CCIT REPORT #746 September 2009

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.