Achievable Key Rates for Universal Simulation of Random Data with Respect to a Set of Statistical Tests

Neri Merhav*

October 9, 2002

Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
merhav@ee.technion.ac.il

Abstract

We consider the problem of universal simulation of an unknown source from a certain parametric family of discrete memoryless sources, given a training vector $\boldsymbol{X}$ from that source and given a limited budget of purely random key bits. The goal is to generate a sequence of random vectors $\{\boldsymbol{Y}_i\}$, all of the same dimension and the same probability law as the given training vector $\boldsymbol{X}$, such that a certain, prescribed set of M statistical tests will be satisfied. In particular, for each statistical test, it is required that for a certain event, $\mathcal{E}_\ell$, $1 \leq \ell \leq M$, the relative frequency $\frac{1}{N}\sum_{i=1}^{N} 1_{\mathcal{E}_\ell}(\boldsymbol{Y}_i)$ ($1_{\mathcal{E}}(\cdot)$ being the indicator function of an event $\mathcal{E}$), would converge, as $N \to \infty$, to a random variable (depending on $\boldsymbol{X}$), that is typically as close as possible to the expectation of $1_{\mathcal{E}_\ell}(\boldsymbol{X})$ with respect to (w.r.t.) the true unknown source, namely, to the probability of the event $\mathcal{E}_\ell$. We characterize the minimum key rate needed for this purpose and demonstrate how this minimum can be approached in principle.

Index Terms: Random number generators, random process simulation, statistical tests, typical sequences.

*This work was done while the author was visiting Hewlett–Packard Laboratories, California, U.S.A.

1 Introduction

Simulation of random processes, or information sources, is about artificial generation of random sequences with a prescribed probability law. This is done by applying a deterministic mapping from a string of purely random (independent, equally likely) bits into sample paths. The simulation problem has applications in speech and image synthesis, texture production (e.g., in image decompression), and generation of noise for purposes of simulating communication systems.

In the last few years, the simulation problem of sources and channels, as well as its interplay with other problem areas in information theory and related fields, have been investigated by a few researchers. Han and Verdú [3] presented the problem of finding the *resolvability* of a random process, namely, the minimum number of random bits required per generated sample, so that the finite dimensional marginals of the generated process converge to those of the desired process w.r.t. a certain distance measure between probability distributions. In [3], it was shown that if this distance measure is the variational distance, the resolvability is given by the *sup–entropy rate*, which coincides with the ordinary entropy rate for stationary and ergodic sources. In [5], a similar problem of channel simulation was studied. In that paper, the focus was on the minimum amount of randomness required in order to implement a good approximation to a conditional distribution corresponding to a given channel (see also [8] for further developments). In [6], the results of [3] were generalized to drop 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 target process, where the fidelity criterion is induced by the accuracy measure. In [2] and [7], specific algorithms for source and channel simulation, respectively, were proposed.

In all these works, the common assumption was that the probability law of the desired process is perfectly known. In a recent paper [4], this assumption was relaxed, and the following universal version of the simulation problem was considered: The target finite–alphabet source P to be simulated is unknown, except for the fact that it belongs to a certain parametric family $\mathcal{P}$, and we are given a training sequence $\boldsymbol{X} = (X_1, \ldots, X_m)$ that has emerged from this unknown source. We are also provided with a key string of k purely

random bits $\boldsymbol{U} = (U_1, \ldots, U_k)$, that is independent of $\boldsymbol{X}$, and our goal is to generate an output sequence $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ $(n \leq m)$ corresponding to the simulated process, that satisfies the following three conditions: (i) The mechanism by which $\boldsymbol{Y}$ is generated can be represented by a deterministic function $\boldsymbol{Y} = \phi(\boldsymbol{X}, \boldsymbol{U})$, where ϕ does not depend on the unknown source P, (ii) the probability law that governs $\boldsymbol{Y}$ is *exactly* the same law P corresponding to $\boldsymbol{X}$ for all $P \in \mathcal{P}$, and (iii) the mutual information $I(\boldsymbol{X}; \boldsymbol{Y})$ is as small as possible, simultaneously for all $P \in \mathcal{P}$.

In this paper, we adopt a similar model setting as in [4], but here (iii) is replaced by another criterion that may be more directly relevant to real–life purposes of the simulation of random processes: Now, instead of considering the behavior of a single generated random vector $\boldsymbol{Y}$ as in [4], we look at many random vectors $\boldsymbol{Y}_i = \phi(\boldsymbol{X}, \boldsymbol{U}_i)$, $i = 1, 2, \ldots, N$, all generated from the same given training vector $\boldsymbol{X}$, and of the same dimension as $\boldsymbol{X}$ (here, denoted by n), but with different (independent) key strings of length k, $\{\boldsymbol{U}_i\}_{i=1}^N$. Denoting by $1_{\mathcal{E}}(\cdot)$ (or, by $1\{\mathcal{E}\}$) the indicator function of an event, and given a set of M events, $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_M$, we would like the relative frequency of each event, $\frac{1}{N}\sum_{i=1}^N 1_{\mathcal{E}_\ell}(\boldsymbol{Y}_i)$, $\ell = 1, 2, \ldots, M$, to converge with probability one (as $N \to \infty$, while M, k, n, and $\boldsymbol{X}$ are held fixed) to a value (depending on $\boldsymbol{X}$) that is typically as close as possible to the expectation of $1_{\mathcal{E}_\ell}(\boldsymbol{X})$, w.r.t. the true underlying source P, that is, to $P\{\mathcal{E}_\ell\}$. The simplest, most common example is the set of $M = (n - r + 1) \cdot |\mathcal{A}|^r$ events $\mathcal{E}_\ell = \{(X_{j+1}, \ldots, X_{j+r}) = (x_1, \ldots, x_r)\}$, where r is a positive integer (typically, much smaller than n), $j = 0, 1, \ldots, (n - r)$, and $(x_1, \ldots, x_r)$ exhausts all possible r–tuples in $\mathcal{A}^r$. One would then like the relative frequency of each such event to come as close as possible to the respective true probability, $P(x_1, \ldots, x_r)$, so as to have a faithful simulation, at least w.r.t. r–th order statistics, namely, marginals of r–vectors.

At this point, a few words are in order with regard to the justification for confining attention to indicator functions of events and their relative frequencies, because one may argue that while this class of tests is important and reasonable on its own right, some statistical tests of interest may involve matching between empirical means and expectations of functions of $\{\boldsymbol{Y}_i\}$ that are more general than indicator functions. For example, in the case of synthesizing textures or, speech–like signals, these functions, say, $\{f_\ell(\boldsymbol{Y})\}_\ell$,[1] may corre-

[1] Here and throughout the sequel, we will use the symbol $\boldsymbol{Y}$ to generically denote a random vector that is generated by the simulation scheme, and which has the same probability law given $\boldsymbol{X}$ as each one of the $\boldsymbol{Y}_i$'s, but the index i is irrelevant to the context.

spond to a certain set of features that the generated data should have, like a prescribed behavior of the empirical autocorrelation vector, where $f_\ell(\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n-\ell} Y_i Y_{i+\ell}$, $\ell = 0, 1, \ldots, r$, corresponding to a certain, desired spectral structure. Nevertheless, almost every function of interest depends on $\boldsymbol{Y}$ only via some "sufficient statistics," $\sigma(\boldsymbol{Y})$ (e.g., the Markov–type in the above example of the autocorrelation test), and hence can be expressed, in the finite alphabet case, as a linear combination over a certain set of indicator functions of events. In particular, if a function, say, f, is such, then $f(\boldsymbol{Y}) = \sum_i \alpha_i 1\{\sigma(\boldsymbol{Y}) = \sigma_i\}$, where $\{\sigma_i\}$ are all possible values of $\sigma(\boldsymbol{Y})$ and $f(\boldsymbol{Y}) = \alpha_i$ wherever $\sigma(\boldsymbol{Y}) = \sigma_i$. Thus, indicator functions can be thought of as 'basis functions' of more general functions, and if their expectations are all approximated adequately by a simulation scheme, the same is true for their linear combinations. Having said that, we are essentially back to the problem of meeting a set of statistical tests w.r.t. indicator functions of events.

Returning then to our problem, let us first confine attention to a single test for the relative frequency of an event, denoted generically by $\mathcal{E}$. First, we observe that since $\{\boldsymbol{Y}_i\}$ are i.i.d. given $\boldsymbol{X}$, the empirical mean of $1_{\mathcal{E}}(\boldsymbol{Y}_i)$ converges almost surely to the *conditional* expected value of $1_{\mathcal{E}}(\boldsymbol{Y})$ given $\boldsymbol{X}$, namely,

$$\boldsymbol{E}\{1_{\mathcal{E}}(\boldsymbol{Y})|\boldsymbol{X}\} = P(\boldsymbol{Y} \in \mathcal{E}|\boldsymbol{X}) = \frac{1}{2^k}\sum_{\boldsymbol{u}\in\{0,1\}^k} 1_{\mathcal{E}}(\phi(\boldsymbol{X}, \boldsymbol{u})) \tag{1}$$

rather than to the desired, unconditional expectation

$$\boldsymbol{E}\{1_{\mathcal{E}}(\boldsymbol{Y})\} = P(\boldsymbol{Y} \in \mathcal{E}) = P(\boldsymbol{X} \in \mathcal{E}), \tag{2}$$

where the second equality is due to the fact that $\boldsymbol{Y}$ is required to obey the same probability law as $\boldsymbol{X}$. Letting ρ denote a certain distortion measure between these two values, a natural objective would be to devise a simulation scheme that minimizes

$$J(\boldsymbol{X};\boldsymbol{Y}) = \boldsymbol{E}\left\{\rho(P(\boldsymbol{Y} \in \mathcal{E}), P(\boldsymbol{Y} \in \mathcal{E}|\boldsymbol{X}))\right\}, \tag{3}$$

where the expectation[2] is over the ensemble of $\boldsymbol{X}$, and where following the common abuse of notation in the information theory literature, here $J(\boldsymbol{X};\boldsymbol{Y})$ is actually a functional of the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$, not $\boldsymbol{X}$ and $\boldsymbol{Y}$ themselves. Note that $J(\boldsymbol{X};\boldsymbol{Y})$ can be thought of as a measure of statistical dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the sense that it vanishes when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are statistically independent and it may be positive otherwise.

[2]Note that $P(\boldsymbol{Y} \in \mathcal{E})$ is a constant.

Thus, $J(\boldsymbol{X};\boldsymbol{Y})$ now replaces the ordinary mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ of [4], as a criterion for good simulation.

Our main results are in characterizing fundamental limitations on simulation performance in the sense of minimizing $J(\boldsymbol{X};\boldsymbol{Y})$ first, for a single statistical test, and then, simultaneously for a set of M such tests. In particular, we will characterize the minimum key rate, $R = k/n$, needed so that $J(\boldsymbol{X};\boldsymbol{Y})$ would be essentially as small as if there was an unlimited supply of random key bits ($k = \infty$). In the case of a single test, this minimum rate depends on the structure of the event $\mathcal{E}$ in a way that will become clear in the sequel. In the more general case of multiple statistical tests, the behavior is as follows: As long as the number of tests, M, grows in at a rate slower than double–exponential rate as a function of n, the key rate needed is essentially the same as for a single test (the most demanding one in the set). If M grows at the double–exponential rate, then some extra key rate will be needed.

There are a few similiarities and differences between the present work and [4]. First, we already mentioned that the main difference is that the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ of [4] is now replaced by $J(\boldsymbol{X};\boldsymbol{Y})$. We are not aware, however, of a way to present $I(\boldsymbol{X};\boldsymbol{Y})$ as a special case of $J(\boldsymbol{X};\boldsymbol{Y})$ for a particular choice of f and ρ that does not depend on the unknown P. Therefore, the criterion of [4] does not seem to be a special case of the criterion $J(\boldsymbol{X};\boldsymbol{Y})$ considered here. Moreover, the results of [4] are not generalizable to continuous–alphabet sources since for these sources, $I(\boldsymbol{X};\boldsymbol{Y}) = \infty$ for any finite number of key bits. By contrast, the approach taken in this paper can be generalized, in principle, to the continuous case. Another difference is the following: The main result in [4] is somewhat pessimistic in that it tells us that in order to keep $I(\boldsymbol{X};\boldsymbol{Y})$ small (sublinear in n), the key rate R must exceed the threshold of H bits/sample, where H is the entropy of the source. Here, on the other hand, as we shall see, the threshold rate needed to satisfy a set of statistical tests will always be less than or equal to the entropy, depending on the tests themselves, as mentioned earlier. Yet another difference is that in [4], the analysis and the results are more refined in the sense that for key rates above the threshold, there is an accurate characterization of the best achievable rate at which $I(\boldsymbol{X};\boldsymbol{Y})/n$ may vanish. There are no parallel results in this paper. Finally, on the more technical side, in [4], there is a distinction between the case where the dimension of $\boldsymbol{Y}$ is the same as the dimension of $\boldsymbol{X}$ and the case where the former is smaller. Here, there is no loss of generality in assuming

the former because for the latter case, one can choose the event $\mathcal{E}$ as being measurable only on a subset of the components of $\boldsymbol{Y}$. The similarities between the two papers are in the proof techniques and in the fact that the proposed simulation schemes are actually the same.

For reasons of simplicity and brevity, our analysis will be carried out under the assumption that $\mathcal{P}$ is the class of all memoryless sources of a given finite alphabet. However, similarly as in [4], our derivations and results extend to more general classes of sources, like exponential families of finite alphabet memoryless sources, Markov sources, finite-state sources, and parametric subfamilies of these classes.

The outline of the paper is as follows: Section 2 is devoted to establish notation conventions and to a more formal description of the problem. In Section 3, we derive a lower bound to the simulation performance w.r.t. a single test, and characterize the key rate needed to come close to this bound. Section 4 extends the setting to to the case of multiple statistical tests. Finally, in Section 5, we summarize our findings and discuss some issues for further research.

2 Notation and Problem Formulation

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets, as well as most of the other sets, will be denoted by caligraphic letters. Similarly, random vectors, their realizations, and their alphabets, will be denoted, respectively, by boldface capital letters, the corresponding boldface lower case letters, and caligraphic letters superscripted by the dimensions. For example, the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, ($n$ – positive integer) may take a specific vector value $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathcal{A}^n$, the nth order Cartesian power of $\mathcal{A}$, which is the alphabet of each component of this vector. The cardinality of a finite set $\mathcal{F}$ will be denoted by $|\mathcal{F}|$.

Let $\mathcal{P}$ denote the class of all discrete memoryless sources (DMSs) with a finite alphabet $\mathcal{A}$, and let P denote a particular DMS in $\mathcal{P}$. For given positive integers n and k, let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, $X_i \in \mathcal{A}$, $i = 1, \ldots, n$, denote an n-vector drawn from P, namely, $\Pr\{X_i = x_i,\ i = 1, \ldots, n\} = \prod_{i=1}^n P(x_i) \stackrel{\triangle}{=} P(\boldsymbol{x})$ for every $(x_1, \ldots, x_n)$, $x_i \in \mathcal{A}$, $i = 1, \ldots, n$. Let $H = -\sum_{x \in \mathcal{A}} P(x) \log P(x)$ denote the entropy of the source P, where here and throughout the sequel, $\log(\cdot) \stackrel{\triangle}{=} \log_2(\cdot)$. For a given positive integer k, let $\boldsymbol{U} = (U_1, \ldots, U_k)$,

$U_i \in \mathcal{B} \stackrel{\triangle}{=} \{0,1\}$, $i = 1, \ldots, k$, denote a string of k random bits, drawn from the binary symmetric source, independently of $\boldsymbol{X}$.

Since we will rely quite heavily on the method of types [1] in this paper, we next describe the notation that will be used in this context: For a given source vector $\boldsymbol{x} \in \mathcal{A}^n$, the empirical probability mass function (EPMF) is the vector $Q_{\boldsymbol{x}} = \{q_{\boldsymbol{x}}(a), a \in \mathcal{A}\}$, where $q_{\boldsymbol{x}}(a)$ is the relative frequency of the letter $a \in \mathcal{A}$ in the vector $\boldsymbol{x}$. The type class $T_{\boldsymbol{x}}$ of a vector $\boldsymbol{x}$ is the set of all vectors $\tilde{\boldsymbol{x}} \in \mathcal{A}^n$ such that $Q_{\tilde{\boldsymbol{x}}} = Q_{\boldsymbol{x}}$. The set of EPMF's of n–vectors will be denoted by $\mathcal{Q}^n$. When we need to attribute a type class to a certain rational PMF $Q \in \mathcal{Q}^n$ rather than to a sequence in $\mathcal{A}^n$, we shall use the notation T_Q. We shall denote by $T_{\boldsymbol{X}}$ the (random) type class of a random vector $\boldsymbol{X}$ drawn from a DMS $P \in \mathcal{P}$.

For two given positive integers, n and k, and a given mapping $\phi : \mathcal{A}^n \times \mathcal{B}^k \to \mathcal{A}^n$, let $\boldsymbol{Y} = \phi(\boldsymbol{X}, \boldsymbol{U})$. Let $W(\boldsymbol{y}|\boldsymbol{x})$ denote the conditional probability of $\boldsymbol{Y} = \boldsymbol{y}$ given $\boldsymbol{X} = \boldsymbol{x}$ corresponding to the channel from $\boldsymbol{X}$ to $\boldsymbol{Y}$ that is induced by ϕ, i.e.,

$$W(\boldsymbol{y}|\boldsymbol{x}) = 2^{-k} |\{\boldsymbol{u} : \ \phi(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{y}\}|. \tag{4}$$

Unless stated otherwise, the expectation operator, denoted by $\boldsymbol{E}\{\cdot\}$, will be understood to be taken w.r.t. the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$.

Let $\mathcal{E} \subseteq \mathcal{A}^n$ be a given event (corresponding to a statistical test), let $\rho : [0,1]^2 \to \mathbb{R}^+$ be a distance (or, distortion) function, convex in its second argument, and satisfying, for every $u \in [0,1]$: (i) $\rho(u,u) = 0$, and (ii) $\rho(u,v)$ is monotonically non-decreasing in v for $v \geq u$ and is monotonically non-increasing in v for all $v \leq u$.[3] For a given ϕ, let

$$P(\boldsymbol{Y} \in \mathcal{E}|\boldsymbol{X} = \boldsymbol{x}\} = \sum_{\boldsymbol{y} \in \mathcal{E}} W(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{2^k} \sum_{\boldsymbol{u} \in \{0,1\}^k} 1_{\mathcal{E}}(\phi(\boldsymbol{x}, \boldsymbol{u})), \tag{5}$$

and define

$$\begin{aligned} J(\boldsymbol{X}; \boldsymbol{Y}) &= \boldsymbol{E}\left\{\rho(P(\boldsymbol{Y} \in \mathcal{E}), P(\boldsymbol{Y} \in \mathcal{E}|\boldsymbol{X}))\right\} \\ &= \sum_{\boldsymbol{x} \in \mathcal{A}^n} P(\boldsymbol{x}) \cdot \rho(P(\boldsymbol{Y} \in \mathcal{E}), P(\boldsymbol{Y} \in \mathcal{E}|\boldsymbol{X} = \boldsymbol{x})). \end{aligned} \tag{6}$$

For a single statistical test, designated by $\mathcal{E}$, and a given ρ, we seek a mapping ϕ that meets the following conditions:

[3] Since both arguments of ρ are probabilities, and in the interesting cases, exponentially small ones, the differences between them would typically be small as well. In such cases, one may let ρ depend on the ratio rather than the difference between its arguments. A reasonable choice of ρ would then be of the form $\rho(u,v) = \rho_0(v/u)$, where $\rho_0 : \mathbb{R}^+ \to \mathbb{R}^+$ is convex and $\rho_0(1) = 0$.

C1. The mapping is independent of P.

C2. For every $P \in \mathcal{P}$ and every $y^n \in \mathcal{A}^n$

$$\Pr\{\boldsymbol{Y} = \boldsymbol{y}\} \stackrel{\triangle}{=} \sum_{\boldsymbol{x}} [W(\boldsymbol{y}|\boldsymbol{x}) \prod_{i=1}^{n} P(x_i)] = P(\boldsymbol{y}) = \prod_{j=1}^{n} P(y_j). \tag{7}$$

C3. The mapping ϕ minimizes $J(\boldsymbol{X};\boldsymbol{Y})$ simultaneously for all $P \in \mathcal{P}$.

In the case of multiple tests, $\{\mathcal{E}_\ell\}_{\ell=1}^{M}$, let $J_\ell(\boldsymbol{X};\boldsymbol{Y})$ be defined as in eq. (6) with $\mathcal{E}$ being replaced by $\mathcal{E}_\ell$, $\ell = 1, 2, \ldots, M$. Then, C3 is replaced by the requirement that $J_\ell(\boldsymbol{X};\boldsymbol{Y})$ are to be minimized, if possible, simultaneously for all $\ell = 1, 2, \ldots, M$ and all $P \in \mathcal{P}$ subject to conditions C1 and C2.

Regarding the asymptotic regime of the length k of the key string, it is common to assume that k grows linearly with n, that is, $k = nR$, where $R \geq 0$ is a constant interpreted as the random-bit rate, i.e., the average number of random bits used per generated symbol of $\boldsymbol{Y}$. However, in our setting, since $\boldsymbol{X}$ is given and fixed throughout the entire experiment of N trials, it makes sense to allow k, and hence also R, depend on $\boldsymbol{X}$ rather than being a constant. In this case, a better notation would be $k(\boldsymbol{X})$ and $R(\boldsymbol{X})$, respectively, or $k(\boldsymbol{x})$ and $R(\boldsymbol{x})$, for a specific vector value of the training vector. However, to avoid cumbersome notation, we will continue to use occasionally the shorthand notations k and R with the understanding that they may depend on $\boldsymbol{x}$.

Finally, we say that a sequence $\{A_n\}$ is of the *exponential order of* $2^{n\Lambda}$ (Λ being a constant) if $\lim_{n\to\infty} \frac{1}{n} \log A_n = \Lambda$. By the same token, we say that $\{A_n\}$ is of the *double–exponential order of* $2^{2^{n\Lambda}}$ if $\lim_{n\to\infty} \frac{1}{n} \log\log A_n = \Lambda$. Similarly, $\{A_n\}$ is said to be *of the double–exponential order of* $2^{-2^{n\Lambda}}$ if $\{1/A_n\}$ is of the double–exponential order of $2^{2^{n\Lambda}}$.

3 A Single Statistical Test

For the case of a single statistical test of the relative frequency of an event $\mathcal{E}$, we begin with a simple lower bound to $J(\boldsymbol{X};\boldsymbol{Y})$ that applies to any simulation scheme that satisfies C1 and C2, with an arbitrarily large number of random key bits. This bound depends only on P, $\mathcal{E}$, and n, not on the particular simulation scheme.

Theorem 1 *Let $\rho(\cdot,\cdot)$ be convex in its second argument. Then, for any simulation scheme ϕ that satisfies C1 and C2,*

$$J(\boldsymbol{X};\boldsymbol{Y}) \geq J_0 \stackrel{\triangle}{=} \boldsymbol{E}\{\rho(P(\mathcal{E}), P(\boldsymbol{X} \in \mathcal{E}|T_{\boldsymbol{X}}))\}$$

$$= \sum_{Q \in \mathcal{Q}^n} P(T_Q) \cdot \rho\left(P(\mathcal{E}), \frac{|T_Q \cap \mathcal{E}|}{|T_Q|}\right). \tag{8}$$

The quantity J_0 manifests the "price of universality," namely, the price that must be paid for the fact that P is unknown and only a finite–length training sequence from P is given. It has nothing to do with the fact that the reservoir of random bits may be limited, a fact which may yield additional cost beyond J_0. Before proving Theorem 1, we pause to provide a simple example for calculating J_0:

Example 1 *Let P be a binary source ($\mathcal{A} = \{0,1\}$) with $p \stackrel{\triangle}{=} P(X_1 = 1)$, and let $\mathcal{E} = \{\boldsymbol{x}: \ x_1 = 1\}$. In this case,*

$$\frac{|T_Q \cap \mathcal{E}|}{|T_Q|} = \frac{\left(\begin{array}{c} n-1 \\ nq_{\boldsymbol{x}}(1)-1 \end{array}\right)}{\left(\begin{array}{c} n \\ nq_{\boldsymbol{x}}(1) \end{array}\right)} = q_{\boldsymbol{x}}(1). \tag{9}$$

Therefore, if ρ is the squared–error distortion measure, then

$$J_0 = \boldsymbol{E}\{[p - q_{\boldsymbol{x}}(1)]^2\} = Var\{q_{\boldsymbol{x}}(1)\} = \frac{p(1-p)}{n}. \tag{10}$$

Proof of Theorem 1. As is shown in [4] (proof of Theorem 1(a) therein), to meet conditions C1 and C2, for any type class T_Q, given the event $\boldsymbol{X} \in T_Q$, the output vector $\boldsymbol{Y}$ must always be uniformly distributed across T_Q. This means that

$$P(\boldsymbol{y}|\boldsymbol{X} \in T_Q) = \left\{ \begin{array}{ll} \frac{1}{|T_Q|} & \boldsymbol{y} \in T_Q \\ 0 & \text{elsewhere} \end{array} \right. \tag{11}$$

and, on the other hand, denoting

$$\mathcal{S}(\boldsymbol{y}, T_Q) = \{(\boldsymbol{x}, \boldsymbol{u}): \ \boldsymbol{x} \in T_Q, \ \phi(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{y}\},$$

we have the following:

$$\begin{aligned} P(\boldsymbol{y}|\boldsymbol{X} \in T_Q) &= P(\phi(\boldsymbol{X}, \boldsymbol{U}) = \boldsymbol{y}|\boldsymbol{X} \in T_Q) \\ &= \sum_{(\boldsymbol{x},\boldsymbol{u}) \in \mathcal{S}(\boldsymbol{y}, T_Q)} \frac{1}{|T_Q|} \cdot \frac{1}{2^{k(\boldsymbol{x})}} \\ &= \frac{1}{|T_Q|} \sum_{(\boldsymbol{x},\boldsymbol{u}) \in \mathcal{S}(\boldsymbol{y}, T_Q)} 2^{-k(\boldsymbol{x})} \end{aligned} \tag{12}$$

which together with eq. (11), implies that for any ϕ that satisfies C1 and C2,

$$\sum_{(\boldsymbol{x},\boldsymbol{u}) \in \mathcal{S}(\boldsymbol{y}, T_Q)} 2^{-k(\boldsymbol{x})} = \left\{ \begin{array}{ll} 1 & \boldsymbol{y} \in T_Q \\ 0 & \text{elsewhere} \end{array} \right. \tag{13}$$

We now derive the lower bound on $J(\boldsymbol{X};\boldsymbol{Y})$ using this fact:

$$\begin{aligned}
J(\boldsymbol{X};\boldsymbol{Y}) &= \sum_{\boldsymbol{x}\in\mathcal{A}^n} P(\boldsymbol{x})\cdot\rho(P(\boldsymbol{Y}\in\mathcal{E}),P(\boldsymbol{Y}\in\mathcal{E}|\boldsymbol{X}=\boldsymbol{x})) \\
&= \sum_{Q\in\mathcal{Q}^n} P(T_Q)\cdot\frac{1}{|T_Q|}\sum_{\boldsymbol{x}\in T_Q}\rho(P(\boldsymbol{Y}\in\mathcal{E}),P(\boldsymbol{Y}\in\mathcal{E}|\boldsymbol{X}=\boldsymbol{x})) \\
&\geq \sum_{Q\in\mathcal{Q}^n} P(T_Q)\cdot\rho\left(P(\boldsymbol{Y}\in\mathcal{E}),\frac{1}{|T_Q|}\sum_{\boldsymbol{x}\in T_Q}P(\boldsymbol{Y}\in\mathcal{E}|\boldsymbol{X}=\boldsymbol{x})\right) \\
&= \sum_{Q\in\mathcal{Q}^n} P(T_Q)\cdot\rho\left(P(\boldsymbol{Y}\in\mathcal{E}),P(\boldsymbol{Y}\in\mathcal{E}|\boldsymbol{X}\in T_Q)\right) \\
&= J(T_{\boldsymbol{X}};\boldsymbol{Y}) \qquad (14)
\end{aligned}$$

where the inequality follows from the assumption on the convexity of $\rho(\cdot,\cdot)$ w.r.t. its second argument and equality is attained if $\{P(\boldsymbol{Y}\in\mathcal{E}|\boldsymbol{X}=\boldsymbol{x})\}_{\boldsymbol{x}\in T_Q}$ are all the same.[4] To complete the proof it remains to show that $P(\boldsymbol{Y}\in\mathcal{E}|\boldsymbol{X}\in T_Q)=P(\boldsymbol{X}\in\mathcal{E}|\boldsymbol{X}\in T_Q)$. Now,

$$\begin{aligned}
P(\boldsymbol{Y}\in\mathcal{E}|\boldsymbol{X}\in T_Q) &= P(\boldsymbol{Y}\in\mathcal{E}\cap T_Q|\boldsymbol{X}\in T_Q) \\
&= \frac{1}{|T_Q|}\sum_{\boldsymbol{x}\in T_Q}\frac{1}{2^{k(\boldsymbol{x})}}\sum_{\boldsymbol{u}}1_{\mathcal{E}\cap T_Q}(\phi(\boldsymbol{x},\boldsymbol{u})) \\
&= \frac{1}{|T_Q|}\sum_{\boldsymbol{y}\in\mathcal{A}^n}\sum_{(\boldsymbol{x},\boldsymbol{u})\in\mathcal{S}(\boldsymbol{y},T_Q)}2^{-k(\boldsymbol{x})}1_{\mathcal{E}\cap T_Q}(\boldsymbol{y}) \\
&= \frac{1}{|T_Q|}\sum_{\boldsymbol{y}\in\mathcal{A}^n}1_{\mathcal{E}\cap T_Q}(y)\sum_{(\boldsymbol{x},\boldsymbol{u})\in\mathcal{S}(\boldsymbol{y},T_Q)}2^{-k(\boldsymbol{x})} \\
&= \frac{1}{|T_Q|}\sum_{\boldsymbol{y}\in\mathcal{A}^n}1_{\mathcal{E}\cap T_Q}(\boldsymbol{y}) \\
&= \frac{1}{|T_Q|}\sum_{\boldsymbol{x}\in\mathcal{A}^n}1_{\mathcal{E}\cap T_Q}(\boldsymbol{x}) \\
&= \frac{1}{|T_Q|}\sum_{\boldsymbol{x}\in T_Q}1_{\mathcal{E}}(\boldsymbol{x}) \\
&= P(\boldsymbol{X}\in\mathcal{E}|\boldsymbol{X}\in T_Q), \qquad (15)
\end{aligned}$$

where the fifth equality follows from eq. (13). This completes the proof of Theorem 1.

Theorem 1 tells us that the best we can do, in order that the Jensen inequality of eq. (14) would come close to equality, is devise a simulation scheme ϕ such that for every T_Q and every $\boldsymbol{x}\in T_Q$,

$$P(\boldsymbol{Y}\in\mathcal{E}|\boldsymbol{X}=\boldsymbol{x})\equiv\frac{1}{2^k}\sum_{\boldsymbol{u}}1_{\mathcal{E}}(\phi(\boldsymbol{x},\boldsymbol{u})) \qquad (16)$$

[4] The inequality $J(\boldsymbol{X};\boldsymbol{Y})\geq J(T_{\boldsymbol{X}};\boldsymbol{Y})$ can be thought of as a certain type of a data processing theorem for $J(\boldsymbol{X};\boldsymbol{Y})$.

would be as close as possible to

$$P(\boldsymbol{X} \in \mathcal{E} | \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{|T_Q|} \sum_{\boldsymbol{x} \in T_Q} 1_{\mathcal{E}}(\boldsymbol{x}) = \frac{|T_Q \cap \mathcal{E}|}{|T_Q|}. \tag{17}$$

However, the reason we need the proximity between (16) and (17) for *every* $\boldsymbol{x}$ is deeper than the aforementioned technical reason of making the inequality of (14) as tight as possible: It is important to remember that the training vector $\boldsymbol{X}$ is generated only once and it remains fixed throught the entire experiment of generating many vectors $\{\boldsymbol{Y}_i\}$, so the law of large numbers for $\{\boldsymbol{Y}_i\}$ applies only w.r.t. the given $\boldsymbol{X}$. Therefore, it is not enough to have a simulation scheme that is good (in the sense of attaining J_0) only over ensemble average of $\boldsymbol{X}$, but one would like to guarantee that it would be good essentially for *every* given, typical $\boldsymbol{X}$. And since P is unknown, this should be true for all type classes $\{T_Q\}$. Therefore, from now on, we focus on simulation schemes for which (16) comes close to (17) for every T_Q and every $\boldsymbol{x} \in T_Q$.

Observe that there are two cases where (16) is exactly equivalent to (17). The first is the case where $T_Q \cap \mathcal{E} = \emptyset$, where both (16) and (17) trivially vanish (the former, for any simulation scheme satisfying C1 and C2, even with no key bits at all). The other case where (16) matches perfectly to (17) is when $k(\boldsymbol{x})$ is at least as large as $\log |T_{\boldsymbol{x}}|$, because in this case, there is enough randomization power to implement a uniform distribution across $T_{\boldsymbol{x}}$ (see also [4]). But this means that the rate

$$R(\boldsymbol{x}) = \frac{k(\boldsymbol{x})}{n} = \frac{\log |T_{\boldsymbol{x}}|}{n},$$

is approximately equal to the empirical entropy, $-\sum_{a \in \mathcal{A}} q_{\boldsymbol{x}}(a) \log q_{\boldsymbol{x}}(a)$, whose expectation, w.r.t. the ensemble of $\{\boldsymbol{X}\}$, tends to the entropy, H. The interesting question is whether we can manage with a smaller bit rate to obtain a good approximation of (17) by (16), and if so, what is the minimum rate.

Intuitively, the answer to the first question is affirmative: The type class should only be populated with sufficiently many points $\{\phi(\boldsymbol{x}, \boldsymbol{u})\}_{\boldsymbol{u} \in \{0,1\}^k}$ so as to have a faithful approximation of the relative number of typical sequences within $\mathcal{E}$. In the following, we try to translate this intuition to a concrete result.

Denoting

$$\hat{H} \equiv \hat{H}(\boldsymbol{x}) \stackrel{\triangle}{=} \frac{1}{n} \log |T_{\boldsymbol{x}}|, \tag{18}$$

and

$$E \equiv E(\boldsymbol{x}) \stackrel{\triangle}{=} \begin{cases} \frac{1}{n}\log|T_{\boldsymbol{x}} \cap \mathcal{E}| & T_{\boldsymbol{x}} \cap \mathcal{E} \neq \emptyset \\ \hat{H} & T_{\boldsymbol{x}} \cap \mathcal{E} = \emptyset \end{cases} \tag{19}$$

we define

$$R_{\mathcal{E}} \equiv R_{\mathcal{E}}(\boldsymbol{x}) \stackrel{\triangle}{=} \hat{H} - E. \tag{20}$$

Following the above discussion, for a given $\Delta \in [0,1)$, we define a simulation scheme ϕ as *Δ–faithful for $\mathcal{E}$ w.r.t. $\boldsymbol{x}$*, if

$$(1-\Delta)\cdot\frac{|T_{\boldsymbol{x}} \cap \mathcal{E}|}{|T_{\boldsymbol{x}}|} \leq \frac{1}{2^k}\sum_{\boldsymbol{u}} 1_{\mathcal{E}}(\phi(\boldsymbol{x},\boldsymbol{u})) \leq (1+\Delta)\cdot\frac{|T_{\boldsymbol{x}} \cap \mathcal{E}|}{|T_{\boldsymbol{x}}|}. \tag{21}$$

We also define a simulation scheme as *Δ–faithful for $\mathcal{E}$* if it is *Δ–faithful for $\mathcal{E}$* w.r.t. every $\boldsymbol{x}$. The next theorem tells us that $R_{\mathcal{E}}(\boldsymbol{x})$ is essentially the minimum key rate, as a function of $\boldsymbol{x}$, required for Δ-faithful simulation w.r.t. all $\boldsymbol{x}$. The asymptotically minimum average rate is, therefore, $\boldsymbol{E}\{R_{\mathcal{E}}(\boldsymbol{X})\}$. Again, this expectation w.r.t. $\boldsymbol{X}$ is meaningful only if it is essentially realized, for large n, by every typical $\boldsymbol{x}$, since the training vector is drawn only once.

Theorem 2 *(a) (Converse part): If $R(\boldsymbol{x}) < R_{\mathcal{E}}(\boldsymbol{x}) - \frac{1}{n}\log(1+\Delta)$, for some $\boldsymbol{x}$, then there exists no simulation scheme that is Δ–faithful for $\mathcal{E}$ w.r.t. that $\boldsymbol{x}$, and hence nor a Δ–faithful scheme for $\mathcal{E}$.*

(b) (Direct part): Let $\epsilon > 0$ be given and let n be sufficiently large. If $R(\boldsymbol{x}) \geq R_{\mathcal{E}}(\boldsymbol{x}) + \epsilon$ for all $\boldsymbol{x}$, there exists a simulation scheme which is Δ-faithful for $\mathcal{E}$, provided that $\Delta \geq 2^{-n\delta}$ for some $\delta < \epsilon/2$.

Before proving this theorem, let us consider the following example for assessing the function $R_{\mathcal{E}}(\boldsymbol{x})$ and its expectation.

Example 2 *Let P be again, the binary memoryless source with $p = P(X_1 = 1)$, as in Example 1. Let now $\mathcal{E} = \{\boldsymbol{x} : \ x_1 = x_2 = \ldots = x_{\lfloor n\lambda \rfloor} = 1\}$, where $\lambda \in (0,1)$. Now, if $nq_{\boldsymbol{x}}(1) < \lfloor n\lambda \rfloor$, then $|T_{\boldsymbol{x}} \cap \mathcal{E}| = 0$, otherwise*

$$|T_{\boldsymbol{x}} \cap \mathcal{E}| = \binom{n-\lfloor n\lambda \rfloor}{nq_{\boldsymbol{x}}(1) - \lfloor n\lambda \rfloor} \approx \exp_2\left\{n(1-\lambda)h\left(\frac{q_{\boldsymbol{x}}(1)-\lambda}{1-\lambda}\right)\right\}, \tag{22}$$

where $h(u) = -u\log u - (1-u)\log(1-u)$, for $u \in [0,1]$, is the binary entropy function. Therefore,

$$R_{\mathcal{E}}(\boldsymbol{x}) \approx \begin{cases} h(q_{\boldsymbol{x}}(1)) - (1-\lambda)h\left(\frac{q_{\boldsymbol{x}}(1)-\lambda}{1-\lambda}\right), & q_{\boldsymbol{x}}(1) > \lambda \\ 0, & q_{\boldsymbol{x}}(1) < \lambda \end{cases} \tag{23}$$

Now, as $n \to \infty$, $q_{\boldsymbol{x}}(1)$ tends to p almost surely, and so,

$$\boldsymbol{E}\{R_{\mathcal{E}}(\boldsymbol{X})\} \to \begin{cases} h(p) - (1-\lambda)h\left(\frac{p-\lambda}{1-\lambda}\right), & p > \lambda \\ 0, & p < \lambda \end{cases} \tag{24}$$

Note that there is a discontinuity at $p = \lambda$.

The remaining part of this section is devoted to the proof of Theorem 2.

Proof. Part (a) is fairly simple. If $T_{\boldsymbol{x}} \cap \mathcal{E} = \emptyset$, there is nothing to prove. Otherwise, note that

$$P(\boldsymbol{Y} \in \mathcal{E} | \boldsymbol{X} = \boldsymbol{x}) = 2^{-k} \sum_{\boldsymbol{u}} 1_{\mathcal{E}}(\phi(\boldsymbol{x}, \boldsymbol{u}))$$

can only take on values that are integer multiples of 2^{-k}, namely, $m \cdot 2^{-k}$ for $m = 0, 1, 2 \ldots$. Now, for $m = 0$, the left inequality in (21) is obviously violated. If $R < R_{\mathcal{E}} - \log(1+\Delta)/n$, the right inequality of (21) is already violated for $m = 1$, let alone larger values of m. This completes the proof of part (a).

Turning to part (b), if $T_{\boldsymbol{x}} \cap \mathcal{E} = \emptyset$, the simulation scheme $\boldsymbol{y} = \boldsymbol{x}$, which requires no key bits at all, satisfies (21) trivially, as mentioned earlier. For the case $T_{\boldsymbol{x}} \cap \mathcal{E} \neq \emptyset$, our proof technique is similar to the one in [4]. Consider mappings of the following structure: List the members of each type class T_Q in a certain order, and for every $\boldsymbol{x} \in T_Q$, let $I(\boldsymbol{x}) \in \{0, 1, \ldots, |T_Q| - 1\}$ denote the index of $\boldsymbol{x}$ within T_Q in this list (starting from zero for the first sequence). Denoting by I^{-1} the inverse map from $\{0, 1, \ldots, |T_Q| - 1\}$ to T_Q, we define

$$\boldsymbol{y} = \phi(\boldsymbol{x}, \boldsymbol{u}) \stackrel{\triangle}{=} I^{-1}\left(I(\boldsymbol{x}) \oplus \sum_{i=1}^{k(\boldsymbol{x})} 2^{i-1} u_i\right), \tag{25}$$

where $\oplus$ denotes addition modulo $|T_Q|$, and the summation over i is taken under the ordinary integer arithmetic (and defined as zero when $k(\boldsymbol{x}) = 0$).

This mapping obviously satisfies condition C1 as it is independent of P. Since $\boldsymbol{X}$ is uniformly distributed within its type class, then so is $\boldsymbol{Y}$ and therefore, ϕ satisfies condition C2 as well. Whether or not such a mapping is Δ–faithful for $\mathcal{E}$ w.r.t. $\boldsymbol{x}$, depends on the ordering, or the permutation corresponding to the ordered list of n-sequences in each of the type classes. There are as many as $|T_Q|!$ different permutations, and we next show that there exists a permutation that induces a Δ–faithful approximation of $|T_Q \cap \mathcal{E}|/|T_Q|$. In fact, we show that the vast majority of permutations of T_Q are such.

First, observe that given $\boldsymbol{x} \in T_Q$, there is a set of $2^k = 2^{nR}$ different sequences $\{\boldsymbol{y}\}$, which we shall denote by $\mathcal{Y}(\boldsymbol{x})$, that are obtained from (25) as $\boldsymbol{u}$ exhausts the key space.

We would like to have a simulation scheme for which the relative frequency of sequences from $\mathcal{E}$ within $\mathcal{Y}(\boldsymbol{x})$ is within a factor of $1 \pm \Delta$ away from the ideal value, $|T_Q \cap \mathcal{E}|/|T_Q|$, simultaneously for all $\boldsymbol{x} \in T_Q$. Let us first upper bound the number of permutations, K, of T_Q such that for a given $\boldsymbol{x}$, $\mathcal{Y}(\boldsymbol{x})$ contains at least

$$L_0 \stackrel{\triangle}{=} (1+\Delta) \cdot \frac{|T_Q \cap \mathcal{E}|}{|T_Q|} \cdot 2^{nR}$$

sequences of $\mathcal{E}$. A straightforward combinatorial consideration implies that this number of permutations is given by

$$K = (|T_Q| - 2^{nR})! \sum_{\ell \geq L_0} \left(\begin{array}{c} 2^{nR} \\ \ell \end{array} \right) \prod_{i=0}^{\ell-1} (|T_Q \cap \mathcal{E}| - i) \prod_{j=0}^{2^{nR}-\ell-1} (|T_Q| - |T_Q \cap \mathcal{E}| - j), \qquad (26)$$

where each summand corresponds to all combinations of 2^{nR} sequences (that form $\mathcal{Y}(\boldsymbol{x})$) such that exactly ℓ members of them are from $\mathcal{E}$ and the factor in front of the summation is the number of permutations of the members of $T_Q \cap [\mathcal{Y}(\boldsymbol{x})]^c$. Equivalently, K can be rewritten as follows:

$$\begin{aligned} K &= (|T_Q| - 2^{nR})! \sum_{\ell \geq L_0} \frac{(2^{nR})!}{\ell!(2^{nR}-\ell)!} \cdot \frac{|T_Q \cap \mathcal{E}|!}{(|T_Q \cap \mathcal{E}| - \ell)!} \cdot \frac{(|T_Q| - |T_Q \cap \mathcal{E}|)!}{(|T_Q| - |T_Q \cap \mathcal{E}| - (2^{nR} - \ell))!} \\ &= |T_Q|! \cdot \frac{\sum_{\ell \geq L_0} \left(\begin{array}{c} |T_Q \cap \mathcal{E}| \\ \ell \end{array} \right) \cdot \left(\begin{array}{c} |T_Q| - |T_Q \cap \mathcal{E}| \\ 2^{nR} - \ell \end{array} \right)}{\left(\begin{array}{c} |T_Q| \\ 2^{nR} \end{array} \right)}. \qquad (27) \end{aligned}$$

Since the first factor of the last expression, $|T_Q|!$, is the total number of permutations of the members of T_Q, the second factor is the fraction of permutations for which $\ell \geq L_0$. We next show that this fraction is doubly exponentially small as a function of n. To this end, we upper bound the numerator and lower bound the denominator of the second factor of right–most side of last equation. The numerator is upper bounded using the fact that for any two nonnegative integers p and q ($q \leq p$):

$$\left(\begin{array}{c} p \\ q \end{array} \right) \leq 2^{ph(q/p)}.$$

Specifically,

$$\begin{aligned} & \sum_{\ell \geq L_0} \left(\begin{array}{c} |T_Q \cap \mathcal{E}| \\ \ell \end{array} \right) \cdot \left(\begin{array}{c} |T_Q| - |T_Q \cap \mathcal{E}| \\ 2^{nR} - \ell \end{array} \right) \\ \leq & \sum_{\ell \geq L_0} \exp_2 \left\{ |T_Q \cap \mathcal{E}| \cdot h\left(\frac{\ell}{|T_Q \cap \mathcal{E}|} \right) \right\} \cdot \exp_2 \left\{ (|T_Q| - |T_Q \cap \mathcal{E}|) \cdot h\left(\frac{2^{nR} - \ell}{|T_Q| - |T_Q \cap \mathcal{E}|} \right) \right\} \end{aligned}$$

$$
\begin{aligned}
&\leq 2^{nR} \max_{\ell \geq L_0} \exp_2 \left\{ |T_Q \cap \mathcal{E}| \cdot h\left(\frac{\ell}{|T_Q \cap \mathcal{E}|}\right) + (|T_Q| - |T_Q \cap \mathcal{E}|) \cdot h\left(\frac{2^{nR} - \ell}{|T_Q| - |T_Q \cap \mathcal{E}|}\right) \right\} \\
&\leq 2^{nR} \cdot 2^{|T_Q| \cdot F} \qquad (28)
\end{aligned}
$$

where $F = \max\{\theta h(\alpha) + (1-\theta)h(\beta)\}$, $\theta \stackrel{\triangle}{=} |T_Q \cap \mathcal{E}|/|T_Q|$, the maximum being over all pairs (α, β) for which $\alpha \geq (1+\Delta)\gamma$ and $\theta\alpha + (1-\theta)\beta = \gamma$, with $\gamma \stackrel{\triangle}{=} 2^{nR}/|T_Q|$. It is easy to show that the function $\theta h(\alpha) + (1-\theta)h((\gamma - \theta\alpha)/(1-\theta))$ is monotonically decreasing in α for $\alpha \geq \gamma$, and so, the maximum defining F is attained for $\alpha = \alpha_0 \stackrel{\triangle}{=} (1+\Delta)\gamma$. Thus, the numerator of the expression at hand is upper bounded by

$$
2^{nR} \cdot \exp_2 \left\{ |T_Q| \cdot [\theta h(\alpha_0) + (1-\theta)h(\beta_0)] \right\},
$$

where $\beta_0 \stackrel{\triangle}{=} (\gamma - \theta\alpha_0)/(1-\theta)$. The denominator, on the other hand, is lower bounded [1] by:

$$
\left(\begin{array}{c} |T_Q| \\ 2^{nR} \end{array} \right) \geq \frac{1}{|T_Q| + 1} \cdot \exp_2\{|T_Q| \cdot h(\gamma)\}. \qquad (29)
$$

When plugging the upper bound on the numerator and the lower bound on the denominator into eq. (27), the exponent of the denominator is subtracted from that of the numerator and we obtain:

$$
\begin{aligned}
\theta h(\alpha_0) + (1-\theta)h(\beta_0) - h(\gamma) &= -\theta D(\alpha_0 \| \gamma) - (1-\theta) D(\beta_0 \| \gamma) \\
&\leq -\theta D(\alpha_0 \| \gamma), \qquad (30)
\end{aligned}
$$

where for $a, b \in [0,1]$, $D(a\|b) \stackrel{\triangle}{=} a \log(a/b) + (1-a)\log[(1-a)/(1-b)]$. It then follows that

$$
K \leq (|T_Q| + 1)! \cdot 2^{nR} \exp_2 \left\{ -|T_Q \cap \mathcal{E}| \cdot D((1+\Delta)\gamma \| \gamma) \right\}. \qquad (31)
$$

To further upper bound K, we next derive a lower bound on $|T_Q \cap \mathcal{E}| \cdot D((1+\Delta)\gamma\|\gamma)$. Using the fact that

$$
\ln(1+u) = -\ln\left(1 - \frac{u}{u+1}\right) \geq \frac{u}{u+1} \quad \forall u > -1, \qquad (32)
$$

we have the following lower bound on the divergence:

$$
\begin{aligned}
D((1+\Delta)\gamma\|\gamma) &= \frac{1}{\ln 2}(1+\Delta)\gamma \ln(1+\Delta) + \\
&\quad \frac{1}{\ln 2}[1-\Delta)\gamma] \ln\left[1 - \frac{\gamma\Delta}{1-\gamma}\right] \\
&\geq \frac{1}{\ln 2}(1+\Delta)\gamma \ln(1+\Delta) - \\
&\quad \frac{1}{\ln 2}[1-(1+\Delta)\gamma] \cdot \frac{\gamma\Delta/(1-\gamma)}{1-\gamma\Delta/(1-\gamma)}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\ln 2}(1+\Delta)\gamma\ln(1+\Delta) - \\
&\quad \frac{1}{\ln 2}[1-(1+\Delta)\gamma]\cdot\frac{\gamma\Delta}{1-(1+\Delta)\gamma} \\
&= \frac{\gamma}{\ln 2}[(1+\Delta)\ln(1+\Delta)-\Delta] \\
&\geq \frac{\gamma\Delta^2}{3\ln 2} \quad \forall \text{ small } \Delta, \qquad (33)
\end{aligned}
$$

where the last line follows from the Taylor series expansion of the function $f(u) = (1+u)\ln(1+u) - u$. Thus, using the definition of γ, we get:

$$
|T_Q\cap\mathcal{E}|\cdot D((1+\Delta)\gamma\|\gamma) \geq \frac{\Delta^2}{3\ln 2}\cdot 2^{\log|T_Q\cap\mathcal{E}|-\log|T_Q|+nR} = \frac{\Delta^2}{3\ln 2}\cdot 2^{n[R-R_{\mathcal{E}}(\boldsymbol{x})]} \geq \frac{\Delta^2 2^{n\epsilon}}{3\ln 2} \qquad (34)
$$

where the last inequality follows from the assumption that $R \geq R_{\mathcal{E}}(\boldsymbol{x}) + \epsilon$.

We conclude that for a given $\boldsymbol{x}\in T_Q$, the number of permutations of T_Q for which L_0 or more members of $\mathcal{Y}(\boldsymbol{x})$ come from $\mathcal{E}$, is upper bounded by:

$$
K \leq (|T_Q|+1)!\cdot 2^{nR}\cdot\exp_2\left\{-\frac{\Delta^2 2^{n\epsilon}}{3\ln 2}\right\}
$$

for large n. As stated in the assertion of part (b), Δ can be chosen to be as small as $2^{-n\delta}$ for any $\delta < \epsilon/2$ and still the r.h.s. of the last inequality would decay in a double–exponential rate. Multiplying this bound by the $|T_Q|$ possible choices of $\boldsymbol{x}$, which is an exponential function of n, we deduce that the total number of permutations that have this property for *some* $\boldsymbol{x}\in T_Q$ is still a doubly exponentially small fraction of the total number of permutations, $|T_Q|!$. This conclusion remains unchanged even after taking into account also the permutations for which

$$
\ell \leq (1-\Delta)\frac{|T_Q\cap\mathcal{E}|}{|T_Q|}\cdot 2^{nR}
$$

whose number is also bounded (similarly) by a doubly exponentially small fraction of $|T_Q|!$. Applying this consideration to all types $\{T_Q\}$, we have a complete simulation scheme that is Δ–faithful scheme for $\mathcal{E}$. In fact, we have shown that the vast majority of schemes of the form (25) are Δ–faithful.

4 Multiple Statistical Tests

We now move on to the more general case, where rather than a single statistical test, we have M tests for the events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_M$, which all have to be simultaneously accommodated in

the sense of Section 3. Interestingly, it turns out that as long as M grows at a rate slower than double–exponential in n, the key rate needed for a given $\boldsymbol{x}$, is simply $\max_\ell R_{\mathcal{E}_\ell}(\boldsymbol{x})$, which is the same rate needed for the single test with the smallest $|T_{\boldsymbol{x}} \cap \mathcal{E}_\ell|$ in the set (cf. Section 3). If, on the other hand, n grows double–exponentially, then the minimum needed key rate increases in a manner that will be detailed next.

For each event $\mathcal{E}_\ell$, let us define (similarly to eq. (19)):

$$E_\ell(\boldsymbol{x}) \stackrel{\triangle}{=} \begin{cases} \frac{1}{n}\log|T_{\boldsymbol{x}} \cap \mathcal{E}_\ell| & T_{\boldsymbol{x}} \cap \mathcal{E}_\ell \neq \emptyset \\ \hat{H}(\boldsymbol{x}) & T_{\boldsymbol{x}} \cap \mathcal{E}_\ell = \emptyset \end{cases} \tag{35}$$

and for every $S \in [0, \hat{H}]$, let $M_S \equiv M_S(\boldsymbol{x})$ denote the number of different intersections $\{T_{\boldsymbol{x}} \cap \mathcal{E}_\ell\}_\ell$ created by the events $\{\mathcal{E}_\ell\}$ for which $E_\ell(\boldsymbol{x}) = S$. Note that $M_S(\boldsymbol{x})$ may be non–zero only over a finite set of values of S, namely, the set $\{\frac{1}{n}\log 1, \frac{1}{n}\log 2, \ldots, \frac{1}{n}\log|T_{\boldsymbol{x}}|\}$. Now, define

$$\Gamma_S(\boldsymbol{x}) \stackrel{\triangle}{=} \begin{cases} -\infty & M_S(\boldsymbol{x}) = 0 \\ 0 & M_S(\boldsymbol{x}) = 1 \\ \frac{1}{n}\log\log M_S(\boldsymbol{x}) & M_S(\boldsymbol{x}) > 1 \end{cases} \tag{36}$$

and the rate–function

$$R_0(\boldsymbol{x}) = \hat{H}(\boldsymbol{x}) + \max_S[\Gamma_S(\boldsymbol{x}) - S]. \tag{37}$$

Observe that the *total* number of different subsets of $T_{\boldsymbol{x}}$, whose sizes are all exactly 2^{nS}, grows at the double–exponential rate of $2^{2^{nS}}$, thus the term $\max_S[\Gamma_S(\boldsymbol{x}) - S]$ is (asymptotically) non–positive, which means that $R_0(\boldsymbol{x})$ essentially never exceeds $\hat{H}(\boldsymbol{x})$ (cf. the discussion after Theorem 1). Observe also that as long as the total number of events M grows at a rate that is slower than double–exponential in n, then $\Gamma_S(\boldsymbol{x})$ is very close to zero for all S with $M_S(\boldsymbol{x}) > 0$, and so, $R_0(\boldsymbol{x})$ is indeed, essentially equal to $\max_\ell R_{\mathcal{E}_\ell}(\boldsymbol{x})$, i.e., the rate for the most demanding event in the set. Simple examples for calculating the function $R_0(\cdot)$ appear in Example 3 below and in the construction described in the proof of Theorem 4 in the sequel.

Our direct theorem for multiple tests is the following:

Theorem 3 *Let $\epsilon > 0$ be given and let n be sufficiently large. If $R(\boldsymbol{x}) \geq R_0(\boldsymbol{x}) + \epsilon$ for all $\boldsymbol{x}$, then there exists a simulation scheme that is Δ–faithful simultaneously for all $\mathcal{E}_\ell$, $\ell = 1, 2, \ldots, M$, provided that Δ is as in Theorem 2.*

Proof of Theorem 3. Recall that in the proof of Theorem 2, we have shown that the fraction of 'bad' permutations of $T_{\boldsymbol{x}}$, in the sense of violating (21) for each $\mathcal{E}_\ell$, is upper bounded by

an expression of the double exponential order of

$$\exp_2\{-2^{n(R-R_{\mathcal{E}_\ell})}\} = \exp_2\{-2^{n[R-(\hat{H}-E_\ell)]}\}.$$

Thus, by the union bound, the fraction of permutations that are 'bad' for at least one of the events in the collection, is upper bounded by an expression with the double–exponential order of

$$\begin{aligned}\sum_S M_S \cdot \exp_2\{-2^{n[R-(\hat{H}(\boldsymbol{x})-S)]}\} &\leq |T_{\boldsymbol{x}}| \cdot \max_S M_S \cdot \exp_2\{-2^{n[R-(\hat{H}(\boldsymbol{x})-S)]}\} \\ &\leq |\mathcal{A}|^n \cdot \exp_2\left\{\max_S \left[2^{n\Gamma_S(\boldsymbol{x})} - 2^{n[R-(\hat{H}(\boldsymbol{x})-S)]}\right]\right\} \quad (38)\end{aligned}$$

which continues to decay with n as long as $R(\boldsymbol{x}) - [\hat{H}(\boldsymbol{x}) - S] \geq \Gamma_S(\boldsymbol{x}) + \epsilon$ for all S, or, equivalently, $R(\boldsymbol{x}) \geq R_0(\boldsymbol{x}) + \epsilon$. This completes the proof of Theorem 3.

The converse part is more involved, both conceptually and technically. Note that one cannot expect a straightforward converse to Theorem 3, asserting that for *any* given collection of events, $\{\mathcal{E}_\ell\}_{\ell=1}^M$, $R \leq R_0(\boldsymbol{x}) - \epsilon$ implies the nonexistence of a Δ–faithful simulation scheme w.r.t. $\boldsymbol{x}$, simultaneously for all $\mathcal{E}_\ell$ in this collection. The reason is that $R_0(\boldsymbol{x})$ does not take much account of the geometry of the constellation of the set of events $\{\mathcal{E}_\ell\}$. Intuitively, if all M events are "sufficiently similar," there might be a simulation scheme that accommodates all of them. To demonstrate that this might be the case, consider the following example.

Example 3 *Let $\mathcal{E}_0$ be an arbitrary subset of some type class T_Q, whose size is $|\mathcal{E}_0| = 2^{nS_0}$, where $S_0 \in (0, H_Q)$ is a constant, H_Q denoting $\frac{1}{n}\log|T_Q|$. Now, for $\ell = 1, \ldots, M$, let $\mathcal{E}_\ell = \mathcal{E}_0 \cup \mathcal{G}_\ell$, where each $\mathcal{G}_\ell$ is a subset of $T_Q - \mathcal{E}_0$ whose size is $\delta \cdot 2^{nS_0}$ for some fixed $\delta << \Delta$. Note that all these events are clearly very 'close' to $\mathcal{E}_0$, and indeed, every simulation scheme that is Δ–faithful for $\mathcal{E}_0$ is also $(\Delta+\delta)$-faithful for $\mathcal{E}_\ell$, $\ell = 1, 2, \ldots, M$. The number of such events is*

$$M = \left(\begin{array}{c} |T_Q| - 2^{nS_0} \\ \delta \cdot 2^{nS_0} \end{array} \right)$$

which is of the double exponential order of $2^{2^{nS_0}}$, and so, for all $\boldsymbol{x} \in T_Q$, $\Gamma_S(\boldsymbol{x})$ is essentially equal to S_0 for $S = S_0$ and is $-\infty$ elsewhere. Therefore, for any $\boldsymbol{x} \in T_Q$, we have

$$R_0(\boldsymbol{x}) \approx H_Q + S_0 - S_0 = H_Q,$$

while the simulation scheme for $\mathcal{E}_0$ needs only about $(H_Q - S_0)$ bits per sample within T_Q. Note that S_0 can be chosen arbitrarily close to H_Q, and so, the gap between these two rates can be nearly as large the one between the two extremes of the full range of rates, $(0, H_Q]$.

In view of this observation, a converse to Theorem 3 can only exist for *some* of the collections $\{\mathcal{E}_\ell\}$, intuitively, those where the events are sufficiently "far apart" from each other.[5] But different collections might, in general, have a different rate function, $R_0(\boldsymbol{x})$. Therefore, in order to state a converse theorem in a coherent manner, we have to consider all collections of M events that share the same rate function $R_0(\boldsymbol{x})$. To ensure that such a converse theorem would be compatible with the direct theorem, we first look at the direct theorem, Theorem 3, in a different manner: Instead of characterizing $R_0(\boldsymbol{x})$ for a given set of events as in Theorem 3, we then go the other way around, namely, we start with a given, arbitrary rate function $R_0(\boldsymbol{x})$ and then characterize the class of collections of events, $\{\mathcal{E}_\ell\}$ whose rate function does not exceed $R_0(\boldsymbol{x})$.

Specifically, fix an arbitrary function $R_0(\boldsymbol{x}) \leq \hat{H}(\boldsymbol{x})$, depending on $\boldsymbol{x}$ only via $T_{\boldsymbol{x}}$. Now, consider the class $\mathcal{M}(R_0)$ of all collections of $\{\mathcal{E}_\ell\}$ with the following property: For every $\boldsymbol{x}$ and every $S \in [0, \hat{H}(\boldsymbol{x})]$,

$$\Gamma_S(\boldsymbol{x}) \leq S - [\hat{H}(\boldsymbol{x}) - R_0(\boldsymbol{x})].$$

Then $\mathcal{M}(R_0)$ is indeed the class of collections of events $\{\mathcal{E}_\ell\}$ for which the rate function does not exceed the given $R_0(\boldsymbol{x})$. Theorem 3 then tells us that if $R(\boldsymbol{x}) \geq R_0(\boldsymbol{x}) + \epsilon$ for all $\boldsymbol{x}$, then for every collection of $\{\mathcal{E}_\ell\}_{\ell=1}^M$ in $\mathcal{M}(R_0)$, there exists a simulation scheme that is Δ–faithful simultaneously for all $\mathcal{E}_\ell$ in this collection of events.

The converse theorem is now the following:

Theorem 4 *Let $\epsilon > 0$, $0 \leq \Delta < 1$, and let n be sufficiently large. If $R(\boldsymbol{x}) < R_0(\boldsymbol{x}) - \epsilon$ for some $\boldsymbol{x}$, then there exists a collection of events $\{\mathcal{E}_\ell\}$ in $\mathcal{M}(R_0)$, for which no simulation scheme is Δ–faithful w.r.t. this $\boldsymbol{x}$, simultaneously for all $\mathcal{E}_\ell$ in that collection.*

In the proof of this theorem, it will actually be shown that not only does a problematic collection in $\mathcal{M}(R_0)$ exist when $R(\boldsymbol{x})$ is below $R_0(\boldsymbol{x})$, but moreover, the *vast majority* of the collections, in a certain class, are such. The remaining part of this section is devoted to the proof of Theorem 4.

[5] In fact, for any reasonable set of tests, the events $\mathcal{E}_\ell$ should be separated and diverse enough in order to cover efficiently a wide variety of patterns.

Proof of Theorem 4. We begin with a generic version of the covering lemma whose original version, due to Csiszár and Körner [1], was stated in the context of type covering. The more general version, stated below, will be used in the sequel in a context that is different than type covering, but the proof is in the same spirit as in [1].

Lemma 1 *Let $\mathcal{F}$ be a finite set and let $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_{m_0}$ be subsets of $\mathcal{F}$ such that $\cup_{i=1}^{m_0} \mathcal{F}_i = \mathcal{F}$, and every $u \in \mathcal{F}$ is a member of at least m of the subsets $\{\mathcal{F}_i\}$. Then, $\mathcal{F}$ can also be covered by some sub–collection of s out of the m_0 subsets $\{\mathcal{F}_i\}$, provided that*

$$|\mathcal{F}| \cdot \left(1 - \frac{m}{m_0}\right)^s < 1.$$

Proof of Lemma 1. Let $\{\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s}\}$ denote an arbitrary sub–collection of s subsets of $\mathcal{F}$, where $i_j \in \{1, 2, \ldots, m\}$, $j = 1, 2, \ldots, s$. The number of elements of $\mathcal{F}$ that are *not* covered by $\{\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s}\}$ is given by

$$N(\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s}) = \sum_{u \in \mathcal{F}} \prod_{j=1}^{s} [1 - 1_{\mathcal{F}_{i_j}}(u)].$$

Now, consider a random selection of $\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s}$, where $\{i_j\}$ are drawn independently, each one with a uniform distribution over $\{1, 2, \ldots, m_0\}$. Then, the expected value of the number of uncovered elements of $\mathcal{F}$, w.r.t. this random selection, is given by

$$\begin{aligned} \boldsymbol{E}\{N(\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s})\} &= \sum_{u \in \mathcal{F}} \boldsymbol{E}\left\{\prod_{j=1}^{s} \left[1 - 1_{\mathcal{F}_{i_j}}(u)\right]\right\} \\ &= \sum_{u \in \mathcal{F}} \prod_{j=1}^{s} \boldsymbol{E}\left\{\left[1 - 1_{\mathcal{F}_{i_j}}(u)\right]\right\} \\ &\leq |\mathcal{F}| \cdot \left(1 - \frac{m}{m_0}\right)^s < 1, \end{aligned} \tag{39}$$

where in the second equality we have used the independence of $\{i_j\}$, and where the last inequality is by the hypothesis of the lemma. Since $N(\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s})$ is an integer valued random variable, then $\boldsymbol{E}\{N(\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s})\} < 1$ implies that there must exist at least one sub–collection $\{\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s}\}$ for which $N(\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s}) = 0$, which means that $\mathcal{F}$ is covered by $\{\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s}\}$. This completes the proof of the Lemma.

Comment: Note that the proof of the lemma also tells us (using the Chebychev inequality) that

$$\Pr\{N(\mathcal{F}_{i_1}, \mathcal{F}_{i_2}, \ldots, \mathcal{F}_{i_s}) \geq 1\} \leq |\mathcal{F}| \cdot \left(1 - \frac{m}{m_0}\right)^s,$$

in other words, the r.h.s. is an upper bound on the fraction of constellations $\{\mathcal{F}_{i_j}\}_{j=1}^{s}$ which do not cover $\mathcal{F}$. If this number is significantly smaller than one, then it can be argued that not only does a cover of s subsets exist, but moreover, most of the constellations of s subsets provide a cover.

Returning to the proof of Theorem 4, let $R_0(\cdot)$ be given, and let $\boldsymbol{x}$ be a vector for which $R(\boldsymbol{x}) \leq R_0(\boldsymbol{x}) - \epsilon$. For a given $0 \leq \Gamma \leq R_0(\boldsymbol{x})$, let

$$E_0 \equiv E_0(\boldsymbol{x}) \stackrel{\triangle}{=} \hat{H}(\boldsymbol{x}) - R_0(\boldsymbol{x}) + \Gamma.$$

Note that $\Gamma \leq E_0 \leq \hat{H}(\boldsymbol{x})$. Consider the collection of all events $\{\mathcal{E}_\ell\}$, each containing exactly 2^{nE_0} members of $T_{\boldsymbol{x}}$. The total number of such events is

$$M_0 \stackrel{\triangle}{=} \left(\begin{array}{c} |T_{\boldsymbol{x}}| \\ 2^{nE_0} \end{array} \right).$$

Now, for every subcollection of size $M = 2^{2^{n\Gamma}} \leq M_0$, the rate function is obviously, $\hat{H}(\boldsymbol{x}) + \max_S[\Gamma_S(\boldsymbol{x}) - S] = \hat{H}(\boldsymbol{x}) + \Gamma - E_0 \equiv R_0(\boldsymbol{x})$, and so, it is a member of $\mathcal{M}(R_0)$. We wish to show that at least one such subcollection of events satisfies the assertion of Theorem 4 provided that $R \leq R_0(\boldsymbol{x}) - \epsilon$.

Now, for a given simulation scheme ϕ and the given $\boldsymbol{x}$, let $\mathcal{Y}(\boldsymbol{x}) \stackrel{\triangle}{=} \{\boldsymbol{y} = \phi(\boldsymbol{x}, \boldsymbol{u}) : \ \boldsymbol{u} \in \{0,1\}^k\}$ (where now ϕ is not necessarily of the form (25)). We next use the covering lemma in the following manner: The set $\mathcal{F}$ of the lemma is the set of all choices of $\mathcal{Y}(\boldsymbol{x})(= u)$ for the given $\boldsymbol{x}$. Each subset $\mathcal{F}_\ell$ is the subset of choices of $\mathcal{Y}(\boldsymbol{x})$ which are 'bad' for the event $\mathcal{E}_\ell$ in the sense that $|\mathcal{Y}(\boldsymbol{x}) \cap \mathcal{E}_\ell|/|\mathcal{Y}(\boldsymbol{x})| \geq (1+\Delta)|T_{\boldsymbol{x}} \cap \mathcal{E}_\ell|/|T_{\boldsymbol{x}}|$. There is a total number of $M_0(= m_0)$ subsets $\{\mathcal{F}_\ell\}$, but if we cover $\mathcal{F}$ with a subcollection of $M(= s)$ such subsets, this means that any choice of $\mathcal{Y}(\boldsymbol{x})$ is 'bad' for at least one $\mathcal{E}_\ell$ corresponding to this subcollection of size M, and so, no scheme would be Δ-faithful w.r.t. $\boldsymbol{x}$ simultaneously for all M members of this subcollection.

To use the lemma then, we first need a lower bound to the analogue of m/m_0, namely, the relative number of sets of size 2^{nE_0} for which a given $\mathcal{Y}(\boldsymbol{x})$ is 'bad'. A simple combinatorial consideration leads to the following expression:

$$\frac{\sum_{\ell \geq L_0} \left(\begin{array}{c} 2^{nR} \\ \ell \end{array} \right) \cdot \left(\begin{array}{c} |T_{\boldsymbol{x}}| - 2^{nR} \\ 2^{nE_0} - \ell \end{array} \right)}{\left(\begin{array}{c} |T_{\boldsymbol{x}}| \\ 2^{nE_0} \end{array} \right)}. \tag{40}$$

where $L_0 = (1+\Delta)2^{nR} \cdot 2^{nE_0}/|T_{\boldsymbol{x}}|$. This is exactly the same kind of expression as the one for K (cf. eq. (27)) that has been handled in the proof of Theorem 2, part (b), except that $|T_Q \cap \mathcal{E}|$ of the former expression is now replaced by 2^{nR}, and 2^{nR} of (27) is in turn replaced by 2^{nE_0}. Accordingly, we now redefine $\gamma = 2^{nE_0}/|T_{\boldsymbol{x}}|$ and $\theta = 2^{nR}/|T_{\boldsymbol{x}}|$ in analogy to the previous derivation. The only difference is that now we need a lower bound on this expression rather than an upper bound. It is easy to verify that all steps of upper bounding K in the proof of Theorem 2 are tight in the double exponential scale. The only two points that are not straightforwardly so are that: (i) the lower bound to the divergence (33) is exponentially tight, and (ii) the term $(1-\theta)D(\beta_0\|\gamma)$, that has been omitted in the second step of eq. (30), is negligible compared to the remaining term $\theta D(\alpha_0\|\gamma)$.

As for (i), the binary divergence defined in the proof of Theorem 2, can easily[6] be shown to be upper bounded by

$$D(a\|b) \leq \frac{(a-b)^2}{b(1-b)\ln 2}.$$

Therefore,

$$\begin{aligned} |T_{\boldsymbol{x}}| \cdot \theta D(\alpha_0\|\gamma) &\leq |T_{\boldsymbol{x}}| \cdot \theta \cdot \frac{\gamma^2\Delta^2}{\gamma(1-\gamma)\ln 2} \\ &= \frac{|T_{\boldsymbol{x}}| \cdot \theta\gamma\Delta^2}{(1-\gamma)\ln 2} \end{aligned} \tag{41}$$

which is again of the exponential order of $\exp\{n[R-(\hat{H}-E_0)]\}$.

Regarding (ii), we have

$$\begin{aligned} |T_{\boldsymbol{x}}| \cdot (1-\theta)D(\beta_0\|\gamma) &= |T_{\boldsymbol{x}}| \cdot (1-\theta)D\left(\gamma(1-\theta(1+\Delta))/(1-\theta)\|\gamma\right) \\ &\leq |T_{\boldsymbol{x}}| \cdot (1-\theta) \cdot \frac{\gamma^2\theta^2\Delta^2}{(1-\theta)^2\gamma(1-\gamma)\ln 2} \\ &= \frac{|T_{\boldsymbol{x}}| \cdot \gamma\theta^2\Delta^2}{(1-\theta)(1-\gamma)\ln 2} \end{aligned} \tag{42}$$

which is of the exponential order of $\exp\{n[2R-(2\hat{H}-E_0)]\}$, and hence always less than or equal to $\exp\{n[R-(\hat{H}-E_0)]\}$ since $R \leq \hat{H}$. Thus, this term does not exponentially dominate the term $|T_{\boldsymbol{x}}| \cdot \theta D(\alpha_0\|\gamma)$.

We conclude that m/m_0 is lower bounded by an expression of the double–exponential order of $\exp_2\{-2^{n[R-(\hat{H}-E_0)]}\}$. Now, $R \leq R_0(\boldsymbol{x}) - \epsilon$ implies that $\Gamma \geq R-(\hat{H}-E_0)+\epsilon$, and so, the covering lemma applies as follows:

$$|\mathcal{F}| \cdot \left(1-\frac{m}{m_0}\right)^s \doteq \left(\begin{array}{c} |T_{\boldsymbol{x}}| \\ 2^{nR} \end{array}\right) \cdot \left[1-\exp_2\{-2^{n[R-(\hat{H}-E_0)]}\}\right]^{\exp_2\{2^{n\Gamma}\}}$$

[6] Use the inequality $\ln u \leq u-1$ for both logarithmic terms.

$$
\begin{aligned}
&\leq |T_{\boldsymbol{x}}|! \cdot \exp\left\{\exp_2\{2^{n\Gamma}\} \ln\left[1 - \exp_2\{-2^{n[R-(\hat{H}-E_0)]}\}\right]\right\} \\
&\leq |\mathcal{A}^n|^{|\mathcal{A}^n|} \cdot \exp\left\{-\exp_2\{2^{n\Gamma}\} \cdot \exp_2\{-2^{n[R-(\hat{H}-E_0)]}\right\} \\
&\leq \exp\left\{(n \ln|\mathcal{A}|) \cdot |\mathcal{A}|^n - \exp_2\{2^{n[R-(\hat{H}-E_0)+\epsilon]} - 2^{n[R-(\hat{H}-E_0)]}\right\} \\
&= \exp\left\{(n \ln|\mathcal{A}|) \cdot |\mathcal{A}|^n - \exp_2\{2^{n[R-(\hat{H}-E_0)]}(2^{n\epsilon}-1)\}\right\} \\
&\leq \exp\left[(n \ln|\mathcal{A}|) \cdot |\mathcal{A}|^n - \exp_2\{(2^{n\epsilon}-1)/(1+\Delta)\}\right] \to 0, \qquad (43)
\end{aligned}
$$

where the notation $\dot{\leq}$ symbolizes the fact that $\frac{1}{n}\log\log(m_0/m) \leq R - (\hat{H} - E_0) + \delta$ for an arbitrarily small $\delta > 0$ and all n sufficiently large, and where in the last inequality, we have assumed that $R \geq \hat{H} - E_0 - \frac{1}{n}\log(1+\Delta)$, as otherwise, by the converse part of Theorem 2, no simulation scheme can be Δ-faithful even for each one of the events *separately*, let alone the whole set of M events simultaneously. Thus, we have proved the existence of a cover of $M = \exp\{2^{n\Gamma}\}$ events $\{\mathcal{E}_\ell\}$ in $\mathcal{M}(R_0)$. Observe also that in this case, the comment following the proof of the covering lemma is in effect, so in fact, we have proved that most subcollections of $M = \exp\{2^{n\Gamma}\}$ in $\mathcal{M}(R_0)$ form covers. This completes the proof of Theorem 4.

5 Conclusion

In this paper, we made an attempt to characterize achievable key rates for universal simulation of random data based on a training vector, where these key rates are allowed to depend on the training vector. For a single test, we have stated and proved a theorem asserting that the rate function $R_{\mathcal{E}}(\boldsymbol{x})$ (defined in eq. (20)) is the minimum key rate needed to guarantee simulation performance that is essentially as good as if there was unlimited supply of key bits.

We then extended the paradigm to that of multiple statistical tests for events $\{\mathcal{E}_\ell\}$, that all have to be accommodated at the same time. For this case, we have characterized the minimum key rate function $R_0(\boldsymbol{x})$ (defined in eq. (37)), that depends on the entire set of events. One interesting conclusion that we have drawn from this result, is that as long as the number of tests is less than double–exponential, the minimum key rate needed for multiple tests remains essentially as small as the rate function of a single test, the most demanding one in the set (namely, $\max_\ell R_{\mathcal{E}_\ell}(\boldsymbol{x})$), independently of the structure of that set.

For the case where the size of the set of events is double–exponential and extra key rate may be needed, our argument regarding the minimality of $R_0(\boldsymbol{x})$ as an achievable key rate, is

somewhat weaker than for a single test, or relatively small sets of tests. Since $R_0(\boldsymbol{x})$ does not depend strongly on the geometry of the set of tests (i.e., the similarities and dissimilarities between the various events in the set), our argument regarding the minimality of $R_0(\boldsymbol{x})$ is, in principle, a worst case (minimax) argument: $R_0(\boldsymbol{x})$ is essentially the minimum key rate for the *worst* constellation of events in a certain class of constellations that we define. In other words, it is the tightest bound that can be obtained among all bounds that depend on $\{\mathcal{E}_\ell\}$ through the same information as $R_0(\boldsymbol{x})$. It is not tight for constellations where the events are too 'close' to each other, but such constellations may not be reasonable anyhow for testing simulated data. In fact, the converse result is stronger than merely a minimax result because, we show that it applies to the vast majority of constellations in a very large class. This happens to be the case because in most of these constellations, the events are fairly 'far apart'.

One interesting direction for further research might be to try to refine the formula of $R_0(\boldsymbol{x})$ so as to incorporate more detailed information about the geometry of the set of events, although a much more detailed geometrical description might be difficult to extract because of the double–exponential number of events.

Acknowledgement

Useful discussions with Marcelo Weinberger and with Tsachy Weissman are acknowledged.

References

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

[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. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. IT–39, no. 3, pp. 752–772, May 1993.

[4] N. Merhav and M. J. Weinberger, "On universal simulation of information sources using training data," to be submitted to *IEEE Trans. Inform. Theory*. Also, available at `www.ee.technion.ac.il/~merhav`.

[5] Y. Steinberg and S. Verdú, "Channel simulation and coding with side information," *IEEE Trans. Inform. Theory*, vol. IT–40, no. 3, pp. 634–646, May 1994.

[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.

[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, September 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, September 2000.