Universal Delay–Limited Simulation[*]

Neri Merhav[†]
Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel
Email: merhav@ee.technion.ac.il

Gadiel Seroussi[‡]
Mathematical Sciences Research Institute
Berkeley, CA 94720, U.S.A.
Email: gadiel@msri.org

Marcelo Weinberger
Hewlett-Packard Laboratories
Palo Alto, CA 94304, U.S.A.
Email: marcelo@hpl.hp.com

***Abstract*— Universal, delay–limited simulation of an unknown information source of a certain parametric family (e.g., the family of memoryless sources or Markov sources of a given order), given a training sequence from that source and a stream of purely random bits, is considered. In the delay–limited setting, the simulation algorithm generates a random sequence sequentially, by delivering one symbol for each training symbol that is made available after a given initial delay, whereas the random bits are assumed to be available on demand. The goal of universal simulation is that the probability law of the generated sequence be identical to that of the training sequence, with minimum mutual information between the random processes generating both sequences. In this paper, the optimal universal delay–limited simulation scheme is characterized, and an upper bound on the expected number of random bits it consumes is presented. As in the non-sequential case, the upper bound is related to the entropy rate of the source. The results are extended to a setting of variable delay.**

Index Terms: Random number generators, random process simulation, universal simulation, mutual information, method of types, enumeration.

I. Introduction

Simulation of random processes is about artificial generation of random data with a prescribed probability law, by using a certain deterministic mapping from a source of purely random (independent, equally likely) bits into sample paths. The simulation problem finds applications in speech and image synthesis, texture reproduction, generation of noise for purposes of simulating communication systems, and cryptography.

The simulation problem of sources and channels has been investigated by several researchers, see, e.g., [1], [2], [3], [4], [5], [6], [7]. In all these works, the common assumption is that the probability law of the desired process is perfectly known. Recently, universal versions of this problem were studied in [8], [9], [10], and [11]. In [8],[9], the assumption of perfect knowledge of the target probability law is relaxed. Specifically, the target source P to be simulated is assumed to belong to a certain parametric family $\mathcal{P}$ (like the family of finite–alphabet memoryless sources, Markov sources of a given order, parametric subsets of these families, etc.) but is otherwise unknown, and a training sequence $x^m = (x_1, \ldots, x_m)$ that has emerged from this source is available. In addition, the simulation scheme is provided with a stream of k purely random bits $u^k = (u_1, \ldots, u_k)$ that are statistically independent of the training sequence. The goal of the simulation schemes in [8], [9] is to generate an output sequence $y^n = (y_1, \ldots, y_n)$, $n \leq m$, corresponding to the simulated process, such that $y^n = \phi(x^m, u^k)$, where ϕ is a deterministic function that does not depend on the unknown source P, and which satisfies the following two conditions:

C1. The probability distribution of the output sequence is *exactly* the n-dimensional marginal of the probability law P corresponding to the training sequence for all $P \in \mathcal{P}$.

C2. The mutual information between the training sequence and the output sequence is as small as possible (or equivalently, under Condition C1, the conditional entropy of the output sequence given the training sequence is as large as possible), simultaneously for all $P \in \mathcal{P}$ (so as to make the generated sample path as "original" as possible).

In [8], the smallest achievable value of the mutual information as a function of n, m, k, and the entropy rate H of the source P is characterized, and simulation schemes that asymptotically achieve these bounds are presented. It is shown in [8] that in order to satisfy Condition C1, it is necessary that the output y^n be a prefix of a sequence y^n having the same *type* [12] as x^m with respect to $\mathcal{P}$ (when $\mathcal{P}$ is the entire class of i.i.d. sources over a finite alphabet, this means that x^m and y^m have the same *composition*, namely yield the same empirical distribution [13]). Moreover, it is shown that for k large enough, the optimal simulation scheme essentially takes the first n symbols of a randomly selected sequence of the same type as x^m.

In [10], the goal was to characterize the minimum key rate required in order to generate a collection of N output sequences $\{(y^m)_i\}_{i=1}^N$, all governed by the same probability law as the given training vector x^m, such that a certain, prescribed set of statistical tests would be satisfied. In [11], x^m is assumed to be an individual sequence not originating from any probabilistic source. Simulation in this setting is based on an extension of the conventional notion of type, referred to in [11] as a *universal type*.

In this paper, we investigate the universal simulation prob-

[*] The material in this paper was presented in part at the 2005 IEEE International Symposium on Information Theory, Adelaide, Australia, September 2005.

[†] This work was done while N. Merhav was visiting Hewlett–Packard Laboratories, Palo Alto, CA, U.S.A.

[‡] This work was essentially done while G. Seroussi was with Hewlett–Packard Laboratories, Palo Alto, CA, U.S.A.

lem, as stated in [8], in a sequential, delay–limited setting. In this setting, upon observing an initial training $(d-1)$-tuple x^{d-1}, where d is some fixed initial delay, the simulation scheme is requested to output one symbol y_t for every additional symbol x_{t+d-1} it observes, $1 \leq t \leq n$. Thus, y^t is a (randomized) function of x^{t+d-1} but, unlike in [8], not of x_{t+d+i}, $i \geq 0$. In order to generate an output sequence satisfying conditions C1–C2 (with $m = n+d-1$), the simulation scheme has also access to a stream of purely random bits $\{u_i\}$ (the *key*), which are available "on demand." This assumption differs from the setting in [8] in that there is no fixed budget of key bits; rather, as in [5] and [11], we will be interested in the *expected* number of key bits that the scheme consumes in order to generate its output, where, here, the expectation is with respect to $\{u_i\}$ and P. Thus, a delay–limited simulation scheme is given by a sequence of conditional probability distributions $\{W_t(y_t|x^{t+d-1}, y^{t-1})\}$. It is well known (cf. [5]) that a corresponding sequence of draws can be implemented with an Elias decoder [14, pp. 479–482]. To output y_t, the Elias decoder is tuned to the distribution $W_t(\cdot|x^{t+d-1}, y^{t-1})$, and uses the random bitstream as its input. As shown in [5], the expected number of key bits that the decoder consumes to (sequentially) produce y^n is upper-bounded by the conditional entropy of the resulting product distribution $W(y^n|x^{n+d-1})$, plus 3 bits (see [11] for algorithmic details).

A special, trivial case of the delay–limited universal simulation problem is obtained when pure sequentiality is required, namely when the allowed delay d is 1, and $\mathcal{P}$ is the entire class of i.i.d. sources over a finite alphabet. In this case, due to the constraint that y^n must be of the same type as x^n in order to satisfy Condition C1, the simulation scheme can only copy the input, and therefore the conditional entropy of the output sequence given the training sequence, vanishes. Thus, the problem becomes interesting as d grows.

As it turns out, for a broad class of families $\mathcal{P}$, the optimal simulation scheme that preserves the probability law (Condition C1) while minimizing the mutual information simultaneously for all $P \in \mathcal{P}$ (Condition C2), takes a form that is reminiscent of an enumerative sequential decoding scheme [15] in which the enumerated set varies as the training data becomes available. For example, when $\mathcal{P}$ is the entire class of i.i.d. sources over a finite alphabet, the simulation scheme draws a symbol a at time t with probability equal to the empirical probability of a in any d-tuple z^d such that $y^{t-1}z^d$ and x^{t+d-1} have the same composition (it is easy to see, by induction, that such a sequence z^d will always exist). In other words, for any symbol a we have

$$W_t(a|x^{t+d-1}, y^{t-1}) = \frac{n_a(x^{t+d-1}) - n_a(y^{t-1})}{d} \qquad (1)$$

where $n_a(v^\ell)$ denotes the number of occurrences of a symbol a in an ℓ–vector v^ℓ. The distribution assigned by the scheme is precisely the one that an enumerative decoder would use to decode y_t sequentially, given that y^{t+d-1} has the same composition as x^{t+d-1}. The corresponding conditional entropy of the output sequence given the entire training sequence, which upper-bounds (up to an additive constant) the expected key length required for implementing the scheme, equals, after normalization by the number of output symbols, the expectation under P of the empirical entropy of d-tuples (and is therefore independent of n and m). By [17], this expectation falls short of the entropy rate H by an $O(1/d)$ term. Since, by Condition C1, the entropy of the output vector is the same as the entropy of a training vector of length n, this term is precisely the normalized mutual information between the input and the output. The above results are actually special cases of those to be presented in Section III for more general parametric families of sources with memory.

In the remainder of this paper, Section II introduces the main concepts and notation. Our main results are then presented in Section III, and extended to the case of arbitrary request schedules (encompassing also the "batch" simulation case) in Section IV.

II. Notation and Problem Formulation

Throughout the paper, random variables will be denoted by capital letters and specific values they may take will be denoted by the corresponding lower case letters. The same convention will apply to random vectors, with an additional superscript denoting their dimension. Thus, x^m, y^n, and u^k will denote specific vector values of the random vectors X^m, Y^n, and U^k, respectively. If the dimension is omitted, random vectors will be denoted in bold. A generic parametric family of sources will be denoted by $\mathcal{P}$, and a particular source in $\mathcal{P}$, defined by a parameter vector θ taking values over some parameter space Ω, will be denoted by P_θ. However, in a context where the parameter value is either fixed or irrelevant, we will omit it, denoting a source in $\mathcal{P}$ simply by P. The (finite) source alphabet is denoted by $\mathcal{A}$.

A finite-state machine (FSM) over a finite state set $\mathcal{S}$ will be identified with its next-state function $g: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$, and will be assumed to start at a given initial state $s_0 \in \mathcal{S}$. A parametric family to which we will refer frequently as an important particular case is the class $\mathcal{F}_{g,s_0}$ of all FSM sources over $\mathcal{A}$ driven by the next-state function g, starting at state s_0, with the parameters given by the transition probabilities. Thus, if $\mathcal{P}$ is a parametric subfamily of $\mathcal{F}_{g,s_0}$ and t is a given positive integer, the probability of a t-vector $x^t = (x_1, x_2, \ldots, x_t)$ drawn from $P \in \mathcal{P}$, $x_i \in \mathcal{A}$, $i = 1, \ldots, t$, is given by

$$\Pr\{X_i = x_i,\ i = 1, \ldots, t\} = \prod_{i=1}^{t} P(x_i|s_{i-1}) \stackrel{\triangle}{=} P(x^t)$$

where $s_0, s_1, \ldots, s_{t-1} \in \mathcal{S}$ denotes the sequence of states assumed by the FSM. We shall define the *type class* [12] T_{x^m} of a vector x^m as the set of all vectors $\tilde{x}^m \in \mathcal{A}^m$ such that $P(\tilde{x}^m) = P(x^m)$ for *every* source $P \in \mathcal{P}$. The set of all type classes of vectors in $\mathcal{A}^m$ will be denoted by $\mathcal{T}^m$. For example, in case $\mathcal{P} = \mathcal{F}_{g,s_0}$ (or if $\mathcal{P}$ is a subclass of $\mathcal{F}_{g,s_0}$ that admits the same *minimal* parametrization as the entire class $\mathcal{F}_{g,s_0}$, but for which Ω is a subset of the entire space), T_{x^m} is the set of all vectors having the same composition as x^m with respect to g [12], [13] (i.e., each state transition occurs as many times in $\tilde{x}^m \in T_{x^m}$ as in x^m, starting from state s_0). Given another sequence $y^n \in \mathcal{A}^n$, $n \le m$, $r = m - n$, let $T_{x^m \setminus y^n} = \{z^r \in \mathcal{A}^r : \; y^n z^r \in T_{x^m}\}$, which is interpreted as a “difference type.” Notice that for $r = 0$, $\mathcal{A}^0 = \{\lambda\}$ (where λ denotes the null string), $|\mathcal{A}^0| = 1$, and thus $|T_{x^m \setminus y^m}| = 1$ if $T_{x^m} = T_{y^n}$ and 0 otherwise. For a family of FSM sources, if the initial state is not assumed to be s_0 but a generic state $s \in \mathcal{S}$, the type of x^m and the difference type will be denoted $T^s_{x^m}$ and $T^s_{x^m \setminus y^n}$, respectively. We will reserve the notation $s_0, s_1, \ldots, s_n$ to denote the state sequence over which the FSM evolves with y^n. Notice that for any sequence $z^r \in T_{x^m \setminus y^n}$, we have $T_{x^m \setminus y^n} = T^{s_n}_{z^r}$.

The probability of every type class $T \in \mathcal{T}^m$ is given by

$$P_\theta(T) \overset{\triangle}{=} \sum_{\tilde{x}^m \in T} P_\theta(\tilde{x}^m) = |T| \cdot P_\theta(x^m) \tag{2}$$

where x^m is a sequence in T and, throughout, $|T|$ denotes the cardinality of T. Next, given some enumeration of $\mathcal{T}^m$, let $T^{(1)}, T^{(2)}, \ldots, T^{(|\mathcal{T}^m|)}$ denote the corresponding type classes. For each j, $1 \le j \le |\mathcal{T}^m|$, $P_\theta(T^{(j)})$ can be regarded as a function of the parameter vector θ defining $P_\theta \in \mathcal{P}$. Following [8], we will assume that the class of sources $\mathcal{P}$ satisfies the following assumption:

A1. *The set* $\{P_\theta(T^{(j)})\}_{j=1}^{|\mathcal{T}^m|}$ *(as functions of* $\theta \in \Omega$*) is linearly independent over* $\mathbb{R}$.

As shown in [8], Assumption A1 is satisfied for a broad class of parametric families, including any i.i.d. exponential family for suitable Ω, or any family $\mathcal{F}_{g,s_0}$.

A simulation scheme with delay limitation d and horizon n consists of a sequence of conditional probability distributions $\{W_t(y_t|x^{t+d-1}, y^{t-1})\}_{t=1}^n$, where x^{n+d-1} is the training sequence and y^n is the output. To generate a sequence of draws y^n distributed accordingly, it will be assumed that a stream $\{U_i\}$ of purely random bits, independent of X^{n+d-1}, is available on demand. The resulting conditional distribution on y^n, which is regarded as a channel, will be denoted by $W(y^n|x^{n+d-1})$, namely

$$W(y^n|x^{n+d-1}) = \prod_{t=1}^{n} W_t(y_t|x^{t+d-1}, y^{t-1}) \,. \tag{3}$$

In the sequel, we alternate freely between the simulation scheme $\{W_t\}$ and the corresponding channel W, referring to $\{W_t\}$ when the emphasis is on sequentiality, and to W when we discuss “batch” properties of the channel.

The conditional entropy achieved by the channel W with the input source P will be denoted $H(Y^n|X^{n+d-1})$. Finally, let $I(X^{n+d-1}; Y^n)$ denote the mutual information between X^{n+d-1} and Y^n that is induced by the source P and the channel W. We seek a delay–limited simulation scheme that meets conditions C1–C2 that were itemized in Section I, for $m = n + d - 1$.

III. Main results

A. The general case

We first state a necessary and sufficient condition for any simulation scheme (not necessarily with a delay limitation) to satisfy Condition C1. Here, a simulation scheme is simply a channel $W(y^n|x^m)$, $m \ge n$.

Lemma 1: Assume $\mathcal{P}$ satisfies Assumption A1. Then, a channel W satisfies Condition C1 if and only if for all sequences $\mathbf{x} \in \mathcal{A}^m$ and $\mathbf{y} \in \mathcal{A}^n$ we have

$$\sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W(\mathbf{y}|\tilde{\mathbf{x}}) = |T_{\mathbf{x} \setminus \mathbf{y}}| \,. \tag{4}$$

Proof. Clearly, Condition C1 is satisfied if and only if for all $\mathbf{y} \in \mathcal{A}^n$ and $P \in \mathcal{P}$ we have

$$\sum_{\mathbf{x} \in \mathcal{A}^m} P(\mathbf{x}) W(\mathbf{y}|\mathbf{x}) = \sum_{T_{\mathbf{x}} \in \mathcal{T}^m} P(\mathbf{x}) \sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W(\mathbf{y}|\tilde{\mathbf{x}}) = P(\mathbf{y}) \,.$$

Now,

$$P(\mathbf{y}) = \sum_{\mathbf{z} \in \mathcal{A}^{m-n}} P(\mathbf{y}\mathbf{z}) = \sum_{T_{\mathbf{x}} \in \mathcal{T}^m} P(\mathbf{x}) |T_{\mathbf{x} \setminus \mathbf{y}}|$$

where the last equality follows from the fact that each type $T_{\mathbf{x}}$ contains $|T_{\mathbf{x} \setminus \mathbf{y}}|$ sequences prefixed by $\mathbf{y}$. Therefore, Condition C1 is satisfied if and only if

$$\sum_{T_{\mathbf{x}} \in \mathcal{T}^m} P(\mathbf{x}) \left[|T_{\mathbf{x} \setminus \mathbf{y}}| - \sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W(\mathbf{y}|\tilde{\mathbf{x}}) \right] = 0 \,.$$

The claim then follows from Assumption A1. □

Notice that Lemma 1 implies that a simulation scheme satisfying Condition C1, when trained with a sequence $\mathbf{x}$, can only output sequences $\mathbf{y}$ such that $T_{\mathbf{x} \setminus \mathbf{y}}$ is nonempty. In other words, $\mathbf{y}$ must be a prefix of a sequence in $T_{\mathbf{x}}$; in [8], such a sequence $\mathbf{y}$ is said to be *feasible* with respect to $\mathbf{x}$. Now, a simulation scheme with delay limitation d and horizon n defines a sequence of simulation schemes that output y^t with training data x^{t+d-1}, $1 \le t \le n$. Clearly, these reduced-horizon schemes preserve the probability law if so does the scheme with horizon n, and therefore also satisfy Condition C1. It then follows that *every* prefix y^t of y^n must be feasible with respect to x^{t+d-1}.

Our main result states that the optimal simulation scheme with delay limitation d and horizon n, in the sense of minimizing the mutual information $I(X^{n+d-1}; Y^n)$ simultaneously for all $P \in \mathcal{P}$ among schemes that preserve the probability law (and, therefore, satisfy the condition (4)), is given by

$$W^*_t(y_t|x^{t+d-1}, y^{t-1}) = \frac{|T_{x^{t+d-1} \setminus y^t}|}{|T_{x^{t+d-1} \setminus y^{t-1}}|} \,. \tag{5}$$

Equation (5) indeed defines a conditional probability distribution on $\mathcal{A}$ since, for any pair of sequences $\mathbf{x}$ and $\mathbf{y}$, the definition of a difference type implies that $\bigcup_{a \in \mathcal{A}} T_{\mathbf{x} \setminus \mathbf{y}a} = T_{\mathbf{x} \setminus \mathbf{y}}$. Specifically, $W^*_t(a|x^{t+d-1}, y^{t-1})$ is the fraction of sequences in $T_{x^{t+d-1} \setminus y^{t-1}}$ starting with a. For the scheme (5) to preserve the probability law, we need the following additional assumption on $\mathcal{P}$:

A2. *If* $T_{\mathbf{y_1}} = T_{\mathbf{y_2}}$ *then for every* $\mathbf{x}$ *we have* $|T_{\mathbf{x} \setminus \mathbf{y_1}}| = |T_{\mathbf{x} \setminus \mathbf{y_2}}|$.

Notice that this assumption holds trivially in the i.i.d. case, since in this case $T_{\mathbf{x} \setminus \mathbf{y}}$ depends on $\mathbf{y}$ only through its type. For a general FSM class, $T_{\mathbf{x} \setminus \mathbf{y}}$ may also depend on the final state to which the FSM evolves with $\mathbf{y}$. However, if $\mathcal{P}$ is a class $\mathcal{F}_{g,s_0}$ parametrized by the transition probabilities, then the assumption still holds since, in this case, two sequences $\mathbf{y_1}$ and $\mathbf{y_2}$ of the same type bring the FSM to the same final state (as the number of transitions between any pair of states is the same for both sequences, with the initial and final states being the only ones for which the nodes in the underlying graph may have an outgoing degree which differs from the incoming degree).

The conditional entropy achieved by the channel W^*, derived from (5) as in (3), will be denoted $H^*(Y^n|X^{n+d-1})$. We have the following optimality result for the delay–limited simulation scheme $\{W_t^*\}$.

Theorem 1: Assume $\mathcal{P}$ satisfies assumptions A1 and A2.

(a) The channel W^* satisfies Condition C1.

(b) For any simulation scheme $\{W_t\}$ with delay limitation d that satisfies Condition C1 we have

$$H(Y^n|X^{n+d-1}) \leq H^*(Y^n|X^{n+d-1}) \tag{6}$$

with equality if and only if $\{W_t\} = \{W_t^*\}$.

Proof. Part (a) follows from showing, by induction in n, that the simulation scheme satisfies the condition (4) of Lemma 1. For $n = 1$ we have, for any $\mathbf{x} \in \mathcal{A}^d$ and $y \in \mathcal{A}$,

$$\sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W^*(y|\tilde{\mathbf{x}}) = \frac{|T_{\mathbf{x}\backslash y}|}{|T_{\mathbf{x}}|} \cdot |T_{\mathbf{x}}| = |T_{\mathbf{x}\backslash y}| \,.$$

Assume now that the condition holds for $n = t-1$ and, for $\mathbf{y} \in \mathcal{A}^t$, let $\mathbf{y} = \mathbf{y}'a$, $a \in \mathcal{A}$. Then, for any $\mathbf{x} \in \mathcal{A}^{t+d-1}$,

$$\begin{aligned} \sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W^*(\mathbf{y}|\tilde{\mathbf{x}}) &= \sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W^*(\mathbf{y}'|\tilde{\mathbf{x}}) \frac{|T_{\tilde{\mathbf{x}}\backslash \mathbf{y}}|}{|T_{\tilde{\mathbf{x}}\backslash \mathbf{y}'}|} \\ &= \frac{|T_{\mathbf{x}\backslash \mathbf{y}}|}{|T_{\mathbf{x}\backslash \mathbf{y}'}|} \sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W^*(\mathbf{y}'|\tilde{\mathbf{x}}) \end{aligned} \tag{7}$$

where the second equality follows from the fact that the difference types depend on $\tilde{\mathbf{x}}$ only through its type. To evaluate the summation on the right-hand side of (7), we first notice that $W^*(\mathbf{y}'|\tilde{\mathbf{x}})$ is independent of the last symbol of $\tilde{\mathbf{x}}$. Thus, we consider the multiset of sequences obtained by deleting the last symbol of each $\tilde{\mathbf{x}} \in T_{\mathbf{x}}$. By Assumption A2, two sequences of the same type in this multiset can be extended in the same number of ways to obtain sequences in $T_{\mathbf{x}}$. Thus, the summation can be broken into a number of partial summations, each over an entire type in $\mathcal{T}^{t+d-2}$, to which the induction hypothesis can be applied. As a result, each partial summation equals the number of sequences $\mathbf{z}$ such that $\mathbf{y}'\mathbf{z}$ ranges over the corresponding type in $\mathcal{T}^{t+d-2}$. Adding these partial contributions, we obtain $|T_{\mathbf{x}\backslash \mathbf{y}'}|$, which by (7) completes the induction step.

As for Part (b), for any simulation scheme $\{W_t\}$ and any $t \geq 1$, we have

$$H(Y_t|X^{t+d-1}, Y^{t-1}) = \sum_{\mathbf{y} \in \mathcal{A}^{t-1}} \sum_{T_{\mathbf{x}} \in \mathcal{T}^{t+d-1}} \sum_{\mathbf{x} \in T_{\mathbf{x}}} P(\mathbf{x}) W(\mathbf{y}|\mathbf{x}) H(Y_t|\mathbf{x}, \mathbf{y}).$$

If W preserves the probability law then it does so for all horizons t, $1 \leq t \leq n$, and also for the marginals. Thus, by Lemma 1, we have $\sum_{\mathbf{x} \in T_{\mathbf{x}}} W(\mathbf{y}|\mathbf{x}) = |T_{\mathbf{x}\backslash \mathbf{y}}| = \sum_{\mathbf{x} \in T_{\mathbf{x}}} W^*(\mathbf{y}|\mathbf{x})$ (as W^* also preserves the law). By Jensen's inequality we then have

$$\begin{aligned} &H(Y_t|X^{t+d-1}, Y^{t-1}) \leq \\ &\sum_{\mathbf{y} \in \mathcal{A}^{t-1}} \sum_{T_{\mathbf{x}} \in \mathcal{T}^{d+t-1}} P(\mathbf{x}) \left[\sum_{\mathbf{x} \in T_{\mathbf{x}}} W^*(\mathbf{y}|\mathbf{x}) \right] \\ &\quad H\left(\frac{1}{|T_{\mathbf{x}\backslash \mathbf{y}}|} \sum_{\mathbf{x} \in T_{\mathbf{x}}} W(\mathbf{y}|\mathbf{x}) W_t(\cdot|\mathbf{x}, \mathbf{y}) \right) \end{aligned} \tag{8}$$

with equality if and only if $W_t(\cdot|\mathbf{x}, \mathbf{y})$ depends on $\mathbf{x}$ only through $T_{\mathbf{x}}$. Again by Lemma 1, the argument of the entropy function is $|T_{\mathbf{x}\backslash \mathbf{y}y_t}|/|T_{\mathbf{x}\backslash \mathbf{y}}|$, which is precisely $W_t^*(y_t|\mathbf{x}, \mathbf{y})$, so the right-hand side of (8) is $H^*(Y_t|X^{t+d-1}, Y^{t-1})$. Equation (6) then follows by the chain rule of conditional entropies and the independence of Y_t from X_{t+i}, $i \geq d$, for both $\{W_t\}$ and $\{W_t^*\}$. To complete the proof, it suffices to show that $\{W_t^*\}$ is the *only* law-preserving scheme that, for every $t \geq 1$, depends on $\mathbf{x}$ only through $T_{\mathbf{x}}$. To this end, let $\{W_t\}$ be any such scheme. Then, with $\mathbf{y} = \mathbf{y}'a$, $a \in \mathcal{A}$, $\mathbf{y} \in \mathcal{A}^t$, and $\mathbf{x} \in \mathcal{A}^{t+d-1}$, we have

$$\begin{aligned} |T_{\mathbf{x}\backslash \mathbf{y}}| &= \sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W(\mathbf{y}|\tilde{\mathbf{x}}) = \sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W(\mathbf{y}'|\tilde{\mathbf{x}}) W_t(a|\tilde{\mathbf{x}}, \mathbf{y}') \\ &= W_t(a|\mathbf{x}, \mathbf{y}') \sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W(\mathbf{y}'|\tilde{\mathbf{x}}) = W_t(a|\mathbf{x}, \mathbf{y}') \, |T_{\mathbf{x}\backslash \mathbf{y}'}| \end{aligned}$$

where the first and the last equalities in the above chain follow from Lemma 1, and the third equality follows from the assumed property of $\{W_t\}$. Therefore,

$$W_t(a|\mathbf{x}, \mathbf{y}') = \frac{|T_{\mathbf{x}\backslash \mathbf{y}}|}{|T_{\mathbf{x}\backslash \mathbf{y}'}|} = W_t^*(a|\mathbf{x}, \mathbf{y}') \,. \qquad \square$$

Notice that the simulation scheme $\{W_t^*\}$ is strictly optimal for *every* value of n, and not merely in an asymptotic sense. Moreover, the simulation scheme is horizon-independent, and therefore it also minimizes $I(X^{t+d-1}; Y^t)$ and preserves the probability law for Y^t for all $t \geq 1$.

For an intuitive interpretation of $\{W_t^*\}$, observe that a draw of a sequence $v_1, v_2, \cdots, v_d$ uniformly at random from a given set can be implemented sequentially by simply drawing each symbol v_i according to the distribution given by the fraction of the sequences in the set which start with $v_1, v_2, \cdots, v_i$ (as $v_1, v_2, \cdots, v_{i-1}$ is already available at the time of the i-th draw). Now, since $W_t^*(a|x^{t+d-1}, y^{t-1})$ is the fraction of sequences in $T_{x^{t+d-1}\backslash y^{t-1}}$ which start with a, then $W_t^*(\cdot|x^{t+d-1}, y^{t-1})$ is the distribution that a scheme drawing uniformly at random from $T_{x^{t+d-1}\backslash y^{t-1}}$ would assign to the *first* draw. In our delay–limited setting, however, a new symbol x_{t+d} is observed after this draw, thus renewing the process. In other words, the scheme aims at outputting a random sequence with the same type as the input, as the latter becomes sequentially available. Another interpretation is that $W_t^*(\cdot|x^{t+d-1}, y^{t-1})$ is the distribution that an enumerative decoder would use to decode y_t sequentially, given that y^{t+d-1} has the same type as x^{t+d-1}.

B. The optimal scheme in the i.i.d. and FSM cases

To investigate the optimal scheme for an FSM family, it is convenient to define, for a given state $s \in \mathcal{S}$ and a given sequence $\mathbf{z} \in \mathcal{A}^d$, a probability distribution $Q_{s,\mathbf{z}}(\cdot)$ on $\mathcal{A}$ given by the fraction of sequences in $T^s_{\mathbf{z}}$ that start with a, namely

$$Q_{s,\mathbf{z}}(a) = \frac{|T^s_{\mathbf{z}\backslash a}|}{|T^s_{\mathbf{z}}|},\ a \in \mathcal{A} \tag{9}$$

(in the i.i.d. case, the distribution will be denoted $Q_{\mathbf{z}}(\cdot)$). Clearly,

$$W^*_t(y_t|x^{t+d-1}, y^{t-1}) = Q_{s_{t-1},\mathbf{z}}(y_t) \tag{10}$$

for any sequence $\mathbf{z}$ in $T^{s_{t-1}}_{x^{t+d-1}\backslash y^{t-1}}$, where we recall that s_{t-1} denotes the state to which the FSM evolves with y^{t-1}. In particular, applying the law preservation property to W^*_1, we have

$$\sum_{\mathbf{z}\in\mathcal{A}^d} P(\mathbf{z}|s) Q_{s,\mathbf{z}}(a) = P(a|s)\,,a \in \mathcal{A}, s \in \mathcal{S}\,.$$

Therefore, $Q_{s,\mathbf{z}}(\cdot)$ is a consistent estimate of $P(\cdot|s)$.

When $\mathcal{P}$ is a class of i.i.d. sources over $\mathcal{A}$ with the parameters given by the symbol probabilities, the size of a type is given by a multinomial coefficient [13], and the ratios defining $Q_{\mathbf{z}}(\cdot)$ take the form (after cancellation of terms) of the empirical probability of a in $\mathbf{z}$, as given in Equation (1), making the computation of the optimal scheme practical. The fact that the estimate $Q_{\mathbf{z}}(\cdot)$ coincides with the maximum-likelihood (ML) probability holds for more general i.i.d. subfamilies (e.g., a subfamily of "symmetric" discrete sources for which symbols are grouped by pairs, with both symbols in a pair having the same probability). However, there exist simple exponential families for which $Q_{\mathbf{z}}(\cdot)$ differs from the ML probability estimate.

When $\mathcal{P} = \mathcal{F}_{g,s_0}$ (or $\mathcal{P}$ is a subclass of $\mathcal{F}_{g,s_0}$ that admits the same *minimal* parametrization as the entire class $\mathcal{F}_{g,s_0}$, but for which Ω is a subset of the entire space), the computation of W^*_t in (10) reduces, by (9), to computing the ratio between the size of an FSM-type class in $\mathcal{T}^{d-1}$ (with initial state $g(s_{t-1}, a)$), and the size of an FSM-type class in $\mathcal{T}^d$ (with initial state s_{t-1}), by application of Whittle's formula for the type size [16]. To state this formula we introduce some further notation.

Let $n_{sa}(\mathbf{u})$ denote the number of occurrences of symbol $a \in \mathcal{A}$ at state s, in a sequence $\mathbf{u}$, where (to simplify the notation) the initial state will be understood from the context. Similarly, let $n_{ss'}(\mathbf{u})$ (where symbol a has been replaced by state s') denote the number of transitions from state s to state s' in $\mathbf{u}$. Further, let $n_{s*}(\mathbf{u}) = \sum_{a\in\mathcal{A}} n_{sa}(\mathbf{u}) = \sum_{s'\in\mathcal{S}} n_{ss'}(\mathbf{u})$ denote the number of occurrences of state s, excluding the occurrence of the final state from the count. The corresponding conditional (empirical) probability distribution is defined as

$$\hat{P}_{\mathbf{u}}(a|s) = \frac{n_{sa}(\mathbf{u})}{n_{s*}(\mathbf{u})}$$

for any state s such that $n_{s*}(\mathbf{u}) > 0$. Since all these counts are invariant over the type classes of $\mathcal{F}_{g,s_0}$, we may abuse the notation by replacing the argument $\mathbf{u}$ with its type T. For a type class T, let Φ^T denote the matrix whose rows and columns are labeled by the states in $\mathcal{S}$, such that the (s, s') entry is the count $n_{ss'}(T)$. Let Ψ^T denote the matrix obtained by subtracting Φ^T, after normalizing every non-zero row, from the $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix, namely

$$\Psi^T_{ss'} = \delta_{ss'} - \frac{n_{ss'}(T)}{n_{s*}(T)}$$

if $n_{s*}(T) > 0$, and $\Psi^T_{ss'} = \delta_{ss'}$ otherwise, where $\delta_{ss'}$ denotes the Kronecker delta function. It can be shown that all entries (s, s') such that $n_{s'*}(T) > 0$, in a fixed row s of Ψ^T, have the same cofactor, independently of s',[1] which we denote $\Psi^T(s)$. Whittle's formula states that

$$|T| = \Psi^T(\sigma) \cdot \prod_{s\in\mathcal{S}} \frac{n_{s*}(T)!}{\prod_{a\in\mathcal{A}} n_{sa}(T)!}$$

where σ denotes the final state for T. Letting $\mathbf{z}$ denote any sequence in $T^{s_{t-1}}_{x^{t+d-1}\backslash y^{t-1}}$, for which the FSM assumes the state sequence $\sigma_0, \sigma_1, \cdots, \sigma_d$ (where $\sigma_0 = s_{t-1}$ and the notation emphasizes the distinction between this state sequence and the sequence $s_0, \cdots, s_n$ induced by y^n), and denoting $T = T^{s_{t-1}}_{\mathbf{z}}$ and $T(a) = T^{s_{t-1}}_{\mathbf{z}\backslash a}$, we then have, by (10) and (9),

$$W^*_t(a|x^{t+d-1}, y^{t-1}) = \frac{\Psi^{T(a)}(\sigma_d)}{\Psi^T(\sigma_d)} \hat{P}_{\mathbf{z}}(a|s_{t-1})\,. \tag{11}$$

The matrices Ψ^T and $\Psi^{T(a)}$ differ in row s_{t-1}. Therefore, $W^*_t(a|x^{t+d-1}, y^{t-1})$ will in general differ from the conditional ML probability estimate $\hat{P}_{\mathbf{z}}(a|s_{t-1})$, except when $s_{t-1} = \sigma_d$ (as all other rows coincide and the cofactors ignore row σ_d). In principle, each draw requires the computation of the cofactor $\Psi^T(\sigma_d)$. Clearly, if this cofactor is computed by expanding the determinant along row s_{t-1}, no further determinants need to be computed to obtain the cofactors $\Psi^{T(a)}(\sigma_d)$, $a \in \mathcal{A}$.[2]

The achievable conditional entropy $H^*(Y^n|X^{n+d-1})$ (which, as discussed in Section I, also determines a bound on the expected number of random bits consumed by the simulator), on the other hand, admits a closed form expression when $\mathcal{P}$ is a general subfamily of $\mathcal{F}_{g,s_0}$, as shown in Theorem 2 below.

Theorem 2: Assume $\mathcal{P}$ is a subfamily of $\mathcal{F}_{g,s_0}$ satisfying assumptions A1 and A2. Then,

$$\frac{1}{n} H^*(Y^n|X^{n+d-1}) = \boldsymbol{E} H(Q_{S,\mathbf{Z}}) \tag{12}$$

[1] Recall that the (s, s')-cofactor is the determinant obtained by deleting row s and column s', with a sign change in case the sum of the row and column indexes in an arbitrary order of $\mathcal{S}$ is odd. If $n_{s*}(T) > 0$ for every $s \in \mathcal{S}$, each row-sum in Ψ^T is 0, and the independence of the cofactors on s' follows easily from the fact that the determinant is unaffected if a column is replaced by the sum of all the columns. Clearly, states for which $n_{s*}(T) = 0$ do not affect the independence of the cofactors on states s' such that $n_{s'*}(T) > 0$.

[2] The ratio of cofactors may be expressed in closed form in some simple cases (e.g., first order binary Markov sources).

where the expectation is with respect to the distribution

$$\Pr\{S=s,\mathbf{Z}=\mathbf{z}\} \;=\; \left[\frac{1}{n}\sum_{t=0}^{n-1}\Pr\{S_t=s\}\right]P(\mathbf{z}|s) \qquad s\in\mathcal{S},\,\mathbf{z}\in\mathcal{A}^d\,. \tag{13}$$

Furthermore,

$$\frac{1}{n}H^*(Y^n|X^{n+d-1})=\boldsymbol{E}\log|T^S_{\mathbf{Z}}|-\boldsymbol{E}\log|T^{S'}_{\mathbf{Z}'}| \tag{14}$$

where the first expectation is as in (13) and the second expectation is with respect to the distribution

$$\Pr\{S'=s',\mathbf{Z}'=\mathbf{z}'\} \;=\; \left[\frac{1}{n}\sum_{t=1}^{n}\Pr\{S_t=s'\}\right]P(\mathbf{z}'|s')\,, \qquad s'\in\mathcal{S},\,\mathbf{z}'\in\mathcal{A}^{d-1}\,. \tag{15}$$

Proof. Since $W_t^*(\cdot|\mathbf{x},\mathbf{y})$ depends on $\mathbf{x}$ only through $T_{\mathbf{x}}$ we have

$$\begin{aligned}&H^*(Y_t|X^{t+d-1},Y^{t-1})=\\&\sum_{\mathbf{y}\in\mathcal{A}^{t-1}}\sum_{T_{\mathbf{x}}\in\mathcal{T}^{d+t-1}}P(\mathbf{x})H(W_t^*(\cdot|\mathbf{x},\mathbf{y}))\sum_{\tilde{\mathbf{x}}\in T}W^*(\mathbf{y}|\tilde{\mathbf{x}})\\&=\sum_{s\in\mathcal{S}}\sum_{\mathbf{y}:s_{t-1}=s}P(\mathbf{y})\sum_{T^s_{\mathbf{z}}\in\mathcal{T}^d}P(\mathbf{z}|s)\;|T^s_{\mathbf{z}}|\;H(Q_{s,\mathbf{z}}(\cdot))\end{aligned}$$

where the last equality follows from Lemma 1, from $T_{x^{t+d-1}\backslash y^{t-1}}=T_{\mathbf{z}}^{s_{t-1}}$, and from (10) and (9). Thus,

$$\begin{aligned}&H^*(Y_t|X^{t+d-1},Y^{t-1})=\\&\sum_{s\in\mathcal{S}}\Pr\{S_{t-1}=s\}\sum_{\mathbf{z}\in\mathcal{A}^d}P(\mathbf{z}|s)H(Q_{s,\mathbf{z}}(\cdot))\,.\end{aligned}$$

Equation (12) then follows from the chain rule of conditional entropies and the independence of Y_t from X_{t+i}, $i\geq d$. To prove Equation (14), we denote $nP_n(s)=\sum_{t=0}^{n-1}\Pr\{S_t=s\}$ and further write

$$\begin{aligned}n\boldsymbol{E}H(Q_{S,\mathbf{Z}}) \;&=\; \sum_{s\in\mathcal{S}}P_n(s)\sum_{T^s_{\mathbf{z}}\in\mathcal{T}^d}P(\mathbf{z}|s)\sum_{a\in\mathcal{A}}|T^s_{\mathbf{z}\backslash a}|\log\frac{|T^s_{\mathbf{z}}|}{|T^s_{\mathbf{z}\backslash a}|}\\&=\;\boldsymbol{E}\log|T^S_{\mathbf{Z}}|-\sum_{s\in\mathcal{S}}P_n(s)\sum_{T^s_{\mathbf{z}}\in\mathcal{T}^d}P(\mathbf{z}|s)\\&\qquad\sum_{a\in\mathcal{A}}|T^s_{\mathbf{z}\backslash a}|\;\log|T^s_{\mathbf{z}\backslash a}|\,.\end{aligned} \tag{16}$$

Denoting with $A(s)$ the summation over $T^s_{\mathbf{z}}$ in (16), we then have

$$\begin{aligned}A(s)&=\sum_{a\in\mathcal{A}}P(a|s)\sum_{T^{g(s,a)}_{\mathbf{z}'}\in\mathcal{T}^{d-1}}P(\mathbf{z}'|g(s,a))|T^{g(s,a)}_{\mathbf{z}'}|\;\log|T^{g(s,a)}_{\mathbf{z}'}|\\&=\sum_{s'\in\mathcal{S}}P(s'|s)\sum_{T^{s'}_{\mathbf{z}'}\in\mathcal{T}^{d-1}}P(\mathbf{z}'|s')|T^{s'}_{\mathbf{z}'}|\;\log|T^{s'}_{\mathbf{z}'}|\end{aligned}$$

which, together with (16), implies

$$\begin{aligned}n\boldsymbol{E}H(Q_{S,\mathbf{Z}})=\boldsymbol{E}\log|T^S_{\mathbf{Z}}|\;-\;&\sum_{s'\in\mathcal{S}}\sum_{\mathbf{z}'\in\mathcal{A}^{d-1}}P(\mathbf{z}'|s')\log|T^{s'}_{\mathbf{z}'}|\\&\sum_{s\in\mathcal{S}}P_n(s)P(s'|s)\,.\end{aligned}$$

Equation (14) then follows by observing that

$$\begin{aligned}\sum_{s\in\mathcal{S}}P_n(s)P(s'|s)\;&=\;\frac{1}{n}\sum_{s\in\mathcal{S}}\sum_{t=0}^{n-1}\Pr\{S_t=s\}P(s'|s)\\&=\;\frac{1}{n}\sum_{t=1}^{n}\Pr\{S_t=s'\}\,.\end{aligned}$$

□

Notice that the source is started at a given initial state, and therefore may not be in stationary mode. However, for a stationary chain, as $n\to\infty$, the expectations on the right-hand sides of (12) and (14) in Theorem 2 are governed, by (13) and (15), by the stationary mode of P.

C. Asymptotic analysis

Next, we study the asymptotic behavior of the (normalized) mutual information achieved by the optimal scheme, namely $[H(X^n)-H^*(Y^n|X^{n+d-1})]/n$ (as X^n and Y^n have the same distribution), as d grows. For the i.i.d. families for which $Q_{\mathbf{z}}(\cdot)$ coincides with the ML probability estimate, by Equation (12) in Theorem 2, the mutual information is the difference between the entropy of n-tuples and n times the expected entropy of the ML probabilities computed over d-tuples. After normalization by n, this difference is the expected divergence between the actual distribution and the ML distribution (over d-tuples), which is shown in [17] to approach $(\mathcal{A}-1)/(2d)$ asymptotically, as d grows. In contrast, consider the following straightforward adaptation of the (non-sequential) simulation scheme of [8] to the delay–limited setting: Upon observing $x^{\ell d}$ and outputting $y^{(\ell-1)d}$, $\ell\geq 1$, the delay–limited scheme outputs a random d-tuple of type $T_{x^{\ell d}\backslash y^{(\ell-1)d}}$. Clearly, this blockwise scheme satisfies the delay constraint, and by the results in [8] it achieves normalized mutual information which is $O((\log d)/d)$.

For i.i.d. families such that $Q_{\mathbf{z}}(\cdot)$ differs from the ML probability estimates, the above asymptotic analysis does not apply. Here, under mild assumptions on $\mathcal{P}$, the asymptotic expansion presented in [8, Equation (A4)] states that

$$\begin{aligned}\boldsymbol{E}\log|T_{\mathbf{Z}}|=d\,H\;-\;&\frac{K'}{2}\log(2\pi d)-\frac{K}{2}\log e\\-\;&\frac{1}{2}\log[\det M(P)]+o(1)\end{aligned}$$

where K is the number of parameters defining $\mathcal{P}$, K' is the lattice dimension of the type classes for $\mathcal{P}$ (in the most common cases, $K'=K$, see [8]), and $M(P)$ is the covariance matrix evaluated for P. Thus, we can use Equation (14) in Theorem 2 to conclude that

$$H-\frac{1}{n}H^*(Y^n|X^{n+d-1})=\frac{K'}{2}\log\frac{d}{d-1}+o(1)\,. \tag{17}$$

Observe that even though we have $\log[d/(d-1)]=O(1/d)$, the asymptotic behavior of the mutual information can be dominated by the $o(1)$ term in (17). Thus, the asymptotic expansion of [8] is not accurate enough to obtain the result.

An $O(1/d)$ vanishing rate can also be obtained in the case $\mathcal{P} = \mathcal{F}_{g,s_0}$, for an irreducible chain.[3] The key observation is that the ratio of cofactors in Equation (11), which is used to compute $\boldsymbol{E}H(Q_{S,\mathbf{Z}})$, equals $1 + O(1/d)$ for those types in which $n_{\sigma a}(\mathbf{z}) = \Theta(d)$ whenever $P(a|\sigma) > 0$ (namely, for all allowable transitions). The reason is that the matrices in the numerator and denominator differ only in row s, by an $O(1/d)$ term for such types, and we can expand the cofactors (which, as shown in the proof of [18, Lemma 3], are bounded away from 0 for such types) by this row. After standard manipulations involving typicality arguments, we then obtain

$$\boldsymbol{E}H(Q_{S,\mathbf{Z}}) = \boldsymbol{E}H(\hat{P}_{\mathbf{Z}}(\cdot|S)) + O(1/d).$$

Using the generalization to Markov sources of the results in [17], given in [19], it can be seen that the normalized mutual information still vanishes at an $O(1/d)$ rate in this case.

D. Key rate

As discussed in Section I, $H^*(Y^n|X^{n+d-1})$ is essentially the expected number of key bits consumed by the optimal scheme per output symbol. For some sample paths, however, the number of key bits may be unbounded, and a hard limit on the key length (resulting in a deviation from the target distribution W^*) would affect exact preservation of the probability law. This behavior differs from the schemes proposed in [8], for which a suboptimal implementation of the target distribution affects the mutual information, but not the output probability law. Such a behavior is achieved in [8] by use of the randomness in the input sequence x^m so that the probability law is preserved for any *given* value of the key. The question of whether a similar idea can be used in the delay–limited setting remains open. However, when $\mathcal{P}$ is the entire class of i.i.d. sources over $\mathcal{A}$, a delay–limited simulation scheme exists that still achieves normalized mutual information which is $O(1/d)$ while satisfying Condition C1 and requiring a limited budget of key bits for *all* sample paths. The reason is that, as observed in Equation (1), all the probabilities can be written as rational numbers with d in the denominator. Thus, if d is a power of 2, the Elias decoder will always terminate after a finite number of input key bits [1]. Now, for an arbitrary d, using $\{W_t^*\}$ with delay limitation d', where d' is the largest power of 2 not larger than d, clearly defines (*a fortiori*) a simulation scheme with delay limitation d. Since $2d' > d$, it follows that the normalized mutual information is still $O(1/d)$.

While, as noted, a hard limit on the key length will in general affect the exact preservation of the probability law, by [5], the probability that the scheme will fail to produce the requested sequence after processing k key bits, $k > H^*(Y^n|X^{n+d-1})$, decays with exponent $k - H^*(Y^n|X^{n+d-1})$. Thus, with probability one it produces an output.

IV. Arbitrary request schedules

The sequential, delay–limited simulation problem discussed so far can be generalized to include, as particular cases, the "batch" setting of [8], as well as a setting in which the rate of production of output samples is smaller than the rate of consumption of training samples. In the generalized setting, a user requests random symbols according to some *arbitrary* schedule, with the only constraint that at any time the total number of requested symbols cannot exceed the number of training symbols observed so far. To this end, the simulation scheme has access to a supply of key bits that are delivered on demand, as needed. Let $\ell_t(x)$ denote the length of the training sequence already observed at the time Y_t is requested. The *instantaneous* delay d_t is then defined as $d_t = \ell_t(x) - t + 1$, and the schedule is given by $\{d_t\}$. Thus, we view this setting as one of varying delay; in the setting discussed so far, we have $d_t = d$ for all t, whereas in the batch case with training sequence x^m, $d_t = m - t + 1$, $1 \le t \le n$. The optimal scheme for this setting turns out to be a straightforward generalization of $\{W_t^*\}$, in which a symbol requested after observation of $\mathbf{x}$ and following the output of $\mathbf{y}$ is drawn with probability

$$W_t^*(a|\mathbf{x}, \mathbf{y}) = \frac{|T_{\mathbf{x}\backslash\mathbf{y}a}|}{|T_{\mathbf{x}\backslash\mathbf{y}}|}\,, \quad a \in \mathcal{A}\,.$$

To prove this result it suffices to follow the proof of Theorem 1 *verbatim*, but with $\mathbf{x} \in \mathcal{A}^{\ell_t(x)}$. Following again *verbatim* the proof of Equation (12) in Theorem 2, with d replaced by d_t, the corresponding conditional entropy in the FSM case takes the form

$$H^*(Y^n|X^m) = \sum_{s\in\mathcal{S}} \sum_{t=0}^{n-1} \Pr\{S_t = s\} \sum_{\mathbf{z}\in\mathcal{A}^{d_t+1}} P(\mathbf{z}|s) H(Q_{s,\mathbf{z}})\,. \tag{18}$$

Clearly, for $d_t = m - t + 1$, this scheme is equivalent to drawing a sequence uniformly at random from T_{X^m} and selecting its prefix of length n. Thus, it indeed coincides with the scheme proposed in [8] for batch simulation with an unlimited budget of key bits, and its conditional entropy takes the form

$$H^*(Y^n|X^m) = \boldsymbol{E}\log|T_{X^m}| - \boldsymbol{E}\log|T_{X^m\backslash Y^n}|$$

which in the FSM case further reduces to

$$H^*(Y^n|X^m) = \boldsymbol{E}\log|T_{X^m}| - \boldsymbol{E}\log|T^S_{Z^{m-n}}|$$

where the second expectation is with respect to the distribution $\Pr\{S = s, \mathbf{Z} = \mathbf{z}\} = \Pr\{S_n = s\}P(\mathbf{z}|s)$.

Another interesting request schedule is given by the case in which, after an initial delay d, each request of a block of r symbols, $0 < r \le d$, is followed by the observation of q new training symbols, $q > r$. The ratio q/r will be denoted by ρ ($\rho > 1$), and will be referred to as the *ratio* of the simulation scheme. Clearly, in this case,

$$\ell_t(x) = q\lfloor\frac{t-1}{r}\rfloor + d\,.$$

The asymptotic behavior of the normalized mutual information $I^*(X^m, Y^n)/n$ achieved by the optimal scheme in this case

[3] If necessary for the chain to remain irreducible, we will assume that certain transition probabilities must be positive. Thus, $\mathcal{P}$ may not be the entire class $\mathcal{F}_{g,s_0}$, but the parameter space Ω is still assumed to be rich enough for the types to be defined by the sequence composition with respect to g.

(where $m = \ell_n(x)$), when $n \to \infty$ and $\mathcal{P}$ is the entire class of i.i.d. sources over $\mathcal{A}$, is studied in the Appendix. Since, by (18), we have

$$I^*(X^m; Y^n) = \sum_{t=1}^{n} \boldsymbol{E}[H - \hat{H}_t] \tag{19}$$

where H denotes the source entropy and $\hat{H}_t$ denotes the (normalized) empirical entropy of d_t-tuples, and $d_t \approx (\rho - 1)t$ in this case, the $O(1/d_t)$ asymptotic behavior of $H - \boldsymbol{E}\hat{H}_t$ [17] suggests that the normalized mutual information will be $O((\log n)/n)$. The analysis in the Appendix indeed shows that

$$\frac{1}{n} I^*(X^m; Y^n) \approx \frac{|\mathcal{A}| - 1}{2(\rho - 1)n} \log \left[1 + \frac{(\rho - 1)n}{d} \right] \tag{20}$$

where it is assumed that $d >> r$ and $n \to \infty$. It is interesting to notice that (20) also captures the $\rho = 1$ regime, which can be seen as a particular case by letting $\rho \to 1$, in which case we obtain

$$\frac{1}{n} I^*(X^m; Y^n) \approx \frac{|\mathcal{A}| - 1}{2d}$$

as expected.

Appendix

In this appendix, we analyze the asymptotic normalized mutual information achieved by the optimal scheme in the setting in which the simulation ratio ρ is larger than 1, when $\mathcal{P}$ is the entire class of i.i.d. sources over $\mathcal{A}$. Since

$$d_t = q \lfloor \frac{t-1}{r} \rfloor + d - t + 1 \tag{A.1}$$

and $d_t(H - \boldsymbol{E}\hat{H}_t)$ tends to $(|\mathcal{A}| - 1)(\log e)/2$ as $d_t \to \infty$ (see [17] and [19, Proposition 5.2] therein), then for given $\epsilon > 0$ and sufficiently large t we have

$$\frac{(1-\epsilon)(|\mathcal{A}| - 1)\log e}{2d_t} < H - \boldsymbol{E}\hat{H}_t < \frac{(1+\epsilon)(|\mathcal{A}| - 1)\log e}{2d_t} \tag{A.2}$$

(alternatively, we can assume that d is large enough for (A.2) to hold for all t). Since (A.2) holds for all but finitely many values of t and we are interested in the asymptotic behavior of $I^*(X^m, Y^n)/n$, it suffices to study the asymptotic behavior of $\sum_{t=1}^{n}(1/d_t)$ as n grows. Clearly,

$$\sum_{i=0}^{\lfloor n/r \rfloor - 1} \sum_{j=1}^{r} \frac{1}{d_{ir+j}} \le \sum_{t=1}^{n} \frac{1}{d_t} \le \sum_{i=0}^{\lfloor n/r \rfloor} \sum_{j=1}^{r} \frac{1}{d_{ir+j}} \tag{A.3}$$

where, by (A.1), $d_{ir+j} = (q - r)i + d - j + 1$. Now, let $f(i)$ denote the inner summation in the rightmost side of (A.3). Since $f(i)$ is non-increasing with i, we have, for all nonnegative integers N,

$$\int_0^{N+1} f(u)du \le \sum_{i=0}^{N} f(i) \le f(0) + \int_0^{N} f(u)du\,.$$

Therefore, by (A.3),

$$\int_0^{\lfloor n/r \rfloor} f(u)du \le \sum_{t=1}^{n} \frac{1}{d_t} \le f(0) + \int_0^{\lfloor n/r \rfloor} f(u)du\,. \tag{A.4}$$

Since $f(0) < r/(d - r + 1)$ and is independent of n, it then suffices to study the asymptotic behavior of the integral in (A.4), which grows without bound as $n \to \infty$. Clearly,

$$\frac{r}{(q-r)u + d} \le f(u) \le \frac{r}{(q-r)u + d + 1 - r}$$

implying

$$\int_0^{\lfloor n/r \rfloor} f(u)du \le \frac{1}{\rho - 1} \ln \left[1 + (\rho - 1) \cdot \frac{r \lfloor n/r \rfloor}{d - r + 1} \right] \tag{A.5}$$

and

$$\int_0^{\lfloor n/r \rfloor} f(u)du \ge \frac{1}{\rho - 1} \ln \left[1 + (\rho - 1) \cdot \frac{r \lfloor n/r \rfloor}{d} \right] \tag{A.6}$$

Equation (20) then follows from (19), (A.2), (A.4), (A.5), (A.6), and from $d >> r$.

Acknowledgment

We would like to thank Álvaro Martín for useful discussions on type sizes in the FSM case.

References

[1] D. E. Knuth and A. Yao, “The complexity of nonuniform random number generation,” in *Algorithms and Complexity, New Directions and Results*, J. F. Traub, Ed. New York: Academic Press, 1976, pp. 357–428.

[2] T. S. Han and S. Verdú, “Approximation theory of output statistics,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 752–772, May 1993.

[3] Y. Steinberg and S. Verdú, “Channel simulation and coding with side information,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 634–646, May 1994.

[4] Y. Steinberg and S. Verdú, “Simulation of random processes and rate-distortion theory,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 63–86, Jan. 1996.

[5] T. S. Han, M. Hoshi, “Interval algorithm for random number generation,” *IEEE Trans. Inform. Theory*, vol. 43, pp. 599–611, March 1997.

[6] T. Uyematsu, F. Kanaya, “Channel simulation by interval algorithm: A performance analysis of interval algorithm,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 2121–2129, Sept. 1999.

[7] K. Visweswariah, S. R. Kulkarni, and S. Verdú, “Separation of random number generation and resolvability,” *IEEE Trans. Inform. Theory*, vol. 46, pp. 2237–2241, Sept. 2000.

[8] N. Merhav and M. J. Weinbeger, “On universal simulation of information sources using training data,” *IEEE Trans. Inform. Theory*, vol. 50, pp. 5–20, Jan. 2004.

[9] N. Merhav and M. J. Weinbeger, Addendum to “On universal simulation of information sources using training data,” *IEEE Trans. Inform. Theory*, vol. 51, pp. 3381–3383, Sept. 2005.

[10] N. Merhav, “Achievable key rates for universal simulation of random data with respect to a set of statistical tests,” *IEEE Trans. Inform. Theory*, vol. 50, pp. 21–30, Jan. 2004.

[11] G. Seroussi, “On universal types,” *IEEE Trans. Inform. Theory*, vol. 52, pp. 171–189, Jan. 2006.

[12] I. Csiszár, “The method of types,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505–2523, Oct. 1998.

[13] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[14] F. Jelinek, *Probabilistic Information Theory.* New York: McGraw-Hill, 1968.

[15] T. M. Cover, “Enumerative source encoding,” *IEEE Trans. Inform. Theory*, vol. 19, pp. 73-77, Jan. 1973.

[16] P. Whittle, “Some distributions and moment formulae for the Markov chain,” *J. Roy. Stat. Soc.*, Ser. B, 17, pp. 235–242, 1955.

[17] B. S. Clarke and A. R. Barron, “Information-theoretic asymptotics of Bayes methods,” *IEEE Trans. Inform. Theory*, vol. 36, pp. 453-471, May 1990.

[18] M. J. Weinberger, N. Merhav, and M. Feder, “Optimal sequential probability assignment for individual sequences,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 384-396, March 1994.

[19] K. Atteson, “The asymptotic redundancy of Bayes rules for Markov chains,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 2104–2109, Sept. 1999.