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 32000, Israel
Email: merhav@ee.technion.ac.il
‡Universidad de la República, Montevideo, Uruguay
Email: gadiel@msri.org
§Hewlett-Packard Laboratories, Palo Alto, CA 94304, 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

The problem of simulating random processes with a prescribed probability law has been extensively investigated, see, e.g., [1], [2], [3], [4], [5], [6], [7]. In all these works, perfect knowledge of the desired probability law is assumed. More recently, universal versions of this problem were studied in [8], [9], [10], and [11]. In [8], [10], 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 [11], 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 [8] and [11].

Specifically, the goal in [8], [10] 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 $P \in \mathcal{P}$ (so as to make the generated sample path as "original" as possible).

In [8], the smallest achievable value of the mutual information as a function of n, 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 [8] 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* [12] 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 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 [13], 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* [14]: 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* schemes. In this paper, we start by addressing the problem of double universality in the simulation setting of [8] when $\mathcal{P}$ is a class of Markov models of unknown (fixed) order. Extensions to the more general tree models [15] are under investigation.

First, we notice that a relaxation of Condition C1 is needed,

* 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.
‡ G. Seroussi is also with visiting Hewlett–Packard Laboratories, Palo Alto, CA, U.S.A.

for otherwise the condition that x^n and y^n be of the same type for *every* Markov order would imply that the two sequences must coincide, leading to a single, trivial simulator.[1] As it turns out, it suffices to allow simulators such that, for each $P \in \mathcal{P}$, Condition C1 is violated only by a fraction of sequences whose total mass (under the simulated probability, or equivalently, under P) is at most a vanishing function $\epsilon(n)$. In fact, a simulator exists such that $\epsilon(n)$ decreases exponentially fast, while achieving per-symbol mutual information which decays essentially as $\frac{m}{2}\frac{\log n}{n}$ for any Markov class $\mathcal{P}$ and any $P \in \mathcal{P}$, where m is the number of parameters corresponding to $\mathcal{P}$. This simulator follows a "plug-in" approach:

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

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., [17], [18], [19], [15], and [20]) both the mass of those sequences violating Condition C1, and the deviation from optimal conditional entropy, can be made negligible.

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 [11], which also extends the conventional notion of type. In the partition of [11], two sequences belong to the same class if and only if their Lempel-Ziv (LZ) parsing [21] 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 [11] 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 [11], 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 [11] 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 [11] by showing 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. Moreover, the class of competing simulators for the converse result turns out to be surprisingly broad. Notice that a "slow" rate of convergence is typical of other applications of the LZ parsing. On the other hand, the improvement has a complexity cost, which we discuss.

In the remainder of this extended abstract, Section II introduces the main concepts and tools. Our results in the stochastic setting are then presented in Section III. In Section IV we study the individual sequence setting.

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 $\mathcal{A}$.

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

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

where we arbitrarily assume $x_0, x_{-1}, \ldots, x_{-k+1}$ to be equal to a fixed symbol in $\mathcal{A}$. The family of Markov sources of order k over $\mathcal{A}$ is denoted $\mathcal{P}_k$. The entropy of n-tuples emitted by P is denoted $H(X^n)$.

The k-th order Markov *type class* [12] $T_k(x^n)$ 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 $P \in \mathcal{P}_k$. 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 [12], [22], 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

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

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 $\hat{P}_T^{(k)}$. Its conditional entropy $\hat{H}_k(x^n)$, namely the k-th order empirical con-

[1] The relaxation of Condition C1 was precisely the motivation for the individual sequence setting of [11]. Relaxation in the stochastic sense discussed here is also discussed in [10] and [16], where universal simulation with a fidelity criterion is studied, in analogy with the (non-universal) scenario of [4].

ditional entropy for x^n, satisfies $n\hat{H}_k(x^n) = -\log \hat{P}_T^{(k)}(x^n)$, where, throughout, logarithms are taken in base 2.

All simulation schemes considered in this paper are assumed to have access to an unlimited budget of random bits, and output sequences y^n of the same length n as the training sequence x^n. 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 $P \in \mathcal{P}_k$. As discussed in Section I, by [8], the optimal scheme that knows k draws y^n uniformly at random from $T_k(x^n)$, with mutual information

$$I(X^n;Y^n) = H(X^n) - \boldsymbol{E} \log |T_k(X^n)| \approx |\mathcal{A}|^k(|\mathcal{A}|-1)\frac{\log n}{2}$$

where the expectation is with respect to P, and the approximation is to the main asymptotic term. Thus, the deviation from the optimal mutual information must be $o(\log n)$ in order to leave the asymptotic behavior unaffected.

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 1 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

$$k(x^n) = \arg\min_{k\geq 0}\{\hat{H}_k(x^n) + |\mathcal{A}|^k f(n)\} \tag{1}$$

where $f(n)$ is a vanishing function of n, and ties are resolved with some fixed policy. For example, $f(n) = (|\mathcal{A}| - 1)(\log n)/(2n)$ corresponds to the asymptotic version of the MDL criterion [13]. 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. To state Lemma 1 we define, for a distribution $P \in \mathcal{P}_k$, the overestimation probability

$$P_{\rm over}(n) \stackrel{\triangle}{=} \Pr(k(X^n) > k)$$

and, similarly, the underestimation probability

$$P_{\rm under}(n) \stackrel{\triangle}{=} \Pr(k(X^n) < k)\,.$$

Lemma 1: For any $k \geq 0$ and any $P \in \mathcal{P}_k$, the estimator of Equation (1) satisfies

(a) $N_{n,k}P_{\rm over}(n)$ vanishes polynomially fast (uniformly in P and k) provided $f(n) > \beta(\log n)/n$ for a sufficiently large constant β.

(b) $P_{\rm under}(n)$ 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) $|\mathcal{A}|^{k(x^n)} = O(1/f(n))$ for any $x^n \in \mathcal{A}^n$.

The proof of parts (a) and (b) is omitted in this extended abstract, as similar results have been shown for variants of this estimator. Part (a) implies, *a fortiori*, a similar convergence for the overestimation probability; this case is handled with the method of types as in [15], with the additional factor $N_{n,k}$ requiring only a larger value of β. Underestimation, on the other hand, is a large deviations event. 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 the penalized maximum-likelihood for model order $k(x^n)$ is not larger than the one for model order 0, which is clearly $O(1)$. The estimate $k(x^n)$ can be obtained in time that is linear in n by use of suffix trees as in [23]. The set of n-tuples x^n such that $k(x^n) = i$ will be denoted $\mathcal{A}_{n,i}$.

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 number of sequences in $\mathcal{M}(x^n)$ is essentially $|T_{k(x^n)}(x^n)|$. By Part (c) of Lemma 1, the remaining sequences are in $\{\mathcal{A}_{n,i}\}_{i>k(x^n)}$. To state the lemma, we define $P_{\rm over}^{(i)}(n)$ as the maximum value of $P_{\rm over}(n)$ over all distributions $P \in \mathcal{P}_i$ such that P is the empirical distribution of a Markov type class of order i and length n. Notice that $P_{\rm over}^{(i)}(n)$ is independent of any probabilistic assumption and, by Part (a) of Lemma 1, it is upper-bounded by a function that decays polynomially fast with n, uniformly in i.

Lemma 2: For any $i \geq 0$, let $T \in \mathcal{T}_i^n$ and assume $T \cap \mathcal{A}_{n,i} \neq \phi$. Then,

$$\frac{|T \cap \mathcal{A}_{n,i}|}{|T|} \geq 1 - N_{n,i}P_{\rm over}^{(i)}(n)\,.$$

Sketch of proof. By Whittle's formula for the size of a type class [24], and lower-bounding the cofactor in the formula as in [25], it is easy to see that $\hat{P}_T^{(i)}(T) \geq 1/N_{n,i}$. Since $\hat{P}_T^{(i)}(\cdot)$ is uniform over T, we then have

$$\frac{1}{N_{n,i}} \leq \hat{P}_T^{(i)}(T \cap \bar{\mathcal{A}}_{n,i})\,\frac{|T|}{|T \cap \bar{\mathcal{A}}_{n,i}|}$$

where the complement of a set $\mathcal{S}$ is denoted $\bar{\mathcal{S}}$. By Lemma 1, Part (c), since $T \cap \mathcal{A}_{n,i}$ is nonempty, we have $k(z^n) \geq i$ for all $z^n \in T$, implying $\hat{P}_T^{(i)}(T \cap \bar{\mathcal{A}}_{n,i}) \leq P_{\rm over}^{(i)}(n)$. □

Lemma 2 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.

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 $\mathcal{M}(x^n)$. Let $Q(\cdot)$ denote the output distribution. Since $y^n \in \mathcal{M}(x^n)$ if and only if $x^n \in \mathcal{M}(y^n)$, we have

$$Q(y^n) = \sum_{x^n \in \mathcal{A}^n} P(x^n) W(y^n|x^n) = \frac{P(\mathcal{M}(y^n))}{|\mathcal{M}(y^n)|} . \quad (2)$$

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

(a) The output distribution satisfies

$$Q(Q(y^n) \neq P(y^n)) \leq P_{\text{under}}(n) .$$

(b) The conditional entropy of the simulator satisfies

$$\begin{aligned} H(Y^n|X^n) &\geq \boldsymbol{E} \log |T_k(X^n)| - nP_{\text{over}} \\ &+ \min_{i \leq k} \log(1 - N_{n,i} P^{(i)}_{\text{over}}(n)) \end{aligned}$$

and $H(Y^n) \leq H(X^n) + P_{\text{under}}(n)[n - \log P_{\text{under}}(n)]$.

By Lemma 1, Part (a) states that the simulator preserves the probability law, except for an exponentially negligible fraction of the output sequences, whereas Part (b) states that, with proper choice of $f(n)$, the conditional entropy deviates from the optimal one (obtained with knowledge of k) by an amount that does not affect the asymptotic behavior of $\boldsymbol{E} \log |T_k(X^n)|$. Part (b) also states that $H(Y^n)$ cannot surpass $H(X^n)$ by more than a negligible amount, implying that no knowledge of k is necessary to achieve the optimal rate of decay of $I(X^n;Y^n)/n$. Thus, the proposed scheme is twice universal. The cost of double universality is a deviation in the probability law for a fraction of the sequences. In principle, it is conceivable that such a deviation, if allowed for a scheme that knows k, would lead to a faster decay of the per-symbol mutual information. We conjecture that this is not the case.

We also observe 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 1, the asymptotic behavior is independent of $f(n)$.

Sketch of proof of Theorem 1. To prove Part (a), notice that if $k(y^n) \geq k$, then $P(y^n) = P(x^n)$ for all $x^n \in \mathcal{M}(y^n)$. Thus, by (2), $Q(y^n) = P(y^n)$. Furthermore, for all $y^n \in \mathcal{A}^n$,

$$\begin{aligned} Q(Q(y^n) \neq P(y^n)) &= 1 - Q(Q(y^n) = P(y^n)) \\ &= 1 - P(Q(y^n) = P(y^n)) \\ &= P(Q(y^n) \neq P(y^n)) \leq P_{\text{under}}(n) . \end{aligned}$$

As for Part (b), the lower bound on the conditional entropy follows from application of Lemma 2 to

$$H(Y^n|X^n) = \sum_{i \geq 0} \sum_{x^n \in \mathcal{A}_{n,i}} P(x^n) \log |T_i(x^n) \cap \mathcal{A}_{n,i}|$$

and the fact that $|T_i(x^n)| \geq |T_k(x^n)|$ for all $i \leq k$. The upper bound on $H(Y^n)$ follows from splitting the summation defining the entropy into two partial summations: one for $k(y^n) \geq k$, for which $Q(y^n) = P(y^n)$, and one for the rest, to which we apply Jensen's inequality. □

From a complexity standpoint, enumeration of the intersection of the type class of x^n for the estimated order with $\mathcal{A}_{n,k(x^n)}$ 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 is drawn. By Lemma 2, with very high probability only one draw will be needed. Another approach consists of modifying the simulation scheme to draw from the type class, instead of drawing from the intersection. Clearly, the conditional entropy can only improve, as we draw from a larger set, and the asymptotic mutual information remains unaffected. On the other hand, we can no longer claim the scheme to preserve the probability law in the strong sense of Part (a) of Theorem 1. However, it can be shown that, with appropriate choice of $f(n)$, for all but a vanishing fraction of the sequences y^n, the ratio $Q(y^n)/P(y^n)$ deviates from 1 by a vanishing quantity. This scheme does not lead to a partition of $\mathcal{A}^n$.

IV. Individual sequences

In this section we analyze the proposed simulation scheme in the individual sequence setting of [11]. Theorem 2 below establishes the statistical similarity between two sequences in the same class $\mathcal{M}(\cdot)$ in the implied partition of $\mathcal{A}^n$.

Theorem 2 (Direct): Let $x^n \in \mathcal{A}^n$ be arbitrary and fix a nonnegative integer j. Let s be an arbitrary string in $\mathcal{A}^j$. Then, for any $y^n \in \mathcal{M}(x^n)$, we have

$$\left| \frac{n_{x^n}(s)}{n} - \frac{n_{y^n}(s)}{n} \right| = O(\sqrt{f(n)}) .$$

Moreover, $n_{x^n}(s) = n_{y^n}(s)$ if $j \leq k(x^n) + 1$.

If $j \leq k(x^n) + 1$ then $n_{x^n}(s) = n_{y^n}(s)$ by the definition of the type classes. For larger values of j, the proof of Theorem 2, which is omitted in this extended abstract due to space limitations, relies on the fact that the occurrence counts for symbols following string s in x^n are "close" to those corresponding to symbols following the suffix of s of length $k(x^n)$, for otherwise the order estimate would have been larger than $k(x^n)$. The same observation applies to y^n, and since the counts for x^n and y^n are identical at order $k(x^n) = k(y^n)$, the result follows by transitivity. Since the criterion for model order selection relies on empirical entropy, an application of Pinsker's inequality is necessary, which explains the $O(\sqrt{f(n)})$ rate.

Theorem 2 corresponds to Property (P1) that was itemized in Section I for the scheme in [11]. 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 [11, 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 $O(\sqrt{f(n)})$, 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 $\Pi(x^n) \neq \Pi(y^n)$ has measure at most $P_{\text{under}}(n)$ 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, given $x^n \in \mathcal{A}_{n,i}$, let a *weakly faithful* simulator $W(Y^n|x^n)$ be only 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. Notice that the condition is required only for model order i, and does not necessarily imply closeness in terms of counts (on the other hand, a scheme that approximately preserves counts will obviously be faithful in this relaxed sense for some function $\gamma(n)$). 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 $W(\mathcal{B}(x^n)|x^n) < \delta(n)$ for some vanishing function $\delta(n)$. The following theorem asserts that no other weakly faithful simulator can achieve a much larger value of the conditional entropy than the proposed one.

Theorem 3 (Converse): Let $W(Y^n|x^n)$ be a weakly faithful simulator, and assume $\log n = o(nf(n))$. Then,

$$H(Y^n|x^n) \leq \log|\mathcal{M}(x^n)| + O\left(\frac{\log n}{f(n)} + n\gamma(n) + n\delta(n)\right).$$

Sketch of proof. For brevity, and to avoid obscuring the main ideas, we assume $\delta(n) = 0$. Let $k(x^n) = i$ and let $N_{n,i}^{(\gamma)}(x^n)$ denote the number of sequences y^n satisfying $\hat{H}_i(y^n) < \hat{H}_i(x^n) + \gamma(n)$ (namely, the potential output sequences). Using classical tools from the method of types, we have

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

Therefore,

$$H(Y^n|x^n) \leq \log N_{n,i}^{(\gamma)}(x^n) \leq n\hat{H}_i(x^n) + n\gamma(n) + \log N_{n,i}.$$

Proceeding as in the proof of Lemma 2 we have $|T_i(x^n)| \geq 2^{n\hat{H}_i(x^n)}/N_{n,i}$. By Lemma 2 we conclude that

$$\begin{aligned} H(Y^n|x^n) &\leq \log|\mathcal{M}(x^n)| + 2\log N_{n,i} + n\gamma(n) \\ &- \log[1 - N_{n,i}P_{\text{over}}^{(i)}(n)]. \end{aligned}$$

Since $\log N_{n,i} \leq (|\mathcal{A}| - 1)|\mathcal{A}|^i \log n$, the result follows from Lemma 1, parts (a) and (d). □

Theorems 2 and 3 unveil the trade-off in the choice of $f(n)$: A larger $f(n)$ implies a slower convergence of the statistics (Theorem 2), but on the other hand it allows a smaller deviation from the performance of a competing simulator (Theorem 3). Unlike the converse in [11], Theorem 3 holds for *every* sequence x^n, and a rate of convergence is provided. The advantages of the type classes $\mathcal{M}(\cdot)$ over those based on LZ parsing, however, 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 [24], 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 elegant parsing process in [11] reduces to independent subsequences of draws at each node of the LZ tree.

REFERENCES

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

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

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

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

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

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

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

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

[9] N. Merhav, "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.

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

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

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

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

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

[15] M. J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Inform. Theory*, vol. 41, pp. 643–652, May 1995.

[16] N. Merhav and M. J. Weinberger, "On universal simulation with a fidelity criterion," Proceedings of the 2006 IEEE International Symposium on Information Theory (ISIT'06), Seattle, WA, July 2006.

[17] 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. IT-35, pp. 1014–1019, Sept. 1989.

[18] M. J. Weinberger, A. Lempel, and J. Ziv, "A sequential algorithm for the universal coding of finite-memory sources," *IEEE Trans. Inform. Theory*, vol. IT-38, pp. 1002–1014, May 1992.

[19] 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.

[20] I. Csiszár and P. C. Shields, "The consistency of the BIC Markov order estimator," *Ann. Statist.*, vol. 28, pp. 1601–1619, 2000.

[21] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.

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

[23] 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.

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

[25] L. B. Boza, "Asymptotically Optimal Tests for Finite Markov Chains," *Ann. Math. Statist.*, vol. 42, pp. 1992–2007, 1971.