Addendum to "On Universal Simulation of Information Sources Using Training Data"

Neri Merhav[*] Marcelo J. Weinberger[†]

May 17, 2005

Abstract

In a recent paper[1] we studied the problem of universal simulation of an unknown information source of a certain parametric family, given a training sequence from that source and given a limited budget of purely random bits. The goal was to generate another random sequence (of the same length or shorter), whose probability law is identical to that of the given training sequence, but with minimum statistical dependency (minimum mutual information) between the input training sequence and the output sequence. In this addendum, we point out to a concrete optimal simulation scheme that is easy to implement, as opposed to the non–constructive existence result in that paper, and we make a number of additional observations on the universal simulation problem.

Index Terms: Random number generators, enumeration, random process simulation, mutual information, typical sequences.

A recent paper,[1] studied the following universal simulation problem: An unknown source P, which is assumed to belong to a certain parametric family $\mathcal{P}$ (like the family of finite–alphabet memoryless sources, Markov sources, finite–state sources, parametric subsets of these families, etc.), is to be simulated. We are given a training sequence $X^m = (X_1, \ldots, X_m)$ that has emerged from this unknown source, as well as a string of k purely random bits $U^k = (U_1, \ldots, U_k)$, that are independent of X^m, and our goal is to generate an output sequence $Y^n = (Y_1, \ldots, Y_n)$, $n \leq m$, corresponding to the simulated process, that satisfies the following three conditions:

C1. The mechanism by which Y^n is generated can be represented by a deterministic function $Y^n = \phi(X^m, U^k)$, where ϕ does not depend on the unknown source P.

[*] N. Merhav is with the Electrical Engineering Department, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel. E-mail address: [merhav@ee.technion.ac.il]. This work was done while N. Merhav was visiting Hewlett–Packard Laboratories, Palo Alto, CA, U.S.A.

[†] M. J. Weinberger is with Hewlett–Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, U.S.A. E-mail address: [marcelo@hpl.hp.com].

[1] N. Merhav and M. J. Weinberger, *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 5–20, January 2004.

C2. The probability distribution of Y^n is *exactly* the n-dimensional marginal of the probability law P corresponding to X^m for all $P \in \mathcal{P}$.

C3. The mutual information $I(X^m; Y^n)$ is as small as possible, simultaneously for all $P \in \mathcal{P}$.

In Subsection 4.2 of the referenced paper, we referred to the case where $n < m$ and the key rate, $R \stackrel{\triangle}{=} k/n$, is finite. Unlike the other cases, for which we were able to demonstrate concrete simulation schemes that satisfy all three conditions, C1–C3, in this case, we only presented a non–constructive existence result in a very large ensemble of schemes (Theorem 3, therein).

The primary purpose of this addendum is to suggest a simple simulation scheme that satisfies the above conditions in the case where $n < m$ as well. In the sequel, lower–case notation such as x^m, y^n, and u^k, will denote specific realizations of the random vectors X^m, Y^n, and U^k, respectively. For a given training sequence x^m, let

$$\phi(x^m, u^k) = \left[J_m^{-1}(J_m(x^m) \oplus \lceil f(u^k) \cdot |T_{x^m}|/2^k \rceil\right]_1^n \tag{1}$$

and

$$\phi'(x^m, u^k) = J_n^{-1}\left(J_n(x^n) + f(u^k) \bmod |T_{x^n}|\right) \tag{2}$$

where J_t (t – positive integer) maps a sequence x^t to the *lexicographic* index within its type class T_{x^t}, $|T_{x^m}|$ is the cardinality of T_{x^m}, f maps u^k to a corresponding integer in $\{0, 1, \ldots, 2^k - 1\}$, $\oplus$ denotes addition modulo $|T_{x^m}|$, J_t^{-1} is the inverse of J_t, and $[\cdot]_1^n$ denotes truncation of an m–sequence to an n–sequence, i.e., elimination of the last $r \stackrel{\triangle}{=} n - m$ symbols. In Eq. (1), $\phi(x^m, u^k)$ is given by the first n symbols of a sequence in T_{x^m}. The key idea is that, as u^k exhausts $\{0, 1\}^k$, the lexicographic indexes of the sequences in T_{x^m} from which the $\phi(x^m, u^k)$ are obtained by the truncation $[\cdot]_1^n$ are uniformly spaced when $|T_{x^m}| \geq 2^k$, with a distance of $|T_{x^m}|/2^k$ between every two consecutive candidate indexes (up to integer truncation errors). If, instead, $|T_{x^m}| < 2^k$, each index is selected $2^k/|T_{x^m}|$ times (again, up to integer truncation errors). In Eq. (2), x^m is first truncated to n bits, and only then a sequence in T_{x^n} is randomly selected. Now, define

$$y^n = \phi^*(x^m, u^k) \stackrel{\triangle}{=} \begin{cases} \phi(x^m, u^k) & \log |T_{x^m}| < mR \\ \phi'(x^m, u^k) & \log |T_{x^m}| \geq mR. \end{cases} \tag{3}$$

As in the paper,[1] the idea is that when R is larger than the entropy rate H of the source, y^n is likely to take the value $\phi(x^m, u^k)$, whereas otherwise it is likely to take the value

$\phi'(x^m,u^k)$. Notice, however, that the test to determine the relation between H and R is based here on x^m, and not just on x^n as in the paper.[1]

The simulation scheme ϕ^* obviously satisfies Condition C1. To see that Condition C2 is also satisfied, notice that with $\mathcal{E} \overset{\triangle}{=} \{x^m : \log|T_{x^m}| \geq mR\}$, we have

$$\Pr\{Y^n = y^n|u^k\} = \sum_{x^m\in\mathcal{E}^c:\phi(x^m,u^k)=y^n} P(x^m) + \sum_{x^m\in\mathcal{E}:\phi'(x^m,u^k)=y^n} P(x^m)\,. \tag{4}$$

For a given u^k, each sequence of the form $y^n z^r$ in $\mathcal{E}^c$ is obtained with ϕ (before truncating z^r) from exactly one sequence $x^m \in \mathcal{E}^c$. In addition, for each sequence y^n there exists exactly one sequence x^n (of the same type) such that $\phi'(x^n z^r, u^k) = y^n$ for any z^r. Therefore, (4) can be written as

$$\Pr\{Y^n = y^n|u^k\} = \sum_{z^r:y^n z^r\in\mathcal{E}^c} P(y^n z^r) + \sum_{z^r:y^n z^r\in\mathcal{E}} P(y^n z^r) = P(y^n)$$

as claimed.

We prove that ϕ^* also satisfies Condition C3. For given sequences x^m and y^n, let $T_{x^m\backslash y^n} = \{z^r : \ y^n z^r \in T_{x^m}\}$. Notice that for all $z^r \in T_{x^m\backslash y^n}$ we have $T_{z^r} = T_{x^m\backslash y^n}$ (the set $T_{x^m\backslash y^n}$ can be viewed as the type of the "difference" between x^m and y^n). In the absence of key rate limitations, it is shown in the paper[1] that the mutual information is minimized when y^n is drawn according to the "channel" $W^*(y^n|x^m) \overset{\triangle}{=} |T_{x^m\backslash y^n}|/|T_{x^m}|$, which is the fraction of sequences in T_{x^m} that start with y^n. However, this "ideal" channel cannot be implemented in general with a limited supply of random bits. Let $W(y^n|x^m)$ denote the channel induced by ϕ^*, i.e.,

$$W(y^n|x^m) = 2^{-k}\sum_{u^k} 1\{y^n = \phi^*(x^m,u^k)\}. \tag{5}$$

For the case in which $R < H$, the analysis in the paper[1] (Eqs. (35) and (36)) remains valid, and the mutual information achieved by the scheme in this case still approaches its optimum value $n(H-R)$. This case is handled with high probability by the scheme ϕ' given in (2), since $\mathcal{E}^c$ is a large deviations event in this case, as shown in Appendix A of the paper.[1] To analyze the case $R > H$, according to the derivation of Eq. (34) in the paper,[1] any channel that satisfies the following requirement asymptotically achieves the minimum attainable mutual information whenever $R > H$: For a given (small) $\epsilon > 0$, let

$$\mathcal{S} \overset{\triangle}{=} \{(x^m,y^n) : \ k \geq \log|T_{x^m}| - \log|T_{x^m\backslash y^n}| + n\epsilon\}.$$

Then, for any $(x^m, y^n) \in \mathcal{S}$,

$$(1 - 2^{-n\epsilon})W^*(y^n|x^m) \leq W(y^n|x^m) \leq (1 + 2^{-n\epsilon})W^*(y^n|x^m). \tag{6}$$

While Lemma 1 in the paper[1] only guarantees the *existence* of such a channel, we now prove that the concrete scheme ϕ given in (1) indeed induces a channel that satisfies this condition. First, note that since J_m is defined by lexicographic ordering, all $|T_{x^m \backslash y^n}|$ members of T_{x^m} that are prefixed by the same y^n are enumerated consecutively. Thus, due to the (roughly) uniform spacing between the possible sequences from which $\phi(x^m, u^k)$ is obtained by truncation as u^k exhausts $\{0,1\}^k$, for $|T_{x^m}| \geq 2^k$, the number $N(x^m)$ of key sequences u^k for which $J_m(x^m) \oplus \lceil f(u^k)|T_{x^m}|/2^k \rceil$ is prefixed by y^n is (within ± 1) the ratio between $|T_{x^m \backslash y^n}|$ and the spacing $|T_{x^m}|/2^k$. For $|T_{x^m}| < 2^k$, in turn, every ℓ consecutive locations are selected $2^k \ell / |T_{x^m}|$ times as candidate output sequences (up to integer truncation errors), so that $N(x^m)$ is in this case $|T_{x^m \backslash y^n}|$ times $2^k/|T_{x^m}|$. Therefore, in any case, $N(x^m) = 2^k W^*(y^n|x^n) \pm 1$, and according to Eq. (5) we have $W(y^n|x^n) = W^*(y^n|x^m) \pm 2^{-k}$. But for $(x^m, y^n) \in \mathcal{S}$ we have $2^{-k} \leq 2^{-n\epsilon} W^*(y^n|x^m)$, and so Eq. (6) indeed holds. Consequently, the scheme ϕ^* satisfies Condition C3 as claimed.

The addition of $J_m(x^m)$ in (1) is aimed at preserving the probability law regardless of inaccuracies in the implementation of the optimal channel W^* due to key rate limitations. If exact implementation of W^* were possible, then it would not be necessary to introduce additional randomness through X^m in order to satisfy Condition C2 and, for a given key value, y^n would depend on x^m only through its type.

In an alternative setting, one can aim at an exact implementation of the channel with an Elias decoder [1, pp. 479–482], using ideas from [2], and upper-bound the *expected* number of random bits that are needed for the decoder to generate n output symbols, where the expectation is with respect to the random key. Specifically,

$$W^*(y^n|x^m) = \prod_{t=1}^{n} W^*(y_t|y^{t-1}, x^m)$$

where

$$W^*(y_t|y^{t-1}, x^m) \triangleq \frac{|T_{x^m \backslash y^t}|}{|T_{x^m \backslash y^{t-1}}|} \tag{7}$$

is clearly the fraction of sequences in $T_{x^m \backslash y^{t-1}}$ that start with y_t, and thus defines a probability distribution on the source alphabet. If $\mathcal{P}$ is the entire class of discrete memoryless sources, then $W^*(\cdot|y^{t-1}, x^m)$ is simply the empirical distribution defined by any sequence in $T_{x^m \backslash y^{t-1}}$. Notice that the sequence of distributions $\{W^*(\cdot|y^{t-1}, x^m)\}_{t=1}^m$ is precisely the one used by an enumerative decoder [3] that decodes a sequence y^m of a *given* type T_{x^m}. To output y_t, the Elias decoder is tuned to the distribution $W^*(\cdot|y^{t-1}, x^m)$, and uses the key as an input bitstream. As shown in [2], the expected number of key bits that the decoder consumes to (sequentially) produce y^n is upper-bounded by the entropy of the distribution $W^*(\cdot|x^m)$, plus 3 bits. The expected value of this upper bound (with respect to X^m and Y^n) is $\boldsymbol{E}\log|T_{x^m}| - \boldsymbol{E}\log|T_{x^m\backslash y^n}| + 3$, which is shown in the paper[1] to be $nH + O(1)$ under mild regularity assumptions. Note that the decoder may require an unbounded number of random input bits. However, the probability that the scheme will fail to produce a sequence y^n after processing k random bits, $k > nH$, decays with exponent $-(k - nH)$. Thus, with probability one it produces an output, although a hard limit on the key rate will in general affect the exact preservation of the probability law if the randomness in X^m is not utilized. An efficient implementation of this approach via arithmetic decoding, with a finite register length [4], will cause additional deviations. The *exact* enumeration in (1) and (2) addresses these issues, at a complexity cost.

This discussion partially applies also to another paper on universal simulation [5], where achievable rates where sought for maintaining probabilities of certain events corresponding to a given set of statistical tests. If there is only one event, or a relatively small number of events such that all members of each event (within the type class of x^m) can be ordered consecutively in the enumeration $J_m(x^m)$ (which may not necessarily be lexicographic), then the simulation algorithm proposed in (3) will give rise to good approximations of the corresponding probabilities. However, there may not be an efficient enumeration algorithm, in general, for an arbitrary set of events.

Our second comment is that this problem setting, in the case $n = m$, has an additional application other than universal simulation, and this is in cryptography: Here, X^n plays the role of the plaintext, U^k is the secret key, and Y^n is the cryptogram. Minimizing the mutual information $I(X^n; Y^n)$ corresponds to maximum equivocation $H(X^n|Y^n)$, which is the classical figure of merit of the cipher system. The fact that Y^n is of the same type class as X^n (which is a necessary condition to maintain Condition C2), corresponds to encryption based on a scrambling method.

Our third and last comment refers to a possible approach of relaxing Condition C2. In [6], the main results of the classical (non–universal) simulation, were extended to relax the requirement of vanishing distances between the probability distributions of the simulated process and the desired process: For a given non-vanishing bound on this distance (defined by several possible accuracy measures), the minimum rate of random bits required is given by the rate-distortion function of the desired process, where the fidelity criterion depends on the accuracy measure. Specifically, one possible distance measure between two distributions P and $\tilde{P}$ is Ornstein's $\bar{\rho}$ distance, which is the minimum expected distortion $(1/n)\sum_{i=1}^{n} E\rho(X_i, Y_i)$ among all joint distributions of (X^n, Y^n) with marginals P and $\tilde{P}$, respectively, where ρ is a given distortion measure. In [6], it is shown that under certain conditions, if the simulated process $\tilde{P}$ is only required to be within $\bar{\rho}$ distance D from the desired P, then the minimum key rate required is given by the rate–distortion function $R(D)$ of the source P with respect to ρ.

This finding has an analogue in the universal simulation problem, where now $R(D)$ (rather than H) becomes the critical key rate beyond which $I(X^m; Y^n)/n$ may vanish when Y^n is distributed according to P_Y such that $\bar{\rho}(P, P_Y) \leq D$. First, we claim that for $R < R(D)$ the normalized mutual information cannot be smaller than $R(D) - R$. This claim follows from the fact that $I(X^m; Y^n) = H(Y^n) - H(Y^n|X^m)$, where $H(Y^n|X^m)$ cannot exceed $k = nR$, and $H(Y^n)$ can be lower-bounded, following [6], by noticing that for any P_Y such that $\bar{\rho}(P, P_Y) \leq D$ (with X^m governed by P and Y^n governed by P_Y), there exists a distribution $W(X^m|Y^n)$ such that $(1/n)\sum_{i=1}^{n} E\rho(X_i, Y_i) \leq D$, so that the corresponding mutual information between X^m and Y^n is lower-bounded by $nR(D)$ and therefore so is $H(Y^n)$. It is also relatively easy to show convergence of the mutual information to zero when $R > R(D)$, as well as the above $R(D) - R$ behavior for $R < R(D)$, although it appears to be much harder now to characterize the exact convergence rate of the mutual information to zero for $R > R(D)$. For example, when $n = m$, adapting ideas from [6] to the universal setting, a universal simulation scheme could be based on a (universal) rate–distortion codebook that covers T_{x^n}. The (at least) $n(R(D) + \epsilon)$ random key bits would be used to select an index in the codebook according to the conditional probabilities of the quantization cells induced by the code. The output would be the reproduction vector corresponding to the selected index.

References

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

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

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

[4] J. Rissanen, "Generalized Kraft inequality and arithmetic coding," *IBM Jl. Res. Develop.*, vol. 20(3), pp. 198–203, May 1976.

[5] 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, no. 1, pp. 21–30, January 2004.

[6] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 63–86, January 1996.