Universal Delay–Limited Simulation

Neri Merhav*[1], Gadiel Seroussi†, and Marcelo J. Weinberger†
*Department of Electrical Engineering, Technion, Haifa 32000, Israel, merhav@ee.technion.ac.il
†Hewlett-Packard Laboratories, Palo Alto, CA 94303, U.S.A., (seroussi,marcelo)@hpl.hp.com

***Abstract*— We consider the problem of universal delay–limited simulation of an unknown information source of a certain parametric family (e.g., the family of memoryless sources or Markov sources), given a training sequence from that source and a stream of purely random bits. 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. We characterize the optimal delay–limited simulation scheme and upper-bound the expected number of random bits it consumes. As in the non-sequential case, this upper bound is related to the entropy rate of the source.**

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]. In all these works, the common assumption is that the probability law of the desired process is perfectly known. Recently, Merhav and Weinberger [7] considered a universal version of this problem for finite–alphabet sources, in which the assumption of perfect knowledge of the target probability law is relaxed. Specifically, the target source P to be simulated is assumed in [7] to belong to a certain parametric family $\mathcal{P}$ (like the family of finite–alphabet memoryless sources, Markov sources of a given order, 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 independent of X^m. The goal of the simulation scheme 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 Y^n is *exactly* the n-dimensional marginal of the probability law P corresponding to X^m for all $P \in \mathcal{P}$.

C2. The mutual information $I(X^m; Y^n)$ is as small as possible, simultaneously for all $P \in \mathcal{P}$ (so as to make the generated sample path Y^n as "original" as possible).

In [7], 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 [7] that in order to satisfy Condition C1, it is necessary that the output Y^n be a prefix of a sequence Y^m having the same *type* [8] 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 [9]). 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. A different perspective on universal simulation is investigated in [10], where x^m is assumed to be an individual sequence not originating from any probabilistic source.

In this paper, we investigate the universal simulation problem 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 depends on X^{t+d-1} but, unlike the setting in [7], it is independent 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 [7] in that there is no fixed budget of key bits; rather, we will be interested in the *expected* number of key bits that the scheme consumes in order to generate its output, where 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 from [1] that a corresponding sequence of draws can be implemented with an Elias decoder [11, 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 [1], 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.

A special, trivial case of the delay–limited universal simulation problem is obtained when pure sequentiality is required,

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

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 $H(Y^n|X^n) = 0$. 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 $I(Y^n; X^{n+d-1})$ simultaneously for all $P \in \mathcal{P}$ (Condition C2) takes a form that is reminiscent of enumerative decoding schemes [12]. 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 $n_a(s)$ denotes the number of occurrences of a in a sequence s. The distribution assigned by the scheme is precisely the one that an enumerative decoder would use to decode y^t given that y^{t+d-1} has the same composition as x^{t+d-1}. The corresponding conditional entropy $H(Y^n|X^{n+d-1})$, 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. By [13], this expectation falls short of the entropy rate H by an $O(1/d)$ term. Since, by C2, $H(Y^n) = H(X^n)$, this term is precisely the normalized mutual information between the input and the output. The above results will actually be extended, in Section III, to a more general setting, where the source P has memory and $\mathcal{P}$ is a *parametric subfamily* of the entire class.

In the remainder of this extended abstract, Section II introduces the main concepts and notation. Our main results are then presented in Section III and discussed 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. 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 space $\mathcal{S}$ will be identified with its next-state function $g : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, and will be assumed to start at a given initial state $s_0 \in \mathcal{S}$. The class of all FSM sources over $\mathcal{A}$ driven by the next-state function g and starting at state s_0 will be denoted by $\mathcal{F}_{g,s_0}$. 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 $s_1, s_2, \ldots, s_t \in \mathcal{S}$ denotes the sequence of states assumed by the FSM.

We shall define the *type class* [8] 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 $\mathcal{P} = \mathcal{F}_{g,s_0}$, T_{x^m} is the set of all vectors having the same composition as x^m with respect to g [8], [9] (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." 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. Notice that for any sequence $z^r \in T_{x^m \setminus y^n}$, $T^{s_n}_{z^r} = T_{x^m \setminus y^n}$, where s_n denotes the state to which the FSM evolves with y^n.

Next, for every type class $T \in \mathcal{T}^m$, we define

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

where x^m is a sequence in T and, throughout, $|T|$ denotes the cardinality of T. 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 [7], 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* θ*) is linearly independent over* Ω.

As shown in [7], 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. 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

$$W(Y^n|X^{n+d-1}) = \prod_{t=1}^{n} W_t(Y_t|X^{t+d-1}, Y^{t-1})\,.$$

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

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 $W(Y^n|X^m)$, $m \geq n$.

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

$$\sum_{\tilde{\mathbf{x}} \in T_{\mathbf{x}}} W(\mathbf{y}|\tilde{\mathbf{x}}) = |T_{\mathbf{x}\backslash\mathbf{y}}| \,. \tag{3}$$

Proof. Clearly, the probability law is preserved if and only if for all $\mathbf{y} \in \mathcal{A}^n$ and $P \in \mathcal{P}$ we have

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

Now, if no sequence in $T_{\mathbf{x}}$ is prefixed by $\mathbf{y}$ we have $|T_{\mathbf{x}\backslash\mathbf{y}}| = 0$; otherwise, $P(\mathbf{x}) = P(\mathbf{y}\mathbf{z})$ where $\mathbf{z}$ runs over all sequences in $|T_{\mathbf{x}\backslash\mathbf{y}}|$. Therefore,

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

and thus the probability law is preserved if and only

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

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 $T_{\mathbf{x}\backslash\mathbf{y}}$ is nonempty. In other words, $\mathbf{y}$ must be a prefix of a sequence in $T_{\mathbf{x}}$ (in [7], 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 simulation schemes that output Y^t with training data X^{t+d-1}, $1 \leq t \leq n$. 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 (3)), is given by

$$W_t^*(y_t|x^{t+d-1}, y^{t-1}) = \frac{|T_{x^{t+d-1}\backslash y^t}|}{|T_{x^{t+d-1}\backslash y^{t-1}}|} \,. \tag{4}$$

Equation (4) 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 $\bigcup_{a \in \mathcal{A}} T_{\mathbf{x}\backslash\mathbf{y}a} = T_{\mathbf{x}\backslash\mathbf{y}}$. For the scheme (4) 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* $|T_{\mathbf{x}\backslash\mathbf{y_1}}| = |T_{\mathbf{x}\backslash\mathbf{y_2}}|$.

Notice that this assumption holds trivially in the i.i.d. case, since in this case $T_{\mathbf{x}\backslash\mathbf{y}}$ depends on $\mathbf{y}$ only through its type. For a general FSM source, $T_{\mathbf{x}\backslash\mathbf{y}}$ may also depend on the final state to which the FSM evolves with $\mathbf{y}$. However, if $\mathcal{P}$ is an entire 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).

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

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

(b) Assume $\mathcal{P}$ is a subfamily of a family $\mathcal{F}_{g,s_0}$ of FSM sources. Given $s \in \mathcal{S}$ and $\mathbf{z} \in \mathcal{A}^d$, let $Q_{s,\mathbf{z}}(\cdot)$ denote a probability distribution on $\mathcal{A}$ defined by

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

Then, for any simulation scheme W_t with delay limitation d that satisfies Condition C1 we have

$$H(Y^n|X^{n+d-1}) \leq n \boldsymbol{E}_{s,\mathbf{z}} H(Q_{s,\mathbf{z}}) \tag{5}$$

where the expectation is with respect to the distribution

$$\Pr\{s, \mathbf{z}\} = \left[\frac{1}{n} \sum_{t=0}^{n-1} \Pr\{s_t = s\} \right] P(\mathbf{z}|s) \tag{6}$$

with equality in (5) if and only if $W_t = W_t^*$.

Sketch of proof. Part (a) follows from showing, by induction in n, that the simulation scheme satisfies the condition (3) 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,

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

where the second equality follows from the fact that the difference types depend on $\tilde{\mathbf{x}}$ only through its type. By Assumption A2, the multiset of sequences obtained by deleting the last symbol of all $\tilde{\mathbf{x}} \in T_{\mathbf{x}}$ is a union of types, and therefore by the induction hypothesis and since $W^*(\mathbf{y}'|\tilde{\mathbf{x}})$ is independent of the last symbol of $\tilde{\mathbf{x}}$, the condition implies that the summation on the right-hand side of (7) is the number of sequences $\mathbf{z}$ such that $\mathbf{y}'\mathbf{z} \in T_{\mathbf{x}}$, namely $|T_{\mathbf{x}\backslash\mathbf{y}'}|$.

As for Part (b), we present the proof for the i.i.d. case only. In this case, it is easy to see that any simulation scheme W_t

satisfies

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

Since, by the i.i.d. assumption, $P(\mathbf{x}) = P(\mathbf{y})P(\mathbf{z})$ for $\mathbf{x} \in T_{\mathbf{yz}}$, we have

$$\sum_{\mathbf{y}\in\mathcal{A}^{t-1}}\sum_{\mathbf{x}\in T_{\mathbf{yz}}}\Pr\{X^{t+d-1}=\mathbf{x},Y^{t-1}=\mathbf{y}\} = P(\mathbf{z})\sum_{\mathbf{y}\in\mathcal{A}^{t-1}}P(\mathbf{y})\sum_{\mathbf{x}\in T_{\mathbf{yz}}}W(\mathbf{y}|\mathbf{x}). \quad (8)$$

If W preserves the probability law then it does so for all horizons t, $1 \le t \le n$, and thus, by Lemma 1, we have $\sum_{\mathbf{x}\in T_{\mathbf{yz}}} W(\mathbf{y}|\mathbf{x}) = |T_{\mathbf{yz}\backslash\mathbf{y}}| = |T_{\mathbf{z}}|$. Therefore, the right-hand side of (8) equals $P(T_{\mathbf{z}})$ and, by Jensen's inequality, we have

$$H(Y_t|X^{t+d-1},Y^{t-1}) \le \sum_{T_{\mathbf{z}}\in\mathcal{T}^d}P(T_{\mathbf{z}})H\left(\frac{1}{P(T_{\mathbf{z}})}\sum_{\mathbf{y}\in\mathcal{A}^{t-1}}\sum_{\mathbf{x}\in T_{\mathbf{yz}}}\Pr\{\mathbf{x},\mathbf{y}\}W_t(\cdot|\mathbf{x},\mathbf{y})\right)$$

with equality if and only if $W_t(\cdot|\mathbf{x},\mathbf{y})$ depends only on $T_{\mathbf{x}\backslash\mathbf{y}}$. Consequently,

$$H(Y_t \mid X^{t+d-1},Y^{t-1}) \le \boldsymbol{E}_{\mathbf{z}}\,H\left(\frac{P(\mathbf{z})}{P(T_{\mathbf{z}})}\sum_{\mathbf{y}\in\mathcal{A}^{t-1}}P(\mathbf{y})\sum_{\mathbf{x}\in T_{\mathbf{yz}}}\Pr\{\mathbf{x},\mathbf{y}\}W(\mathbf{y}y_t|\mathbf{x})\right)$$

where the expectation is on d-tuples and with respect to P. Again by Lemma 1,

$$H(Y_t|X^{t+d-1},Y^{t-1}) \le \boldsymbol{E}_{\mathbf{z}}H\left(\frac{1}{|T_{\mathbf{z}}|}\sum_{\mathbf{y}\in\mathcal{A}^{t-1}}P(\mathbf{y})|T_{\mathbf{yz}\backslash\mathbf{y}y_t}|\right) = \boldsymbol{E}_{\mathbf{z}}H\left(\frac{|T_{\mathbf{z}\backslash\cdot}|}{|T_{\mathbf{z}}|}\right).$$

Using the chain rule of conditional entropies it follows that

$$H(Y^n|X^{n+d-1}) \le n\boldsymbol{E}_{\mathbf{z}}H\left(\frac{|T_{\mathbf{z}\backslash\cdot}|}{|T_{\mathbf{z}}|}\right)$$

as claimed (for the i.i.d. case). The probability assignment of the proposed scheme depends only on $T_{\mathbf{x}\backslash\mathbf{y}}$ and therefore it achieves maximum conditional entropy. Moreover, consider any law-preserving scheme W that depends on $\mathbf{x}$ only through $T_{\mathbf{x}\backslash\mathbf{y}}$. Then, with $\mathbf{y} = \mathbf{y}'a$, $a \in \mathcal{A}$, Lemma 1 implies

$$|T_{\mathbf{x}\backslash\mathbf{y}}| = \sum_{\tilde{\mathbf{x}}\in T_{\mathbf{x}}}W(\mathbf{y}|\tilde{\mathbf{x}}) = \sum_{\tilde{\mathbf{x}}\in T_{\mathbf{x}}}W(\mathbf{y}'|\tilde{\mathbf{x}})W_t(a|\tilde{\mathbf{x}},\mathbf{y}') = W_t(a|\mathbf{x},\mathbf{y}')\sum_{\tilde{\mathbf{x}}\in T_{\mathbf{x}}}W(\mathbf{y}'|\tilde{\mathbf{x}}). \quad (9)$$

Using Assumption A2 as in the proof of Part (a), the summation in the right-hand side of (9) equals $|T_{\mathbf{x}\backslash\mathbf{y}'}|$, and therefore $W_t(a|\mathbf{x},\mathbf{y}') = W_t^*(a|\mathbf{x},\mathbf{y}')$. Thus, W_t^* is the *only* law-preserving scheme whose assignment depends on $\mathbf{x}$ through $T_{\mathbf{x}\backslash\mathbf{y}}$. □

Notice that the simulation scheme $\{W_t^*\}$ is strictly optimum 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 $t \ge 1$.

IV. Discussion

When $\mathcal{P} = \mathcal{F}_{g,s_0}$, W_t^* takes the form of an empirical distribution. Since $T_{x^{t+d-1}\backslash y^{t-1}}$ is an FSM-type in $\mathcal{T}^d$ (but with initial state s_{t-1}), we can apply Whittle's formula [15] to show that $W_t^*(a)$ is the empirical conditional probability of a given s_{t-1} in any d-tuple in $T_{x^{t+d-1}\backslash y^{t-1}}$ (see (1) for the i.i.d. case). Intuitively, the scheme aims at outputting a sequence with the same composition as the input, as the latter becomes sequentially available. Similarly, the conditional entropy in Part (b) of the theorem takes the form of an expected empirical entropy. For a stationary chain, the distribution (6) under which the expectation is taken tends, as $n \to \infty$, to the stationary distribution of P (notice that the source is started at a given initial state, and therefore may not be in stationary mode). Thus, the mutual information achieved by the optimal scheme, namely $H(X^n) - H(Y^n|X^{n+d-1})$ (as X^n and Y^n have the same distribution), is the difference between the entropy of n-tuples and n times the expected empirical entropy computed over d-tuples. In the i.i.d. case, after normalization by n, this difference is the expected divergence between the actual distribution and the empirical distribution (over d-tuples), whose asymptotic behavior with d is shown in [13] to be $O(1/d)$ (this result is generalized to Markov sources in [14]). In contrast, consider the following straightforward adaptation of the (non-sequential) simulation scheme of [7] to the delay–limited setting: Upon observing X^{kd} and outputting $Y^{(k-1)d}$, $k \ge 1$, the delay–limited scheme outputs a random d-tuple of type $T_{X^{kd}\backslash Y^{(k-1)d}}$. Clearly, this blockwise scheme satisfies the delay constraint, and by the results in [7] it achieves a mutual information whose asymptotic behavior with d is $O((\log d)/d)$.

When $\mathcal{P}$ is an arbitrary parametric subfamily of $\mathcal{F}_{g,s_0}$, the above interpretation as empirical distributions may no longer hold. However, at least for exponential families, the asymptotic expansion of the type size presented in [7] can be used to show that the minimum mutual information still vanishes at an $O(1/d)$ rate.

As discussed in Section I, $\boldsymbol{E}_{s,\mathbf{z}}H(Q_{s,\mathbf{z}})$ is also 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 [7], 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 [7] 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 is under

investigation. 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 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 in this case all the probabilities can be written as rational numbers with d in the denominator (see (1)), due to the interpretation as empirical probabilities. Thus, if d is a power of 2, the Elias decoder will always terminate after a finite number of input key bits. 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 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 [1], the probability that the scheme will fail to produce the requested sequence after processing k key bits, $k > n\boldsymbol{E}_{s,\mathbf{z}}H(Q_{s,\mathbf{z}})$, decays with exponent $k - n\boldsymbol{E}_{s,\mathbf{z}}H(Q_{s,\mathbf{z}})$. Thus, with probability one it produces an output.

The sequential, delay–limited simulation setting discussed so far can be generalized to include the "batch" setting of [7] as a particular case. In the generalized setting, a user requests random symbols according to some *arbitrary* schedule, with the only constraint that at any time the overall number of requested symbols cannot exceed the number of training symbols seen 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 is a straightforward generalization of W^*, in which a symbol requested after observation of $\mathbf{x}$ and following the output of $\mathbf{y}$ is drawn with probability

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

It is easy to see that for $d_t = m - t + 1$, this scheme indeed coincides with the scheme proposed in [7] for batch simulation with an unlimited budget of key bits. The corresponding conditional entropy takes the form

$$H(Y^n|X^m) = \sum_{s\in\mathcal{S}}\sum_{t=0}^{n-1} \Pr\{s_t = s\} \sum_{\mathbf{z}\in\mathcal{A}^{d_{t+1}}} P(\mathbf{z}|s) H(Q_{s,\mathbf{z}})\,. \tag{10}$$

Again, for $d_{t+1} = m - t$, using the chain rule of conditional entropies it can be seen that (10) indeed coincides with the result in [7].

REFERENCES

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

[2] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. IT–39, no. 3, pp. 752–772, May 1993.

[3] Y. Steinberg and S. Verdú, "Channel simulation and coding with side information," *IEEE Trans. Inform. Theory*, vol. IT–40, no. 3, 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, no. 1, pp. 63–86, January 1996.

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

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

[7] N. Merhav and M. J. Weinbeger, "On universal simulation of information sources using training data," *IEEE Trans. Inform. Theory*, vol. IT–50, no. 1, pp. 5–20, January 2004.

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

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

[10] G. Seroussi, "On universal types," *Proc. of 2004 IEEE Intern'l Symp. on Inform. Theory (ISIT'04)*, p. 223, Chicago, USA, June/July 2004.

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

[12] T. M. Cover, "Enumerative source encoding," *IEEE Trans. Inform. Theory*, vol. 19, pp. 73-77, January 1973.

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

[14] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 2104–2109, September 1999.

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