ISIT'2013 1569716391

Universal variable to fixed-length random number generators for finite memory sources*

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

Marcelo J. Weinberger
Center for Science of Information
West Lafayette, IN, USA
Email: marcwein@ieee.org

***Abstract*—We study variable to fixed-length random number generators (VFRs) that input a variable number of symbols from a finite memory source of arbitrary order and unknown parameters, and output a number uniformly distributed in $\{0, 1, \ldots, M-1\}$ for arbitrary fixed M. We further require that the VFR output be uniformly distributed also if an arbitrary bound N is imposed on the input length, at the cost of a positive probability of the VFR terminating with no output (failing). We characterize the essentially unique optimal VFR, which minimizes both the expected input length and the failure probability for all truncation levels N. We precisely characterize, up to an additive constant, the expected input length of the optimal VFR, which includes a model cost term similar to those encountered in universal data compression and universal simulation.**

I. Introduction

Procedures for transforming non-uniform random sources into uniform ("perfectly random") ones have been a subject of great interest in statistics, information theory, and computer science for decades, going back to at least [1]. For the purposes of this paper, a *(fair) random number generator* (RNG) is a deterministic procedure that takes, as input, samples from a random process over a finite alphabet $\mathcal{A}$, and generates, as output, an integer r that is uniformly distributed in some range $0 \leq r < M$. In the *variable to fixed-length RNG* (in short, VFR) variant of interest here, M is an arbitrary but fixed integer, and the length, n, of the input sequence consumed by the VFR is a random variable. When $M = p^m$, the output r can be regarded as the outcome of m independent tosses of a *fair p-sided coin* (or *die*); when $p = 2$, it is often said, loosely, that the RNG generates m *random bits*. VFRs were studied in [1], [2], [3] (with emphasis on the case $M = 2$ and Bernoulli sources) and more recently, in more generality, in [4], [5]. The dual *fixed to variable-length* RNG problem has also received much attention, starting from Elias's optimal solution for Bernoulli sources [6] (efficiently implemented in [7]), and is the subject of a companion paper [8].

A VFR is said to be *universal* in a class of processes $\mathcal{P}$ if it produces a uniformly distributed output for any process in the class. We are interested in universal VFRs that minimize the expectation of the input length n. Although, in principle, n is unbounded, we are also interested in *truncated VFRs* (TVFRs). A TVFR either produces a uniformly distributed output on an input of length $n \leq N$, for some fixed N, or *fails* (producing no output). We require VFRs to produce uniform outputs, while admitting some failure probability, *at all truncation levels N*. In that sense, our notion of a universal VFR is stricter than the one in the earlier literature (cf. [1], [2], [3], [4]), where generally no conditions are posed on the truncated VFRs. The stricter notion may be useful in practical applications, where there is likely to be some prior knowledge or requirement of a minimal value of the source entropy. If the VFR has not produced an output long enough after the entropy estimate indicates that (with very high probability) it should have, the whole system function should probably become suspect. With the stricter definition, the input length threshold can be set arbitrarily, while preserving perfect uniformity of the VFR output. As it turns out, the average input length penalty incurred by the more restrictive definition on the optimal VFR is negligible relative to the main asymptotic term.

In this paper, we study universal VFRs in the class of all k-th order *finite memory (Markov)* processes, for arbitrary k. We characterize the essentially unique optimal VFR, which minimizes both the expected input length and the failure probability, at all truncation levels N (and, thus, also asymptotically). We precisely characterize, up to an additive constant, the expected input length of this VFR, and show that it exhibits a second order term of the same form as the "model cost" term found in other problems in information theory, such as universal lossless compression or universal simulation. This term is proportional to $\log \log M$ (in a sense, the logarithm of the "length" of the output), and to the number of free statistical parameters of the input model class. We also show that the failure probability of the optimal VFR decays exponentially fast. The optimal VFR coincides with the optimal "even" procedure of [2] for Bernoulli processes and $M = 2$. A universal VFR that, although sub-optimal for all N, attains the same asymptotic performance as our scheme for Bernoulli processes and $M = 2^m$ is described in [5] (without an analysis of the second order term). Extensions of our results to FSM sources are discussed in the full paper.

II. Definitions and preliminaries

Let $\mathcal{A}$ be a finite alphabet of size $\alpha = |\mathcal{A}|$. We denote a sequence $x_i x_{i+1} \ldots x_j$ over $\mathcal{A}$ by x_i^j, with x_1^j also denoted x^j. We let $x^* \in \mathcal{A}^*$ denote a generic string of unspecified length. For sets $U, V \subseteq \mathcal{A}^*$, we let $U \cdot V = \{uv | u \in U,\ v \in V\}$, and, for an integer $M > 0$, we let $[M] = \{0, 1, \ldots, M-1\}$. For

* Work done while the authors were with Hewlett-Packard Laboratories, Palo Alto, CA, USA.

integers t and u, we write $t \equiv u \pmod{M}$ if $M|(u-t)$, and $t = u \bmod M$ if $t \equiv u \pmod{M}$ and $0 \le t < M$.

A k-th order *finite memory (Markov)* process P over the alphabet $\mathcal{A}$ is defined by a set of α^k conditional probability mass functions $p(\cdot|s) : \mathcal{A} \to [0,1]$, $s \in \mathcal{A}^k$, where $p(a|s)$ denotes the probability of the process emitting a immediately after having emitted the k-tuple s. The latter is referred to as a *state* of the process, and we assume for simplicity a fixed but arbitrary initial state $x_{-k+1}^0 = s_0$.[1] Let $c \in \mathcal{A}$ be a fixed symbol. We denote by $\mathbf{p}$ the vector $\mathbf{p} = [\,p(a|s)\,]_{a \in \mathcal{A}\setminus\{c\}, s \in \mathcal{A}^k}$, and by Ω its domain of definition. The $K = (\alpha-1)\alpha^k$ components of $\mathbf{p}$ form a set of free statistical parameters that completely specify a k-th order process P. For simplicity, we further assume that all conditional probabilities $p(a|s)$ are nonzero.

The *type class* of x^n with respect to the family $\mathcal{P}_k$ of all k-th order finite memory processes is defined as the set

$$T(x^n) = \{\, y^n \in \mathcal{A}^n \,|\, P(x^n) = P(y^n) \;\; \forall P \in \mathcal{P}_k \,\}. \tag{1}$$

Let $n_s^{(a)}(x^n)$ denote the number of occurrences of a following s in x^n, and let $\mathbf{n}(x^n)$ denote the vector of α^{k+1} integers $n_s^{(a)}(x^n)$ ordered according to some fixed convention. It is well known that $T(x^n)$ is equivalently characterized as

$$T(x^n) = \{\, y^n \in \mathcal{A}^n \,|\, \mathbf{n}(x^n) = \mathbf{n}(y^n) \,\}. \tag{2}$$

The vector $\mathbf{n}(x^n)$ is referred to as the *type* of x^n, and we refer to n as the length of the type. The set of all type classes for sequences of length n is denoted $\mathcal{T}_n$.

The order k of the Markov process is assumed known and fixed throughout. Type classes of finite memory processes (and of broader model families) have been studied extensively (see, e.g., [9] and references therein). In particular, the cardinality of a type class is explicitly characterized by *Whittle's formula* [10]. This formula also allows for the efficient enumeration of the type class, namely, the computation of the index of a given sequence in its class, and the derivation of a sequence from its index, by means of enumeration methods such as those described in [11].[2] These enumerations are a key component of the RNG procedures discussed in this paper.

III. Universal variable-to-fixed length RNGs

A. Formal definition

A (prefix-free, complete) *dictionary* is a (possibly infinite) set $\mathcal{D} \subseteq \mathcal{A}^*$ of finite sequences, associated with the set of leaves of a complete α-ary tree (i.e., a rooted tree where each node is either a leaf, or the parent of α children nodes). A *variable-to-fixed length random number generator* (VFR) is a triplet $\mathcal{V} = (\mathcal{D}, \Phi, M)$ where $\mathcal{D}$ is a dictionary, M is a fixed positive integer, and Φ a function $\Phi : \mathcal{D} \to [M]$. For $N \ge 1$, the *restriction* to level N of $\mathcal{D}$ is

$$\mathcal{D}_N = \{\, x^n \in \mathcal{D} \,|\, n \le N \,\}.$$

[1] Other possible initial state assumptions are discussed in the full paper.

[2] In this context, "efficient" means computable in polynomial time. Although further complexity optimizations are outside the scope of this paper, various tools developed for similar problems in the literature would be applicable also here. See, e.g., the efficient implementation of Elias's scheme in [7], and references therein.

Associated with $\mathcal{D}_N$ is a *failure set* $\mathcal{E}_N$, defined as

$$\mathcal{E}_N = \{\, x^N \in \mathcal{A}^N \;\mid\; x^N \text{ has no prefix in } \mathcal{D}_N \,\}.$$

The strings in $\mathcal{D}_N \cup \mathcal{E}_N$ are identified with the leaves of a finite complete tree, which is the truncation to depth N of the tree underlying $\mathcal{D}$.

The VFR $\mathcal{V}$ generates random numbers from a process $P = X^\infty$ by reading symbols from a realization of P until a string x^n in $\mathcal{D}$ is reached, at which point $\mathcal{V}$ outputs $\Phi(x^n)$. The *truncated VFR* (TVFR) $\mathcal{V}_N = (\mathcal{D}_N, \Phi, M)$, operates similarly, except it restricts the length of the input string to $n \le N$, so that Φ is applied only to strings in $\mathcal{D}_N$, and the input may reach strings $x^N \in \mathcal{E}_N$, in which case $\mathcal{V}_N$ *fails* and outputs nothing.

A VFR $\mathcal{V} = (\mathcal{D}, \Phi, M)$ is *perfect* for P if for every $n \ge 1$, either $\mathcal{D}_n$ is empty, or $\Phi(x^*)$, conditioned on $x^* \in \mathcal{D}_n$, is uniformly distributed in $[M]$; $\mathcal{V}$ is *universal* in a class of processes $\mathcal{P}$ if it is perfect for all $P \in \mathcal{P}$. We extend these definitions also to TVFRs, with the lengths of the sequences considered upper-bounded by some N. The figure of merit for a perfect VFR is its *expected dictionary length*, which, relative to an input process P, is defined as

$$L_P(\mathcal{D}) = \lim_{N \to \infty} L_P(\mathcal{D}_N)\,, \tag{3}$$

where

$$L_P(\mathcal{D}_N) = \sum_{n=1}^{N} \sum_{x^n \in \mathcal{D}_N} nP(x^n) + NP(\mathcal{E}_N) \tag{4}$$

is the corresponding figure of merit for the truncated VFRs. The expected dictionary length measures the amount of "raw" random data that the VFR consumes in order to produce a perfectly uniform distribution on $[M]$. For a family of processes $\mathcal{P}$, we are interested in universal VFRs that approach the minimal expected length simultaneously for all $P \in \mathcal{P}$, either in a pointwise sense, i.e., minimizing $L_P(\mathcal{D}_N)$ for all N, or asymptotically, i.e., minimizing $L_P(\mathcal{D})$. Secondary objectives are the minimization of the failure probability $P(\mathcal{E}_N)$, and polynomial running time VFR implementations.

B. Necessary and sufficient conditions for universality

Consider a VFR $\mathcal{V} = (\mathcal{D}, \Phi, M)$. For $r \in [M]$, define

$$\Phi^{-1}(r) = \{\, x^* \in \mathcal{D} \mid \Phi(x^*) = r \,\}.$$

Also, for a type class T, let $\mathcal{D}(T) = \mathcal{D} \cap T$. The following necessary and sufficient condition for universality is similar to conditions previously derived for problems in universal simulation [12] and fixed to variable-length RNGs [13].

Lemma 1: For every $P \in \mathcal{P}_k$, except possibly for a subset of volume zero in the parameter space Ω, if $\mathcal{V} = (\mathcal{D}, \Phi, M)$ is a perfect VFR for P, then, for every n and every $T \in \mathcal{T}_n$, there exists an integer $j \ge 0$ such that

$$|\Phi^{-1}(r) \cap T| = j \tag{5}$$

for all $r \in [M]$ and, thus, $|\mathcal{D}(T)| = jM$. In particular, (5) must hold for universal VFRs. Conversely, given a dictionary $\mathcal{D}$ such that M divides $|\mathcal{D}(T)|$ for all $T \in \mathcal{T}_n$ and all n, there exists a function Φ such that $(\mathcal{D}, \Phi, M)$ is a universal VFR.

Lemma 1 implies that our universal VFRs are akin to the *even procedures* discussed in [2] and [3]. In our case, the

Input: Integers $M \geq 2,\ N \geq 1$.
Output: TVFR $\mathcal{V}_N^* = (\mathcal{D}_N^*, \Phi^*, M)$.

1) Set $n = 1$, $\mathcal{D}_N^* = \emptyset$.
2) For each type class $T \in \mathcal{T}_n$, do:
   a) Let $j = \left\lfloor |A_{\mathcal{D}_N^*}(T)|/M \right\rfloor$. Select any subset of jM sequences from $A_{\mathcal{D}_N^*}(T)$, and add them to $\mathcal{D}_N^*$.
   b) Let $\mathcal{I}(y^n)$ denote the index of $y^n \in \mathcal{D}_N^*(T)$ in some ordering of $\mathcal{D}_N^*(T)$. Define
   $$\Phi^*(y^n) = \mathcal{I}(y^n) \mod M\,, \quad y^n \in \mathcal{D}_N^*(T).$$
3) Set $n \leftarrow n+1$. If $n \leq N$, go to Step 2. Otherwise, **stop**.

Fig. 1. Procedure G1: Greedy TVFR construction.

necessity of the condition (5) stems from the requirement of perfection at every truncation level N. When this requirement is relaxed, the condition need no longer hold, as evidenced by some of the procedures presented in [2] and [3].

Notice that the condition on universality in Lemma 1 depends only on the *sizes* of the sets $\mathcal{D}(T)$, but not on their composition. Clearly, the same holds for the expected length and the failure probability of a (restricted) dictionary, since sequences of the same type have the same length and probability. We conclude that the main properties of interest for a VFR are fully determined by the *type profile* of its dictionary, namely, the sequence of numbers $\{\, |\mathcal{D}(T)| \,\}_{T\in\mathcal{T}_n, n\geq 1}$.

IV. A "GREEDY" UNIVERSAL VFR

A. Construction and pointwise optimality

We describe the construction of a universal VFR. The construction is "greedy," in the sense that at every point, it tries to add to the dictionary as many sequences as allowed by the necessary condition of Lemma 1. In this sense, the procedure can be seen as a counterpart, for VFRs, to Elias's scheme [6] for fixed to variable length RNGs.

For a prefix-free set of sequences $\mathcal{D}$, let $A_{\mathcal{D}}(T)$ denote the subset of sequences from a type class T that have no proper prefix in $\mathcal{D}$. Consider a process in which $\mathcal{D}$ is formed by successively adding sequences of increasing length, type by type. Then, $A_{\mathcal{D}}(T)$ represents those sequences in T that, at some point in this process, were *available* for inclusion in $\mathcal{D}$ without breaking the prefix condition. Further, let $R_{\mathcal{D}}(T) = A_{\mathcal{D}}(T) \setminus \mathcal{D}(T)$, namely those sequences from T that were available for inclusion, but were not included, in $\mathcal{D}$.

Procedure G1 in Fig. 1 shows the construction of a greedy TVFR $\mathcal{V}_N^* = (\mathcal{D}_N^*, \Phi, M)$. The VFR $\mathcal{V}^* = (\mathcal{D}^*, \Phi, M)$ is then obtained by letting $\mathcal{D}^* = \bigcup_{N\geq 1} \mathcal{D}_N^*$. The procedure is presented as a characterization of $\mathcal{V}^*$, rather than as a computational device. An effective, sequential implementation of $\mathcal{V}^*$ will be presented in Subsection IV-B.

Theorem 1: $\mathcal{V}_N^* = (\mathcal{D}_N^*, \Phi^*, M)$ is a universal TVFR.

Theorem 1 is a direct consequence of the construction process and of Lemma 1. In contrast, showing that $\mathcal{V}_N^*$ is also pointwise optimal requires a series of lemmas that we present next and whose proof is given in the full paper.

Lemma 2: Let $\mathcal{D}$ be the dictionary of a universal VFR. Then, for every type class T, we have

$$|A_{\mathcal{D}}(T)| \equiv |R_{\mathcal{D}}(T)| \equiv |T| \pmod M. \tag{6}$$

We say that a type class $T\in\mathcal{T}_n$ is *underrepresented* in a dictionary $\mathcal{D}$ if $|R_{\mathcal{D}}(T)|\geq M$. By construction, no type class is underrepresented in $\mathcal{D}^*$ implying, by Lemma 2,

$$|R_{\mathcal{D}^*}(T)| = |T| \bmod M\,. \tag{7}$$

The following converse statement also holds.

Lemma 3: Let $(\mathcal{D}, \Phi, M)$ be a universal VFR such that no type class $T \in \mathcal{T}_n$, $n \geq 1$, is underrepresented in $\mathcal{D}$. Then, $|\mathcal{D}(T)| = |\mathcal{D}^*(T)|$ for all T.

Lemma 4: Let $\mathcal{V} = (\mathcal{D}, \Phi, M)$ be a universal VFR such that there exists a type class $T\in\mathcal{T}_n$ that is underrepresented in $\mathcal{D}$. Then, for any $N > n$, the TVFR $\mathcal{V}_N$ can be transformed into a universal TVFR $\mathcal{V}'_N = (\mathcal{D}'_N, \Phi', M)$ such that $|R_{\mathcal{D}'_N}(T)| < M$ and $L_P(\mathcal{D}'_N) < L_P(\mathcal{D}_N)$ for all $P \in \mathcal{P}_k$.

The following theorem establishes the pointwise optimality of $\mathcal{V}_N^*$ and the uniqueness of the optimal type profile for a universal VFR.

Theorem 2: Let $\mathcal{V} = (\mathcal{D}, \Phi, M)$ be a universal VFR. Then, for every $N \geq 1$, we have $L_P(\mathcal{D}_N^*) \leq L_P(\mathcal{D}_N)$ and $P(\mathcal{E}_N^*) \leq P(\mathcal{E}_N)$ for all $P \in \mathcal{P}_k$. Moreover, if $|\mathcal{D}(T)| \neq |\mathcal{D}^*(T)|$ for any n and $T\in\mathcal{T}_n$, then $L_P(\mathcal{D}_N^*) < L_P(\mathcal{D}_N)$ for all $N > n$ and all $P \in \mathcal{P}_k$.

B. Sequential implementation of $\mathcal{V}^*$

Our implementation builds on a decomposition of $T \in \mathcal{T}_n$ of the form

$$T = \textstyle\bigcup_{u\in\mathcal{A}^\ell} S_u(T) \tag{8}$$

where $1 \leq \ell \leq n$ and

$$S_u(T) = \{\, x^n \,|\, x^n \in T\,,\ x_{n-k-\ell+1}^{n-k} = u \,\}\,. \tag{9}$$

Since sequences in T coincide in the last k symbols, a nontrivial decomposition of the form (8) requires, as in (9), examination of more than k symbols in a suffix of x^n. Define

$$S_u^-(T) = \{\, x^{n-\ell} \,|\, x^n \in S_u(T) \,\}\,,\ u \in \mathcal{A}^\ell.$$

Lemma 5: If $S_u^-(T)$ is not empty then $S_u^-(T) \in \mathcal{T}_{n-\ell}$.

Since the sets $S_u^-(T)$ are (shorter) type classes, a recursion can be devised as follows. First, clearly, if $x^n\in A_{\mathcal{D}}(T)$ then $x^{n-1}\in A_{\mathcal{D}}(T(x^{n-1}))$. Hence, if by the decomposition (8) we have $x^n\in S_a(T)$, $a\in\mathcal{A}$, we must also have $x^{n-1}\in R_{\mathcal{D}}(S_a^-(T))$. Let

$$R_{\mathcal{D}}^+(S_a^-(T)) = \left\{\, x^n \in T \;\middle|\; x^{n-1} \in R_{\mathcal{D}}(S_a^-(T)) \,\right\}\,.$$

Lemma 6: For a dictionary $\mathcal{D}$ and any $T \in \mathcal{T}_n$, we have

$$A_{\mathcal{D}}(T) = \bigcup_{a\in\mathcal{A}} R_{\mathcal{D}}^+(S_a^-(T))\,, \quad |A_{\mathcal{D}}(T)| = \sum_{a\in\mathcal{A}} |R_{\mathcal{D}}(S_a^-(T))|\,. \tag{10}$$

Procedure G2 in Fig. 2 describes the proposed sequential implementation of $\mathcal{V}^*$. The procedure can easily be modified to implement the TVFR $\mathcal{V}_N^*$, for arbitrary N, with a possible failure exit.

The procedure relies on an alphabetic enumeration of $A_{\mathcal{D}}(T)$, $T{=}T(x^n)$, which is based, in turn, on (10). We assume a total (alphabetic) order $<$ of the elements of $\mathcal{A}$; for the purpose of comparing sequences of length n, symbols in x^n decrease in significance from x_n down to x_1. We

Input: Sequence $x_1x_2x_3\cdots x_n\cdots$, integer M.
Output: Number $r\in[M]$.

1) Set $\mathcal{I}_R=0,\ n=0,\ \mathbf{n}(x^n)=\mathbf{0}$.
2) Set $n=n+1$, read x_n, update $\mathbf{n}(x^n)$, and let $T=T(x^n)$.
3) Compute $|R_{\mathcal{D}}(S_a^-(T))|=|S_a^-(T)| \bmod M$ for each $a\in\mathcal{A}$.
4) Compute $\mathcal{I}_A=\sum_{a<x_{n-k}}|R_{\mathcal{D}}(S_a^-(T))|+\mathcal{I}_R$.
5) Compute $|A_{\mathcal{D}}(T)|=\sum_{a\in\mathcal{A}}|R_{\mathcal{D}}(S_a^-(T))|$.
6) Set $j=\big\lfloor |A_{\mathcal{D}}(T)|/M\big\rfloor$.
7) If $\mathcal{I}_A<jM$ then **output** $r=\mathcal{I}_A \bmod M$ and **stop**. Otherwise, set $\mathcal{I}_R=\mathcal{I}_A-jM$ and **go to** Step 2.

Fig. 2. Procedure G2: Sequential implementation of $\mathcal{V}_N^*$.

assume, recursively, that each set $R_{\mathcal{D}}(S_a^-(T))$ that contributes to $A_{\mathcal{D}}(T)$ per (10) has been enumerated alphabetically. In particular, we assume that after processing x^{n-1}, we have the index $\mathcal{I}_R(x^{n-1})$ of x^{n-1} in $R_{\mathcal{D}}(S_{x_{n-k}}^-(T))$. Since all the sequences in T coincide in their last k symbols, if $y^n\in T$ and $y_{n-k}<x_{n-k}$, then $y^n<x^n$ in the alphabetical order. Therefore, by (10), the index of x^n in $A_{\mathcal{D}}(T)$ is given by

$$\mathcal{I}_A(x^n)=\sum_{a<x_{n-k}}|R_{\mathcal{D}}(S_a^-(T))|+\mathcal{I}_R(x^{n-1})\,.$$

The sizes $R_{\mathcal{D}}(S_a^-(T))$, $a\in\mathcal{A}$, can be obtained from (7), by means of Whittle's formula [10] applied to $S_a^-(T)$. The type $\mathbf{n}^{(a)}$ associated with $S_a^-(T)$, which is required to evaluate Whittle's formula, is easily obtained from the type $\mathbf{n}(x^n)$. All the computations run in time polynomial in n.

V. Performance

We study the asymptotic performance of universal VFRs in terms of expected dictionary length and failure probability as $M\to\infty$. To this end, for sufficiently large N, we derive a lower bound on $L_P(\mathcal{D}_N)$ for any restricted dictionary $\mathcal{D}_N$ derived from a universal VFR and we show that the bound is achievable. For the achievability result we will not use the optimal universal VFR $\mathcal{V}^*$, but a different VFR, for which the analysis is simpler.[3] Both this analysis and the lower bound are rooted in the source coding literature, with special focus on [14]. We also show that the failure probability decays with N exponentially fast.

For $P\in\mathcal{P}_k$, let $H(P)$ denote its entropy rate, given by

$$H(P)=-\sum_{s\in\mathcal{A}^k}P(s)\sum_{a\in\mathcal{A}}p(a|s)\log p(a|s)$$

where $P(s)$ denotes the stationary probability of state s. We will say that a sequence x^* is δ-bounded if, for every $a\in\mathcal{A}$ and $s\in\mathcal{A}^k$, we have

$$n_s^{(a)}(x^*)>\delta\sum_{b\in\mathcal{A}}n_s^{(b)}(x^*)$$

(i.e., counts grow linearly). For a dictionary $\mathcal{D}$, and $P\in\mathcal{P}_k$, let

$$H_{\mathcal{D}}(P)\triangleq-\sum_{x^*\in\mathcal{D}}P(x^*)\log P(x^*)\,.$$

Our results will be based on the following key lemma.

[3]The situation is akin to lossless source coding, for which the entropy bound is shown to be achievable with, say, the Shannon code, rather than with the (optimal) Huffman code.

Lemma 7: Let $P\in\mathcal{P}_k$ and let $\mathcal{D}$ be a dictionary.

(i) If for every $x^*\in\mathcal{D}$ we have $|T(x^*)|\ge M$, then

$$H_{\mathcal{D}}(P)\ge\log M+(K/2)\log\log M+O(1)\,. \quad (11)$$

(ii) If there exist positive constants δ and C such that for every δ-bounded sequence $x^*\in\mathcal{D}$ we have $|T(x^*)|<CM$, and δ is sufficiently small (depending on P), then

$$H_{\mathcal{D}}(P)\le\log M+(K/2)\log\log M+O(1)\,. \quad (12)$$

We sketch the ideas behind the proof of Lemma 7, focusing, for clarity, on the memoryless case (the ideas extend to sources with memory, covered in the full paper), and assuming that the conditions for both (11) and (12) hold simultaneously (proving equality). To obtain this estimate, we use the universal probability assignment $Q(x^*)$ on $\mathcal{A}^*$ given by a uniform mixture over $\mathcal{P}_0$, for which it is readily verified that

$$Q(x^n)=\Big(|T(x^n)|\tbinom{n+\alpha-1}{\alpha-1}\Big)^{-1}\,, \quad (13)$$

and write

$$H_{\mathcal{D}}(P)=-\sum_{x^*\in\mathcal{D}}P(x^*)\log\frac{P(x^*)}{Q(x^*)}+\sum_{x^*\in\mathcal{D}}P(x^*)\log\frac{1}{Q(x^*)}\,. \quad (14)$$

Since, by our assumption on $\mathcal{D}$, we have $\log|T(x^*)|=\log M+O(1)$, since the length of a typical sequence satisfying this assumption is $\Theta(\log M)$ (as the type size is exponential in the sequence length), and since $K=\alpha-1$ for a memoryless source, (13) and Stirling's approximation imply that the value of the second summation in (14) is $\log M+K\log\log M+O(1)$. The first summation, on the other hand, is the divergence between P and Q, as distributions over the complete set $\mathcal{D}$. This divergence is well characterized when $\mathcal{D}=\mathcal{A}^n$ for some n, whereas here, $\mathcal{D}$ is a set of sequences of varying length. However, it is not hard to see that the summation can be "sandwiched" between those corresponding to sets of sequences of fixed length $\Theta(\log M)$, and apply the divergence estimate in [15] (extended to Markov sources in [16]) to conclude that the value of the first summation in (14) is $(K/2)\log\log M+O(1)$, completing the proof sketch.

To apply Lemma 7 to our problem, we use a variant of "Massey's leaf-node theorem," which in the memoryless case simply states that $H_{\mathcal{D}}(P)=H(P)L_P(\mathcal{D})$ and was extended to stationary Markov sources in [14]. In our case, we need further arguments (provided in the full paper) to deal with our choice of a fixed initial state.

A. Lower bound

The lower bound on $L_P(\mathcal{D}_N)$ for universal VFRs, stated in Theorem 3 below, follows from Lemma 1, (11), our variant of Massey's leaf-node theorem, and typicality arguments.

Theorem 3: Let $\mathcal{V}=(\mathcal{D},\Phi,M)$ be a universal VFR. Then, for every $P\in\mathcal{P}_k$ and sufficiently large N, we have

$$L_P(\mathcal{D}_N)\ge\frac{\log M+(K/2)\log\log M+O(1)}{H(P)}\,. \quad (15)$$

The term $(K/2)\log\log M$ in (15) resembles a typical "model cost" term in universal lossless compression. Although Theorem 3 is stated for a universal VFR, it applies, like Lemma 1, also to perfect VFRs for "almost all" $P\in\mathcal{P}_k$.

1) Set $i = 0$, $\tilde{\mathcal{D}}^* = \emptyset$, and let $\Delta_0 = \tilde{\mathcal{D}}$. For $x^* \in \tilde{\mathcal{D}}$, define $\mathcal{G}_0(x^*) = \tilde{\mathcal{D}}(T(x^*))$.
2) Set $\Delta_{i+1} = \emptyset$. For each $\mathcal{G} = \mathcal{G}_i(x^*)$, $x^* \in \Delta_i$, do:
   a) Let $m = |\mathcal{G}| \bmod M$, and let U be a set of m sequences from $\mathcal{G}$. Add $\mathcal{G} \setminus U$ to $\tilde{\mathcal{D}}^*$.
   b) If $m > 0$, let s_f be the common final state of all sequences in U. Add $U \cdot \tilde{\mathcal{D}}\,(s_f, \lceil M/m \rceil)$ to Δ_{i+1}.
3) If $\Delta_{i+1} = \emptyset$, **stop**. Otherwise, for each $x^* \in \Delta_{i+1}$, let x^ℓ be its prefix in Δ_i, and define
$$\mathcal{G}_{i+1}(x^*) = \{\, y^* \in T(x^*) \cap \Delta_{i+1} \mid y^\ell \in \mathcal{G}_i(x^\ell) \,\}\,.$$
Increment i, and go to Step 2.

Fig. 3. Construction of the universal VFR $\tilde{\mathcal{D}}^*$.

While perfect VFRs for arbitrary $P \in \mathcal{P}_0$ that do not incur the "model cost" term are described in [4], these VFRs are not required to be perfect at all truncation levels. Since $(\log M + O(1))/H(P)$ is a lower bound for any perfect VFR (shown in [4] for $k{=}0$; follows from our variant of Massey's theorem for $k{>}0$), the cost of maintaining perfection under truncation, in either the universal or individual process cases, is at least $(K/2)\log\log M$. The question of whether the bound (15) applies to universal VFRs on which no truncation requirements are posed remains open.

B. Achievability

In order to use Lemma 7 to establish the achievability of the bound of Theorem 3, we need a universal VFR such that the size of the type class of each δ-bounded sequence in its dictionary is at most CM for some constant C (that will depend on δ); typicality arguments are used again to deal with the rest of the sequences. Such a bound on the type class size does not appear to follow easily from the definition of the optimal VFR $\mathcal{V}^*$ since, in principle, the construction may require $|T| > CM$ for any constant C to guarantee $|A_{\mathcal{D}^*}(T)| \geq M$. The sought for universal VFR will make use of an auxiliary dictionary $\tilde{\mathcal{D}}$, given by

$$\tilde{\mathcal{D}} = \{\, x^* \;\big|\; |x^*| > k,\, |S_u(T(x^*))| \geq M \;\; \forall u \in \mathcal{A}^{k+1}, \text{ and no } x^{**} \prec x^* \text{ has these properties} \,\}\,, \quad (16)$$

where $\prec$ denotes the proper prefix relation. Thus, $\tilde{\mathcal{D}}$ grows until the first time *each* component $S_u(T(x^*))$ of $T(x^*)$ is large enough. The "stopping set" S defined in [5, Section IV] is the special case of $\tilde{\mathcal{D}}$ for the class of Bernoulli sources.

Lemma 8: For all $x^* \in \tilde{\mathcal{D}}$ we have $|\tilde{\mathcal{D}}(T(x^*))| \geq M$. In addition, for every $\delta > 0$ there exists a constant C such that every δ-bounded sequence $x^* \in \tilde{\mathcal{D}}$ satisfies $|T(x^*)| < CM$.

Lemma 8, (12), and Massey's leaf-node theorem, guarantee that $L_P(\tilde{\mathcal{D}})$ attains a value of the form of the right-hand side of (15). However, $\tilde{\mathcal{D}}$ need not satisfy the condition of Lemma 1, and, thus, cannot be relied upon as the basis of a universal VFR. In the construction of the dictionary $\tilde{\mathcal{D}}^*$ of a universal VFR, we explicitly denote $\tilde{\mathcal{D}}$ by $\tilde{\mathcal{D}}(s_0, M)$, since we will rely on dictionaries $\tilde{\mathcal{D}}(s, \ell)$ where the value of the initial state s will *not* be fixed at the same s_0 throughout, and the value of the threshold ℓ used in (16) may differ from M.

The iterative construction of $\tilde{\mathcal{D}}^*$ is shown in Figure 3. In the ith iteration of the construction, Δ_i denotes a set of sequences that are still pending processing (i.e., either inclusion in $\tilde{\mathcal{D}}^*$, or extension), starting with $\tilde{\mathcal{D}}$. Sequences in Δ_i are collected into groups $\mathcal{G}_i(x^*)$, where the latter consists of all the pending sequences of the same type as x^*, and whose prefixes in prior iterations were also of the same type. The dictionary $\tilde{\mathcal{D}}^*$ is built up, in Step 2a, of sets of sizes divisible by M, consisting of sequences of the same type. Thus, M divides $|\tilde{\mathcal{D}}^*(T_n)|$ for all n and all type classes T_n, so that Lemma 1 guarantees the existence of a universal VFR based on $\tilde{\mathcal{D}}^*$. The remaining m sequences are recursively extended by "hanging," in Step 2b, dictionaries $\tilde{\mathcal{D}}\,(s_f, \lceil M/m \rceil)$. Thus, the new set Δ_{i+1} contains m copies of type classes of sizes at least M/m. The following property can be shown to follow from the iterative process.

Lemma 9: For every $P \in \mathcal{P}_k$ we have
$$L_P(\tilde{\mathcal{D}}^*) - L_P(\tilde{\mathcal{D}}) = O(1).$$

The achievability of the bound (11) follows:

Theorem 4: For every $P \in \mathcal{P}_k$, the universal VFR based on $\tilde{\mathcal{D}}^*$ satisfies
$$L_P(\tilde{\mathcal{D}}^*) = \frac{\log M + (K/2)\log\log M + O(1)}{H(P)}$$
and $P(\mathcal{E}_N)$ decays exponentially fast with N.

By Theorem 2, the statement of Theorem 4 also applies to the optimal universal VFR $\mathcal{V}^*$.

Acknowledgment. Thanks to Erik Ordentlich for useful discussions.

References

[1] J. V. Neumann, "Various techniques used in connection with random digits," *Nat. Bur. Standards, Appl. Math Series*, vol. 12, pp. 36–38, 1951.
[2] W. Hoeffding and G. Simons, "Unbiased coin tossing with a biased coin," *Ann. Math. Statist.*, vol. 41, pp. 341–352, 1970.
[3] Q. F. Stout and B. Warren, "Tree algorithms for unbiased coin tosssing with a biased coin," *Ann. Probab.*, vol. 12, pp. 212–222, 1984.
[4] T. S. Han and M. Hoshi, "Interval algorithm for random number generation," *IEEE Trans. Inform. Theory*, vol. 43, pp. 599–611, 1997.
[5] H. Zhou and J. Bruck, "A universal scheme for transforming binary algorithms to generate random bits from loaded dice," *ArXiv:1209.0726 [cs.IT]*, Sep. 2012.
[6] P. Elias, "The efficient construction of an unbiased random sequence," *Ann. Math. Statist.*, vol. 43, pp. 865–870, 1972.
[7] B. Y. Ryabko and E. Matchikina, "Fast and efficient construction of an unbiased random sequence," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1090–1093, 2000.
[8] G. Seroussi and M. J. Weinberger, "Twice-universal fixed to variable-length random number generators for finite memory sources," ISIT'13.
[9] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505–2523, 1998.
[10] P. Whittle, "Some distribution and moment formulae for the Markov chain," *J. Roy. Statist. Soc. Ser. B*, vol. 17, no. 3, pp. 235–242, 1955.
[11] T. M. Cover, "Enumerative source encoding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 73–77, 1973.
[12] N. Merhav and M. J. Weinberger, "On universal simulation of information sources using training data," *IEEE Trans. Inform. Theory*, vol. 50, pp. 5–20, 2004.
[13] S. il Pae and M. C. Loui, "Randomizing functions: Simulation of a discrete probability distribution using a source of unknown distribution," *IEEE Trans. Inform. Theory*, vol. 52, pp. 4965–4976, 2006.
[14] T. J. Tjalkens and F. M. Willems, "Variable-to-fixed length codes for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 246–257, 1987.
[15] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, 1990.
[16] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2104–2109, 1999.