Estimation of the Likelihood for Context Set Models

Hee Sun Kim* Zsolt Talata†

Abstract

Finitely-valued stationary time series are described by the collection of the conditional probabilities of the possible values given the infinite pasts. The concept of context is extended to be an arbitrary part – not necessarily a continuous ending – of the past that determines the transition probability. The context set model of the time series consists of the collection of all contexts and the corresponding transition probabilities. The likelihood is estimated from a sample using a double mixture over the possible models and their parameters. An optimality of the estimator is proved and an algorithm is shown to calculate the estimator in reasonable time despite the very large number of possible models.

Key Words: context set; context tree; Markov chain; time series; statistical estimation; double mixture

1. Source Coding

The problem of lossless source coding in information theory can be formulated in statistical terms. An information source emitting symbols from a finite alphabet is equivalent to a time series $X_1, X_2, \ldots$ taking values from a finite set A. Then a message, a length-n sequence of emitted symbols, is equivalent to a statistical sample $x_1, \ldots, x_n = x_1^n$. Codes assign binary strings to messages. The length of the binary string is the code length. If the distribution Q of the source is known, there are methods to construct codes approaching the ideal code length $-\log Q(x_1^n)$. If the distribution of the source is not known, it is estimated from the message so that the obtained coding distribution P_C can be used to construct codes approaching the code length $-\log P_C(x_1^n)$. This estimation problem is equivalent to the statistical estimation of the likelihood $Q(x_1^n)$ by a distribution $P_C(x_1^n)$ from the sample x_1^n. The (parameter) redundancy is the difference $-\log P_C(x_1^n) + \log Q(x_1^n)$, which is equivalent to a log-loss function.

For example, for i.i.d. processes the likelihood is $Q(x_1^n) = \prod_{a\in A} Q(a)^{N_n(a)}$, where $N_n(a)$ is the number of occurrences of $a \in A$ in x_1^n. The maximum likelihood estimate of $Q(a)$ is $N_n(a)/n$, but the maximum likelihood

$$\mathrm{ML}(x_1^n) = \prod_{a\in A} \left(\frac{N_n(a)}{n}\right)^{N_n(a)}$$

is not a possible coding distribution $P_C(x_1^n)$, because $\sum_{x_1^n} \mathrm{ML}(x_1^n)$ is not necessarily 1. Instead, the normalized maximum likelihood

$$\mathrm{NML}(x_1^n) = \frac{\mathrm{ML}(x_1^n)}{\sum_{x_1^n} \mathrm{ML}(x_1^n)}$$

is used, that can be shown to minimize the worst-case redundancy. The normalized maximum likelihood is not sequentially computable as the sample size n increases. The Krichevsky-

*Department of Mathematics, University of Kansas, 1460 Jayhawk Boulevard, Lawrence, Kansas 66045
†Department of Mathematics, University of Kansas, 1460 Jayhawk Boulevard, Lawrence, Kansas 66045

Trofimov distribution

$$\mathrm{KT}(x_1^n) = \frac{\prod_{a:N_n(a)\geq 1} \left(N_n(a)-\frac{1}{2}\right)\left(N_n(a)-\frac{3}{2}\right)\cdots\frac{1}{2}}{\left(n-1+\frac{|A|}{2}\right)\left(n-2+\frac{|A|}{2}\right)\cdots\frac{|A|}{2}}$$

sums up to 1 over x_1^n and has a nearly minimal redundancy, but it is also sequentially computable as

$$\mathrm{KT}(x_1^{n+1}) = \frac{N_n(x_{n+1})+\frac{1}{2}}{n+\frac{|A|}{2}}\,\mathrm{KT}(x_1^n).$$

This paper considers the above source coding problem for processes with memory.

2. Context Set Model

A stationary ergodic source over a finite alphabet A has finite memory if the conditional probability of the next symbol x_i given the infinite past $\ldots x_{i-2}x_{i-1}$ depends only on a finite number k of preceding symbols $x_{i-k}\ldots x_{i-1}$. Then the process is said to be a Markov source of order k. The Krichevsky-Trofimov (KT) distribution [4] tailored to these processes can be used in arithmetic coding procedures [6]. The coding redundancy, the cost of using arithmetic coding, is negligible to the parameter redundancy, the cost of not knowing the actual distribution. Thus, the KT distribution for messages $x_1^n \in A^n$ provides a universal code for Markov sources of order k as its worst case maximum redundancy is $\frac{1}{2}(|A|-1)|A|^k \log n + \mathcal{O}(1)$, which is the smallest possible, up to an additive constant [2].

The tree source [12, 13] allows the number k of the relevant preceding symbols to vary with the past $\ldots x_{i-2}x_{i-1}$. The strings of these relevant suffixes $s = s_{-l(s)}\ldots s_{-1}$ of the past are called contexts [5] and their lengths $l(s)$ may be substantially shorter than the Markov order. The set of all contexts for a source can be represented by a tree graph, and is called context tree. The KT distribution tailored to a context tree $\mathcal{S}$ improves [2] the worst case maximum redundancy to $\frac{1}{2}(|A|-1)|\mathcal{S}|\log n + \mathcal{O}(1)$ for the processes with context tree $\mathcal{S}$ as the number of contexts $|\mathcal{S}|$ may be smaller than $|A|^k$.

Given universal codes for countable number of models, weighting these coding distributions is known to provide a twice-universal code [8] over all models. If only the maximum memory length k of the source is known but its context tree is not, then calculating a mixture of the coding distributions over all possible context trees in a direct way would be infeasible because of the large number of possible context trees. The Context Tree Weighting (CTW) method [15] finds the mixture distribution in an efficient way and provides a code with the worst case maximum redundancy upper bounded by $\Gamma_k(\mathcal{S})+|\mathcal{S}|\gamma\left(n/|\mathcal{S}|\right)+2$ for all tree sources with context lengths at most k, in case $|A|=2$. Here, $\gamma(z)=\frac{1}{2}\log z+1$ if $z\geq 1$ and $\gamma(z)=z$ if $z<1$, and $\Gamma_k(\mathcal{S})=|\mathcal{S}|-1+|\{\text{internal nodes of } \mathcal{S}\}|$. The model redundancy $\Gamma_k(\mathcal{S})$, the cost of not knowing the actual context tree $\mathcal{S}$, is negligible to the parameter redundancy.

If the maximum memory length k of the source is not known, an extension of the CTW method is available for binary sources [14]. The extended method also allows the source to have infinite memory. The symbols whose context is not available in the message remain uncoded, that introduces a starting redundancy $\Delta_{\mathcal{S}}(x_1^n)$. The method provides a code with the worst case maximum redundancy upper bounded by

$$2|\mathcal{S}|-1+|\mathcal{S}|\gamma\left(\frac{n-\Delta_{\mathcal{S}}(x_1^n)}{|\mathcal{S}|}\right)+\Delta_{\mathcal{S}}(x_1^n)+2$$

for all tree sources.

The context tree model is a parsimonious parametrization of the Markov model as it merges a set of k-length pasts $\mathcal{S}(s_{-k'},\dots,s_{-1}) = \{s_{-k},\dots,s_{-1} : s_i \in A, -k \le i \le -k'-1\}$, $k' \le k$, together to their common suffix $s' = s_{-k'},\dots,s_{-1}$ if $Q(a|s)$, $s \in \mathcal{S}(s_{-k'},\dots,s_{-1})$, are equal for all $a \in A$, where Q denotes the transition probability. More efficient parametrization than the tree source can be achieved [12] by merging arbitrary pasts if they share the same transition distribution. Universal codes can be obtained efficiently by weighting the KT coding distribution over all models in this model class [16], called Class I. The parameter redundancy for these models may be much less than for the context tree models, but the model redundancy, the cost of not knowing the model, and the computational complexity are much larger because of the large number of possible models. To find a better trade-off, Class II models only allow successive splittings of the lexicographically ordered k-length pasts [16]. Furthermore, Class III models allow sequential merging of pasts that differ only in the value s_i of the arbitrary coordinate $1 \le i \le k$ [16].

The extended context tree model [9] adds a "don't care" symbol to A. A don't care symbol in the i'th coordinate of a k-length context s represents that $Q(a|s)$ does not depend on the value s_i. Dropping the don't care symbols from the end of the k-length contexts [10] makes the model equivalent to the context tree model with extending the alphabet by the don't care symbol and allows the application of the CTW method.

In this paper, we consider a new model that allows to conditionally drop $j+r$ consecutive coordinates of the k-length strings if the proceeding symbol is not in T, for some $\emptyset \subset T \subset A$, when the transition probability does not actually depend on the values of these coordinates. That is, it allows to merge a set of k-length pasts $\mathcal{S}(s_{-k},\dots,s_{-m-j-r-1}, s_{-m},\dots,s_{-1}) = \{s_{-k},\dots,s_{-1} : s_i \in A, -m-j-r \le i \le -m-1\}$ together for each $s_{-m-j-r-1} \notin T$ and all possible $s_{-k},\dots,s_{-m-j-r-2}$ if $Q(a|s)$, $s \in \mathcal{S}(s_{-k},\dots,s_{-m-j-r-1}, s_{-m},\dots,s_{-1})$, are equal for each $a \in A$, $s_{-m-j-r-1} \notin T$ and $s_{-k},\dots,s_{-m-j-r-2}$. For the proceeding symbols in T, only the j consecutive coordinates of the k-length strings are dropped. That is, it merges the set of k-length pasts $\mathcal{S}(s_{-k},\dots,s_{-m-j-1},s_{-m},\dots,s_{-1}) = \{s_{-k},\dots,s_{-1} : s_i \in A, -m-j \le i \le -m-1\}$ together for each $s_{-m-j-1} \in T$ and all possible $s_{-k},\dots,s_{-m-j-2}$ if $Q(a|s)$, $s \in \mathcal{S}(s_{-k},\dots,s_{-m-j-1},s_{-m},\dots,s_{-1})$, are equal for each $a \in A$, $s_{-m-j-1} \notin T$ and $s_{-k},\dots,s_{-m-j-2}$. This merging may be applied successively, even to the sets $\{s_{-k},\dots,s_{-m-j-r-2} : s_j \in A, -k \le i \le -m-j-r-2\}$ and $\{s_{-m},\dots,s_{-1} : s_i \in A, -m \le i \le -1\}$. The above model can describe processes whose transition distribution is determined if a certain symbol appears in a certain position of the past, and then it does not depend on the symbols between that position and the present. Some of the subsequent results were also presented at the IEEE International Symposium on Information Theory in 2016. Such models are natural in some disciplines, for example, in bioinformatics [7, 1].

Since arbitrary coordinates of the strings of the past symbols may be dropped, these relevant parts of the pasts are not necessarily consecutive sequences of symbols. Such broken strings are called stringoids. The collection of these stringoid contexts are called context set.

In case $r = 0$, the above procedure drops j consecutive coordinates unconditionally on the value s_{-m-j-1}. Models with arbitrary subsets T of A are considered, except $T = \emptyset$ and $T = A$ because these lead to the case $r = 0$. Dropping the coordinates from $-k$ to $-k'-1$ recovers context trees. Dropping coordinates not necessarily from the ending of the k-length pasts is covered by the extended context tree model.

Example 1. Let $A = \{0,1\}$ and the process be a Markov source of order 4. The set of 4-length pasts is $\mathcal{C}_0 = \{0000, 1000, 0100, 1100, 0010, 1010, 0110, 1110, 0001, 1001, 0101, 1101, 0011, 1011, 0111, 1111\}$. Suppose that the transition probabilities are equal over each of the subsets $\{0000, 0100\}$, $\{1000, 1100\}$, $\{0001, 1001, , 0101, 1101\}$, $\{0010, 1010,$

$0011, 1011\}$, and $\{0110, 1110, 0111, 1111\}$. The context tree model merges $\{0010, 1010\}$ to 010, $\{0110, 1110\}$ to 110, $\{0001, 1001, , 0101, 1101\}$ to 01, $\{0011, 1011\}$ to 011, and $\{0111, 1111\}$ to 111. Thus, the context tree consists of the 9 contexts $\mathcal{C}_0^{(1)} = \{0000, 1000,$ $0100, 1100, 010, 110, 01, 011, 111\}$, see Fig. 1.

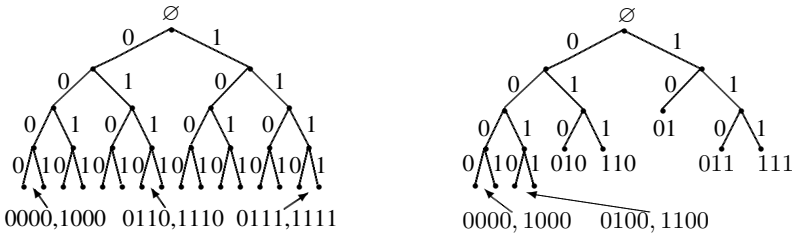

Figure 1: Graphs of $\mathcal{C}_0$ (left) and $\mathcal{C}_0^{(1)}$ (right).

The extended context tree model further merges $\{0000, 0100\}$ to the stringoid 000 with the coordinate set $\{-4, -2, -1\}$, as the coordinate -3 is dropped. In Fig. 2 (left), the notation $0\alpha00$ indicates the dropped coordinate. The set $\{1000, 1100\}$ is merged similarly, and the extended context tree $\mathcal{C}_0^{(2)}$ consists of 7 contexts. The context set model, allows to conditionally drop the coordinate -1 if the symbol at coordinate -2 is 1, as the transition probabilities are equal for $s_{-3} = 0$ and $s_{-3} = 1$, respectively, if $s_{-2} = 1$. Here, $r = 1$ and $h = 0$. Hence, the context set $\mathcal{C}_0^{(3)}$ is further reduced to 5 contexts, see Fig. 2. □

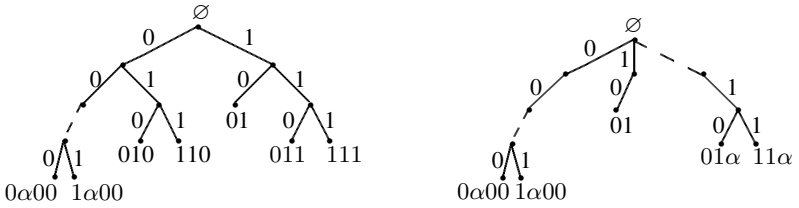

Figure 2: Graphs of $\mathcal{C}_0^{(2)}$ (left) and $\mathcal{C}_0^{(3)}$ (right).

In this paper, a universal code for the context set model class is achieved by weighting the KT distributions determined by the context sets. The coding distribution is obtained by calculating a weighted probability recursively for the possible stringoids. For extended context trees, a weighted probability is calculated for the possible strings composed from the alphabet A and the don't care symbol. These extended strings can be organized into a tree graph, whose nodes are identified with the extended strings and the children of a node are the extended strings obtained by attaching one more symbol (from A or a don't care symbol) to the string of the parent node. The weighted probability at a node is calculated from the KT distribution determined by the extended string of the node and from the weighted probabilities at the children nodes [10]. In case of context sets, the stringoids could be similarly organized to a tree graph, although not each node of such graph would be identified by a stringoid, because a stringoid's last coordinate cannot be a dropped coordinate but an extended string's last symbol can be a don't care symbol. Moreover, the weighted probability at a node identified by a stringoid is calculated from the KT distribution determined by the stringoid and from the weighted probabilities not only at the children but also at many descendant. The values of the weighted probability at these descendant are used multiple times as well.

The context set model allows to conditionally drop coordinates of strings. Namely, for the strings $s_{-k} \cdots s_{-1}$ with the same values $s_{-m}, \ldots, s_{-1}$, the coordinates $-m - r, \ldots, -m - 1$ may be dropped if $s_{-k} \cdots s_{-m-r-1} \in B$ and then are not dropped if

$s_{-k}\cdots s_{-m-r-1} \notin B$, where B is a subset of $(k-m-r)$-length strings. In case of the extended context tree model, the coordinates may be dropped or not uniformly for all $s_{-k}\cdots s_{-m-r-1}$. In case of Class III models (see above) the coordinates may be dropped or not individually for each $s_{-k}\cdots s_{-m-r-1}$. Hence, the context set model class is considerably larger than the generalized context tree model class but still allows to define a weighting method with a computational complexity significantly less than that for the Class III models [16]. In this paper, we prove that the presented universal code can be computed in a time polynomial in the message length. This version of the paper assumes that $r = 0$ or $r = 1$ in order to simplify the notation and the presentation of the results. Our results do not assume a known maximum memory length of the source and allow infinite contexts. The complete proofs of all of the results given in this paper are contained in [3].

3. Known Context Set

For a finite set A, $|A|$ denotes its cardinality. A *stringoid* composed from A is defined as $s \in A^{\mathcal{I}}$, where $\mathcal{I}$ is an arbitrary subset of the negative integers $-\mathbb{N}$. $\mathcal{I}(s)$ denotes the index set I of the stringoid. The projection of the stringoid to the coordinates with indices in an interval $[i,j]$ is denoted by s_i^j, in particular, the i-coordinate is $s_i \in A$. The number of coordinates of s is $|\mathcal{I}(s)|$, and the length of s is $l(s) = -\inf \mathcal{I}(s)$.

The empty stringoid is denoted by $\varnothing$, its length is $l(s) = 0$. Two stringoids s and u are *coherent*, denoted by $s \curlyvee u$, if $s_i = u_i$ for all $i \in \mathcal{I}(s) \cap \mathcal{I}(u)$. The composition of two stringoids s and u with disjoint index sets is $su \in A^{\mathcal{I}(s)\cup\mathcal{I}(u)}$, where $(su)_i = s_i$ if $i \in \mathcal{I}(s)$ and $(su)_i = u_i$ if $i \in \mathcal{I}(u)$.

Let $X = \{\, X_i, -\infty < i < \infty \,\}$ be a stationary ergodic process with each random variable X_i taking values from the finite set A. For a stringoid $s \in A^{\mathcal{I}}$, write $Q(s) = \mathrm{Prob}\{X_i = s_i, i \in \mathcal{I}\}$ and, if $Q(s) > 0$, $Q(a|s) = \mathrm{Prob}\{X_0 = a | X_i = s_i, i \in \mathcal{I}\}$.

Definition 1. *A stringoid $s \in A^{\mathcal{I}}$ is a* context *for the process X if $Q(s_{-i}^{-1}) > 0$ ($i = 1, 2, \ldots$) and*

$$\mathrm{Prob}\left\{X_0 = a | X_{-\infty}^{-1} = x_{-\infty}^{-1}\right\} = Q(a|s) \quad \textit{for all } a \in A,$$

whenever s is coherent with the semi-infinite sequence $x_{-\infty}^{-1} \in A^{-\mathbb{N}}$. A stringoid s with $l(s) < \infty$ and $Q(s) = 0$ is also a context.

Notice that if s is a context, then any $u \curlyvee s$ with $\mathcal{I}(u) \supseteq \mathcal{I}(s)$ is a context, too. Motivated by the definition of context, a context s can be called minimal if no $u \curlyvee s$ with $\mathcal{I}(u) \subset \mathcal{I}(s)$ is a context. For a semi-infinite sequence $x_{-\infty}^{-1}$, however, the minimal context $s \curlyvee x_{-\infty}^{-1}$ is not necessarily unique.

Definition 2. *A collection $\mathcal{C}_0$ of contexts is a* context set *of the process X if it satisfies*

1. *for each semi-infinite sequence $x_{-\infty}^{-1}$ there exists one and only one context $s \in \mathcal{C}_0$ with $s \curlyvee x_{-\infty}^{-1}$*
2. *for any contexts s and u in $\mathcal{C}_0$ with $s_{-l}^{-1} = u_{-l}^{-1}$ satisfying $-l-1 \in \mathcal{I}(s)$ and $-l-1 \notin \mathcal{I}(u)$, it follows that $-l-2 \in \mathcal{I}(s)$, $-l-2 \in \mathcal{I}(u)$, and $s_{-l-2} \neq u_{-l-2}$.*

Clearly, no two contexts in $\mathcal{C}_0$ may be coherent. Moreover, $\mathcal{C}_0$ exists for the process X, but it is not unique. Note that if we selected one minimal context $s \curlyvee x_{-\infty}^{-1}$ for each semi-infinite sequence $x_{-\infty}^{-1}$, the collection of such minimal contexts would not necessarily be a context set. The distribution Q of the process X is determined by the context set $\mathcal{C}_0$ and the parameters $Q_{\mathcal{C}_0} = \{Q(a|s) : s \in \mathcal{C}_0,\, Q(s_{-i}^{-1}) > 0 \text{ for all } i \le l(s)\}$. Denote $d(\mathcal{C}_0)$ the depth of the context set $\mathcal{C}_0$: $d(\mathcal{C}_0) = \sup\{l(s) : s \in \mathcal{C}_0\}$. Note that $d(\mathcal{C}_0)$ may be infinite.

Remark 1. Definition 2 implies that a *graph* representing a context set $\mathcal{C}_0$ can be obtained using the rooted graphs $\mathcal{F}(j)$ and $\mathcal{G}(k,T)$, where $j \geq 0, k \geq 0$ and $\emptyset \subset T \subset A$, see Fig. 3. In particular, any $\mathcal{C}_0$ can be obtained by setting $\mathcal{C}_0(0) = \{\varnothing\}$ and consecutively applying the following procedure: $\mathcal{C}_0(l+1)$ is the disjoint union of $\mathcal{C}_0(l)$ and $\mathcal{F}(j)$ or $\mathcal{G}(k,T)$ for some j, k, T with a leaf of $\mathcal{C}_0(l)$ identified by the root of $\mathcal{F}(j)$ or $\mathcal{G}(k,T)$. For example, $\mathcal{C}_0^{(3)}$ in Example 1, $\mathcal{C}_0(1) = \mathcal{G}(0,\{0\})\dot{\cup}\mathcal{C}_0(0)$, $\mathcal{C}_0(2) = \mathcal{F}(1)\dot{\cup}\mathcal{C}_0(1)$ with $00 \in \mathcal{C}_0(1)$ identified by the root of $\mathcal{F}(1)$, and $\mathcal{C}_0(3) = \mathcal{F}(0)\dot{\cup}\mathcal{C}_0(2)$ with $1\alpha \in \mathcal{C}_0(2)$ identified by the root of $\mathcal{F}(0)$, see Fig. 4.

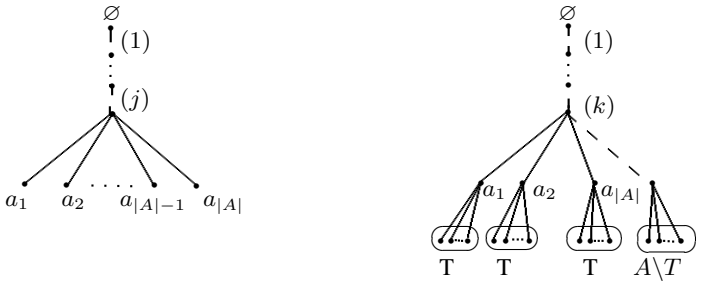

Figure 3: The rooted graphs $\mathcal{F}(j)$ (left) and $\mathcal{G}(k,T)$ (right).

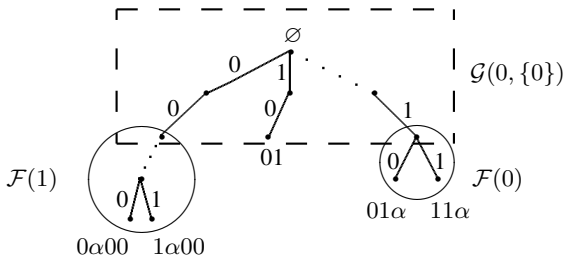

Figure 4: Graph of $\mathcal{C}_0(3)$.

In this paper, a universal code will be introduced for context sets. The pointwise redundancy of a code C for a message $x_1^n \in A^{\{1,2,\ldots,n\}}$ is

$$R_{L_C,Q}(x_1^n) = L_C(x_1^n) + \log Q(x_1^n), \tag{1}$$

where L_C is the code length.

Definition 3. *The redundancy (1) is the sum of the* coding redundancy

$$R_{L_C,P_C}(x_1^n) = L_C(x_1^n) + \log P_C\left(x_1^n\right),$$

where $P_C\left(x_1^n\right)$ is a coding distribution used to generate the code C, the model redundancy

$$R_{P_C,P_{\mathcal{C}_0}}(x_1^n) = -\log P_C\left(x_1^n\right) + \log P_{\mathcal{C}_0}\left(x_1^n\right),$$

where $P_{\mathcal{C}_0}\left(x_1^n\right)$ is a coding distribution for known $\mathcal{C}_0$ and unknown parameters $Q_{\mathcal{C}_0}$, and the parameter redundancy

$$R_{P_{\mathcal{C}_0},Q}(x_1^n) = -\log P_{\mathcal{C}_0}\left(x_1^n\right) + \log P_{\mathcal{C}_0,Q_{\mathcal{C}_0}}\left(x_1^n\right),$$

where $P_{\mathcal{C}_0,Q_{\mathcal{C}_0}}\left(x_1^n\right) = Q\left(x_1^n\right)$ is the coding distribution for known $\mathcal{C}_0$ and known $Q_{\mathcal{C}_0}$.

Here, the code redundancy can be upper bounded by 2 for, for example, arithmetic coding [6]. *If the context set is known but the parameters are unknown,* the Krichevsky-Trofimov [4] distribution can be used as follows.

Note that in this case $P_{\mathcal{C}_0}(x_1^n)$ is used for $P_C(x_1^n)$ and $R_{P_C,P_{\mathcal{C}_0}}(x_1^n) = 0$, and hence $R_{L_C,Q}(x_1^n) = R_{L_C,P_C}(x_1^n) + R_{P_{\mathcal{C}_0},Q}(x_1^n)$ denoted by $R_{L_{\mathcal{C}_0},Q}(x_1^n)$. In Section 4, if both the context set and the parameters are unknown, the Context Set Weighting method is presented and proved to provide a code C with an upper bound on the model redundancy $R_{P_C,P_{\mathcal{C}_0}}(x_1^n)$ not exceeding the order of the upper bound on the parameter redundancy $R_{P_{\mathcal{C}_0},Q}(x_1^n)$ shown below.

For a stringoid $s \in A^I$ and a letter $a \in A$, $N_n(s,a)$ denotes the number of occurrences of s. For a message $x_1^n \in A^{\{1,2,\cdots,n\}}$, $\Delta_{\mathcal{C}_0}(x_1^n)$ denotes the number of symbols x_t in x_1^n for which there is no $s \in \mathcal{C}_0$ such that $s_j = x_{t+j}$ for all $j \in \mathcal{I}(s)$. Consequently, $\Delta_{\mathcal{C}_0}(x_1^n) = n - \sum_{s\in\mathcal{C}_0} N_n(s)$. Note that $\Delta_{\mathcal{C}_0}(x_1^n)$ depends on both the context set $\mathcal{C}_0$ and the message x_1^n, and $\Delta_{\mathcal{C}_0}(x_1^n) \le d(\mathcal{C}_0)$.

Definition 4. *Given a message $x_1^n \in A^{\{1,2,\ldots,n\}}$, the Krichevsky-Trofimov probability for the context set $\mathcal{C}_0$ is*

$$P_{\mathcal{C}_0}(x_1^n) = \frac{1}{|A|^{\Delta_{\mathcal{C}_0}(x_1^n)}} \prod_{s\in\mathcal{C}_0} KT_{x_1^n}(s)$$

where $KT_{x_1^n}(s)$ is the Krichevsky-Trofimov probability [4] assigned to $\{x_i : l(s) < i \le n, s_j = x_{i+j}$ for all $j \in \mathcal{I}(s)\}$, the set of symbols in the message following the stringoid s.

The redundancy of the corresponding code satisfies

$$R_{L_{\mathcal{C}_0},Q}(x_1^n) \le \frac{|A|-1}{2} |\{s \in \mathcal{C}_0 : l(s) < n\}| \log\left(\frac{n - \Delta_{\mathcal{C}_0}(x_1^n)}{|\{s \in \mathcal{C}_0 : N_n(s) \ge 1|} + \frac{|A|-1}{2}\right) + |\{s \in \mathcal{C}_0 : l(s) < n\}| \log \frac{\pi^{\frac{1}{2}}}{\Gamma(\frac{|A|}{2})} + \Delta_{\mathcal{C}_0}(x_1^n) + c \quad (2)$$

for all $x_1^n \in A^{\{1,2,\ldots,n\}}$, where Γ is Gamma function and c is a constant [11].

4. Context Set Weighting

In this section, the Context Set Weighting method is introduced to define a coding distribution for unknown context sets. For $1 \le L < D$, let $\mathcal{B}_D^L$ be the set of stringoids s with $0 \le l(s) \le D$ such that for all stringoid s with $L-1 \le l(s) \le D$, $\{k \in -\mathbb{N} : -l(s) \le k \le -L+1\}$ is a subset of $\mathcal{I}(s)$. The weighted probability of the stringoid $s \in \mathcal{B}_D^L$ is determined recursively in $l(s)$, starting from D. In particular, $P_{w,x_1^n}^{L,D}(s)$ with $L-1 \le l(s) \le D-1$ is calculated from the values $P_{w,x_1^n}^{L,D}(s')$, $l(s') = l(s)+1$, and the value $P_{w,x_1^n}^{L,D}(s)$ with $l(s) \le L-2$ is calculated from the values $P_{w,x_1^n}^{L,D}(s')$, $l(s) < l(s') \le L$. Regarding a message x_1^n as a sequence proceeded by the unknown past $\ldots\epsilon\epsilon\epsilon$, the quasi-stringoid ϵs occurs if $0 \le l(s) < n$ and $s_i = x_{l(s)+1+i}$ for all $i \in \mathcal{I}(s)$. In that case the KT value is $\frac{1}{|A|}$, which is assigned to the weighted probability $P_{w,x_1^n}^{L,D}(\epsilon s)$. Similarly, the quasi-stringoid $\epsilon\alpha^h s$ occurs if $0 \le h + l(s) < n$ and $s_i = x_{h+l(s)+1+i}$ for all $i \in \mathcal{I}(s)$. The weights of the weighted probability depend on the length of s and the sum of them is one. The latter is implied by $\frac{1}{2} = \frac{1}{2^p} + \sum_{\emptyset\subset T\subset A} \frac{1}{2^{p(|T|+1)}}$, that follows from (5) of Definition 5.

Definition 5. *Given a message $x_1^n \in A^{\{1,2,\ldots,n\}}$ and an arbitrary $1 \le L < D$, for $s \in \mathcal{B}_D^L$ define the* weighted probability $P_{w,x_1^n}^{L,D}(s)$

$$
\begin{aligned}
&= KT_{x_1^n}(s) && \text{if } l(s) = D \\
&= \frac{1}{2} KT_{x_1^n}(s) + \frac{1}{2} P_{w,x_1^n}^{L,D}(\epsilon s) \prod_{a \in A^{\{-l(s)-1\}}} P_{w,x_1^n}^{L,D}(as) && \text{if } L-1 \le l(s) \le D-1 \\
&= \frac{1}{2} KT_{x_1^n}(s) + \frac{1}{(L-l(s))2^p} \left[\sum_{j=0}^{L-l(s)-1} \prod_{h=0}^{j} P_{w,x_1^n}^{L,D}\left(\epsilon \alpha^h s\right) \prod_{a \in A^{\{-l(s)-j-1\}}} P_{w,x_1^n}^{L,D}(as) \right] \\
&\quad + \frac{1}{L-l(s)-1} \sum_{k=0}^{L-l(s)-2} \sum_{V \subset A^{\{-l(s)-k-2\}}} \frac{1}{2^{p(|V|+1)}} \prod_{h=0}^{k+1} P_{w,x_1^n}^{L,D}\left(\epsilon \alpha^h s\right) \\
&\quad \times \prod_{a \in A^{\{-l(s)-k-1\}}} \prod_{b \in V} P_{w,x_1^n}^{L,D}(bas) \prod_{\bar{b} \in A^{\{-l(s)-k-2\}} \setminus V} P_{w,x_1^n}^{L,D}\left(\bar{b}s\right) && \text{if } 0 \le l(s) \le L-2
\end{aligned}
$$

where $0 < |V| < |A|$. p is the unique positive solution of the equation $2^{p-1} = (1+2^{-p})^{|A|} - 2^{-p|A|}$, and for a quasi-stringoid $\epsilon\alpha^h s$,

$$
P_{w,x_1^n}^{L,D}\left(\epsilon \alpha^h s\right) = \begin{cases} \frac{1}{|A|} & \text{if } s_i = x_{h+l(s)+1+i} \text{ for all} \\ & \quad i \in \mathcal{I}(s) \text{ and } l(s)+h < n \\ 1 & \text{otherwise.} \end{cases}
$$

Remark 2. The set $\mathcal{B}_D^L$ can be represented by a graph and the calculation of the weighted probability can be regarded as a sequential procedure over the nodes of the graph. For each node s with $0 \le l(s) \le L-2$, the computation uses the values from the related graphs $\mathcal{F}(j)$, for all $0 \le j \le L-l(s)-1$, and $\mathcal{G}(k,T)$, for all $0 \le k \le L-l(s)-2$ and $\emptyset \subset T \subset A$, whose root is s, see Remark 1. That is, such s has $L - l(s) + (L-l(s)-1)(2^{|A|}-2)$ number of related graphs. In addition, the computation uses the values of $j+1$ and $k+2$ number of quasi-stringoids for the above related graphs, respectively. For each node s with $L-1 \le l(s) \le D-1$, the computation uses the values only from the related graph $\mathcal{C}^A(0)$, whose root is s and from one quasi-stringoid, ϵs.

Definition 6. *The* coding distribution *for a message x_1^n is defined as the weighted probability for the empty stringoid $\varnothing$ with $D = n$ and $1 \le L < D$, i.e., $P_C(x_1^n) = P_{w,x_1^n}^{L,n}(\varnothing)$.*

A main result of this paper is the following bound on the redundancy of the above code determined by the Context Set Weighting method. Let Γ_D^L denote the collection of all possible context sets $\mathcal{C}_0'$ which satisfy $d(\mathcal{C}_0') \le D$ and $\{k \in -\mathbb{N} : -l(s) \le k \le -L+1\} \subseteq \mathcal{I}(s)$ for each $s \in \mathcal{C}_0'$ with $l(s) \ge L-1$. That is, the indices of the stringoids in the context sets $\mathcal{C}_0'$ may be missing only up to the depth $L-1$. For any context set $\mathcal{C}_0$, there is a unique $\mathcal{C}_0' \in \Gamma_\infty^L$ satisfying that for each $s \in \mathcal{C}_0$, all the stringoids s' with $\mathcal{I}(s') = \mathcal{I}(s) \cup \{k \in -\mathbb{N} : -l(s) \le k \le -L+1\}$ and $s' \curlyvee s$ belong to $\mathcal{C}_0'$. The above $\mathcal{C}_0'$ is denoted by $\mathcal{C}_0^L$. That is, $\mathcal{C}_0^L$ completes the stringoids in $\mathcal{C}_0$ by substituting all possible coordinates at the missing indices below the depth $L-1$. For any $\mathcal{C}_0^L \in \Gamma_\infty^L$, its truncation at level D is denoted by $\mathcal{C}_0^L|_D$. Clearly, $\mathcal{C}_0^L|_D \in \Gamma_D^L$ for any $\mathcal{C}_0$.

Theorem 1. *Given a message $x_1^n \in A^{\{1,2,\ldots,n\}}$, the model redundancy $R_{P_C,P_{\mathcal{C}_0}}(x_1^n)$ of the code C provided by the Context Set Weighting method does not exceed the order of the*

upper bound (2) of the parameter redundancy. In particular, for any $L > 1$,

$$R_{P_C,P_{\mathcal{C}_0}}(x_1^n) \le \left|\mathcal{C}_0^L|_{n-1}\right| \left(\frac{p+\log L}{|A|-1}+1\right) + \frac{|A|-1}{2}|\{s \in \mathcal{C}_0^L : l(s) < n\}| \log\left(\frac{n-\Delta_{\mathcal{C}_0^L}(x_1^n)}{|\{s \in \mathcal{C}_0^L : N_n(s) \ge 1|} + \frac{|A|-1}{2}\right) + |\{s \in \mathcal{C}_0^L : l(s) < n\}| \log \frac{\pi^{\frac{1}{2}}}{\Gamma(\frac{|A|}{2})}. \quad (3)$$

In particular, if $L \ge d(\mathcal{C}_0)$

$$R_{P_C,P_{\mathcal{C}_0}}(x_1^n) \le |\mathcal{C}_0| \left(\frac{p+\log L}{|A|-1}+1\right). \quad (4)$$

Remark 3. In bound (3), $\mathcal{C}_0$ is any context set of the process. The bound (2) on the parameter redundancy suggests considering the context set with the smallest $\left|\mathcal{C}_0^L|_{n-1}\right|$. By the comments after Definition 2, such a context set may not be well-defined and seems difficult to be identified. The bound (3) holds for the smallest $\left|\mathcal{C}_0^L|_{n-1}\right|$ possible for the process.

By Definition 3, a bound on the redundancy $R_{L_C,Q}(x_1^n)$ of the code C determined by the Context Set Weighting method can be obtained as a sum of (2) and (3). The bound shows that the Context Set Weighting method provides a universal code as $(1/n)\max_{x_1^n} R_{L_C,Q}(x_1^n) \to 0$ $(n \to \infty)$ if $|\mathcal{C}_0| < \infty$.

The Context Set Weighting method suggests that the coding distribution is a mixture of the Krichevsky-Trofimov distributions over all possible $\mathcal{C} \in \Gamma_D^L$.

Theorem 2. *For any $x_1^n \in A^{\{1,2,\ldots,n\}}$, the coding distribution $P_C(x_1^n)$ is a mixture of Krichevsky-Trofimov distributions over all $\mathcal{C} \in \Gamma_n^L$,*

$$P_C(x_1^n) = \sum_{\mathcal{C}' \in \Gamma_n^L} 2^{-\Lambda_{\mathcal{C}'}} P_{\mathcal{C}'}(x_1^n)$$

for some $\Lambda_{\mathcal{C}'}$ with $\sum_{\mathcal{C}' \in \Gamma_n^L} 2^{-\Lambda_{\mathcal{C}'}} = 1$.

Finally, we show an algorithm with practical computational complexity to calculate the value of the coding distribution provided by the Context Set Weighting method for a message x_1^n. In particular, choosing $L = \mathcal{O}(\log n)$ the computational complexity is polynomial in n.

Theorem 3. *Given a message $x_1^n \in A^{\{1,2,\ldots,n\}}$, the coding distribution $P_C(x_1^n)$ provided by the Context Set Weighting method can be computed with $\mathcal{O}(n^3 2^L)$ number of computations and storing $\mathcal{O}(n^2 2^L)$ data.*

Acknowledgment

Research was supported in part by the ARO under Grant 65386-MA-II and the NSF under Grant DMS 1407819.

References

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

[2] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial,* Hanover, MA: now, 2004.

[3] H.S. Kim and Zs. Talata, "Context Set Weighting Method," *Manuscript,* 33 pp, 2017.

[4] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory,* vol. 27, pp. 199–207, Mar. 1981.

[5] J. Rissanen, "A universal data compression system," *IEEE Trans. Inform. Theory,* vol. 29, pp. 656–664, Sep. 1983.

[6] J. Rissanen, and G. G. Langdon, Jr., "Universal Modeling and Coding," *IEEE Trans. Inform. Theory,* vol. IT-27, pp. 12–23, Jan. 1981.

[7] S. Robin, F. Rodolphe, and S. Schbath, *DNA, Words and Models.* New York: Cambridge, 2005.

[8] B. Ya. Ryabko, "Twice-universal coding," *Probl. Peredachi Inform.,* vol. 20, No. 3, pp. 24–28, 1984.

[9] J. Suzuki, "A CTW scheme for some FSM models," In *IEEE Int. Symp. on Inform. Theory,* p. 389, Whistler, British Columbia, Canada, 1995.

[10] P.A.J. Volf and F.M.J. Willems, "Context-tree weighting for extended tree sources," In *Symp. on Inform. Theory in the Benelux,* vol. 17, pp. 95–101, Enschede, The Netherlands, May 30-31 1996.

[11] Y. M. Shtarkov, T. J. Tjalkens, and F. M. J. Willems, "Multialphabet weighting universal coding of context tree sources," *Probl. Inform. Trans.,* vol. 33, No. 1, 1997.

[12] M. J. Weinberger, A. Lempel, and J. Ziv, "A sequentional algorithm for the universal coding of finite memory sources," *IEEE Trans. Inform. Theory,* vol. 38, pp. 1002–1014, May 1992.

[13] M. J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Inform. Theory,* vol. 41, pp. 643–652, May 1995.

[14] F. M. J. Willems, "The context-tree weighting method: Extensions," *IEEE Trans. Inform. Theory,* vol. 44, pp. 792–798, Mar. 1998.

[15] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inform. Theory,* vol. 41, pp. 653–664, May 1995.

[16] F. M. J. Willems,Y. M. Shtarkov, and T. J. Tjalkens,"Context weighting for general finite-context sources," *IEEE Trans. Inform. Theory,* vol. 42, No. 5, Sep. 1996.