CCIT Report #636 August 2007

Universal Simulation with Fidelity Criteria*

Neri Merhav†
Department of Electrical Engineering
Technion – I.I.T., Haifa 32000, Israel
Email: merhav@ee.technion.ac.il

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

***Abstract*— We consider the problem of universal simulation of a memoryless source (with some partial extensions to Markov sources), based on a training sequence emitted from the source. The objective is to maximize the conditional entropy of the simulated sequence given the training sequence, subject to a certain distance constraint between the probability distribution of the output sequence and the probability distribution of the input, training sequence. We derive, for several distance criteria, single–letter expressions for the maximum attainable conditional entropy as well as corresponding universal simulation schemes that asymptotically attain these maxima.**

Index Terms: Universal simulation, distance measures, generalized divergence, $\bar{\rho}$–distance, ϵ–contaminated model.

I. Introduction

Simulation of a source means artificial production of random data with some probability law, by using a certain device that is fed by a sequence of purely random bits. Simulation of sources and channels is a problem that has been studied in a series of works, see, e.g., [7], [15], [16], [17] and references therein. In all these works, it was assumed that the probability law of the desired process is perfectly known.

More recently, a universal version of this problem was studied in [12], [13] (see also [10]), where the assumption of perfect knowledge of the target probability law was relaxed. Instead, the target source P to be simulated was assumed in [12] to belong to a certain parametric family $\mathcal{P}$, but is otherwise unknown, and a training sequence $X^m = (X_1, \ldots, X_m)$, that has emerged from this source, is available. In addition, the simulator is provided with a sequence of ℓ random key bits $U^\ell = (U_1, \ldots, U_\ell)$, which is independent of X^m. The goal of the simulation scheme in [12] was to generate an output sequence $Y^n = (Y_1, \ldots, Y_n)$, $n \leq m$, corresponding to the simulated process, such that $Y^n = \psi(X^m, U^\ell)$, where ψ is a deterministic function that does not depend on the unknown source P, and which satisfies the following two conditions: (i) the probability distribution of Y^n is *exactly* the n-dimensional marginal of the probability law P corresponding to X^m for all $P \in \mathcal{P}$, and (ii) the mutual information $I(X^m; Y^n)$ is as small as possible, or equivalently (under (i)), the conditional entropy $H(Y^n|X^m)$ is as large as possible, simultaneously for all $P \in \mathcal{P}$ (so as to make the generated sample path Y^n as "original" as possible). In [12], the smallest achievable value of the mutual information (or, the largest conditional entropy) was characterized, and simulation schemes that asymptotically achieve these bounds were presented (see also [13]). It turns out that for these optimal schemes, for ℓ large enough, the normalized mutual information asymptotically vanishes. In [11], the same simulation problem was studied in the regime of a delay–limited system, in which the simulator produces output samples on–line, as the training data is fed into the system sequentially. The cost of limited delay was characterized and a strictly optimum simulation system was proposed. A different perspective on universal simulation was investigated in [14], where x^m was assumed to be an individual sequence not originating from any probabilistic source.

In this work, we extend the scope of the universal simulation problem in another direction, namely, relaxing the requirement of *exact* preservation of the probability law at the output of the simulator. In particular, we study the best achievable tradeoff between the performance of the simulation scheme and the distance (measured in terms of a certain metric) between the probability law of the output and that of the input. Observe that when the probability law of the simulated sequence is not constrained to be identical to that of the training sequence, the criteria $\min I(X^m; Y^n)$ and $\max H(Y^n|X^m)$ are no longer equivalent. While the former criterion aims at weak dependency, it should be emphasized that, for a large enough key rate, *vanishing* normalized mutual information was shown to be achievable with exact preservation of the probability law [12]. Therefore, under the $\min I(X^m; Y^n)$ criterion, the main objective of a relaxation of this requirement is to save on the key rate necessary for the normalized mutual information to vanish, as studied in [13] in the context of the $\bar{\rho}$–distance between probability distributions.[1] On the other hand, the asymptotic performance as given by the $\max H(Y^n|X^m)$ criterion (as a measure of the "originality" or the "diversity" of the typical sample paths generated by the simulator), on which we focus in this paper, can potentially benefit from the proposed relaxation.

For the class of discrete memoryless sources (DMSs), we derive single-letter formulas for the maximum achievable conditional entropy subject to various distance constraints

* The material in this paper was presented in part at the 2006 IEEE International Symposium on Information Theory, Seattle, WA, USA, July 2006.
† This work was done while N. Merhav was visiting Hewlett–Packard Laboratories, Palo Alto, CA, U.S.A.

[1] In addition, it is conceivable that, by deviating from the input probability law, a faster vanishing rate for the normalized mutual information is possible. However, this aspect of the problem is not discussed in [13].