ISIT'2013 1569718665

Twice-universal fixed to variable-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 fixed to variable-length random number generators (FVRs) that input a fixed number of symbols from a finite memory source of arbitrary order and unknown parameters, and output a number uniformly distributed in $\{0, 1, \ldots, M-1\}$, where M is also random. We review Elias's FVR in the context of the method of types, and show that it remains universal and optimal in the broad class of k-th order finite memory processes. We precisely characterize, up to an additive constant, the expected output length of the optimal FVR, and show that it includes a model cost term similar to those encountered in universal data compression and universal simulation. We further define twice-universal FVRs, which produce quasi-uniform distributions when the input is a finite memory source of unknown order and parameters. We propose a twice-universal FVR whose expected output length is the same, up to an additive constant, as that of an optimal FVR constructed with knowledge of the order k, with the distance of the output to a uniform distribution vanishing exponentially fast with the input length.**

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 (almost) uniformly distributed in some range $0 \leq r < M$. In the *fixed to variable-length* (in short, FVR) variant of interest here, the input sequence is of fixed length n, and M is a random variable; the output r is (almost) uniform conditioned on M. When M is restricted to the form $M = p^m$, 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*. FVRs have been studied extensively, in various universal (e.g. [2]) and non-universal (e.g., [3]) settings. In the computer science literature, FVRs are referred to as *randomness extractors* (see, e.g., [4] and references therein). The dual *variable to fixed-length RNG* problem has also received attention, starting from [1], and is the subject of a companion paper [5].

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

We focus on k-th order *finite memory (Markov)* processes, where the order k is arbitrary.[1] An FVR is said to be *universal* in the class of k-th order processes (with k fixed and source parameters unknown) if it produces uniformly distributed outputs for any process in the class. We are interested in universal FVRs that maximize the expectation of $\log M$,[2] the *output length*. An FVR is *twice universal* in the class of *all* finite memory processes (with both k and the process parameters unknown) if its output approaches a uniform distribution for every process in the class (the formal definition of distribution proximity is provided in Section IV). The relaxation of the uniformity requirement is necessary for a meaningful statement of the twice-universal setting, as will follow from the characterization of universal FVRs.

In this paper, we first review results on universal FVRs. In [2], Elias presented a universal FVR procedure for Bernoulli processes and binary outputs. The procedure is optimal in expected output length, pointwise for every input block size n. An efficient implementation is described in [6], and a generalization to first order processes for any $\mathcal{A}$ is presented in [7].[3] We show that Elias's procedure, when studied in the more general context of the *method of types* [8], is applicable, almost verbatim and with a uniform treatment, to broader classes of processes, and, in particular, to the class of k-th order finite memory processes for any value of k and any finite alphabet $\mathcal{A}$, while retaining its universality and pointwise optimality properties. We precisely characterize, up to an additive constant, the expected output length of the procedure in the more general setting. The estimate shows that FVRs exhibit a "model cost" term proportional to the number of free statistical parameters in the model class, as do universal compression and universal simulation. However, somewhat

[1] Extensions to broader model classes, such as *finite state machine* models, are discussed in the full paper.

[2] Logarithms are to base 2, unless specified otherwise.

[3] Clearly, the class of first order finite memory processes for all finite $\mathcal{A}$ is equivalent to the class of k-th order processes, if we interpret a process P of order k over $\mathcal{A}$ as a process P' of order 1 over $\mathcal{A}^k$. However, this interpretation is wasteful, since very few state transitions would have nonzero probability in this case. State transition information for P' (whether probabilities or empirical counts) could be condensed in a $|\mathcal{A}|^k \times k$ matrix, whereas a generic scheme for first order processes over $\mathcal{A}^k$ would allocate and manipulate a full $|\mathcal{A}|^k \times |\mathcal{A}|^k$ matrix.

surprisingly, we observe that this model cost is incurred, for almost all processes P in the class, even if the FVR is designed to produce uniform outputs only for P. Thus, in the case of FVRs, the model cost is not necessarily interpreted as a "cost of universality," as is the case in the other mentioned problems.

After reviewing universal FVRs, we present a twice-universal FVR, also inspired on Elias's scheme, but based on the partition, presented in [9], of the space $\mathcal{A}^n$ into classes that generalize the notion of type class for the twice-universal setting. We show that the expected output length of the twice-universal FVR is the same, up to an additive constant, as that of a universal FVR designed for a known order k, with the distance of the output to a uniform distribution vanishing exponentially fast with the input length.

II. Definitions and preliminaries

Let $\mathcal{A}$ be a finite alphabet of size $\alpha = |\mathcal{A}|$. We denote a sequence $x_i x_{i+1} \dots x_j$ over $\mathcal{A}$ by x_i^j, with x_1^j also denoted x^j. For a positive integer M, we denote by $[M]$ the set $\{0, 1, \dots, M-1\}$.

A k-th order *finite memory (Markov)* process P over $\mathcal{A}$ is defined by a set of α^k conditional probability mass functions $p(\cdot|s) : \mathcal{A} \to [0,1]$, $s \in \alpha^k$, 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 s_0.[4] 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

$$T^{(k)}(x^n) = \{\, y^n \in \mathcal{A}^n \mid P(x^n) = P(y^n) \ \ \forall P \in \mathcal{P}_k \,\} \quad (1)$$

with the superscript omitted when clear from the context (e.g., in a setting where k is given). It is well known that $T^{(k)}(x^n)$ is equivalently characterized as the set of sequences with the same occurrence counts of $(k+1)$-tuples as x^n (the set of counts being also referred to as the *type* of x^n). The set of all type classes of length n with respect to order k is denoted $\mathcal{T}_n^{(k)}$. Again, the superscript is omitted when clear from the context.

Type classes of finite memory processes (and of broader model families) have been studied extensively (see, e.g., [8] 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].[5]

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

III. Universal fixed-to-variable length RNGs

Let $\mathbb{N}$ and $\mathbb{N}^+$ denote, respectively, the nonnegative and positive integers. An FVR $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$ is defined by an integer n, a *target set* $\mathbb{N}_t \subseteq \mathbb{N}^+$ such that $1 \in \mathbb{N}_t$, and functions $\rho : \mathcal{A}^n \to \mathbb{N}$, $\mathcal{M} : \mathcal{A}^n \to \mathbb{N}_t$, such that $\rho(x^n) \in [\mathcal{M}(x^n)]$. The *output length* of $\mathcal{F}_n$ on input x^n is defined as $\log \mathcal{M}(x^n)$. Thus, the function $\mathcal{M}$ determines the range of the output and the output length, while the function ρ determines the output random number itself. When the goal is to generate fair p-sided coin tosses, we choose

$$\mathbb{N}_t = \{\, p^i \mid i \geq 0 \,\}, \quad p \geq 2\,. \quad (2)$$

An FVR $\mathcal{F}_n$ is *perfect* for $P \in \mathcal{P}_k$ if $\rho(x^n)$, conditioned on $\mathcal{M}(x^n) = M$, is uniformly distributed in $[M]$; $\mathcal{F}_n$ is *universal* in $\mathcal{P}_k$ if it is perfect for all $P \in \mathcal{P}_k$. The *expected output length* of $\mathcal{F}_n$ with respect to P is

$$L_P(\mathcal{M}) \triangleq \mathbf{E}_P \log \mathcal{M}(X^n) = \sum_{x^n \in \mathcal{A}^n} P(x^n) \log \mathcal{M}(x^n)\,, \quad (3)$$

where $\mathbf{E}_P$ denotes expectation with respect to P.

Notice that our setting is slightly more general than the usual one for FVRs in the literature, where the condition (2) for some p is generally assumed in advance. The broader setting will allow us to better highlight the essence of the optimal solutions, as well as connections to related problems in information theory.

A. Necessary and sufficient conditions for universality

Consider functions $g(\mathbf{p}) = \sum_{T \in \mathcal{T}_n} g_T P(T)/|T|$, where the g_T are integers, and $P(T)$ is the total probability of the type class T for a parameter $\mathbf{p} \in \Omega$. These functions are multivariate polynomials in the components of $\mathbf{p}$. Let

$$G = \left\{\, g(\mathbf{p}) \;\middle|\; |g_T| \leq |T|,\ g_T \neq 0 \text{ for some } T \in \mathcal{T}_n \,\right\}.$$

It is known [12][9] that the type probabilities $P(T)$, as functions of $\mathbf{p}$, are linearly independent over the reals. Thus, no $g \in G$ is identically zero. Let Ω_0 denote the set of all vectors $\mathbf{p}$ such that $g(\mathbf{p}) = 0$ for some $g \in G$. It is readily verified that Ω_0 has volume zero in Ω. We write, loosely, $P \in \Omega_0$ when $P \in \mathcal{P}_k$ has parameter $\mathbf{p} \in \Omega_0$.

The following condition for perfection is similar to, albeit stronger than, conditions previously derived for problems in universal simulation [12][9] and universal FVRs [13][7].

Lemma 1: Let $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$ be an FVR satisfying the following condition: for $T \in \mathcal{T}_n$ and every $M \in \mathbb{N}_t$, the number of sequences $x^n \in T$ such that $\rho(x^n) = r$ is the same for all $r \in [M]$ (in particular, the number of sequences $x^n \in T$ such that $\mathcal{M}(x^n) = M$ is a multiple of M). Then,

[5] 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., [6] and references therein.

$\mathcal{F}_n$ is universal in $\mathcal{P}_k$. If $\mathcal{F}_n$ *does not* satisfy the condition, and $\mathcal{F}_n$ is perfect for some $P \in \mathcal{P}_k$, then $P \in \Omega_0$.

Sketch of proof: Let χ denote the set of sequences x^n such that $\mathcal{M}(x^n) = M$ and $\rho(x^n) = r$. Since sequences of the same type are equiprobable, we have

$$\Pr\Big(\mathcal{M}(X^n) = M,\, \rho(X^n) = r\Big) = \sum_{T \in \mathcal{T}_n} \frac{|\chi \cap T|}{|T|} P_{\mathbf{p}}(T)\,, \quad (4)$$

where we use $P_{\mathbf{p}}$ instead of P to emphasize the dependence of the probability on the parameter vector $\mathbf{p} \in \Omega$. If the condition of the lemma holds, then the right-hand side of (4) is independent of r and, thus, $\mathcal{F}_n$ is universal. Conversely, if $\mathcal{F}_n$ is perfect for some $\mathbf{p} \in \Omega$, we have, for any $r, r' \in [M]$,

$$\sum_{T \in \mathcal{T}_n} \frac{|\chi \cap T| - |\chi' \cap T|}{|T|} P_{\mathbf{p}}(T) = 0\,, \quad (5)$$

where χ' denotes the corresponding set for r'. Clearly, if the condition of the lemma does not hold, the expression on the left-hand side of (5), viewed as a multivariate polynomial in the components of $\mathbf{p}$, belongs to G. Thus, by the definition of Ω_0, we must have $\mathbf{p} \in \Omega_0$. ∎

Corollary 1: An FVR is universal if and only if it is perfect for some $P \in \Omega \setminus \Omega_0$.

B. Elias's procedure and generalizations

The Elias procedure becomes strikingly simple if we choose $\mathbb{N}_t = \mathbb{N}^+$ (i.e., no restrictions such as (2) on the ranges of the generated random numbers). For a sequence $x^n \in \mathcal{A}^n$, and $T = T(x^n)$, let $\mathcal{I}_T(x^n)$ denote the index of x^n in an enumeration of T. The following procedure defines a very simple FVR $\mathcal{F}_n^* = (\mathbb{N}^+, \rho^*, \mathcal{M}^*)$.

Procedure E1: Given an input sequence x^n, let $\mathcal{M}^*(x^n) = |T(x^n)|$, and $\rho^*(x^n) = \mathcal{I}_T(x^n)$.

It is straightforward to verify that $\mathcal{F}_n^*$ satisfies the condition of Lemma 1 and is, thus, universal. The following theorem shows that $\mathcal{F}_n^*$ attains the maximum possible expected output length of a universal FVR for n and all $P \in \mathcal{P}_k$.

Theorem 1: If $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$ is universal then, for *any* $P \in \mathcal{P}_k$,

$$L_P(\mathcal{M}) \le L_P(\mathcal{M}^*) = \mathbf{E}_P \log |T(X^n)|\,. \quad (6)$$

Proof: The equality is straightforward from the definition of Procedure E1. It follows from Lemma 1 and Corollary 1 that $\mathcal{M}(x^n) \le |T(x^n)|$ for all $x^n \in \mathcal{A}^n$, implying the inequality. ∎

The term on the right-hand side of (6) was precisely estimated in [12] by analyzing the expectation of Whittle's formula, and obtaining

$$\mathbf{E}_P \log |T(X^n)| = H_n(P) - (K/2)\log n + O(1)\,, \quad (7)$$

where $H_n(P)$ denotes the entropy of the marginal $P(X^n)$, and $K = (\alpha - 1)\alpha^k$.

The term $(K/2)\log n$ on the right hand side of (7) resembles a typical "model cost" term in universal lossless compression. By Theorem 1 and Corollary 1, this term determines the rate at which the expected output length of $\mathcal{F}_n^*$ approaches (from below) $H_n(P)$, which sets an upper bound on the convergence rate for *any* universal FVR.

Input: Sequence $x^n \in \mathcal{A}^n$.
Output: Pair (r, M), $M \in \mathbb{N}_t$, $r \in [M]$.

1) Let $\mu = |T(x^n)|$, $r_0 = \mathcal{I}_T(x^n)$.
2) Repeat forever:
   a) Let $M = \lfloor \mu \rfloor_{\mathbb{N}_t}$.
   b) If $r_0 < M$ then output (r_0, M) and **Stop.**
   c) Let $\mu = \mu - M$, $r_0 = r_0 - M$.

Fig. 1. Procedure E2: Generalized Elias procedure ($\mathcal{F}_n^{**}$).

Procedure E1 is similar to a *universal enumerative encoder*, a two-part universal lossless compressor for the class $\mathcal{P}_k$. The encoder differs from the FVR in that it outputs, together with r and in lieu of M, an efficient description of $T(x^n)$. It is known (see, e.g., [14]) that $K \log n + O(1)$ bits are sufficient for this description, resulting in an overall expected code length of

$$H_n(P) + \frac{K}{2}\log n + O(1),$$

which is optimal, up to an additive constant, for any universal lossless compressor for the class $\mathcal{P}_k$. The rate of convergence to the entropy is the same as for FVRs, but convergence, in this case, is from above. It follows from these observations that a universal lossless compressor *cannot be* a universal FVR for $\mathcal{P}_k$ (and vice versa). The estimate (7) was also used, in [12], to derive a model cost for the problem of universal simulation of sources in $\mathcal{P}_k$.

We now shift our attention to arbitrary target sets $\mathbb{N}_t$, including the ones in Elias's original scheme. For any $M \in \mathbb{N}^+$, let

$$\lfloor M \rfloor_{\mathbb{N}_t} = \max\{\, j \in \mathbb{N}_t \,|\, j \le M \,\}\,.$$

Let c be a constant, $c \ge 1$. We say that $\mathbb{N}_t$ is *c-dense* if for any $M \in \mathbb{N}^+$, we have $M \le c\,\lfloor M \rfloor_{\mathbb{N}_t}$. For example, $\mathbb{N}^+$ is 1-dense, and the target set in (2) (used in Elias's procedure for fair p-sided coins) is p-dense.

Procedure E2 in Fig. 1 defines a FVR $\mathcal{F}_n^{**} = (\mathbb{N}_t, \rho^{**}, \mathcal{M}^{**})$ (we recall that $\mathcal{I}_T(x^n)$ denotes the index of x^n in an enumeration of $T(x^n)$). It is readily verified that, since $1 \in \mathbb{N}_t$, Procedure E2 always stops, and its output has the desired form. Also, $\mathcal{F}_n^{**}$ satisfies the condition of Lemma 1 and is, thus, universal.

We refer to Procedure E2 as "greedy," since, in Step 2c, it always chooses to reduce μ by the largest possible element of $\mathbb{N}_t$. The procedure trivially coincides with Procedure E1 when $\mathbb{N}_t = \mathbb{N}^+$. When $\mathbb{N}_t$ is of the form (2), the procedure coincides with Elias's original scheme in [2], suitably extended to finite-memory sources of arbitrary order, and arbitrary p.

Remark 1: Procedure E2 can be regarded as generating a partition of $\mathcal{A}^n$ into classes, with the size of each class belonging to $\mathbb{N}_t$. This partition is a refinement of the original partition into type classes, so that all the sequences in a class

are still equiprobable for all $P \in \mathcal{P}_k$. Then, the output is obtained by applying Procedure E1 to the refined partition.

Theorem 2: If $\mathbb{N}_t$ is c-dense, the expected output length of $\mathcal{F}_n^{**}$ for $P \in \mathcal{P}_k$ is

$$L_P(\mathcal{M}^{**}) = L_P(\mathcal{M}^*) + O(1)\,. \tag{8}$$

Sketch of proof: The key observation in the proof is that for any distribution $\{\mu_1, \mu_2, \ldots, \mu_m\}$ with entropy H, if $\mu_i \geq \gamma \sum_{j=1}^{i} \mu_j$ for all i and some γ, $0 < \gamma < 1$, then $H \leq h(\gamma)/\gamma$, where $h(\cdot)$ denotes the binary entropy function. This claim can be proved by induction on m.

Now, let T denote an arbitrary type class, with $|T| = \mu$, and let $M_1, M_2, \ldots, M_m$ denote the integers in $\mathbb{N}_t$ determined by the decomposition of μ implicit in Procedure E2, sorted from smallest to largest. Since the sequences in a type class are equiprobable, the expectation of $\log \mathcal{M}^{**}(X^n)$ conditioned on T is $\mu^{-1} \sum_{i=1}^{m} M_i \log M_i$. Next, we apply our key observation to the distribution given by $\mu_i = M_i/\mu$ which, by the density assumption, satisfies the required condition with $\gamma = c^{-1}$. The claim of the theorem then follows from taking expectation over the type classes. ∎

It follows from Theorem 2 and Theorem 1 that $\mathcal{F}_n^{**}$ is optimal, up to an additive constant, among all FVRs for the same target set $\mathbb{N}_t$. Furthermore, by (7), an additive constant affects only third order terms in the asymptotic expansion of the expected output length. The following theorem shows that when $\mathbb{N}_t$ is of the form (2), $\mathcal{F}_n^{**}$ is in fact the optimal FVR for $\mathbb{N}_t$. This result was proved for $k{=}0$ in [13], and for $k{=}1$ in [7]. In fact, once the basic properties of type classes are established, the result and proof are rather insensitive to the order k.

Theorem 3: Let $\mathbb{N}_t = \{p^i \,|\, i \geq 0\}$ for some integer $p \geq 2$, let $\mathcal{F}_n{=}(\mathbb{N}_t, \rho, \mathcal{M})$ be a universal FVR, and consider $\mathcal{F}_n^{**}$ with target set $\mathbb{N}_t$. Then, for any n and any $P \in \mathcal{P}_k$, we have

$$L_P(\mathcal{M}) \leq L_P(\mathcal{M}^{**})\,.$$

IV. Twice-universal FVRs

In this section, we assume that the order k of the Markov source is not known, yet we want to produce a universal FVR whose model cost is not larger (up to lower order terms) than the one we would incur had the value of k been given. To this end, as mentioned in Section I, we need to relax our requirement of a uniformly distributed output since, by Theorem 1 and (7), an FVR that is universal in $\mathcal{P}_k$ would incur too high a model cost for any order $k' < k$.[6] We assume throughout that $\mathbb{N}_t$ is c-dense.

Letting $Q_M(r)$ denote the output probability of $r{\in}[M]$, $M \in \mathbb{N}_t$, conditioned on $\mathcal{M}(x^n){=}M$, for an FVR $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$, the distance of $\mathcal{F}_n$ to uniformity is measured by

$$D(\mathcal{F}_n) \triangleq \sum_{M \in \mathbb{N}_t} \frac{\Pr(\mathcal{M}(X^n) = M)}{M} \sum_{r,r' \in [M]} |Q_M(r) - Q_M(r')|\,. \tag{9}$$

[6]Of course, application of Procedure E2 with k replaced with a slowly growing function of n leads, for n sufficiently large, to a perfect FVR for any (fixed, but arbitrary) Markov order. However, the model cost incurred does not meet our demands.

It is easy to see that, for any distribution $R(\cdot)$ with support $\mathcal{B}$,

$$\sum_{x \in \mathcal{B}} \left| R(x) - \frac{1}{|\mathcal{B}|} \right| \leq \frac{1}{|\mathcal{B}|} \sum_{x,y \in \mathcal{B}} |R(x) - R(y)|\,.$$

In particular, the inner summation in (9) is lower-bounded by $M \sum_{r \in [M]} |Q_M(r) - 1/M|$. Therefore, our measure of uniformity is more demanding than the weighted L_1 measure used in [3]. Notice that, as in [3], the measure (9) is *unnormalized*. We aim at FVRs for which $D(\mathcal{F}_n)$ vanishes exponentially fast with n.

As in [9], our twice-universal FVR will rely on the existence of Markov order estimators with certain consistency properties, which are specified in Lemma 2 below. For concreteness, we will focus on a specific estimator, namely a penalized maximum-likelihood estimator that, given a sample x^n from the source, chooses order $k(x^n)$ such that

$$k(x^n) = \arg\min_{k \geq 0} \left\{ \hat{H}_k(x^n) + \alpha^k f(n) \right\} \tag{10}$$

where $\hat{H}_k(x^n)$ denotes the k-th order empirical conditional entropy for x^n, $f(n)$ is a vanishing function of n, ties are resolved, e.g., in favor of smaller orders, and it is assumed that the fixed string determining the initial state is as long as needed (e.g., a semi-infinite all-zero string). For example, $f(n) = (\alpha - 1)(\log n)/(2n)$ corresponds to the asymptotic version of the MDL criterion. The estimate $k(x^n)$ can be obtained in time that is linear in n by use of suffix trees. The set of n-tuples x^n such that $k(x^n) = i$ will be denoted $\mathcal{A}_i^n$. To state Lemma 2 we define, for a distribution $P \in \mathcal{P}_k$, the overestimation probability

$$P_{\mathrm{o/e}}(n) \triangleq \Pr(k(X^n) > k)$$

and, similarly, the underestimation probability

$$P_{\mathrm{u/e}}(n) \triangleq \Pr(k(X^n) < k)\,.$$

Lemma 2: For any $k \geq 0$ and any $P \in \mathcal{P}_k$, the estimator of Equation (10) satisfies

(a) $(n{+}1)^{\alpha^{k+1}} P_{\mathrm{o/e}}(n)$ vanishes polynomially fast (uniformly in P and k) provided $f(n) > \beta(\log n)/n$ for a sufficiently large constant β.

(b) $P_{\mathrm{u/e}}(n)$ vanishes exponentially fast.

Following [9], we consider a partition of $\mathcal{A}^n$ in which the class of x^n, denoted $U(x^n)$, is given by

$$U(x^n) \triangleq T^{(k(x^n))}(x^n) \cap \mathcal{A}^n_{k(x^n)}\,. \tag{11}$$

Thus, two sequences are in the same class if they estimate the same Markov order and are in the same type class with respect to the estimated order. Our twice-universal FVR, $\mathcal{F}_n^{(\mathrm{TU})} = (\mathbb{N}_t, \rho^{(\mathrm{TU})}, \mathcal{M}^{(\mathrm{TU})})$, is given by replacing, in Procedure E2, $T^{(k)}(x^n)$ with $U(x^n)$ and $\mathcal{I}_T(x^n)$ with the index of x^n in an enumeration of $U(x^n)$.

Theorem 4: For $P \in \mathcal{P}_k$, the FVR $\mathcal{F}_n^{(\mathrm{TU})}$ satisfies $D(\mathcal{F}_n^{(\mathrm{TU})}) \leq 2P_{\mathrm{u/e}}$, and, for a suitable choice of $f(n)$ in (10), its expected output length $L_P(\mathcal{M}^{(\mathrm{TU})})$ satisfies

$$L_P(\mathcal{M}^{(\mathrm{TU})}) - L_P(\mathcal{M}^*) = O(1) \tag{12}$$

provided $\mathbb{N}_{\rm t}$ is c-dense.

By Lemma 2, Theorem 4 states that the distance of $\mathcal{F}_n^{(\rm TU)}$ to uniformity is exponentially small whereas, by (6) and (7), its expected output length is essentially the same as that of $\mathcal{F}_n^*$. It should be pointed out, however, that Theorem 4 falls short of stating that the cost of twice-universality in terms of expected output length is asymptotically negligible. The reason is that, in principle, it could be the case that by allowing a small deviation from uniformity, as we do, we open the door for schemes that (with knowledge of k) produce an output significantly larger than $\mathcal{F}_n^*$. We conjecture that, just as in twice-universal simulation [9], this is not the case.

One problem in the implementation of $\mathcal{F}_n^{(\rm TU)}$ is that it requires an efficient enumeration of $U(x^n)$. Such an enumeration appears to be elusive. Instead, the following FVR can be efficiently implemented: Compute $k(x^n)$ and apply Procedure E2 with $k = k(x^n)$. A variant of the proof of Theorem 4 shows that the output length of this scheme still satisfies (12), whereas its distance to uniformity is upper-bounded by $4(P_{\rm u/e} + P_{\rm o/e})$. By Lemma 2, this means that a suitable choice of $f(n)$ still guarantees vanishing distance, but we can no longer claim it to be exponentially small.

Sketch of proof of Theorem 4: Let $\mathcal{U}$ denote the set of classes in the refinement of the partition (11) determined by Procedure E2 (see Remark 1), and let $\mathcal{U}_M$ denote the subset of $\mathcal{U}$ formed by classes of size $M \in \mathbb{N}_{\rm t}$. For $U \in \mathcal{U}_M$, let $\rho_U^{-1}(r)$ denote the unique sequence in U such that $\rho^{(\rm TU)}(\rho_U^{-1}(r)) = r$. Let $Q(r, M)$ denote the probability that $\mathcal{M}^{(\rm TU)}(x^n) = M$ and $\rho^{(\rm TU)}(x^n) = r$, $M \in \mathbb{N}_{\rm t}$, so that

$$Q_M(r) = \frac{Q(r, M)}{\sum_{r\in[M]} Q(r, M)} = \frac{Q(r, M)}{\Pr(\mathcal{M}(X^n) = M)} \,.$$

Clearly,

$$Q(r, M) = \sum_{U\in\mathcal{U}_M} P(\rho_U^{-1}(r)) \,.$$

By (9),

$$\begin{aligned} D(\mathcal{F}_n^{(\rm TU)}) &= \sum_{M\in\mathbb{N}_{\rm t}} \frac{1}{M} \sum_{r,r'\in[M]} |Q(r, M) - Q(r', M)| \\ &\le \sum_{M\in\mathbb{N}_{\rm t}} \frac{1}{M} \sum_{U\in\mathcal{U}_M} \sum_{r,r'\in[M]} |P(\rho_U^{-1}(r)) - P(\rho_U^{-1}(r'))| \end{aligned}$$

which, given the existence of a one-to-one correspondence between $U \in \mathcal{U}_M$ and $[M]$, takes the form

$$D(\mathcal{F}_n^{(\rm TU)}) \le \sum_{M\in\mathbb{N}_{\rm t}} \frac{1}{M} \sum_{U\in\mathcal{U}_M} \sum_{u,v\in U} |P(u) - P(v)| \,. \qquad (13)$$

Now, since U is a subset of a type class $T \in \mathcal{T}_n^{(k(x^n))}$, we have $P(u) = P(v)$ for all $u, v \in U$ whenever $k(x^n) \ge k$. In addition, it can be shown (e.g., by induction on $|\mathcal{B}|$) that, for any distribution $R(\cdot)$ on a set containing $\mathcal{B}$,

$$\sum_{u,v\in\mathcal{B}} |R(u) - R(v)| \le 2(|\mathcal{B}| - 1)R(\mathcal{B}) \,.$$

Therefore, letting $\mathcal{U}_M^{\rm u/e}$ denote the subset of $\mathcal{U}_M$ formed by all the classes such that $k(x^n) < k$, (13) implies

$$D(\mathcal{F}_n^{(\rm TU)}) \le \sum_{M\in\mathbb{N}_{\rm t}} \frac{2(M-1)}{M} \sum_{U\in\mathcal{U}_M^{\rm u/e}} P(U) \le 2P_{\rm u/e}$$

as claimed. As for the expected output length, lower-bounding by 0 the output length produced by sequences which are not in $\mathcal{A}_k^n$, and noting that the claim of Theorem 2 is valid not only for expectations conditioned on a type (as implicit in its proof), but also when conditioning on subsets of types, we have

$$L_P(\mathcal{M}^{(\rm TU)}) \ge \sum_{T\in\mathcal{T}_n^{(k)}} P(T \cap \mathcal{A}_k^n) \log |T \cap \mathcal{A}_k^n| + O(1) \,.$$

By [9, Lemma 1], the number of sequences in a type class T that estimate order k is $|T| - o(1)$ for suitable choices of $f(n)$, provided that at least one sequence in T estimates order k (i.e., almost all the sequences in the type class estimate the right order). Therefore,

$$L_P(\mathcal{M}^{(\rm TU)}) \ge \mathbf{E}_P \log |T_k(X^n)| - n(P_{\rm u/e} + P_{\rm o/e}) + O(1) \,.$$

The claim then follows from Lemma 2. ∎

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] P. Elias, "The efficient construction of an unbiased random sequence," *Ann. Math. Statist.*, vol. 43, pp. 865–870, 1972.

[3] S. Vembu and S. Verdú, "Generating random bits from an arbitrary source: fundamental limits," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1322–1332, 1995.

[4] R. Shaltiel, "An introduction to randomness extractors," in *ICALP (2)*, 2011, pp. 21–41.

[5] G. Seroussi and M. Weinberger, "Universal variable to fixed-length random number generators for finite memory sources," submitted, ISIT'13.

[6] 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.

[7] H. Zhou and J. Bruck, "Efficient generation of random bits from finite state Markov chains," *IEEE Trans. Inform. Theory*, vol. 58, pp. 2490–2506, 2012.

[8] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[9] A. Martín, N. Merhav, G. Seroussi, and M. J. Weinberger, "Twice-universal simulation of markov sources and individual sequences," *IEEE Trans. Inform. Theory*, vol. 56, pp. 4245–4255, 2010.

[10] P. Whittle, "Some distribution and moment formulae for the Markov chain," *J. Roy. Statist. Soc. Ser. B*, vol. 17, 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] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, pp. 384–396, 1994.