected. The only thing that changes is the cost of a model which is \( 2|S| - 1 \) bits for all \( S \) now.

**Theorem 2:** For any tree source with unknown model \( S \) and unknown parameter vector \( \Theta_S \), the individual redundancies with respect to \( (S, \Theta_S) \) are upper-bounded by

\[
\rho(x^T | x_\infty \infty, S, \Theta_S) \triangleq L(x^T) - \log \frac{1}{P_\infty(x^T)} \\
\leq 2|S| - 1 + |S| \gamma \left( T - \Delta_S(x^T) \right) \\
+ \Delta_S(x^T) + 2,
\]

(18)

for all \( x^T \in \{0, 1\}^T \), for all sequences \( x_\infty \infty = \cdots x_{t-1} x_0 \) of past symbols, if we use coding distribution (17).

**IV. ACHIEVING ENTROPY FOR ARBITRARY STATIONARY ERGODIC SOURCES**

Now that we can use the context-tree weighting algorithm for arbitrary-depth tree sources and without having access to the past symbols \( \cdots x_{t-1} x_0 \), we can show that this method achieves entropy for arbitrary stationary and ergodic sources.

**Theorem 3:** For any binary stationary and ergodic source

\[
\lim_{T \to \infty} \frac{L(x^T)}{T} \leq H_\infty(X) \text{ with probability one}
\]

(19)

if we use the coding method and distribution (17) presented in the previous section.

**Proof:** We start this proof with the statement

\[
\tilde{P}_w, \infty (x^T) = \tilde{P}_w, \infty (x^T) \geq 2^{1 - 2|S| - \Delta_S(x^T)} \\
\prod_{s \in S} P_a(s | x_s^T, x_{s'}^T)
\]

(20)

which holds for all tree models \( S \). This inequality is identical to the statement (14), however, for the infinite-depth context-tree weighting method the cost of a model \( S \) is \( 2|S| - 1 \) instead of \( \Gamma_0(S) \). If we combine this with the arithmetic coding result (see [8, Theorem 1]) and use the fact that

\[
a_s(x^T) \leq a_s(x^T | x_0^{\infty})
\]

and

\[
b_s(x^T) \leq b_s(x^T | x_0^{\infty})
\]

we obtain lower bounds on the codeword length, one for each tree model \( S \).

\[
L(x^T) \leq \log \frac{1}{P_\infty(x^T)} + 2
\]

\[
= \log \prod_{s \in S} P_a(s | x_s^T, x_{s'}^T) + 2|S| - 1 + \Delta_s(x^T) + 2
\]

\[
\leq \log \prod_{s \in S} P_a(s | x_s^T, x_{s'}^T) + \Delta_s(x^T) + 2
\]

(21)

For the terms \( P_a(\cdot, \cdot) \) we now apply the lower bound [8, eq. (10)] which states (for \( a + b \geq 1 \)) that

\[
P_a(a, b) \geq \frac{1}{\sqrt{a + b}} \left( \frac{a}{a + b} \right)^a \left( \frac{b}{a + b} \right)^b.
\]

(22)

For \( a = b = 0 \) we have that \( P_a(a, b) = 1 \). Denoting \( a_s(x^T | x_0^{\infty}) \) by \( a_s \) and \( b_s(x^T | x_0^{\infty}) \) by \( b_s \), this leads to the upper bound

\[
\log \prod_{s \in S} P_a(a_s, b_s) \leq \sum_{s \in S} \left( a_s \log \frac{a_s + b_s}{a_s} + b_s \log \frac{a_s + b_s}{b_s} \\
+ \frac{1}{2} \log (a_s + b_s) + 1 \right)
\]

\[
= \sum_{s \in S} \left( a_s \log \frac{a_s + b_s}{a_s} + b_s \log \frac{a_s + b_s}{b_s} \right)
\]

\[
+ \sum_{s \in S} \left( \frac{1}{2} \log (a_s + b_s) + 1 \right)
\]

\[
\leq \sum_{s \in S} \left( a_s \log \frac{a_s + b_s}{a_s} + b_s \log \frac{a_s + b_s}{b_s} \right) + |S| \gamma \left( T | S \right)
\]

(23)

Note that by convention \( 0 \log(0) = 0 \). The last inequality follows from arguments as in [8, eq. (23)]. If we now combine (21) and (23) we obtain

\[
L(x^T) \leq \sum_{s \in S} \left( a_s \log \frac{a_s + b_s}{a_s} + b_s \log \frac{a_s + b_s}{b_s} \right)
\]

\[
+ |S| \gamma \left( T | S \right) + 2|S| - 1 + \Delta_S(x^T) + 2.
\]

(24)

Now assume that \( d = 0, 1, 2, \ldots \) and let \( S \) be determined by \( d \). For \( d = 0 \) let \( S = \{ \lambda \} \), i.e., a (memoryless) tree with only a root node. For \( d = 1, 2, \ldots \), take \( S = \{0, 1\}^n \), i.e., a full \( d\)-th-order Markov tree with depth \( d \).
Fix some $d$: Then for $T \geq |S| = 2^d$, and noting that $\Delta_S (x^T_I) \leq d$, we have that

$$L(x^T_I) \leq \sum_{s \in S} \left( a_s \log \frac{a_s + b_s}{a_s} + b_s \log \frac{a_s + b_s}{b_s} \right) + 2^{d-1} \log \frac{T}{2^2} + 3 \cdot 2^d + d + 1. \tag{25}$$

Note that $a_s$ and $b_s$ are now the number of zeros respectively ones in $x^T_I$ that follow context $s \in \{0, 1\}^d$.

Next let

$$c(T, d) \triangleq 2^{d-1} \log \frac{T}{2^2} + 3 \cdot 2^d + d + 1. \tag{26}$$

Then, from this definition and (25), we obtain that

$$\frac{L(x^T_I)}{T} \leq \sum_{s \in S} \left( -a_s \log \frac{a_s}{T} - b_s \log \frac{b_s}{T} \right)$$

$$- \left( -a_s \log \frac{a_s}{T} - b_s \log \frac{b_s}{T} \right) + c(T, d). \tag{27}$$

From the ergodic theorem (see, e.g., Shields [4]), since the actual source is stationary and ergodic, we know for $s \in \{0, 1\}^d$ that

$$\lim_{T \to \infty} \frac{a_s}{T} = P_s(X_k^T = s, X_{k+1} = 0)$$

and

$$\lim_{T \to \infty} \frac{b_s}{T} = P_s(X_k^T = s, X_{k+1} = 1) \tag{28}$$

with probability one. Moreover, for all $d = 0, 1, 2, \ldots$

$$\lim_{T \to \infty} \frac{c(T, d)}{T} = 0. \tag{29}$$

Therefore, with probability one

$$\limsup_{T \to \infty} \frac{L(x^T_I)}{T} \leq H(X_1, X_2, \ldots, X_d, X_{d+1})$$

$$\leq H(X_1, X_2, \ldots, X_d)$$

$$= H(X_{d+1}|X_1, X_2, \ldots, X_d). \tag{30}$$

This bound holds for all $d = 0, 1, 2, \ldots$. Since

$$\lim_{d \to \infty} H(X_{d+1}|X_1, X_2, \ldots, X_d) = H_\infty(X)$$

we may conclude that for all binary stationary and ergodic sources the codeword length $L(x^T_I)$ divided by the sequence length $T$ is, with probability one, not larger than the entropy $H_\infty(X)$ of the source.

**V. SOME REMARKS**

Coding schemes for the class of stationary sources were probably first studied by Shiharkov and Babkin [5]. They showed, using a combinatorial approach, that over this class the average codeword length converges to the source entropy.

We have shown here that the extended version of the context-tree weighting algorithm achieves entropy for all binary stationary and ergodic sources in the sense that with probability one the compression ratio

$$c(T, d) \leq \log \frac{T}{2^2} + 3 \cdot 2^d + d + 1.$$

This bound holds for all $d = 0, 1, 2, \ldots$. Since

$$\lim_{d \to \infty} \log \frac{T}{2^2} + 3 \cdot 2^d + d + 1 = \infty$$

we may conclude that for all binary stationary and ergodic sources the codeword length $L(x^T_I)$ divided by the sequence length $T$ is, with probability one, not larger than the source entropy $H_\infty(X)$ for $T \to \infty$. A similar result was proved for the Ziv–Lempel incremental parsing procedure (tree algorithm) presented in [10].

Moreover, Ziv and Lempel showed that for any finite-state code the achievable compression ratio for an individual infinite-length sequence is lower-bounded by the limit of the empirical normalized block entropy $H(x_1, x_2, \ldots, x_d)/|S|$ of this sequence for $d \to \infty$ (see [10, Theorem 3]). The extended context-tree weighting algorithm achieves a compression ratio which is upper-bounded by the limit of the empirical conditional entropy $H(x_{d+1}|x_1, x_2, \ldots, x_d)$ for $d \to \infty$ of the individual infinite length source sequence. Both the empirical normalized block entropy and the empirical conditional entropy have the same limit so the method presented here is optimal in this sense. This also holds for the Ziv–Lempel incremental parsing method (see [10, Theorem 2]).

The storage complexity of the method that we have described in this manuscript turned out to be not larger than linear in the source sequence length $T$. It should be mentioned that the storage complexity of the Ziv–Lempel incremental parsing algorithm is smaller, and behaves roughly like $T/\log(T)$. For the computational complexity the comparison is similar. The Ziv–Lempel method visits a new node in the dictionary tree for each processed source symbol, while for the context-tree weighting method it is necessary to go from the root of the context tree to a leaf for each processed symbol. Wyner and Ziv [9] showed that $\log(t)$ divided by the number of nodes that are visited in the context tree for processing $x_t$ converges in probability to $H_\infty(X)$. Stronger results appear in Ornstein and Weiss [3] and Szpakowski [6].

Although it is possible to extend the context-tree weighting method to the nonbinary case, we emphasize that here only coding for binary sources is considered.

**APPENDIX**

**WEIGHTING OVER THE BINARY-TREE MODELS**

We will show here that, although the context tree $\hat{T}_D$ is in principle ternary, $\hat{P}_s(x^T_I)$ is a weighting over coding distributions $P_s'(x^T_I | S)$ for all binary-tree models $S \in C_D$. To see this, first consider, e.g., a binary-tree model $S = \{0, 1, 0, 1\}$ (see Fig. 4). A ternary-tree model that coexists with this binary model $S$ is, e.g.,

$$S' = \{00, 00, 0, 00, 01, 00, 01, 00, 01\}.$$ A ternary model $S'$ coexists with binary model $S$ if $S \subset S'$. The cost $\Gamma_D(S)$ of the ternary model $S'$ with respect to class $C_D$ of ternary models is defined as

$$\Gamma_D(S) = \frac{1}{2} |S| - 1 + |\{s: s \in S, |s| \neq D\}| \tag{31}$$

where it is assumed that $S \in C_D$. Analogous to the binary case (see [8, Lemma 2]) we can show that $\hat{P}_s(x^T_I)$ is a weighting over all ternary models $S$. The weight of a ternary model $S$ is $2^{-|S|} \rho(S)$. We can, therefore, write

$$\hat{P}_s(x^T_I) = \sum_{S \in C_D} 2^{-|S|} \rho(S) \prod_{s \in S} P_s'(a_s(x^T_I) | D), b_s(x^T_I) | D)$$

$$= \sum_{S \in C_D} \sum_{S' \supseteq S} 2^{-|S'|} \rho(S')$$

$$\cdot 2^{-\Delta(S')} \prod_{s \in S} P_s'(a_s(x^T_I), b_s(x^T_I))$$

$$= \sum_{S \in C_D} 2^{-|S|} \rho(S) P_s'(x^T_I | S). \tag{32}$$
Note that there is only one (underlying) binary model $\mathcal{S}$ that ternary model $\mathcal{S}$ can coexist with. The weights of all ternary models $\mathcal{S} \in \mathcal{C}_2$ that have underlying model $\mathcal{S}$ (i.e., models $\mathcal{S} \rightarrow \mathcal{S}$) sum up to $2^{-1}p(\mathcal{S})$. Furthermore, observe that
\[
\prod_{s \in \mathcal{S}} P_r(\hat{a}_{s}, \hat{b}_{s}) = 2^{-\Delta_{\mathcal{S}}(x_r^T)}
\]
since the nodes $s$ in this product must accommodate the $\Delta_{\mathcal{S}}(x_r^T)$ missing contexts. Finally, (1) is used to obtain the last equality.

Equation (32) can be used to give an alternative proof of (15). To see this note that
\[
\hat{P}_r(x_r^T) \geq 2^{-1\mathcal{P}(s)} P_r(x_r^T | S).
\]

ACKNOWLEDGMENT

The author wishes to thank Associate Editor M. Feder and a reviewer for the comments which led to improvements of the present correspondence.

REFERENCES


