Twice-Universal Simulation of Markov Sources and Individual Sequences*

Álvaro Martín†, Neri Merhav‡, Gadiel Seroussi§, and Marcelo J. Weinberger¶
†Instituto de Computación, Universidad de la República, Montevideo, Uruguay
Email: almartin@fing.edu.uy
‡Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa, Israel
Email: merhav@ee.technion.ac.il
§Hewlett-Packard Laboratories, Palo Alto, CA, U.S.A., and Universidad de la República, Uruguay
Email: gseroussi@ieee.org
¶Hewlett-Packard Laboratories, Palo Alto, CA, U.S.A.
Email: marcelo@hpl.hp.com

***Abstract*— The problem of universal simulation given a training sequence is studied both in a stochastic setting and for individual sequences. In the stochastic setting, the training sequence is assumed to be emitted by a Markov source of unknown order, extending previous work where the order is assumed known and leading to the notion of twice-universal simulation. A simulation scheme, which partitions the set of sequences of a given length into classes, is proposed for this setting and shown to be asymptotically optimal. This partition extends the notion of type classes to the twice-universal setting. In the individual sequence scenario, the same simulation scheme is shown to generate sequences which are statistically similar, in a strong sense, to the training sequence, for statistics of any order, while essentially maximizing the uncertainty on the output.**

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. It 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], [8]. In all these works, perfect knowledge of the desired probability law is assumed. Universal simulation was introduced in [9] and versions of this problem were studied in [10], [11], [12], [13], [14]. In [9], the target source P to be simulated is assumed to belong to a certain parametric family $\mathcal{P}$ (like the family of finite–alphabet Markov sources of a given order) but is otherwise unknown, and a training sequence $x^\ell = (x_1, \ldots, x_\ell)$ that has emerged from P is available. In [12], x^ℓ is assumed to be an individual sequence not originating from any probabilistic source. In both cases, the simulation schemes are also provided with a stream of r purely random bits $u^r = (u_1, \ldots, u_r)$ that are statistically independent of the training sequence. While, as explained below, the goals of the simulation schemes differ in each case, this paper can be viewed as extending the results of both [9] and [12].

Specifically, the goal in [9] is to generate an output sequence $y^n = (y_1, \ldots, y_n)$, $n \leq \ell$, corresponding to the simulated process, such that $y^n = \phi(x^\ell, u^r)$, 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 $P \in \mathcal{P}$.

Condition C1 states that the simulated sequence is a sample of the same process as the training sequence, universally in $\mathcal{P}$. Condition C2 guarantees that the generated sample path is as "original" as possible, namely, with as small a statistical dependence as possible on the training sequence (as opposed to the case in which $Y^n = X^n$, which obviously satisfies Condition C1). For example, in a texture reproduction scenario, this condition would help to avoid undesired periodicities if a texture is generated by appending various sequences Y^n generated from a single X^m (we refer to [9] for further motivation of these conditions).

In [9], the smallest achievable value of the mutual information as a function of n, ℓ, r, and the entropy rate H of the source P is characterized, and simulation schemes that asymptotically achieve these bounds are presented. For a broad class of families $\mathcal{P}$, it is shown in [9] that in order to satisfy Condition C1, it is necessary that the output y^n be a prefix of a sequence y^ℓ having the same *type* [15] as x^ℓ with respect to $\mathcal{P}$. Moreover, it is shown that for r large enough, the optimal simulation scheme essentially takes the first n symbols of a randomly selected sequence of the same type as x^ℓ. For

* The material in this paper was presented in part at the 2007 IEEE International Symposium on Information Theory, Nice, France, July 2007.
† Work supported by grant CSIC 05 PDT 63.
‡ This work was done while N. Merhav was visiting Hewlett–Packard Laboratories, Palo Alto, CA, U.S.A.

unlimited r and $n = \ell$ (which will be our assumption in the rest of this paper), the resulting optimal mutual information between X^n and Y^n, after normalization, vanishes with n as $\frac{m}{2}\frac{\log n}{n}$, where m is the number of free parameters defining $\mathcal{P}$.

The above rate prompts similar "model cost" issues as the universal source coding problem [16], in the sense that the larger the class $\mathcal{P}$, the larger the cost of universality (which in data compression takes the form of an analogous rate of convergence to the source entropy). A natural question that has then been asked in data compression is that of *double universality* [17] or *hierarchical universality* [18]: Assuming a nested family of model classes (e.g., Markov models of different orders), is it possible to achieve the optimal convergence rate corresponding to the *smallest* class containing the *actual* source, without prior knowledge of the class? The answer to this question is well known to be positive, giving rise to the notion of *twice-universal* (or hierarchically universal) schemes. In this paper, we start by addressing the problem of hierarchical universality in the simulation setting of [9] when $\mathcal{P}$ is a class of Markov models of unknown (fixed) order k, denoted $\mathcal{P}_k$.

First, we notice that (as discussed in [13]), families of Markov models are indeed among those requiring that x^ℓ and y^ℓ be of the same type in order for Condition C1 to be satisfied. Since, in the unknown model order setting, the types of x^ℓ and y^ℓ must be the same for *every* Markov order, the two sequences must then coincide, leading to a single, trivial simulator. Thus, a relaxation of Condition C1 is necessary for the problem to become meaningful.[1] As it turns out, it suffices to allow simulators such that, for every k and $P \in \mathcal{P}_k$, Condition C1 is violated only by a fraction of sequences whose total mass (under the simulated probability, or equivalently, under P) is upper-bounded by a vanishing function $\delta(n)$. In fact, a simulator exists such that $\delta(n)$ decreases exponentially fast, while achieving per-symbol mutual information which decays essentially as $\frac{m}{2}\frac{\log n}{n}$ for any Markov order k and any $P \in \mathcal{P}_k$, where m is the number of parameters corresponding to $\mathcal{P}_k$. This simulator follows a "plug-in" approach:

a. From x^n, estimate an order $k(x^n)$ of the Markov source;
b. Draw uniformly at random from the set of sequences having the same Markov type of order $k(x^n)$ as x^n and for which the estimated order is also $k(x^n)$.

We show that the total mass of the sequences which do not satisfy Condition C1 is upper-bounded by the probability of underestimating the model order, whereas the conditional entropy achieved by this scheme differs from the one achieved by the optimal scheme that knows the "true" order by a quantity that depends on the overestimation probability. With a proper choice of the order estimator (in the spirit of those used in, e.g., [19], [20], [21], [22], [23]) both the mass of those sequences violating Condition C1, and the deviation from optimal conditional entropy, can be made negligible. While this comparison is with a scheme that fully satisfies Condition C1, we further show that such a relaxation of the condition can only produce a negligible decrease in the achievable mutual information in the known order case. Therefore, under the relaxed criterion, the proposed scheme is asymptotically optimal in that it achieves essentially the same performance as if the model order were known. In that sense, the scheme is hierarchically universal.

The above simulation scheme is based on a partition of the set of n-tuples, where two sequences are in the same class if and only if they both estimate the same Markov order, and have the same Markov type for that order. This partition is in the same spirit as the one giving rise to the simulation scheme in [12], which also extends the conventional notion of a type. In the partition of [12], two sequences belong to the same class if and only if their Lempel-Ziv (LZ) parsing [24] yields the same tree. Any pair of sequences that belong to the same class in this partition has the following property, which parallels conventional types in an individual sequence setting:

P1. For any fixed integer j, the L_1 distance between the empirical distributions of j-tuples corresponding to the two sequences is a vanishing function of n.

The rate of convergence of the L_1 distance demonstrated in [12] is $O(1/\log n)$. It is easy to see that Property P1 implies that, for any fixed Markov source, the normalized logarithm of the ratio between the probabilities of two sequences in the same class is also $O(1/\log n)$, provided the sequences have positive probability. In [12], a sequence of length n is said to be a *faithful* reproduction of another sequence of the same length if the pair satisfies Property P1. It is further claimed that, for simulation purposes, faithfulness parallels Condition C1 in an individual sequence setting. Thus, the simulator that draws a sequence uniformly at random from the (LZ-based) class of the training sequence x^n is a faithful simulator. Moreover, it is shown in [12] that no other faithful simulator can produce significantly more uncertainty than the proposed one, in the spirit of Condition C2.

In this paper, we extend the results of [12] in two directions. First, we show that the equivalence classes defined for the twice-universal simulation scheme for Markov sources possess similar properties in the individual sequence setting as those shown for the LZ parsing-based scheme, but the distance between empirical distributions (as defined in Property P1) exhibits a faster convergence rate. Second, we formulate a converse similar to the one presented in [12], but that applies to a broad family of simulators, which includes both the one proposed here and the LZ-based one. This converse unveils the essence of the universal simulation problem in an individual sequence setting: Find a partition of the sequence space into a relatively small (sub-exponential) number of classes such that all the sequences in a class have approximately uniform probability (as per Property P1). Notice that a "slow" rate of convergence is typical of other applications of the LZ parsing. On the other hand, our improvement has a complexity cost, which we discuss.

The rest of this paper is organized as follows. Section II

[1] The relaxation of Condition C1 was precisely the motivation for the individual sequence setting of [12]. Relaxation in the stochastic sense discussed here is also discussed in [11] and [14], where universal simulation with a fidelity criterion is studied, in analogy with the (non-universal) scenario of [5].

introduces the main concepts and tools. Our results in the stochastic setting are then presented in Section III, whereas the individual sequence setting is studied in Section IV.

II. Preliminaries

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^n and y^n will denote specific values of the random vectors X^n and Y^n, respectively. The (finite) source alphabet will be denoted by $\mathcal{A}$, with its cardinality, $|\mathcal{A}|$, denoted α.

A Markov source P of order k over $\mathcal{A}$, with transition probabilities $P(a_{k+1}|a_k, a_{k-1}, \ldots, a_1)$, $a_i \in \mathcal{A}$, $i = 1, \ldots, k+1$, draws a sequence x^n with probability

$$P(x^n) = \prod_{i=1}^{n} P(x_i|x_{i-1}, x_{i-2}, \ldots, x_{i-k}) \tag{1}$$

where we arbitrarily assume a fixed string $x_{-k+1}, x_{-1}, \ldots, x_0$ determining the initial state. We assume that k is the minimum possible order, in the sense that no integer $k' < k$ can replace k in (1). The family of Markov sources of order k over $\mathcal{A}$ is denoted $\mathcal{P}_k$. Thus, a source in $\mathcal{P}_k$ is defined by a parameter vector $\theta \in [0,1]^{\alpha^k(\alpha-1)}$ (excluding parameter vectors that correspond to a lower order); such a source P will sometimes be denoted by P_θ. The entropy of n-tuples emitted by P is denoted $H(X^n)$.

The k-th order Markov *type class* [15] $T_k(x^n)$ of a sequence x^n is the set of all sequences $\tilde{x}^n \in \mathcal{A}^n$ such that $P(\tilde{x}^n) = P(x^n)$ for *every* source $P \in \mathcal{P}_i$, $0 \le i \le k$. The set of all k-th order Markov type classes of sequences in $\mathcal{A}^n$ will be denoted by $\mathcal{T}_k^n$, with $|\mathcal{T}_k^n| = N_{n,k}$. Clearly, $T_k(x^n)$ is the set of all sequences having the same composition as x^n with respect to the k-th order Markov model [15], [25], i.e., each state transition occurs as many times in $\tilde{x}^n \in T_k(x^n)$ as in x^n, starting from the fixed initial state $(x_{-k+1}, x_{-k+2}, \ldots, x_0)$. Equivalently, the type is given by the number of occurrences in x^n of each string $s \in \mathcal{A}^{k+1}$, denoted $n_{x^n}(s)$, namely

$$n_{x^n}(s) = |\{i : 0 < i \le n, (x_{i-k}, \ldots, x_{i-1}, x_i) = s\}| \tag{2}$$

where $|\cdot|$ denotes cardinality. Thus, the k-th order empirical Markov source defined by the transition counts of x^n depends on x^n only through $T_k(x^n)$=T, and is denoted $\hat{P}_T^{(k)}$. It corresponds to the maximum-likelihood estimate, and its transition probabilities are given by

$$\hat{p}(a|s) = \frac{n_{x^n}(sa)}{n'_{x^n}(s)},\ s \in \mathcal{A}^k,\ a \in \mathcal{A}$$

where $n'_{x^n}(s) = \sum_{a \in \mathcal{A}} n_{x^n}(sa)$ (which may defer from $n_{x^n}(s)$ by one unit as it corresponds to the number of occurrences of the string s in $x_0, x_1, \ldots, x_{n-1}$, rather than in x^n) and $\hat{p}(a|s)$ is defined only for s such that $n'_{x^n}(s) > 0$. The conditional entropy $\hat{H}_k(x^n)$ of this distribution, namely the k-th order empirical conditional entropy for x^n, is given by

$$\hat{H}_k(x^n) = \sum_{s \in \mathcal{A}^k} \sum_{a \in \mathcal{A}} \frac{n_{x^n}(sa)}{n} \log \frac{n'_{x^n}(s)}{n_{x^n}(sa)} \tag{3}$$

and satisfies $n\hat{H}_k(x^n) = -\log \hat{P}_T^{(k)}(x^n)$, where, throughout, logarithms are taken in base 2, and we adopt the conventions $0 \log 0 = 0$ and $0 \log(0/0) = 0$. For $P_\theta \in \mathcal{P}_i$, $0 \le i \le k$, the probability of a type class $T \in \mathcal{T}_k^n$ is given by

$$P_\theta(T) \stackrel{\triangle}{=} \sum_{\tilde{x}^n \in T} P_\theta(\tilde{x}^n) = |T| \cdot P_\theta(x^n) \tag{4}$$

where x^n is any sequence in T. In the sequel, we will make extensive use of the well-known properties of Markov types summarized in Lemma 1 below. While stronger versions of these properties can be derived, the claims in Lemma 1 are sufficient for our purposes in this paper.

Lemma 1:

(a) $N_{n,k} \le (n+1)^{\alpha^{k+1}}$.

(b) $\sum_{x^n \in \mathcal{A}^n} 2^{-n\hat{H}_k(x^n)} = \sum_{T \in \mathcal{T}_k^n} \hat{P}_T^{(k)}(T) \le N_{n,k}$.

(c) For every type $T \in \mathcal{T}_k^n$, $\hat{P}_T^{(k)}(T) \ge (n+1)^{-\alpha^{k+1}}$.

Proof. Part (a) is an obvious consequence of the characterization of types in terms of sequence composition with respect to the k-th order Markov model. The equality in Part (b) follows from breaking the summation into type classes, whereas the inequality follows from the fact that each term in the summation over the types is at most 1. As for Part (c), we apply Equation (4) with $P_\theta(x^n) = \hat{P}_T^{(k)}(x^n)$. The size of a type class is given by Whittle's formula [28], which consists of the product of α^k multinomial coefficients (one per state) and a cofactor. As shown in [29, page 1996], the cofactor is lower-bounded by $n^{-\alpha^k+1}$, whereas each multinomial coefficient is lower-bounded, using Stirling's formula, by the maximum-likelihood probability of the sub-sequence of symbols occurring at the corresponding state, divided by $\sqrt{(2\pi n)^{\alpha-1}}$. Multiplying by $\hat{P}_T^{(k)}(x^n)$, $x^n \in T$, it follows that

$$\hat{P}_T^{(k)}(T) \ge (2\pi n)^{-\alpha^k(\alpha-1)/2} n^{-\alpha^k+1}\,.$$

The claimed bound follows by noticing that $2\pi n < (n+1)^2$ provided $n \ge 5$. □

Denoting with $T^{(j)}$, $1 \le j \le N_{n,k}$, the type classes in $\mathcal{T}_k^n$, $P_\theta(T^{(j)})$ can be regarded as a function of the parameter vector θ. A key property of the family $\mathcal{P}_k$ in the context of universal simulation is that the set $\{P_\theta(T^{(j)})\}_{j=1}^{N_{n,k}}$, as functions of $\theta \in \Omega$ (where Ω is any subset of $[0,1]^{\alpha^k(\alpha-1)}$ with positive measure), is linearly independent over $\mathbb{R}$ (see [13] for a discussion on this property for Markov models). Our converse in the stochastic setting will make use of the following equivalent property, which follows immediately from the characterization of linear independence in terms of *Casorati determinants* given in, e.g., [26, Chapter 14, Lemma 1].

Lemma 2: Given any subset Λ of $[0,1]^{\alpha^k(\alpha-1)}$ with positive measure, there exist $N_{n,k}$ parameter values $\theta_1, \ldots, \theta_{N_{n,k}} \in \Lambda$ such that the matrix $\{P_{\theta_i}(T^{(j)})\}_{i,j=1}^{N_{n,k}}$ is nonsingular.

Throughout this paper, all simulation schemes $y^n = \phi(x^\ell, u^r)$ assume $n=\ell$, and the key u^r is assumed to be an unlimited stream of random bits, u^∞. The resulting conditional distribution on y^n given x^n is regarded as a channel, denoted $W(y^n|x^n)$, with entropy $H(Y^n|x^n)$. In case x^n is assumed to emerge from a probabilistic source $P \in \mathcal{P}_k$, the conditional entropy achieved by the channel W and the mutual information between X^n and Y^n that is induced by P and W will be denoted $H(Y^n|X^n)$ and $I(X^n;Y^n)$, respectively. In this case, we seek a simulation scheme that, without knowledge of k (which may take on any nonnegative integer value), achieves essentially the same mutual information as the optimal universal scheme that knows k (Condition C2 in Section I), while deviating from Condition C1 for just a negligible fraction of the sequences, for any $P \in \mathcal{P}_k$.

In both the stochastic and the individual sequence setting, our simulation scheme will rely on the existence of Markov order estimators with certain properties, which are specified in Lemma 3 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

$$k(x^n) = \arg\min_{k\geq 0}\{\hat{H}_k(x^n) + \alpha^k f(n)\} \tag{5}$$

where $f(n)$ 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 [16]. In the classical estimation problem, $f(n)$ governs the trade-off between the probabilities of underestimating and overestimating the model order. In the simulation problem for individual sequences, $f(n)$ will be shown to govern a trade-off between faithfulness and entropy of the simulator. The estimate $k(x^n)$ can be obtained in time that is linear in n by use of suffix trees as in [27]. The set of n-tuples x^n such that $k(x^n)=i$ will be denoted $\mathcal{A}_i^n$. To state Lemma 3 we define, for a distribution $P \in \mathcal{P}_k$, the overestimation probability

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

and, similarly, the underestimation probability

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

Lemma 3: For any $k \geq 0$ and any $P \in \mathcal{P}_k$, the estimator of Equation (5) 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 provided $f(n) = o(1)$.
(c) If $z^n \in T_{k(x^n)}(x^n)$ then $k(z^n) \geq k(x^n)$.
(d) $\alpha^{k(x^n)} = O(1/f(n))$ for any $x^n \in \mathcal{A}^n$.

Proof. Part (a) is handled with the method of types as in [22]. A rough bounding procedure (which requires a larger value of β) is given next for completeness:

$$\begin{aligned}
P_{\mathrm{o/e}}(n) &\leq \sum_{i>k}\sum_{x^n\in\mathcal{A}_i^n} 2^{-n\hat{H}_k(x^n)} \\
&\leq \sum_{i>k}\sum_{x^n\in\mathcal{A}_i^n} 2^{-n[\hat{H}_i(x^n)+(\alpha^i-\alpha^k)f(n)]} \\
&\leq \sum_{i=k+1}^{n} n^{-\beta(\alpha^i-\alpha^k)} \sum_{x^n\in\mathcal{A}^n} 2^{-n\hat{H}_i(x^n)} \\
&\leq \sum_{i=k+1}^{n} n^{-\beta(\alpha^i-\alpha^k)} N_{n,i}
\end{aligned}$$

where in the first inequality we upper-bound $P(x^n)$ with the maximum-likelihood probability, the second inequality follows from the definition of $\mathcal{A}_i^n$ and (5), in the third one we apply the condition on $f(n)$ and we extend the inner summation to all sequences x^n, and in the fourth one we use Lemma 1, Part (b). The claim then follows from Lemma 1, Part (a), from observing that the largest term in the summation is the one corresponding to $i = k+1$, and that a suitable choice of β will result in a polynomial decay even after multiplication of $P_{\mathrm{o/e}}(n)$ by $(n+1)^{\alpha^{k+1}}$. The exponential decay in Part (b), on the other hand, follows from the fact that underestimation is a large deviations event. The complete proof is omitted, since similar results have been shown for several variants of this estimator (see, e.g., [23]). Part (c) is an obvious consequence of the fact that $\hat{H}_i(x^n) = \hat{H}_i(z^n)$ for all $i \leq k(x^n)$. Finally, Part (d) follows from the fact that, by the definition of $k(x^n)$ in (5),

$$\hat{H}_{k(x^n)}(x^n) + \alpha^{k(x^n)} f(n) \leq \hat{H}_0(x^n) + f(n) = O(1)$$

since $f(n) = o(1)$ and $\hat{H}_0(x^n) = O(1)$. □

We consider the simulation scheme that, given a training sequence x^n, draws y^n uniformly at random from the set

$$\mathcal{M}(x^n) \stackrel{\triangle}{=} T_{k(x^n)}(x^n) \cap \mathcal{A}^n_{k(x^n)}\,.$$

A key lemma in the analysis of this simulation scheme, for both the stochastic and the individual sequence setting, states that for any x^n the set $\mathcal{M}(x^n)$ comprises all but a negligible fraction of the sequences in $T_{k(x^n)}(x^n)$. By Part (c) of Lemma 3, the remaining sequences are in $\{\mathcal{A}_i^n\}_{i>k(x^n)}$. To state the lemma, we define

$$\mathsf{P}_{\mathrm{o/e}}^{(i)}(n) \stackrel{\triangle}{=} \max_{P=P_T, T\in\mathcal{T}_i^n} P_{\mathrm{o/e}}(n) \tag{6}$$

namely, $\mathsf{P}_{\mathrm{o/e}}^{(i)}(n)$ is the maximum value of $P_{\mathrm{o/e}}(n)$ over all empirical distributions of Markov type classes of order i and length n. Notice that $\mathsf{P}_{\mathrm{o/e}}^{(i)}(n)$ is a deterministic function of i and n, independent of any underlying probability law. Since Part (a) of Lemma 3 holds uniformly in P and k, it follows from (6) that $(n+1)^{\alpha^{i+1}}\mathsf{P}_{\mathrm{o/e}}^{(i)}(n)$ is upper-bounded, for a suitable choice of $f(n)$, by a function that decays polynomially fast with n, uniformly in i.

Lemma 4: For any $i \geq 0$, let $T \in \mathcal{T}_i^n$ and assume $T \cap \mathcal{A}_i^n \neq \phi$. Then,

$$\frac{|T \cap \mathcal{A}_i^n|}{|T|} \geq 1 - (n+1)^{\alpha^{i+1}} \mathsf{P}_{\mathrm{o/e}}^{(i)}(n) \,.$$

Proof. By Lemma 1, Part (c), and since $\hat{P}_T^{(i)}(\cdot)$ is uniform over T, we have

$$(n+1)^{-\alpha^{i+1}} \leq \hat{P}_T^{(i)}(T) = \hat{P}_T^{(i)}(T \cap \bar{\mathcal{A}}_i^n) \, \frac{|T|}{|T \cap \bar{\mathcal{A}}_i^n|}$$

where the complement of a set $\mathcal{S}$ is denoted $\bar{\mathcal{S}}$. By Lemma 3, Part (c), since $T \cap \mathcal{A}_i^n$ is nonempty, we have $k(z^n) \geq i$ for all $z^n \in T$. Therefore,

$$\hat{P}_T^{(i)}(T \cap \bar{\mathcal{A}}_i^n) \leq \sum_{r>i} \hat{P}_T^{(i)}(\mathcal{A}_r^n) \,. \tag{7}$$

Since $\hat{P}_T^{(i)} \in \mathcal{P}_i$ (i.e., the order of the empirical distribution is not smaller than i, for otherwise no sequence in T would have estimated order i), the summation in the right-hand side of (7) is the overestimation probability for $P = \hat{P}_T^{(i)}$, which, by (6), is upper-bounded by $\mathsf{P}_{\mathrm{o/e}}^{(i)}(n)$. □

Lemma 4 is valid for any model order i and any type class containing sequences that do estimate order i, regardless of any probabilistic assumption. It should be noticed, however, that the assumption of equal weight for counting all sequences in T can be regarded as implicitly implying that these sequences are drawn from a Markov source of order i or less. Notice also that, as stated in the proof of the lemma, types T such that $\hat{P}_T^{(i)} \notin \mathcal{P}_i$ (e.g., for $i > 0$, the type of order i of the all-zero sequence) do not contain sequences that estimate order i.

III. The stochastic setting

Theorem 1 below states the properties, in the stochastic setting, of the simulator that draws y^n uniformly at random from $\mathcal{M}(x^n)$. For this simulator, $W(y^n|x^n) = 1/|\mathcal{M}(x^n)|$ if $y^n \in \mathcal{M}(x^n)$ and $W(y^n|x^n) = 0$ otherwise. Since $y^n \in \mathcal{M}(x^n)$ if and only if $x^n \in \mathcal{M}(y^n)$ ($\mathcal{M}(\cdot)$ is a partition of $\mathcal{A}^n$ and $\mathcal{M}(x^n) = \mathcal{M}(y^n)$), the output distribution $Q(\cdot)$ satisfies

$$\begin{aligned} Q(y^n) &= \sum_{x^n \in \mathcal{A}^n} P(x^n) W(y^n|x^n) \\ &= \sum_{x^n \in \mathcal{M}(y^n)} \frac{P(x^n)}{|\mathcal{M}(y^n)|} = \frac{P(\mathcal{M}(y^n))}{|\mathcal{M}(y^n)|} \,. \end{aligned} \tag{8}$$

Theorem 1: For any $k \geq 0$ and any $P \in \mathcal{P}_k$ we have

(a) The output distribution satisfies

$$Q(Q(y^n) \neq P(y^n)) \leq P_{\mathrm{u/e}}(n) \,.$$

(b) The conditional entropy of the simulator satisfies

$$\begin{aligned} H(Y^n|X^n) &\geq \boldsymbol{E} \log |T_k(X^n)| - n P_{\mathrm{o/e}}(n) \log \alpha \\ &+ \min_{i \leq k} \log[1 - (n+1)^{\alpha^{i+1}} \mathsf{P}_{\mathrm{o/e}}^{(i)}(n)] \end{aligned}$$

where the expectation is with respect to P, and its entropy satisfies

$$H(Y^n) \leq H(X^n) + P_{\mathrm{u/e}}(n)[n \log \alpha - \log P_{\mathrm{u/e}}(n)] \,.$$

By Part (b) of Lemma 3, Part (a) of the theorem states that the proposed simulator preserves the probability law, except for a set of exponentially decaying probability of the outcomes of the simulation. In addition, Part (b) states that, with proper choice of $f(n)$,

$$I(X^n; Y^n) \leq H(X^n) - \boldsymbol{E} \log |T_k(X^n)| + o(1) \tag{9}$$

for the proposed scheme. As shown in [9] and discussed in Section I, the mutual information of the optimal scheme that knows k (and preserves the probability law), which draws y^n uniformly at random from $T_k(x^n)$, is

$$H(X^n) - \boldsymbol{E} \log |T_k(X^n)| \approx \alpha^k \frac{\alpha - 1}{2} \log n \tag{10}$$

where the approximation is to the main asymptotic term. Thus, the asymptotic behavior is unaffected by the addition of the $o(1)$ term in the proposed scheme. However, the scheme of [9] is optimal for *exact* preservation of the probability law, and therefore does not yet establish a converse theorem for the relaxed version of Condition C1.

Notice that the choice of $f(n)$ governs the tension between preservation of the probability law (which is only affected by underestimation) and conditional entropy (which is reduced by overestimation). However, as long as $f(n) > \beta(\log n)/n$, as stated in Lemma 3, the asymptotic behavior is independent of $f(n)$.

Proof of Theorem 1. To prove Part (a), notice that if $k(y^n) \geq k$, then $P(y^n) = P(z^n)$ for all $z^n \in \mathcal{M}(y^n)$. Thus, by (8), $Q(y^n) = P(y^n)$. Hence,

$$P(Q(y^n) \neq P(y^n)) \leq P(k(y^n) < k) = P_{\mathrm{u/e}}(n) \,. \tag{11}$$

Furthermore, for all $y^n \in \mathcal{A}^n$,

$$\begin{aligned} Q(Q(y^n) \neq P(y^n)) &= 1 - Q(Q(y^n) = P(y^n)) \\ &= 1 - P(Q(y^n) = P(y^n)) \end{aligned}$$

which, together with (11), proves the claim.

As for Part (b),

$$\begin{aligned} H(Y^n|X^n) &= \sum_{i \geq 0} \sum_{x^n \in \mathcal{A}_i^n} P(x^n) \log |T_i(x^n) \cap \mathcal{A}_i^n| \\ &\geq \sum_{i=0}^{k} \sum_{x^n \in \mathcal{A}_i^n} P(x^n) \Big[\log |T_i(x^n)| \\ &+ \log[1 - (n+1)^{\alpha^{i+1}} \mathsf{P}_{\mathrm{o/e}}^{(i)}(n)] \Big] \\ &\geq \boldsymbol{E} \log |T_k(X^n)| \\ &- \sum_{i>k} \sum_{x^n \in \mathcal{A}_i^n} P(x^n) \log |T_k(x^n)| \\ &+ \min_{i \leq k} \log[1 - (n+1)^{\alpha^{i+1}} \mathsf{P}_{\mathrm{o/e}}^{(i)}(n)] \\ &\geq \boldsymbol{E} \log |T_k(X^n)| \\ &- n(\log \alpha) \sum_{i>k} \sum_{x^n \in \mathcal{A}_i^n} P(x^n) \\ &+ \min_{i \leq k} \log[1 - (n+1)^{\alpha^{i+1}} \mathsf{P}_{\mathrm{o/e}}^{(i)}(n)] \end{aligned}$$

where the first inequality follows from Lemma 4, the second inequality follows from the fact that $|T_i(x^n)| \geq |T_k(x^n)|$ for all $i \leq k$, and the third inequality follows from upper-bounding the type class sizes for $i > k$ with α^n. The claimed bound on the conditional entropy then follows from the definition of $P_{\mathrm{o/e}}(n)$. To upper-bound $H(Y^n)$, we observe that

$$\begin{aligned} H(Y^n) &= -\sum_{y^n:Q(y^n)=P(y^n)} P(y^n)\log P(y^n) \\ &\quad -\sum_{y^n:Q(y^n)\neq P(y^n)} Q(y^n)\log Q(y^n) \\ &\leq H(X^n) + Q(Q(y^n)\neq P(y^n)) \\ &\quad \cdot \log \sum_{y^n:Q(y^n)\neq P(y^n)} \frac{1}{Q(Q(y^n)\neq P(y^n))} \\ &\leq H(X^n) + P_{\mathrm{u/e}}(n)\log \sum_{y^n\in\mathcal{A}^n} \frac{1}{P_{\mathrm{u/e}}(n)} \\ &= H(X^n) + P_{\mathrm{u/e}}(n)\log \frac{\alpha^n}{P_{\mathrm{u/e}}(n)} \end{aligned}$$

where the first inequality follows from Jensen's inequality, and the second inequality follows from Part (a) and assumes $P_{\mathrm{u/e}}(n) < 1/e$. □

From an algorithmic perspective, the enumeration of the intersection $\mathcal{M}(x^n)$ of $T_{k(x^n)}(x^n)$ and $\mathcal{A}^n_{k(x^n)}$, on which the implementation of the draw is based, may be a challenging problem. We can circumvent the problem by drawing uniformly at random from the type class, until a sequence that estimates the same order as x^n is drawn. By Lemma 4, with very high probability, only one draw will be needed. We can also consider the simulation scheme that simply draws from the type class $T_{k(x^n)}(x^n)$, rather than from the intersection $\mathcal{M}(x^n)$. While it cannot be claimed that such a scheme preserves the probability law in the strong sense of Part (a) of Theorem 1, it can be shown that, with an appropriate choice of $f(n)$,

$$Q\left(\left|\frac{Q(y^n)}{P(y^n)} - 1\right| > \epsilon(n)\right) < \delta(n)$$

where $\epsilon(n)$ and $\delta(n)$ are vanishing functions. Clearly, the conditional entropy of this simulator satisfies, *a fortiori*, Part (b) of Theorem 1, as it draws from a larger set, whereas we can upper-bound $H(Y^n)$ so that the upper bound (9) on the mutual information still holds. However, such a scheme does not yield a partition of $\mathcal{A}^n$ and *does not* satisfy our claims in the individual sequence setting.

Next, we establish a converse theorem, for which we need to show that a deviation in the probability law as the one allowed for the scheme of Theorem 1 cannot open the possibility for a faster decay of the per-symbol mutual information in a competing scheme that knows k. This fact is established in Theorem 2 below.

Theorem 2 (Converse): Let Ω denote an open subset of $[0,1]^{\alpha^k(\alpha-1)}$. Let $\delta_\theta(n)$ be a set of functions indexed by θ such that

$$\tau_\theta \stackrel{\triangle}{=} \liminf_{n\to\infty} \frac{-1}{n}\log\delta_\theta(n) > 0$$

and assume that for every $\theta\in\Omega$ there exists an open neighborhood Λ_θ containing θ over which the infimum of τ_θ is positive. Let W be any simulation scheme such that its output distribution Q_θ satisfies, for every $\theta \in \Omega$, $Q_\theta(Q_\theta(y^n) \neq P_\theta(y^n)) \leq \delta_\theta(n)$. Then, for every $\theta\in\Omega$,

$$I(X^n;Y^n) \geq H(X^n) - \boldsymbol{E}\log|T_k(X^n)| - o(1)$$

where the $o(1)$ term may not be uniform in θ.

The conditions on the set of functions $\delta_\theta(n)$ in Theorem 2 essentially state that W is any simulation scheme such that the probability of the outcomes for which the probability law is not preserved is exponentially small, and that the bound $\delta_\theta(n)$ on this probability is well-behaved (for example, continuity of $\delta_\theta(n)$ as a function of θ would suffice). Notice that, in particular, the upper bound $P_{\mathrm{u/e}}(n)$ on the probability of the sequences which are not law preserving for the proposed scheme satisfies these conditions (for sets Ω for which $P_{\mathrm{u/e}}(n)$ is well-defined, namely sets which do not include parameter vectors corresponding to distributions in $\mathcal{P}_{k'}$, $k'<k$). Thus, by the asymptotic approximation in (10), the theorem states that no asymptotically significant improvement of $I(X^n;Y^n)$ can be obtained by letting the output probability deviate from the input one by an amount such as the one allowed in the direct (Theorem 1).

Proof of Theorem 2. Let $B_\delta(\theta)$ denote the set of sequences y^n for which $Q_\theta(y^n)\neq P_\theta(y^n)$. We have

$$\begin{aligned} H(Y^n) &\geq H(X^n) + \sum_{y^n\in B_\delta(\theta)} P_\theta(y^n)\log P_\theta(y^n) \\ &\geq H(X^n) - P_\theta(B_\delta(\theta))\log\frac{|B_\delta(\theta)|}{P_\theta(B_\delta(\theta))} \end{aligned}$$

where the second inequality follows from Jensen's inequality. In addition, for all $\theta\in\Omega$, we have

$$P_\theta(B_\delta(\theta)) = Q_\theta(B_\delta(\theta)) \leq \delta_\theta(n) \tag{12}$$

which, assuming $\delta_\theta(n) < 1/e$ and bounding $|B_\delta(\theta)|$ with α^n, implies

$$H(Y^n) \geq H(X^n) - n\delta_\theta(n)\log\alpha + \delta_\theta(n)\log\delta_\theta(n)\,. \tag{13}$$

Also,

$$\begin{aligned} H(Y^n|X^n) &\leq H(Y^n|T_k(X^n)) \\ &= \sum_{T\in\mathcal{T}_k^n} P_\theta(T) \sum_{y^n\in\mathcal{A}^n} W(y|T)\log\frac{1}{W(y|T)}\,. \end{aligned}$$

Splitting the summation in y^n according to whether y^n belongs to T or not and applying Jensen's inequality to each partial summation, we obtain

$$\begin{aligned} H(Y^n|X^n) &\leq \sum_{T\in\mathcal{T}_k^n} P_\theta(T)\left[W(T|T)\log\frac{|T|}{W(T|T)}\right. \\ &\quad + \left.[1-W(T|T)]\log\frac{\alpha^n-|T|}{1-W(T|T)}\right]\,. \end{aligned}$$

Denoting $w_T = W(T|T)$ and letting $h(\cdot)$ denote the binary entropy function, we obtain

$$\begin{aligned} H(Y^n|X^n) &\leq \sum_{T\in\mathcal{T}_k^n} P_\theta(T)[w_T \log|T| \\ &+ (1-w_T)\log(\alpha^n - |T|) + h(w_T)] \\ &\leq \boldsymbol{E}\log|T_k(X^n)| \\ &+ \sum_{T\in\mathcal{T}_k^n} P_\theta(T)[n(1-w_T)\log\alpha + h(w_T)]\,. \end{aligned}$$

This bound, together with (13), yields

$$\begin{aligned} I(X^n;Y^n) &\geq H(X^n) - \boldsymbol{E}\log|T_k(X^n)| \\ &- n\delta_\theta(n) + \delta_\theta(n)\log\delta_\theta(n) \qquad (14) \\ &- \sum_{T\in\mathcal{T}_k^n} P_\theta(T)[n(1-w_T)\log\alpha + h(w_T)]\,. \end{aligned}$$

Since $\delta_\theta(n)$ decays exponentially fast, the proof is complete if we show that the summation on the right-hand side of (14) also vanishes with n. To this end, it suffices to prove that $n(1-w_T)$ vanishes (and, hence, $h(w_T)$ also vanishes) for every type T such that the maximum-likelihood estimate $\hat{\theta}_T$ of θ, obtained from a sequence of type T, belongs to Λ_θ. This observation follows from the fact that, for the other types, $\hat{\theta}_T \notin \Lambda_\theta$ is a large deviations event and therefore its probability decays exponentially fast (depending on the sets Λ_θ, this decay may not be uniform in θ).[2] We next analyze w_T for types T such that $\hat{\theta}_T \in \Lambda_\theta$.

First, notice that we can assume $\Lambda_\theta \subseteq \Omega$ for all $\theta \in \Omega$ (otherwise, we can replace Λ_θ with the interior of its intersection with Ω). For a given type T such that $\hat{\theta}_T \in \Lambda_\theta$, consider an open neighborhood Λ_T of $\hat{\theta}_T$, $\Lambda_T \subseteq \Lambda_\theta$. By Lemma 2, there exist parameter values $\theta_1, \ldots, \theta_{N_{n,k}} \in \Lambda_T$ such that the matrix $\{P_{\theta_i}(T^{(j)})\}_{i,j=1}^{N_{n,k}}$ is nonsingular. By (12), we have $P_{\theta_i}(B_\delta(\theta_i)) \leq \delta_{\theta_i}(n)$, $i = 1, 2, \ldots, N_{n,k}$. By the definition of Λ_θ, $\inf_{\theta' \in \Lambda_\theta} \tau_{\theta'} > 0$, so that the functions $\delta_{\theta_i}(n)$ are, in turn, upper-bounded by a single exponentially decaying function. Now, since Λ_T can be made arbitrarily small, the parameters θ_i are arbitrarily close to $\hat{\theta}_T$, and it can be assumed that no component of $\hat{\theta}_T$ (transition probability) has a positive value while the corresponding component of a vector in Λ_T is zero. Hence, by the continuity of $P_\theta(\cdot)$ (as a function of θ), there exists another exponentially decaying function $\delta'_\theta(n)$ such that

$$\hat{P}_T^{(k)}(B_\delta(\theta_i)) = P_{\hat{\theta}_T}(B_\delta(\theta_i)) \leq \delta'_\theta(n)$$

$i = 1, 2, \ldots, N_{n,k}$.[3] Letting $B_T = \cup_i B_\delta(\theta_i)$, a union bound yields

$$\hat{P}_T^{(k)}(B_T) \leq N_{n,k}\delta'_\theta(n)\,. \qquad (15)$$

[2]Of course, since θ is arbitrary in Ω, our proof will actually apply to any type T such that $\hat{\theta}_T \in \Omega$.

[3]E.g., we can pick Λ_T such that the positive transition probabilities of $\hat{\theta}_T$ are within a factor e^ϵ of the corresponding component in each vector in Λ_T, so that the probability of an event under $\hat{\theta}_T$ is upper-bounded by $e^{n\epsilon}$ times the corresponding probability under each θ_i, and pick ϵ sufficiently small.

Thus,

$$\frac{|B_T \cap T|}{|T|} = \frac{\hat{P}_T^{(k)}(B_T \cap T)}{\hat{P}_T^{(k)}(T)} \leq \frac{\hat{P}_T^{(k)}(B_T)}{\hat{P}_T^{(k)}(T)} \leq (n+1)^{2\alpha^{k+1}}\delta'_\theta(n) \qquad (16)$$

where the second inequality follows from (15) and Lemma 1, parts (a) and (c). Now, for any P and any $y^n \in \mathcal{A}^n$, we have

$$Q(y^n) = \sum_{j=1}^{N_{n,k}} P(T^{(j)})W(y^n|T^{(j)})\,. \qquad (17)$$

In particular, for $P = P_{\theta_i}$ and $y^n \in G_T \cap T$, where G_T denotes the complement of B_T, $Q_{\theta_i}(y^n) = P_{\theta_i}(y^n)$ for all values of i. Thus, since y^n has type T, (17) takes the form

$$\frac{P_{\theta_i}(T)}{|T|} = \sum_{j=1}^{N_{n,k}} P_{\theta_i}(T^{(j)})W(y^n|T^{(j)}),\ i = 1, 2, \ldots, N_{n,k}\,. \qquad (18)$$

Since our choice of the parameters θ_i was dictated by Lemma 2, the system of equations (18) (in the $N_{n,k}$ variables $W(y^n|T^{(j)})$) yields a unique solution, and hence $W(y^n|T) = 1/|T|$ for all $y^n \in G_T \cap T$. It follows that

$$w_T = \sum_{y^n\in T} W(y^n|T) \geq \sum_{y^n\in G_T\cap T} W(y^n|T) = \frac{|G_T \cap T|}{|T|}$$

or, by (16),

$$1 - w_T \leq (n+1)^{2\alpha^{k+1}}\delta'_\theta(n)\,.$$

Thus, we conclude that $n(1-w_T)$ indeed vanishes. □

Notice that in the proof of Theorem 2 we have shown that, when a deviation in the probability law is allowed for a set of sequences $B_\delta(\theta)$ of exponentially vanishing probability, $1 - W(T|T)$ is exponentially small for types T such that the maximum-likelihood estimate $\hat{\theta}_T \in \Omega$. This is a generalization of the result in [9] stating that when perfect preservation of the probability law is required, $W(T|T) = 1$ for every type (since a law-preserving simulator can only output sequences from the same type as x^n). Both the law-preserving result and its relaxation follow essentially from the linear independence of the set of functions $\{P_\theta(T^{(j)})\}_{j=1}^{N_{n,k}}$, as functions of $\theta \in \Omega$.

IV. The individual sequence setting

In this section we analyze the proposed simulation scheme in the individual sequence setting of [12]. By drawing uniformly at random a sequence from $\mathcal{M}(x^n)$ (where x^n is now an individual sequence not originating from any probabilistic source), we have $H(Y^n|x^n) = \log|\mathcal{M}(x^n)|$. To complete our "direct" result we establish, in Theorem 3 below, the statistical similarity between two sequences in the same class $\mathcal{M}(\cdot)$ in the implied partition of $\mathcal{A}^n$.

Theorem 3 (Direct): Let $x^n \in \mathcal{A}^n$ be arbitrary and fix a nonnegative integer j. Then, for any $y^n \in \mathcal{M}(x^n)$, we have

$$\sum_{s\in\mathcal{A}^j}\left|\frac{n_{x^n}(s)}{n} - \frac{n_{y^n}(s)}{n}\right| = O\left(\sqrt{\left[f(n) + \frac{1}{n^2}\right]\alpha^j}\right)\,.$$

Moreover, if $j \le k(x^n)+1$ then $n_{x^n}(s) = n_{y^n}(s)$ for all $s \in \mathcal{A}^j$.

Proof. Let $i = k(x^n)$. By (5), for every $r > 0$, we have

$$\hat{H}_i(x^n) - \hat{H}_{i+r}(x^n) \le \alpha^i(\alpha^r - 1) f(n)\,. \tag{19}$$

Using Equation (3), and since, by (2), for every string s we have $\sum_{u\in\mathcal{A}^r} n_{x^n}(us) = n_{x^n}(s)$, the left-hand side of (19) takes the form

$$\begin{aligned}\hat{H}_i(x^n) - \hat{H}_{i+r}(x^n) &= \sum_{u\in\mathcal{A}^r}\sum_{s\in\mathcal{A}^i}\sum_{a\in\mathcal{A}} \frac{n_{x^n}(usa)}{n} \\ &\cdot \log \frac{n_{x^n}(usa)\, n'_{x^n}(s)}{n'_{x^n}(us)\, n_{x^n}(sa)} \\ &= D\left(\hat{P}^{(1)}_{x^n}(\cdot)\,||\,\hat{P}^{(2)}_{x^n}(\cdot)\right)\end{aligned}$$

where $\hat{P}^{(1)}_{x^n}(\cdot)$ and $\hat{P}^{(2)}_{x^n}(\cdot)$ are probability mass functions over $\mathcal{A}^{i+r+1}$ defined by $\hat{P}^{(1)}_{x^n}(usa) \stackrel{\triangle}{=} n_{x^n}(usa)/n$ and

$$\hat{P}^{(2)}_{x^n}(usa) \stackrel{\triangle}{=} \frac{n'_{x^n}(us)\, n_{x^n}(sa)}{n'_{x^n}(s)\, n}$$

with $\hat{P}^{(2)}_{x^n}(usa) = 0$ if $n'_{x^n}(s) = 0$. By Pinsker's inequality we then have

$$\sum_{w\in\mathcal{A}^{r+i+1}} \left|\hat{P}^{(1)}_{x^n}(w) - \hat{P}^{(2)}_{x^n}(w)\right| \le \sqrt{2(\ln 2)[\hat{H}_i(x^n) - \hat{H}_{i+r}(x^n)]}$$

which, together with (19) yields

$$\sum_{w\in\mathcal{A}^{r+i+1}} \left|\hat{P}^{(1)}_{x^n}(w) - \hat{P}^{(2)}_{x^n}(w)\right| \le \sqrt{2(\ln 2) f(n)\alpha^i(\alpha^r-1)}\,. \tag{20}$$

Now, given $y^n \in \mathcal{M}(x^n)$, let

$$\begin{aligned}\Delta_r &\stackrel{\triangle}{=} \sum_{w\in\mathcal{A}^{r+i+1}} \left|\frac{n_{x^n}(w)}{n} - \frac{n_{y^n}(w)}{n}\right| \\ &= \sum_{w\in\mathcal{A}^{r+i+1}} \left|\hat{P}^{(1)}_{x^n}(w) - \hat{P}^{(1)}_{y^n}(w)\right|\,.\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}\Delta_r \le \sum_{w\in\mathcal{A}^{r+i+1}} \Bigg(&\left|\hat{P}^{(1)}_{x^n}(w) - \hat{P}^{(2)}_{x^n}(w)\right| \\ &+ \left|\hat{P}^{(2)}_{x^n}(w) - \hat{P}^{(2)}_{y^n}(w)\right| + \left|\hat{P}^{(1)}_{y^n}(w) - \hat{P}^{(2)}_{y^n}(w)\right|\Bigg)\,.\end{aligned} \tag{21}$$

We upper-bound the first term in the summation in (21) using (20) which, since $k(x^n) = k(y^n)$, also applies to the third term. As for the second term, since $y^n \in T_i(x^n)$, we have $n'_{x^n}(s) = n'_{y^n}(s)$ and $n_{x^n}(sa) = n_{y^n}(sa)$ for all $s \in \mathcal{A}^i$ and $a \in \mathcal{A}$. Therefore,

$$\begin{aligned}\Delta_r &\le 2\sqrt{2(\ln 2) f(n)\alpha^i(\alpha^r-1)} \\ &+ \sum_{u\in\mathcal{A}^r}\sum_{s\in\mathcal{A}^i} \left|\frac{n'_{x^n}(us)}{n} - \frac{n'_{y^n}(us)}{n}\right| \sum_{a\in\mathcal{A}} \frac{n_{x^n}(sa)}{n'_{x^n}(s)} \\ &= 2\sqrt{2(\ln 2) f(n)\alpha^i(\alpha^r-1)} \\ &+ \sum_{w\in\mathcal{A}^{r+i}} \left|\frac{n'_{x^n}(w)}{n} - \frac{n'_{y^n}(w)}{n}\right|\,.\end{aligned} \tag{22}$$

By the definition of the counts $n'_{x^n}(w)$ and $n'_{y^n}(w)$, and noticing that $x_{-(r+i-1)},\ldots,x_0 = y_{-(r+i-1)},\ldots,y_0$, we have, for all $w \in \mathcal{A}^{r+i}$,

$$\begin{aligned}n'_{x^n}(w) - n'_{y^n}(w) &= n_{x^n}(w) - n_{y^n}(w) \\ &- \mathbf{1}(w = x_{n+1-r-i},\ldots,x_n) \\ &+ \mathbf{1}(w = y_{n+1-r-i},\ldots,y_n)\end{aligned}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function of the specified event. Therefore,

$$\sum_{w\in\mathcal{A}^{r+i}} |n'_{x^n}(w) - n'_{y^n}(w)| \le \sum_{w\in\mathcal{A}^{r+i}} |n_{x^n}(w) - n_{y^n}(w)| + 2$$

implying, by (22),

$$\Delta_r \le 2\sqrt{2(\ln 2) f(n)\alpha^i(\alpha^r-1)} + \frac{2}{n} + \Delta_{r-1}\,. \tag{23}$$

Since $y^n \in T_i(x^n)$, we have $\Delta_0 = 0$, so that the recurrence in (23) yields

$$\Delta_r \le 2\sqrt{2(\ln 2) f(n)}\,\alpha^{i/2} \sum_{m=1}^{r} \sqrt{\alpha^m - 1} + \frac{2r}{n}\,. \tag{24}$$

The summation in (24) is $O(\alpha^{r/2})$, implying

$$\Delta_r = O\left(\sqrt{\left[f(n) + \frac{1}{n^2}\right]\alpha^{k+r}}\right)\,. \tag{25}$$

The claim of the theorem for $j > k(x^n)$ follows from taking $r = j - i - 1$ in (25). The claim that $n_{x^n}(s) = n_{y^n}(s)$ for $j \le k(x^n)+1$ and all $s \in \mathcal{A}^j$ is an immediate consequence of the fact that $y^n \in T_{k(x^n)}(x^n)$. □

Theorem 3 corresponds to Property P1 that was itemized in Section I for the scheme in [12]. With a proper choice of $f(n)$, the preservation of empirical probabilities (or degree of "faithfulness") within $\mathcal{M}(x^n)$ is stronger than the one claimed for the LZ parsing-based types, for which the convergence is $O(1/\log n)$. As in [12, Corollary 1], for any fixed Markov measure $\Pi \in \mathcal{P}_k$, if $y^n \in \mathcal{M}(x^n)$ then $(1/n)|\log(\Pi(x^n)/\Pi(y^n))|$ is also $O(\sqrt{[f(n)+1/n^2]})$, provided both $\Pi(x^n)$ and $\Pi(y^n)$ are positive. Moreover, the set of sequences x^n for which there exists a sequence $y^n \in \mathcal{M}(x^n)$ such that $\Pi(x^n) \ne \Pi(y^n)$ has measure at most $P_{\mathrm{u/e}}(n)$ under Π. Thus, for "most" sequences x^n, $\Pi(x^n) = \Pi(y^n)$ for all $y^n \in \mathcal{M}(x^n)$.

Yet, the entropy $H(Y^n|x^n) = \log|\mathcal{M}(x^n)|$ of the proposed simulator is essentially optimal when compared to any competing faithful simulator, even if we are extremely "generous" in the definition of faithfulness for the competitor, provided the type classes are defined for an estimator such that $\log n = o(nf(n))$. Specifically, let a *weakly faithful* simulator $W(Y^n|x^n)$ be only constrained to output sequences y^n such that, for any fixed nonnegative integer i, $\hat{H}_i(y^n) < \hat{H}_i(x^n) + \gamma_i(n)$, where $\{\gamma_i(n)\}_{i\ge 0}$ is any family of vanishing functions of n (not necessarily uniformly in i). Notice that the condition does not necessarily imply closeness in terms of counts (on the other hand, a scheme that approximately preserves counts in the sense of Theorem 3 will obviously

be faithful in this relaxed sense for some family $\{\gamma_i(n)\}_{i\geq 0}$). Moreover, we further relax this condition by assuming that a set $\mathcal{B}(x^n)$ of potential output sequences y^n may not satisfy it, with $W(\mathcal{B}(x^n)|x^n) < \varsigma(n)$ for some vanishing function $\varsigma(n)$. Theorem 4 below asserts that, given *any* partition of the sequence space into a sub-exponential number of classes, for most sequences x^n (under any stationary ergodic measure), the conditional entropy achieved by a weakly faithful simulator cannot be substantially larger than the logarithm of the size of the class of x^n. Our converse will then follow as a corollary to Theorem 4.

Theorem 4: Let $\mathcal{C}(x^n)$ denote any partition of $\mathcal{A}^n$ into $N(\mathcal{C})$ classes, where $\log N(\mathcal{C}) = o(n)$, and let $W(Y^n|x^n)$ be a weakly faithful simulator. Then,

$$\limsup_{n\to\infty} \frac{H(Y^n|x^n) - \log|\mathcal{C}(x^n)|}{n} \leq 0$$

almost surely under any stationary ergodic measure μ.

Proof. Consider the probability distribution $\tilde{P}(x^n) = [N(\mathcal{C})\,|\mathcal{C}(x^n)|]^{-1}$ over $\mathcal{A}^n$. By the sample converse to the noiseless Source Coding Theorem [30, Theorem 3.1] (see also [31]),

$$\liminf_{n\to\infty} \frac{-\log \tilde{P}(X^n)}{n} \geq \mathcal{H}$$

in a set of μ-volume one, where $\mathcal{H}$ denotes the entropy rate of the given measure μ. Since $\log N(\mathcal{C}) = o(n)$, we then have

$$\liminf_{n\to\infty} \frac{\log|\mathcal{C}(x^n)|}{n} \geq \mathcal{H} \tag{26}$$

in a set of μ-volume one.

Now, to upper-bound the conditional entropy $H(Y^n|x^n)$ of the given weakly faithful simulator, define, for every $i \geq 0$,

$$\mathcal{G}_i^{(\gamma)}(x^n) \stackrel{\triangle}{=} \{y^n \in \mathcal{A}^n : \hat{H}_i(y^n) < \hat{H}_i(x^n) + \gamma_i(n)\}$$

and $\mathcal{G}^{(\gamma)}(x^n) = \cap_i \mathcal{G}_i^{(\gamma)}(x^n)$. Observe that

$$\begin{aligned} H(Y^n|x^n) = & - \sum_{y^n\in\mathcal{G}^{(\gamma)}(x^n)} W(y^n|x^n)\log W(y^n|x^n) \\ & - \sum_{y^n\in\mathcal{B}(x^n)} W(y^n|x^n)\log W(y^n|x^n)\,. \end{aligned}$$

Thus, by Jensen's inequality, letting $N_i^{(\gamma)}(x^n) = |\mathcal{G}_i^{(\gamma)}(x^n)|$, for every fixed i we have

$$\begin{aligned} H(Y^n|x^n) &\leq \log N_i^{(\gamma)}(x^n) \\ &+ W(\mathcal{B}(x^n)|x^n)\log\frac{|\mathcal{B}(x^n)|}{W(\mathcal{B}(x^n)|x^n)} \\ &\leq \log N_i^{(\gamma)}(x^n) + \varsigma(n)[-\log\varsigma(n) + n\log\alpha] \end{aligned}$$

where the last inequality assumes $\varsigma(n) < 1/e$. To upper-bound $N_i^{(\gamma)}(x^n)$ we observe that, by Lemma 1, Part (b),

$$N_{n,i} \geq \sum_{y^n\in\mathcal{A}^n} 2^{-n\hat{H}_i(y^n)} \geq N_i^{(\gamma)}(x^n)2^{-n[\hat{H}_i(x^n)+\gamma_i(n)]}\,.$$

Therefore, for any fixed i, we have

$$\begin{aligned} H(Y^n|x^n) &\leq n\hat{H}_i(x^n) + n\gamma_i(n) + \log N_{n,i} \\ &+ \varsigma(n)[-\log\varsigma(n) + n\log\alpha]\,. \end{aligned} \tag{27}$$

Since $\gamma_i(n) = o(1)$, $\varsigma(n) = o(1)$, and, by Lemma 1, Part (a), $N_{n,i}$ grows polynomially fast with n, it follows that

$$\limsup_{n\to\infty}\left[n^{-1}H(Y^n|x^n) - \hat{H}_i(x^n)\right] \leq 0$$

for any fixed i. The theorem then follows from (26) and the fact that $\lim_i \lim_n \hat{H}_i(x^n) = \mathcal{H}$ *a.s.* [32]. □

Corollary 1 (Converse): For any weakly faithful simulator

$$\limsup_{n\to\infty} \frac{H(Y^n|x^n) - \log|\mathcal{M}(x^n)|}{n} \leq 0$$

almost surely under any stationary ergodic measure, provided that $\log n = o(nf(n))$.

Proof. By Theorem 4, it suffices to show that $\log N(\mathcal{M}) = o(n)$ for $f(n)$ satisfying the condition in the corollary. Let $K_n = \max\{k(x^n) : x^n \in \mathcal{A}^n\}$ and $S_n = \alpha^{K_n}$. Clearly,

$$N(\mathcal{M}) \leq \sum_{i=0}^{K_n} N_{n,i} < (K_n+1)(n+1)^{\alpha S_n}\,.$$

By Lemma 3, Part (d), $S_n = O(1/f(n))$, implying $\log N(\mathcal{M}) = O((\log n)/f(n))$. Since $\log n = o(nf(n))$, we conclude that $\log N(\mathcal{M}) = o(n)$. □

The converse in [12] (Theorem 6 therein) is also an immediate consequence of Theorem 4 since, by [12, Corollary 5], the logarithm of the number of LZ parsing-based classes is $O(n/(\log n))$. The key is to find a partition of the sequence space into a sub-exponential number of classes such that the sequences in a class satisfy Property P1. In fact, more obvious simulation schemes possess the same properties in the individual sequence setting: consider, for example, the simulator that draws uniformly at random from $T_{g(n)}(x^n)$, where $g(n)\to\infty$ and $\alpha^{g(n)} = o(n/(\log n))$. For fixed j and sufficiently large n, two sequences in the same type class clearly yield identical occurrence counts for every j-tuple (*exact* preservation in Property P1), whereas, by the slow growth of $g(n)$, the converse also holds.

While all these simulators exhibit similar properties for arbitrary sequences (in the sense of Theorem 3 and Corollary 1), when the sequence is typical of a Markov source of order k, the proposed simulator is superior in that, as $k(x^n)$ converges to k, it eventually draws from a larger set (by Lemma 4, it essentially draws from $T_k(x^n)$). In contrast to the hierarchical universality demonstrated in the stochastic setting, though, the converses derived from Theorem 4 are unable to capture such nuances, due to their asymptotic nature. In fact, second order improvements derived from model size are negligible compared to the loss in conditional entropy caused by the fact that these schemes draw from a single type class: indeed, the faithfulness requirement of Property P1 does not imply that y^n be of the same type as x^n, as is (essentially) the case

in the stochastic setting (one could, as in [14], draw from neighboring classes).

For this reason, while it is possible to further refine the type selection by use of tree models [22], as studied in [33], the potential advantages are not captured by any converse. Moreover, the advantage of tree models in terms of type size is an open problem due to the lack of a result analogous to Lemma 4. We conjecture that such a result indeed holds. On the other hand, it is shown in [33] that tree models offer an additional advantage in terms of Property P1 if we measure the difference between the empirical probabilities with the L_∞, rather than the L_1 distance. Indeed, for Markov models, a j-tuple may exist for which this difference is still $O(\sqrt{[f(n)+1/n^2]\alpha^j})$. Instead, the finer discrimination of tree models, which can grow unbalanced to fit the data faster with different Markov orders for different contexts, guarantees that this difference is $O(\sqrt{f(n)}j^{3/2})$ for *every* j-tuple. An even slower growth of the L_∞ distance, linear in j, has been shown for the LZ parsing-based types [12] (as noted, the convergence in n is slower).

An alternative converse, which holds for *every* sequence x^n (rather than for a volume-one set) and which provides a rate of convergence, can be proved for the proposed scheme. The competing set of simulators for which such a converse applies, however, is weaker. The idea is to modify the definition of weak faithfulness so that, with input x^n and $i=k(x^n)$, a weakly faithful simulator is constrained to output sequences y^n such that $\hat{H}_i(y^n) < \hat{H}_i(x^n)+\gamma(n)$, where $\gamma(n)$ is some vanishing function of n. While the condition is now required only for model order $k(x^n)$ (and is trivially satisfied by the proposed scheme), the fact that, in general, $k(x^n)$ depends on x^n, is akin to requiring uniformity in the vanishing rate of the family $\{\gamma_i(n)\}_{i\geq 0}$ in the original condition. Applying Equation (27) for $i=k(x^n)$ and $\gamma_i(n)=\gamma(n)$, and upper-bounding $\hat{H}_i(x^n)$ and $N_{n,i}$ using Lemma 1, parts (a) and (c), we obtain

$$\begin{aligned} H(Y^n|x^n) &\leq \log|T_i(x^n)| + 2\alpha^{i+1}\log(n+1) + n\gamma(n) \\ &+ \varsigma(n)[-\log\varsigma(n) + n\log\alpha] \\ &\leq \log|\mathcal{M}(x^n)| + 2\alpha^{i+1}\log(n+1) + n\gamma(n) \\ &- \log[1-(n+1)^{\alpha^{i+1}}\mathsf{P}_{\mathrm{o/e}}^{(i)}(n)] \\ &+ \varsigma(n)[-\log\varsigma(n) + n\log\alpha] \end{aligned}$$

where the second inequality follows from Lemma 4. Thus, by Lemma 3, parts (a) and (d),

$$H(Y^n|x^n) \leq \log|\mathcal{M}(x^n)| + O\Big(\frac{\log n}{f(n)} + n\gamma(n) + n\varsigma(n)\Big) \tag{28}$$

for *all* sequences x^n. Equation (28) unveils the trade-off in the choice of $f(n)$: A larger $f(n)$ implies a slower convergence of the statistics (Theorem 3), but on the other hand it allows a smaller deviation from the performance of a competing simulator. For a Markov source of order k, $k(x^n)$ converges to k almost surely, and therefore, for typical sequences, the existence of the function $\gamma(n)$ is not a stronger requirement than the existence of the family $\{\gamma_i(n)\}_{i\geq 0}$.

The above advantages of the type classes $\mathcal{M}(\cdot)$ over those based on LZ parsing have a complexity cost. Indeed, even if the draw from $\mathcal{M}(x^n)$ is implemented by drawing uniformly at random from $T_{k(x^n)}(x^n)$ until a sequence that estimates order $k(x^n)$ is picked, enumeration of $T_{k(x^n)}(x^n)$ is more cumbersome than enumeration of the LZ parsing-based type class. The reason is linked to the cofactor in Whittle's formula [28], which reflects the fact that enumeration of Markov types does not reduce to independent enumerations of the state sub-sequences for memoryless types. In contrast, the scheme in [12] reduces to independent sub-sequences of draws at each node of the LZ tree.

Acknowledgment

We would like to thank Ronny Roth for useful discussions on the linear independence of type probabilities as functions of the parameter.

References

[1] R. M. Gray, "Time-invariant trellis encoding of ergodic discrete-time sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 23, pp. 71–83, Jan. 1977.

[2] 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.

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

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

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

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

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

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

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

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

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

[13] N. Merhav, G. Seroussi, and M. J. Weinberger, "Universal delay-limited simulation," *IEEE Trans. Inform. Theory*, vol. 54, pp. 5525–5533, Dec. 2008.

[14] N. Merhav and M. J. Weinberger, "On universal simulation with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 55, pp. 292–302, Jan. 2009.

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

[16] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. 30, pp. 629–636, July 1984.

[17] B. Ryabko, "Twice-universal coding," *Problems of Information Transmission*, vol. 20, pp. 173–177, July/September 1984.

[18] M. Feder and N. Merhav, "Hierarchical universal coding," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1354–1364, Sept. 1996.

[19] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1014–1019, Sept. 1989.

[20] M. J. Weinberger, A. Lempel, and J. Ziv, "A sequential algorithm for the universal coding of finite-memory sources," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1002–1014, May 1992.
[21] L. Finesso, "Estimation of the order of a finite Markov chain," in *Recent Advances in the Mathematical Theory of Systems, Control, and Network Signals, Proc. MTNS-91* (K. Kimura and S. Kodama, eds.), Mita Press, 1992, pp. 643–645.
[22] M. J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Inform. Theory*, vol. 41, pp. 643–652, May 1995.
[23] I. Csiszár and P. C. Shields, "The consistency of the BIC Markov order estimator," *Ann. Statist.*, vol. 28, pp. 1601–1619, 2000.
[24] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. 24, pp. 530–536, Sept. 1978.
[25] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[26] J. Aczél and J. Dhombres, *Functional Equations in Several Variables* (series: Encyclopedia of Mathematics and its Applications). Cambridge, New York: Cambridge University Press, 2008.
[27] A. Martín, G. Seroussi, M. J. Weinberger, "Linear Time Universal Coding and Time Reversal of Tree Sources via FSM Closure," *IEEE Trans. Inform. Theory*, vol. 50, pp. 1442–1468, July 2004.
[28] P. Whittle, "Some distributions and moment formulae for the Markov chain," *J. Roy. Stat. Soc.*, Ser. B, 17, pp. 235–242, 1955.
[29] L. B. Boza, "Asymptotically Optimal Tests for Finite Markov Chains," *Ann. Math. Statist.*, vol. 42, pp. 1992–2007, 1971.
[30] A. B. Barron, *Logically Smooth Density Estimation*. PhD thesis, Stanford University, Stanford, CA, 1985.
[31] J. C. Kieffer, "Sample Converses in Source Coding Theory," *IEEE Trans. Inform. Theory*, vol. 37, pp. 263–268, March 1991.
[32] P. H. Algoet and T. M. Cover, "A sandwich proof of the Shannon-McMillan-Breiman Theorem," *The Annals of Probability*, vol. 16, pp. 899–909, 1988.
[33] A. Martín, *Tree Models: Algorithms and Information-Theoretic Properties*. PhD thesis, Universidad de la República, Montevideo, Uruguay, 2009.