On Universal Simulation of Information Sources Using Training Data

Neri Merhav[1], Marcelo J. Weinberger
Information Theory Research Group
HP Laboratories Palo Alto
HPL-2002-263
September 20th , 2002*

E-mail: merhav@ee.technion.ac.il, marcelo@hpl.hp.com

random number generators, random process simulation, mutual information, typical sequences

We consider the problem of universal simulation of an unknown random process, or 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 is 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. We derive lower bounds on the mutual information that are shown to be achievable by conceptually simple algorithms proposed herein. We show that the behavior of the minimum achievable mutual information depends critically on the relative amount of random bits and on the lengths of the input sequence and the output sequence.

While in the ordinary (non-universal) simulation problem, the number of random bits per symbol must exceed the entropy rate H of the source in order to simulate it faithfully, in the universal simulation problem considered here, faithful preservation of the probability law is not a problem, yet the same minimum rate of H random bits per symbol is still needed to essentially eliminate the statistical dependency between the input sequence and the output sequence. The results are extended to more general information measures.

* Internal Accession Date Only Approved for External Publication
[1] With the Electrical Engineering Department, Technion, Israel Institute of Technology, Technion City, Haifa, 32000, Israel. This work was done while visiting Hewlett-Packard Laboratories, Palo Alto, USA

On Universal Simulation of Information Sources Using Training Data

Neri Merhav* Marcelo J. Weinberger†

September 16, 2002

Abstract

We consider the problem of universal simulation of an unknown random process, or 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 is 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. We derive lower bounds on the mutual information that are shown to be achievable by conceptually simple algorithms proposed herein. We show that the behavior of the minimum achievable mutual information depends critically on the relative amount of random bits and on the lengths of the input sequence and the output sequence.

While in the ordinary (non–universal) simulation problem, the number of random bits per symbol must exceed the entropy rate H of the source in order to simulate it faithfully, in the universal simulation problem considered here, faithful preservation of the probability law is not a problem, yet the same minimum rate of H random bits per symbol is still needed to essentially eliminate the statistical dependency between the input sequence and the output sequence. The results are extended to more general information measures.

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

*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. 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 Introduction

Simulation of random processes, or information sources, is about artificial generation of random data with a prescribed probability law, by using a certain deterministic mapping from a source of purely random (independent equally likely) bits into sample paths. The simulation problem finds its applications in speech and image synthesis, texture production (e.g., in image compression), and generation of noise for purposes of simulating communication systems.

In the last decade, the simulation problem of sources and channels, as well as its relation to other problem areas in information theory, has been investigated by several researchers. In particular, Han and Verdú [12] posed 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. It was shown in [12] that if convergence is defined in terms of variational distance, the resolvability is given by the *sup–entropy rate*, which coincides with the ordinary entropy rate in the stationary ergodic case. In [14], the dual problem of channel simulation was studied, where 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 [17] for further developments). In [15], the results of [12] 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. In [11] and [16], concrete algorithms for source simulation 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 this paper, we relax the assumption of perfect knowledge of the target probability law and we focus on the following universal version of the simulation problem for finite–alphabet sources: The target source P to be simulated, 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 unknown, but we are given a training sequence $X^m = (X_1, \ldots, X_m)$ that has emerged from this unknown

source. We are also provided with 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 $Y^n = (Y_1, \ldots, Y_n)$, 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.

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 $P \in \mathcal{P}$.

Condition C1 means that Y^n is actually a randomized function of X^m, where the amount of randomization is given by the number k of random bits available. We can think of this as a conditional distribution of Y^n given X^m corresponding to a *channel* of limited randomness from X^m to Y^n. Note that every channel of this type is characterized by the property that its conditional probabilities are all integer multiples of 2^{-k}. This implies that the conditional entropy $H(Y^n|X^m = x^m)$ of Y^n given any realization $x^m = (x_1, \ldots, x_m)$ of X^m, cannot exceed k, and so neither can $H(Y^n|X^m)$.

In Condition C2, we require exact preservation of the probability law P, which is different from the ordinary simulation problem, where only a good approximation of P is required. Note that for $n \leq m$, this requirement can always be satisfied even when $k = 0$, trivially by letting $Y^n = X^n = (X_1, \ldots, X_n)$. For $n > m$, it is impossible to meet Condition C2 no matter how large k may be.[1] Therefore, we will always assume that $n \leq m$. Note that the "plug–in" approach (which is perhaps the first idea one may think of in the context of this universal simulation problem), where P is first estimated from X^m and then used instead of the true P as in ordinary simulation, may not meet Condition C2.

It should be pointed out that a more general setting of the problem could have been formalized, where Condition C2 is relaxed and a certain level of tolerance in approximating P is allowed (as is done in non-universal simulation). Nonetheless, we preferred, in this initial work on the universal simulation problem, to focus our study only on tradeoffs

[1] Had it been possible to generate sequences longer than m with perfect preservation of the unknown P, we could, for example, estimate P with arbitrarily small error, which is contradictory to any nontrivial lower bound on the estimation error corresponding to m observations.

among the mutual information, the amount of randomness, and the richness of the class of sources, leaving the additional factor of the approximation tolerance, perhaps for future work. We will only comment here that if such a tolerance is allowed and the divergence $D(P\|P')$ is chosen to be the distance measure between the desired probability law P and the approximate one P' (of the simulated process), then in the case where $\mathcal{P}$ is the class of finite-alphabet memoryless sources (as well as in some other parametric families), it is possible to make $I(X^m;Y^n)=0$, simply by taking P' to be an appropriate mixture of all members of $\mathcal{P}$ (independently of X^m), thus making $D(P\|P')=O(\log n)$, like in universal coding [5].

Condition C3 is actually meant to avoid an uninteresting trivial solution like $Y^n=X^n$, that we mentioned earlier. Loosely speaking, we would like the sample path Y^n that we generate to be as 'original' as possible, namely, with as small a statistical dependence as possible on the input training sequence X^m. There are a few reasons why small dependence between the two sequences is desirable. For example, one of the applications of the universal simulation problem considered here is lossy compression of textures: Since the exact details of texture are immaterial to the human eye and only the 'statistical behavior' is important, a plausible approach to texture compression is to compress (or even transmit uncodedly) a relatively short sample texture segment (training sequence), and let the decoder synthesize 'statistically similar' patterns, instead of the missing texture segments, by repetitive application of the function ϕ (of course, with different k-vectors of random bits every time). Condition C3 may help to maintain the regular structure of the texture and to avoid undesired periodicities that may appear if each Y^n depends strongly on X^m. (A similar comment applies also to voice coding methods of speech signals, e.g., linear predictive coding and its variants.) Another reason for which one may be interested in Condition C3 is that if we use the same training vector X^m to generate many output vectors (again, by repetitive use of ϕ), as may be the case in experimental simulation of algorithms and systems, then weak dependence between the training sequence and each output sequence would yield weak dependence among the different output sequences, which is obviously desirable.

The goal of minimizing $I(X^m;Y^n)$ *simultaneously* for all sources in $\mathcal{P}$, as expressed in Condition C3, might seem somewhat too ambitious at first glance. However, as we show in the sequel, this in fact can be done. At this point, one may argue that the choice of the mutual information as a measure of dependence between two random variables is

somewhat arbitrary as there are many possible measures of dependence. Indeed, as we show in Section 5, our results apply not only to the ordinary mutual information, but also to a much wider class of dependency measures, namely, the class of generalized mutual information measures proposed by Ziv and Zakai [18],[19].

Our main results in this paper, are in characterizing the smallest achievable value of the mutual information as a function of n, m, k, and the entropy rate H of the source P, and in demonstrating simulation schemes that asymptotically achieve these bounds. In some special cases, the simulation schemes are strictly optimum, namely, they meet the corresponding lower bound precisely (not merely asymptotically) and simultaneously for all sources in the class $\mathcal{P}$. It turns out that the asymptotically achievable lower bound on $I(X^m;Y^n)$, which we will denote here by $I_{\min}$, has at least four types of asymptotic behavior, depending on the relations between n, m, k, and H. Specifically, letting R denote the random bit rate k/n (relative to the number of *output* symbols), and assuming, for simplicity, that $\mathcal{P}$ is the class of all memoryless sources with a finite alphabet $\mathcal{A}$, we have the following: If $R < H$, then $I_{\min} = n(H-R)$, i.e., there is linear growth with n, independently of m (of course, as long as $m \geq n$). If $R > H$, then the important factor is the growth rate of m relative to n: If $m = n$ or if $m/n \searrow 1$, then $I_{\min}$ grows logarithmically with n, according to $\frac{|\mathcal{A}|-1}{2}\log n$ for large n (which means that the normalized mutual information $I_{\min}/n$ tends to zero), whereas if $m/n \to C$, where C is a constant strictly larger than unity, then $I_{\min}$ tends to the constant $\frac{|\mathcal{A}|-1}{2}\log\frac{C}{C-1}$. Finally, if $m/n \to \infty$, our lower bound vanishes as $n \to \infty$, and its achievability is shown when $\log m = o(n)$. The last case is largely equivalent to the case where P is known.

The above results will actually be shown in a more general setting, where $\mathcal{P}$ is a *parametric subfamily* of the class of all memoryless sources of a given finite alphabet $\mathcal{A}$. In this setting, the role of $|\mathcal{A}|$ is played by a quantity related to the number of free parameters defining the class, with a possible reduction in the 'cost of universality' (i.e., the minimum achievable mutual information), depending on the richness of the family. Moreover, as we demonstrate in Section 5, our derivations and results extend quite straightforwardly to sources with memory, such as specific classes of Markov and finite-state sources.

Notice that, according to our results, the number of random bits needed to essentially remove the dependency between X^m and Y^n grows with the output length n, while we would like to make the input length m as large as possible in order to faithfully represent

the characteristics of the data. Thus, in applications where it is sufficient to simulate a sequence shorter than the available training data (e.g., the removal of a small object in an image where the background is a large texture), one would indeed use $n < m$, rather than generating the maximum possible number m of samples for which P can still be preserved.

The outline of the remaining part of this paper is as follows: In Section 2, we give our notation conventions and a formal description of the problem. In Section 3, we focus on the case $n = m$, which is presented first due to its simplicity and its special properties. In Section 4, we study the case $n < m$, first for an unlimited supply of random bits ($k = \infty$, or equivalently, $R = \infty$), and then for finitely many random bits. Finally, in Section 5, we discuss the aforementioned extensions.

2 Notation, Preliminaries, 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 some other sets, will be denoted by calligraphic letters. Similarly, random vectors, their realizations, and their alphabets, will be denoted, respectively, by capital letters, the corresponding lower case letters, and calligraphic letters, all superscripted by their dimensions. For example, the random vector $X^m = (X_1, \ldots, X_m)$, (m – positive integer) may take a specific vector value $x^m = (x_1, \ldots, x_m)$ in $\mathcal{A}^m$, the mth order Cartesian power of $\mathcal{A}$, which is the alphabet of each component of this vector. For $i \leq j$ (i, j – integers), x_i^j will denote the segment $(x_i, \ldots, x_j)$, where for $i = 1$ the subscript will be omitted.

Let $\mathcal{P}$ denote a parametric subfamily of the class of all discrete memoryless sources (DMSs) with a finite alphabet $\mathcal{A}$, and let d denote the number of parameters describing $\mathcal{P}$. A particular DMS in $\mathcal{P}$, defined by a d-dimensional parameter vector θ taking values over some parameter space Ω, will be denoted by P_θ. However, in a context where the parameter value is either fixed or irrelevant, we will omit it, denoting a source in $\mathcal{P}$ simply by P. For a given positive integer m, let $X^m = (X_1, X_2, \ldots, X_m)$, $X_i \in \mathcal{A}$, $i = 1, \ldots, m$, denote an m-vector drawn from P, namely, $\Pr\{X_i = x_i,\ i = 1, \ldots, m\} = \prod_{i=1}^m P(x_i) \stackrel{\triangle}{=} P(x^m)$ for every $(x_1, \ldots, x_m)$, $x_i \in \mathcal{A}$, $i = 1, \ldots, m$. 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 $U^k = (U_1, \ldots, U_k)$, $U_i \in \mathcal{B} \stackrel{\triangle}{=} \{0, 1\}$, $i = 1, \ldots, k$, denote a string of k random bits, drawn from the binary symmetric source, independently of X^m.

We shall define the *type class* T_{x^m} of a vector x^m as the set of all vectors $\tilde{x}^m \in \mathcal{A}^m$ such that $P(\tilde{x}^m) = P(x^m)$ simultaneously for *every* source $P \in \mathcal{P}$. Accordingly, we shall denote by T_{X^m} the (random) type class of a random vector X^m drawn from a DMS $P \in \mathcal{P}$. The set of all type classes of vectors in $\mathcal{A}^m$ will be denoted by $\mathcal{T}^m$, and its cardinality by $N(\mathcal{P}, m)$. For example, in case $\mathcal{P}$ is the entire class of DMSs ($d = |\mathcal{A}| - 1$), the type class T_{x^m} coincides with the set of all vectors having the same empirical probability mass function (EPMF) as x^m, where the EPMF is the vector $Q_{x^m} = \{Q_{x^m}(a), a \in \mathcal{A}\}$ and $Q_{x^m}(a)$ is the relative frequency of the letter $a \in \mathcal{A}$ in the vector x^m. In this case, $N(\mathcal{P}, m) = (m + |\mathcal{A}| - 1)!/[(|\mathcal{A}| - 1)!m!]$. The type classes in this special case will be referred to as *elementary* type classes. Notice that for any family $\mathcal{P}$, the type classes are given by unions of elementary type classes. As a result, we will rely quite heavily on the method of types [8]. Similar notations will be used for types of sequences y^n, with m, x, and X being replaced by n, y and Y, respectively.

Next, for every type class $T \in \mathcal{T}^m$, we define

$$P_\theta(T) \stackrel{\triangle}{=} \sum_{\tilde{x}^m \in T} P_\theta(\tilde{x}^m) = |T| \cdot P_\theta(x^m) \tag{1}$$

where x^m is a sequence in T. Given some enumeration of $\mathcal{T}^m$, let $T^{(1)}, T^{(2)}, \ldots, T^{(N(\mathcal{P},m))}$ denote the corresponding type classes. For each j, $1 \le j \le N(\mathcal{P}, m)$, $P_\theta(T^{(j)})$ can be regarded as a function of the d-dimensional parameter vector θ defining $P_\theta \in \mathcal{P}$. In the sequel, we will assume that the class of sources $\mathcal{P}$ satisfies the following assumption:

A1. The set $\{P_\theta(T^{(j)})\}_{j=1}^{N(\mathcal{P},m)}$ *(as functions of* θ*) is linearly independent over* Ω.

Assumption A1 is satisfied for a broad class of parametric families, including any exponential family when Ω contains an open subset of $\mathbb{R}^d$. In this case, the probability of a sequence x^m takes the form

$$\begin{aligned} P_\theta(x^m) &= \prod_{i=1}^m P_\theta(x_i) = \prod_{i=1}^m \exp_2\{<\theta, \tau(x_i)> - \psi(\theta)\} \\ &= \exp_2\{m[<\theta, \boldsymbol{\tau}(x^m)> - \psi(\theta)]\} \end{aligned} \tag{2}$$

where the d-dimensional vector $\boldsymbol{\tau}(x^m) \stackrel{\triangle}{=} m^{-1} \sum_{i=1}^m \tau(x_i)$ is a *minimal* sufficient statistic [7, Chapter 2], $<u, v>$ denotes the inner product of the vectors u and v, and $\psi(\theta)$ guarantees that the probabilities sum up to unity. By the definition of a type class and the minimality of the sufficient statistic, $T_{x^m} = T_{\tilde{x}^m}$ if and only if $\boldsymbol{\tau}(x^m) = \boldsymbol{\tau}(\tilde{x}^m)$. Thus, for any $T^{(j)} \in \mathcal{T}^m$,

$$P_\theta(T^{(j)}) = |T^{(j)}| \exp_2\{m[<\theta, \boldsymbol{\tau}_j> - \psi(\theta)]\}$$

where $\boldsymbol{\tau}_j$ is the value of the sufficient statistic for any $x^m \in T^{(j)}$. Since a *finite* set of exponential functions is linearly independent over any open set, the validity of Assumption A1 follows, as it suffices to consider the collection $\{e^{m<\theta,\boldsymbol{\tau}_j>}\}_{j=1}^{N(\mathcal{P},m)}$, in which any two functions are distinct, over an open subset of Ω. In case the entries of $\boldsymbol{\tau}_j$ are integers, after reparametrization, this is in fact a collection of (multivariate) monomials. As a simple example, consider the case in which $\mathcal{P}$ is the Bernoulli family, parametrized by the probability p of a one, and $\Omega = (0,1)$. Let $T^{(j)}$ denote the type of a sequence containing $j-1$ ones, $1 \le j \le n+1$. Here, taking $\log\frac{p}{1-p}$ as the new parameter, the functions $P_\theta(T^{(j)})$ are monomials of degree $j-1$, which are linearly independent over Ω.

The model of (2) covers families of practical interest, which have more structure than the entire class of DMSs. For example, we can think of a family of 'symmetric' discrete sources for which the symbols are grouped by pairs (with the possible exception of one symbol in case of odd cardinality), with both symbols in a pair having the same probability. In practice, such sources can result from quantizing a symmetric density. With three quantization regions, this is a single-parameter ternary source ($d=1$) in which two of its symbols have equal probabilities. More generally, we can think of probabilities that are proportional to $e^{\theta\beta_h}$, $h=0,1,2$, where θ is a scalar parameter and β_0,β_1,β_2 are fixed real numbers. Clearly, in this case, the type classes are given by sets of sequences such that $\sum_{h=0}^{2} n_h\beta_h = \beta$, where n_0,n_1,n_2 are the number of symbol occurrences and β is some constant. Notice that when the constants β_h are positive integers (as in the particular case of *geometric* distributions), the type classes are larger than the elementary type classes. However, when the ratio $(\beta_1-\beta_0)/(\beta_2-\beta_0)$ is irrational (that is, in geometric terms, the differences are *incommensurable*), all the sequences in a type must have exactly the same values for n_h. Therefore, the type classes coincide with the elementary type classes. It will be shown in Subsection 3.2 that, in the context of universal simulation, the richness of the family is determined, in fact, by the structure of the type classes (rather than by the number of parameters).

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

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

The expectation operator, denoted $\boldsymbol{E}\{\cdot\}$, will be understood to be taken with respect to (w.r.t.) the joint distribution $P \times W$ of (X^m, Y^n). The notation $\boldsymbol{E}_\theta\{\cdot\}$ will also be used, in case we want to emphasize the dependency on the parameter θ.

Finally, let $I(X^m; Y^n)$ denote the mutual information between X^m and Y^n that is induced by the source P and the channel W (or, equivalently, the mapping ϕ).

We seek a mapping ϕ that meets conditions C1–C3 that were itemized in the Introduction, and are re–stated here more formally:

C1. The mapping is independent of P.

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

$$\Pr\{Y^n = y^n\} \stackrel{\triangle}{=} \sum_{x^m} W(y^n|x^m) \prod_{i=1}^{m} P(x_i) = P(y^n) = \prod_{j=1}^{n} P(y_j). \tag{3}$$

C3. The mapping ϕ minimizes $I(X^m; Y^n)$ simultaneously for all $P \in \mathcal{P}$.

When referring to the special case of an unlimited supply of random bits ($k = \infty$), it will be understood that every channel W from X^m to Y^n is implementable. In this case, we shall no longer mention the mapping ϕ explicitly, but only the channel W. As mentioned in the Introduction, in the general case, we shall 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 Y^n.

3 The Case $n = m$

We begin with the special case where $n = m$, which is easier and for which the results are more explicit and stronger than for the case $n < m$ that will be treated in Section 4.

3.1 Main Theorem for the Case $n = m$

Our first theorem gives a lower bound, which is achievable within a fraction of a bit for every n by a certain mapping ϕ^*. An asymptotic analysis will later show that, in fact, for $R \neq H$, the gap between the lower bound and the upper bound vanishes exponentially fast.[2] Our upper bound is based on the following explicit construction of ϕ^*: Given x^n, enumerate all

[2]The fact that our asymptotic analysis covers the cases $R < H$ and $R > H$, but not $R = H$, is analogous to performance analyses in source coding, for compression ratios either strictly smaller or strictly larger than H, and in channel coding, where the information rate is never assumed to be *exactly* equal to channel capacity.

members of T_{x^n} in an arbitrary order, for which the index of $\tilde{x}^n \in T_{x^n}$ (starting from zero for the first sequence) is denoted by $J_n(\tilde{x}^n) \in \{0, 1, \ldots, |T_{x^n}| - 1\}$, and define

$$y^n = \phi^*(x^n, u^k) \stackrel{\triangle}{=} J_n^{-1}\left(J_n(x^n) \oplus \sum_{i=1}^{k} 2^{i-1} u_i\right) \tag{4}$$

where J_n^{-1} denotes the inverse map from $\{0, 1, \ldots, |T_{x^n}| - 1\}$ to T_{x^n}, $\oplus$ denotes addition modulo $|T_{x^n}|$, and the sum over i is taken under the ordinary integer arithmetic. This scheme attempts at randomly selecting a sequence in T_{x^n} as uniformly as possible. Notice that if T_{x^n} contains less than 2^k elements, multiple sequences u^k may map to the same sequence y^n, but for two given sequences y^n, the corresponding numbers of originating sequences u^k will differ by at most one. This 'wrapping around' provides better resolution for achieving a uniform conditional probability assignment (and, as a result, smaller mutual information) than the alternative of discarding those random bits in excess of $\lceil \log |T_{x^n}| \rceil$ bits.[3] The complexity of this simulation scheme is determined by the type class enumeration step. In the case of elementary type classes, lexicographic enumeration can be efficiently done, as shown, e.g., in [6]. In the general case, the problem can often be reduced to the enumeration of elementary type classes, or to an equally efficient scheme based on the general formula given in [6, Proposition 2] for the enumeration of generic subsets of $\mathcal{A}^n$.

Theorem 1 *Let $\mathcal{P}$ satisfy Assumption A1. Then,*

(a) For every mapping ϕ that satisfies conditions C1 and C2,

$$I(X^n; Y^n) \geq nH - \boldsymbol{E} \min\{nR, \log |T_{X^n}|\}. \tag{5}$$

(b) The mapping ϕ^* defined by eq. (4) satisfies conditions C1 and C2 and yields

$$I(X^n; Y^n) \leq nH - \boldsymbol{E} \min\{nR, \log |T_{X^n}|\} + \sigma \tag{6}$$

where $\sigma \stackrel{\triangle}{=} 1 - \log e + \log \log e \approx 0.086$.

Comment: Interestingly, the gap σ coincides with the constant that appears in the Gallager upper bound on the redundancy of the Huffman code [10]. Here, σ is an upper bound on the difference between the entropies of a dyadic distribution with highest entropy on an arbitrary alphabet, and the uniform distribution on that alphabet.

[3] Another alternative would be to use only $\lfloor \log |T_{x^n}| \rfloor$ bits, covering every sequence in T_{x^n} *at most* once, thus achieving a uniform distribution over a possibly *smaller* set. However, the resulting mutual information would be, in general, larger.

The remaining part of this subsection is devoted to the proof of Theorem 1.

Proof. Our first step will be to show that Y^n *must* be of the same type class as X^n in order to satisfy Condition C2. To this end, we first observe that every channel W from X^n to Y^n induces a channel $\tilde{W}$ from T_{X^n} to T_{Y^n}, defined in the following manner: Given an input type class T_{x^n}, select X^n under a uniform distribution within T_{x^n}, then apply the channel W from X^n to Y^n, and finally, extract the type T_{Y^n} of the resulting Y^n. Stated mathematically,

$$\tilde{W}(T_{y^n}|T_{x^n}) = \frac{1}{|T_{x^n}|} \sum_{\tilde{x}^n \in T_{x^n}} \sum_{\tilde{y}^n \in T_{y^n}} W(\tilde{y}^n|\tilde{x}^n). \tag{7}$$

Clearly, by the definitions (1) and (7), Condition C2 implies that for each type class $T \in \mathcal{T}^n$, the set of constraints

$$\sum_{T' \in \mathcal{T}^n} P(T')\tilde{W}(T|T') = P(T), \quad \forall P \in \mathcal{P}$$

must hold. Equivalently,

$$\sum_{T' \neq T} P(T')\tilde{W}(T|T') + P(T)[\tilde{W}(T|T) - 1] = 0, \quad \forall P \in \mathcal{P}\,. \tag{8}$$

For each $T \in \mathcal{T}^n$, we may think of (8) as a linear combination of the $N(\mathcal{P}, n)$ type probabilities $P(T')$, which must be identically zero for all $P \in \mathcal{P}$, where the coefficients are $\{\tilde{W}(T|T')\}_{T' \neq T}$ and $\tilde{W}(T|T) - 1$. Thus, by Assumption A1, we must have

$$\tilde{W}(T_{y^n}|T_{x^n}) = \tilde{W}_I(T_{y^n}|T_{x^n}) \stackrel{\triangle}{=} \begin{cases} 1 & T_{y^n} = T_{x^n} \\ 0 & T_{y^n} \neq T_{x^n} \end{cases} \tag{9}$$

as claimed.

Now, to prove part (a),

$$I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n) = nH - H(Y^n|X^n) \tag{10}$$

where the second equality follows from the fact that, by Condition C2, the probability law of Y^n is according to the DMS P. Thus, to obtain a lower bound on $I(X^n; Y^n)$, we need an upper bound on $H(Y^n|X^n) = \sum_{x^n} P(x^n)H(Y^n|X^n = x^n)$. Now, given $X^n = x^n$, Y^n may take no more than $2^k = 2^{nR}$ values, and, as we have just shown, they all must lie within T_{x^n}. Therefore, $H(Y^n|X^n = x^n) \leq \min\{nR, \log|T_{x^n}|\}$, which completes the proof of part (a) by taking the expectation w.r.t. P.

Turning now to the proof of part (b), the mapping ϕ^* obviously satisfies Condition C1 as it is independent of P. Since X^n is uniformly distributed within its type class T_{X^n}, then

so is $Y^n = \phi^*(X^n, U^k)$ and therefore, ϕ^* satisfies Condition C2 as well. Finally, to upper-bound the mutual information achieved by ϕ^*, notice that if $2^{nR} \leq |T_{x^n}|$, Y^n is uniformly distributed in a set of size 2^{nR}, and so $H(Y^n|X^n = x^n) = nR$. If, instead, $2^{nR} > |T_{x^n}|$, consider the integer division $2^{nR} = \alpha_1|T_{x^n}| + \alpha_2$, where $\alpha_1 \geq 1$ and $0 \leq \alpha_2 < |T_{x^n}|$. Clearly, there are α_2 sequences y^n for which the conditional probability $W^*(y^n|x^n)$ given by this scheme is $2^{-nR}(\alpha_1 + 1)$, whereas for the other sequences, $W^*(y^n|x^n) = 2^{-nR}\alpha_1$. It is easy to see that the corresponding conditional entropy takes the form

$$H(Y^n|X^n = x^n) = \log|T_{x^n}| - \Delta(\alpha_1, \alpha_2) \tag{11}$$

where

$$\Delta(\alpha_1, \alpha_2) \stackrel{\triangle}{=} 2^{-nR}\alpha_2(\alpha_1 + 1)\log(1 + \frac{1}{\alpha_1}) + \log(1 - 2^{-nR}\alpha_2) \tag{12}$$

and achieves its maximum value σ for $\alpha_1 = 1$ and $\alpha_2 = 2^{nR}[1 - (\log e)/2]$. Thus, in any case,

$$H(Y^n|X^n = x^n) \geq \min\{nR, \log|T_{x^n}| - \sigma\}$$

and, therefore,

$$H(Y^n|X^n) \geq \boldsymbol{E}\min\{nR, \log|T_{X^n}|\} - \sigma$$

thus completing the proof of part (b). □

3.2 Asymptotic Behavior of the Bounds

We would now like to evaluate the behavior of the lower bound given in part (a) of Theorem 1, as well as of the gap between this bound and the performance of the proposed scheme. It will be shown that the lower bound on the 'per-symbol dependence' $I(X^n; Y^n)/n$ between the two sequences is vanishing with n when $R > H$, but tends to $H - R$ when $R < H$. On the other hand, as we shall see, the rate of decay in the case $R > H$ is not too fast, so the gap between the upper bound and the lower bound, which is at most σ, is asymptotically insignificant in any case. Furthermore, we will show that, in fact, this gap vanishes exponentially fast for $R \neq H$.

We start with the lower bound, and let us first assume that $k = \infty$, or, equivalently, $R = \infty$. In this case, the lower bound boils down to

$$\begin{aligned} I(X^n; Y^n) &\geq nH - \boldsymbol{E}\{\log|T_{X^n}|\} \\ &= -\sum_{x^n} P(x^n)[\log P(x^n) + \log|T_{x^n}|] \end{aligned}$$

$$= -\sum_{x^n} P(x^n) \log P(T_{x^n})$$
$$= -\sum_{T_{x^n}} P(T_{x^n}) \log P(T_{x^n}) \stackrel{\triangle}{=} H(T_{X^n}), \tag{13}$$

that is, the entropy of the randomly selected type class under P. Since each type class is given by a union of elementary type classes, their number grows no faster than polynomially with n, and therefore it is already clear that $H(T_{X^n})$ cannot grow faster than logarithmically with n. As a result, according to (13), the normalized mutual information will tend to zero at least as fast as $(\log n)/n$.

A more refined asymptotic expansion of $H(T_{X^n})$ is provided in Appendix A via a local limit theorem for lattice random vectors [3, Chapter 5],[4], which requires some additional assumptions on the family $\mathcal{P}$. Basically, we will assume that $\mathcal{P}$ is an exponential family of the form (2), with some mild constraints on Ω. The reason for focusing on exponential families, apart from their importance and simplicity, is that they satisfy the following properties:

P1. There exists a lattice random vector $\boldsymbol{L}(X^n)$ with rank d' for some $d' \geq d$, such that $\boldsymbol{L}(X^n)$ is a one-to-one function of the random variable T_{X^n}, and can be written as a sum of n independently identically distributed (i.i.d.) random variables, denoted L_i, $i = 1, 2, \cdots, n$ (each independent of n).

P2. For each sequence x^n, the expectation of $\boldsymbol{L}(X^n)$ w.r.t. $P_{\hat{\theta}(x^n)}$, where $\hat{\theta}(x^n)$ denotes the maximum-likelihood estimate (MLE) of the parameter θ (which is assumed to exist and to belong to the closure of Ω), is $\boldsymbol{L}(x^n)$.

As discussed in Appendix A, Property P1 holds for any $\mathcal{P}$ such that there exists a one-to-one function of T_{X^n} over $\mathbb{R}^d$, which can be written as a sum of n i.i.d. random variables. For the exponential families (2), this function is clearly given by the sufficient statistics $n\boldsymbol{\tau}(X^n) = \sum_{i=1}^n \tau(X_i)$. Moreover, assuming that for every sequence x^n the equation $\boldsymbol{\tau}(x^n) = \nabla\psi(\theta)$ on θ has a solution in the closure of Ω, Property P2 will follow from the fact that the expectation of $\boldsymbol{\tau}(X^n)$ under $\hat{\theta}(x^n)$ is precisely $\boldsymbol{\tau}(x^n)$, and the property is inherited by the associated lattice random vector $\boldsymbol{L}(X^n)$. As further discussed in Appendix A, the rank d' is minimal, in the sense that for some d'-vector v such that $\Pr\{L_i = v\} > 0$, without loss of generality, $u - v$ spans $\mathbb{Z}^{d'}$ as u ranges over all d'-vectors with positive probability. Our technical assumptions on the exponential families, which ensure the validity of properties P1

and P2, are summarized as follows:

Assumption A2

a. *The parameter space* Ω *is an open subset of* $\mathbb{R}^d$.

b. *For any sequence* x^n, *the MLE* $\hat{\theta}(x^n)$ *of the parameter* θ *exists and belongs to the closure of* Ω.

c. *The covariance matrix* $M(\theta)$ *of the random variables* L_i *is nonsingular over the parameter space* Ω.

Assumption A2c. may require the deletion of singularity points of $M(\theta)$ from the parameter space. For example, in case $\mathcal{P}$ is the entire class of DMSs, $\det M(\theta) = \prod_{a\in\mathcal{A}} P_\theta(a)$. Thus, the symbol probabilities are assumed to be strictly positive. While in this and other simple cases $d' = d$, the possible difference between these two dimensions is discussed in Appendix A. Here, we will only mention that, roughly speaking, while d is the dimension of the parameter space, d' conveys, in a sense, the dimensionality of the type class. An example of disagreement between these two dimensions was provided in Section 2: a ternary source with a scalar parameter ($d = 1$) and symbol probabilities taking the form $C(\theta)\cdot e^{\theta\beta_h}$, $h = 0, 1, 2$, where θ is the parameter, $C(\theta)$ is a normalizing factor, and the ratio $(\beta_1-\beta_0)/(\beta_2-\beta_0)$ is irrational. As discussed in Section 2, the type classes coincide with the elementary type classes, and the 'effective number of parameters' d' is $|\mathcal{A}| - 1 = 2$, rather than just one.

Under Assumption A2, it is shown in Appendix A that for any $P_\theta \in \mathcal{P}$,

$$nH - \boldsymbol{E}_\theta\{\log|T_{X^n}|\} = H(T_{X^n}) = \frac{d'}{2}\log(2\pi n) + \frac{d}{2}\log e + \frac{1}{2}\log[\det M(\theta)] + o(1)\,. \qquad (14)$$

Thus, the price of universality is dominated by the term $\frac{d'}{2}\log n$, which for the case $d' = d$ parallels the universal lossless source coding problem (see, e.g., [13]). In particular, for the entire class of DMSs with positive probabilities, the asymptotic expansion (14) can alternatively be obtained from Stirling's formula.

To evaluate the lower bound of Theorem 1 for a finite value of R, we need to evaluate the expression

$$\boldsymbol{E}\min\{nR, \log|T_{X^n}|\} = nR\cdot\sum_{x^n\in\mathcal{D}} P(x^n) + \sum_{x^n\in\mathcal{D}^c} P(x^n)\log|T_{x^n}| \qquad (15)$$

where $\mathcal{D} \stackrel{\triangle}{=} \{x^n : \log|T_{x^n}| \geq nR\}$ and $\mathcal{D}^c$ is the complement of $\mathcal{D}$. Since $n^{-1}\log|T_{X^n}|$ is close to H with high probability (see Appendix A), it turns out that the bound behaves differently when $R < H$ and when $R > H$. In the first case, $\mathcal{D}^c$ is a large deviations event and the first term is dominant, whereas in the second case $\mathcal{D}$ is a large deviations event and the second term is dominant. As a result, the lower bound is approximately equal to $n(H-R)$ for $R < H$ and to $H(T_{X^n})$ for $R > H$ (again, see Appendix A for the detailed analysis).

This asymptotic behavior of the lower bound already guarantees that the gap between the lower bound and the upper bound is immaterial for large n, as it is at most σ (the case $R = H$ has not been analyzed explicitly, but the mutual information in this case cannot be smaller than the one obtained with $R > H$). Moreover, we will show that, in fact, this gap vanishes exponentially fast for $R \neq H$. To this end, we first notice that, for the scheme of Theorem 1, eqs. (11) and (15) imply that

$$\boldsymbol{E}\min\{nR, \log|T_{X^n}|\} - H(Y^n|X^n) = \sum_{x^n \in \mathcal{D}^c} P(x^n)\Delta(\alpha_1,\alpha_2) \leq 2^{-nR+1} \sum_{x^n \in \mathcal{D}^c} P(x^n)|T_{x^n}| \tag{16}$$

where the new upper bound on $\Delta(\alpha_1,\alpha_2)$ (used in lieu of the constant σ) follows from (12) by discarding the second (negative) term in the right-hand side, and using the inequality $(x+1)\log(1+1/x) < 2$ for $x \geq 1$. Since $|T_{x^n}| < 2^{nR}$ for $x^n \in \mathcal{D}^c$, and $\mathcal{D}^c$ is a large deviations event for $R < H$, the gap vanishes exponentially fast in this case. For $R > H$, we upper-bound $|T_{x^n}|$ by further classifying the sequences $x^n \in \mathcal{D}^c$ as belonging to the set $\mathcal{D}_\delta \stackrel{\triangle}{=} \{x^n : \ |\log|T_{x^n}| - nH| > n\delta\}$, which is shown in Lemma A.1 (Appendix A) to have exponentially vanishing probability for any $\delta > 0$, or to its complement $\mathcal{D}_\delta^c$. For the sequences in $\mathcal{D}_\delta^c$, $|T_{x^n}| < 2^{n(H+\delta)}$, whereas for the other sequences in $\mathcal{D}^c$, $|T_{x^n}| < 2^{nR}$ by the definition of the set. As a result, we obtain the upper bound

$$\boldsymbol{E}\min\{nR, \log|T_{X^n}|\} - H(Y^n|X^n) \leq 2(\Pr\{\mathcal{D}_\delta\} + 2^{-n(R-H-\delta)})$$

which, choosing $\delta < R - H$, vanishes exponentially fast.

4 The Case $n < m$

We now turn to the case of $n < m$, but we first assume that $R = \infty$. This assumption will be dropped later.

4.1 Unlimited Supply of Random Bits

We shall say that an output sequence y^n is *feasible* w.r.t. a given input sequence x^m, if y^n is a prefix of a sequence in T_{x^m}, in other words, if there exists a sequence z^r, $r \stackrel{\triangle}{=} m-n$, such that the concatenation $y^n z^r$ belongs to the same type class as x^m. Note that since, for all $P \in \mathcal{P}$, $P(x^m) = P(y^n)P(z^r) = P(\tilde{y}^n)P(z^r)$ for *every* $\tilde{y}^n \in T_{y^n}$, the feasibility of y^n w.r.t. x^m depends only on the type classes of these sequences. Thus, we shall also say that T_{y^n} is feasible w.r.t. T_{x^m}. Furthermore, the type class of z^r is fully determined by x^m and y^n and, conversely, $y^n \tilde{z}^r \in T_{x^m}$ for every $\tilde{z}^r \in T_{z^r}$. In addition, since $P(y^n z^r) = P(z^r y^n)$ for all $P \in \mathcal{P}$, z^r is also feasible w.r.t. x^m.

Let the conditional distribution (or, channel) $W^* : X^m \to Y^n$ be defined such that Y^n consists of the first n coordinates of a randomly selected member of T_{X^m} (under the uniform distribution). This channel is mathematically given by

$$W^*(y^n|x^m) = \begin{cases} \frac{|T_{z^r}|}{|T_{x^m}|} & \text{if } y^n \text{ is feasible w.r.t. } x^m \text{ with } y^n z^r \in T_{x^m} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

In (17), the denominator in the first line expresses the uniform distribution over T_{x^m}, and at the numerator, $|T_{z^r}|$ is the number of members of T_{x^m} whose first n coordinates agree with a given string y^n. Thus, W^* indeed sums up to unity for every x^m.

Let $I^*(X^m;Y^n)$ denote the mutual information between the random variables X^m and Y^n induced by the input distribution of X^m and the channel W^*. Our first result in this section tells us that, under Assumption A1, $I^*(X^m;Y^n)$ is a lower bound on $I(X^m;Y^n)$ for any channel satisfying conditions C1 and C2, and that the channel W^* which *precisely* achieves this bound, indeed satisfies these conditions.

Theorem 2 *Let $\mathcal{P}$ satisfy Assumption A1. Then,*

(a) For every channel $W : X^m \to Y^n$ that satisfies conditions C1 and C2,

$$I(X^m;Y^n) \geq I^*(X^m;Y^n). \tag{18}$$

(b) The channel W^* satisfies conditions C1 and C2, and hence it satisfies also Condition C3.

Comments:

(a) The theorem tells us that the best one can do is to randomly select a sequence with uniform distribution across T_{x^m}, and then truncate it to the suitable dimension. Another option that one might think of is the other way around: first truncate and then apply a random selection within the type class of the truncated sequence. While this option also satisfies conditions C1 and C2, it is not difficult to show that it gives a larger value of the mutual information than the option proposed in Theorem 2.

(b) If $|T_{x^m}| = 2^{k(x^m)}$ is an integer power of 2, the uniform distribution over T_{x^m} that is required for implementing W^* can be achieved with $k(x^m)$ random bits. However, in any other case, $R = \infty$ is required.

Proof. As in the proof of Theorem 1, an arbitrary given channel $W : X^m \to Y^n$ induces a channel $\tilde{W}$ from T_{X^m} to T_{Y^n} via a uniform distribution over T_{X^m}, namely,

$$\tilde{W}(T_{y^n}|T_{x^m}) = \frac{1}{|T_{x^m}|} \sum_{\tilde{x}^m \in T_{x^m}} \sum_{\tilde{y}^n \in T_{y^n}} W(\tilde{y}^n|\tilde{x}^m). \tag{19}$$

In particular, since $W^*(y^n|x^m)$ depends on x^m and y^n only through the respective type classes, we have

$$\tilde{W}^*(T_{y^n}|T_{x^m}) = |T_{y^n}|W^*(y^n|x^m). \tag{20}$$

Again, Condition C2 implies that for each y^n, the set of constraints

$$\sum_{T' \in \mathcal{T}^m} P(T')\tilde{W}(T|T') = |T_{y^n}|P(y^n), \quad \forall P \in \mathcal{P} \tag{21}$$

must hold. Now, with $r = m - n$,

$$P(y^n) = \sum_{T_{z^r} \in \mathcal{T}^r} P(y^n z^r)|T_{z^r}| = \sum_{T_{x^m} : y^n \text{ feasible w.r.t. } x^m} P(T_{x^m}) \cdot \frac{|T_{z^r}|}{|T_{x^m}|}. \tag{22}$$

Thus, by (21) and (22), for each class type $T \in \mathcal{T}^n$ and all $P \in \mathcal{P}$ we have the constraint

$$\sum_{T' : T \text{ not feasible w.r.t. } T'} P(T')\tilde{W}(T|T') + \sum_{T' : T \text{ feasible w.r.t. } T'} P(T') \left[\tilde{W}(T|T') - \frac{|T| \cdot |T_{z^r}|}{|T'|}\right] = 0. \tag{23}$$

Using Assumption A1 as in the proof of Theorem 1, it follows that $\tilde{W}^*(T_{y^n}|T_{x^m})$ is the only channel satisfying the set of constraints (23).

Now, to prove part (a), we observe that $T_{X^m} \to X^m \to Y^n \to T_{Y^n}$ is a Markov chain, and so, by the data processing theorem, for any channel W we have

$$I(X^m; Y^n) \geq I(T_{X^m}; T_{Y^n}) \tag{24}$$

where the underlying channel in the right hand side is $\tilde{W}$, induced by W according to (19). Since $\tilde{W}^*$ is the only induced channel satisfying conditions C1 and C2, it suffices to prove that the mutual information achieved by this channel (with the input distribution of T_{X^m}) is $I^*(X^m; Y^n)$. In other words, we need to show that W^* achieves equality in (24), which follows from the following chain of equalities:

$$I^*(X^m; Y^n) = \boldsymbol{E} \log \frac{W^*(Y^n|X^m)}{P(Y^n)} = \boldsymbol{E} \log \frac{|T_{Y^n}| \cdot W^*(Y^n|X^m)}{P(T_{Y^n})} = \boldsymbol{E} \log \frac{\tilde{W}^*(T_{Y^n}|T_{X^m})}{P(T_{Y^n})} \tag{25}$$

where the first equality follows from the definition of I^* and the last equality follows from (20).

To prove part (b), notice that Condition C1 is obviously satisfied by W^*, Condition C2 is easily seen to hold by observing that $\tilde{W}^*$ is a solution to the system (21), dividing by $|T_{y^n}|$, and using (20), and Condition C3 follows from the fact that W^* provides the lower bound of part (a) of the theorem. This completes the proof of Theorem 2. □

Asymptotic Behavior of the Lower Bound

We now evaluate the behavior of $I^*(X^m; Y^n)$. We have

$$\begin{aligned} I^*(X^m; Y^n) &= \boldsymbol{E} \log \frac{W^*(Y^n|X^m)}{P(Y^n)} \\ &= \boldsymbol{E} \log \frac{|T_{Z^r}|}{|T_{X^m}| \cdot P(Y^n)} \\ &= nH + \boldsymbol{E} \log |T_{Z^r}| - \boldsymbol{E} \log |T_{X^m}| \end{aligned} \tag{26}$$

where it is assumed that for a given x^m, $|T_{z^r}| = 0$ for infeasible y^n.

As shown in Appendix A, under Assumption A2,

$$\boldsymbol{E}_\theta \log |T_{X^m}| = mH - \frac{d'}{2} \log(2\pi m) - \frac{d}{2} \log e - \frac{1}{2} \log[\det M(\theta)] + o(1) \tag{27}$$

and, since Z^r is drawn according to P (by symmetry with the drawing of Y^n), a similar expression applies to $\boldsymbol{E}_\theta \log |T_{Z^r}|$ with m replaced by r (assuming that r is also large). On substituting these expansions into eq. (26), we get, with $r = m - n$,

$$I^*(X^m; Y^n) = \frac{d'}{2} \log \frac{m}{m-n} + o(1)\,. \tag{28}$$

We learn from this expression that if m grows linearly with n and the ratio m/n tends to a constant $C > 1$, then $I^*(X^m;Y^n)$ tends also to a constant, $\frac{d'}{2}\log\frac{C}{C-1}$ (compare with the logarithmic behavior of Section 3). Moreover, if $m/n \to \infty$, then $I^*(X^m;Y^n) \to 0$.

It should be kept in mind that this asymptotic analysis holds under the assumption that both m and r are large. Thus, the case $n = m$ (where $r = 0$), treated in Section 3, is not obtained as a special case. Nonetheless, Theorem 2, as stated, without explicit computation of $I^*(X^m;Y^n)$, is still free of this assumption and hence includes the case $n = m$ (and $R = \infty$) as a special case. Furthermore, the case $m/n \searrow 1$ (i.e., r large but $r = o(n)$) yields the same behavior as Theorem 1.

4.2 Limited Budget of Random Bits

We now move on to the more general case, where both $n < m$ and $R < \infty$, which turns out to be surprisingly more involved than the special cases that have been studied thus far, ($n = m, R \leq \infty$) and ($n < m$, large enough R). The main difficulty essentially lies in the fact that it is hard to construct a concrete simulation scheme that implements W^* (or at least, approximates it faithfully) with a limited number of random bits, even if this number is larger than nH. As we shall see, not only the results will now be merely asymptotic, but we will only be able to prove the *existence* of good simulation schemes (in the sense of Condition C3), without any explicit construction. Nonetheless, as is typical to many information–theoretic existence proofs (especially those that involve random coding arguments), it will become apparent (cf. Appendix B) that not only do good schemes exist, but moreover, *almost* every scheme in a very large family that we shall define, is good. Consequently, a randomly selected scheme in this family will be good with very high probability. Moreover, it will also emerge from the proof that if the parameter space is known to be limited to sources such that $H \cdot \limsup_{n\to\infty}(m/n) < R$, then *every* simulation scheme in the family is good, so that the proof is actually constructive.

Basically, the idea behind the proof is to demonstrate that, at least for pairs (x^m, y^n) that are typical to sources for which $H < R$, W^* can be implemented with very high accuracy by *some* simulation scheme in the family, provided that $\log m = o(n)$. The case $H > R$ is handled separately by another scheme which is conceptually simple, and the choice between the two schemes is carried out by a decision rule that attempts to determine whether $H < R$ or $H > R$, based on the input sequence.

Our main result in this paper is given in Theorem 3 below. It states that the lower bound, $n(H-R)$ for $R<H$, and $I^*(X^m;Y^n)$ for $R>H$, is achievable within an asymptotically negligible term, as long as the growth of m with n is sub-exponential (namely, $\log m = o(n)$).

Theorem 3 *Let $\mathcal{P}$ satisfy Assumption A1. Then,*

(a) For every mapping ϕ that satisfies conditions C1 and C2,

$$I(X^m;Y^n) \geq I_{\min}(X^m;Y^n) \stackrel{\triangle}{=} \begin{cases} n(H-R) & R<H \\ I^*(X^m;Y^n) & R>H. \end{cases} \tag{29}$$

(b) If $\log m = o(n)$, then there exists a mapping ϕ^* that satisfies conditions C1 and C2 and for which

$$I(X^m;Y^n) \leq I_{\min}(X^m;Y^n) + \zeta_n$$

where ζ_n vanishes exponentially rapidly.

(c) If, in addition, $\mathcal{P}$ satisfies Assumption A2, then ζ_n is negligible relative to $I_{\min}(X^m;Y^n)$ provided that the sequence $\{m/n\}$ is bounded.

Comment: Part (c) of Theorem 3 excludes the cases in which m/n grows without bound. While in those cases the first term in the right-hand side of (28) vanishes at the same rate as n/m, which dominates over ζ_n due to the assumption $\log m = o(n)$, the remaining $o(1)$ term may actually determine the asymptotic behavior.

The remaining part of this section is devoted to the proof of Theorem 3.

Proof. The proof of part (a) is very simple. For $R<H$, we obviously have

$$I(X^m;Y^n) = H(Y^n) - H(Y^n|X^m) = nH - H(Y^n|X^m) \geq nH - nR$$

where the inequality follows from the fact that $H(Y^n|X^m) \leq nR$ when the number of random bits is limited by nR. For $R>H$, we use the lower bound of Theorem 2, which does not make any assumption on R. This completes the proof of part (a).

To prove part (b), we consider mappings of the following structure (whose description is similar to that of the mapping ϕ^* in Section 3): List the members of every type class T_{x^m} in a certain order, and for every $\tilde{x}^m \in T_{x^m}$, let $J_m(\tilde{x}^m) \in \{0,1,\ldots,|T_{x^m}|-1\}$ denote the index of $\tilde{x}^m$ within T_{x^m} in this list (starting from zero for the first sequence). Denoting by J_m^{-1} the inverse map from $\{0,1,\ldots,|T_{x^m}|-1\}$ to T_{x^m}, define

$$y^n = \phi(x^m,u^k) \stackrel{\triangle}{=} \left[J_m^{-1}\left(J_m(x^m) \oplus \sum_{i=1}^{k} 2^{i-1}u_i \right) \right]_1^n \tag{30}$$

where $\oplus$ denotes addition modulo $|T_{x^m}|$, the sum over i is taken under the ordinary integer arithmetic, and the operator $[\cdot]_1^n$ forms an n-vector by taking the first n coordinates of an m-vector (e.g., $[x^m]_1^n = x^n = (x_1, \ldots, x_n)$).

This mapping obviously satisfies Condition C1 as it is independent of P. Since X^m is uniformly distributed within its type class T_{X^m}, then so is $Y^n = \phi(X^m, U^k)$ and therefore, ϕ satisfies Condition C2 as well.

Whether or not such a mapping meets also Condition C3, depends on the ordering, or the permutation corresponding to the ordered list of m-sequences in each of the type classes. There are as many as $\prod_{T\in\mathcal{T}^m} |T|!$ different combinations of such permutations across all type classes and each such combination corresponds to a different scheme ϕ in the family of schemes, denoted $\mathcal{F}$, that we now consider. The following lemma, whose proof appears in Appendix B, guarantees, as explained earlier, that there exists a scheme in this family that induces a very good approximation to $W^*(y^n|x^m)$, at least for a subset $\mathcal{S}$ of pairs (x^m, y^n) that are typical to sources for which $H < R$. This is the key property that is needed to achieve asymptotically the lower bound of part (a), and hence to essentially satisfy Condition C3. Moreover, for a subset of sequences x^m that are typical to sources for which $H \cdot \limsup_{n\to\infty}(m/n) < R$, *every* simulation scheme in the family induces such an approximation.

Lemma 1 *Assume that* $\log m = o(n)$. *Let* R *be given and let* W^* *be defined as in eq. (17). For every* $\epsilon > 0$ *and all sufficiently large* n, *there exists a permutation of each* T_{x^m}, *such that the conditional probability distribution* W *induced by eq. (30), satisfies*

$$(1 - 2^{-n\epsilon})W^*(y^n|x^m) \le W(y^n|x^m) \le (1 + 2^{-n\epsilon})W^*(y^n|x^m) \tag{31}$$

for every $(x^m, y^n) \in \mathcal{S}$, *where*

$$\mathcal{S} \stackrel{\triangle}{=} \{(x^m, y^n) : \ nR \ge \log|T_{x^m}| - \log|T_{z^r}| + 3n\epsilon\}.$$

Moreover, for every x^m *such that* $\log|T_{x^m}| < n(R - \epsilon)$, *every permutation of* T_{x^m} *satisfies (31).*

Let $\tilde{\phi}$ be defined by eq. (30), where the permutation of the members in each type class satisfies Lemma 1. Consider now the following simulation scheme:

$$y^n = \phi^*(x^m, u^k) \stackrel{\triangle}{=} \begin{cases} \tilde{\phi}(x^m, u^k) & \log|T_{x^n}| < nR \\ J_n^{-1}\left(J_n(x^n) \oplus \sum_{i=1}^{nR} 2^{i-1}u_i\right) & \log|T_{x^n}| \ge nR \end{cases} \tag{32}$$

where J_n corresponds to an arbitrary enumeration of each type class (as in the definition of the mapping of Theorem 1). The idea is that when $H < R$, the probability that $\log|T_{x^n}| < nR$ is high for large n, and then y^n is likely to take the value of $\tilde{\phi}(x^m, u^k)$. If, on the other hand, $H > R$, then y^n takes the alternative value defined in eq. (32) with high probability. Notice that in this case, we first truncate x^m to n bits, and only then we choose a sequence in T_{x^n} randomly.

We next analyze the performance of the scheme proposed in eq. (32). Define $\mathcal{E} = \{x^m : \log|T_{x^n}| \geq nR\} \times \mathcal{A}^n$. Assume first that $H < R$. In the following derivation, we shall make use of the following simple fact: If $\delta \in [0,1)$, $u_0 \in (0,1]$, and $u_0(1-\delta) \leq u \leq u_0(1+\delta)$, then

$$\begin{aligned} u\log u &\leq u\log[u_0(1+\delta)] \\ &= u\log u_0 + u\log(1+\delta) \\ &\leq (1-\delta)u_0\log u_0 + u_0(1+\delta)\log(1+\delta). \end{aligned} \tag{33}$$

Now, the mutual information associated with the simulation scheme defined in eq. (32) is upper-bounded as follows:

$$\begin{aligned} I(X^m;Y^n) &= nH + \sum_{x^m,y^n} P(x^m)W(y^n|x^m)\log W(y^n|x^m) \\ &\leq nH + \sum_{(x^m,y^n)\in\mathcal{E}^c\cap\mathcal{S}} P(x^m)W(y^n|x^m)\log W(y^n|x^m) \\ &\leq nH + \sum_{(x^m,y^n)\in\mathcal{E}^c\cap\mathcal{S}} P(x^m)[(1-2^{-n\epsilon})W^*(y^n|x^m)\log W^*(y^n|x^m) + \\ &\quad W^*(y^n|x^m)(1+2^{-n\epsilon})\log(1+2^{-n\epsilon})] \\ &= nH + \sum_{x^m,y^n} P(x^m)[(1-2^{-n\epsilon})W^*(y^n|x^m)\log W^*(y^n|x^m) + \\ &\quad W^*(y^n|x^m)(1+2^{-n\epsilon})\log(1+2^{-n\epsilon})] - \\ &\quad \sum_{(x^m,y^n)\in\mathcal{E}\cup\mathcal{S}^c} P(x^m)[(1-2^{-n\epsilon})W^*(y^n|x^m)\log W^*(y^n|x^m) + \\ &\quad W^*(y^n|x^m)(1+2^{-n\epsilon})\log(1+2^{-n\epsilon})] \\ &\leq nH - (1-2^{-n\epsilon})H_{PW^*}(Y^n|X^m) + \\ &\quad (1+2^{-n\epsilon})\log(1+2^{-n\epsilon}) + \Pr\{\mathcal{E}\cup\mathcal{S}^c\}\cdot m\log|\mathcal{A}| \\ &\leq I^*(X^m;Y^n) + 2^{-n\epsilon}n\log|\mathcal{A}| + \\ &\quad (1+2^{-n\epsilon})\log(1+2^{-n\epsilon}) + \Pr\{\mathcal{E}\cup\mathcal{S}^c\}\cdot m\log|\mathcal{A}| \end{aligned}$$

$$\stackrel{\triangle}{=} \quad I^*(X^m;Y^n)+\zeta_n \tag{34}$$

where in the second inequality we have used eqs. (32) and (33), $H_{PW^*}(Y^n|X^m)$ denotes the conditional entropy of Y^n given X^m induced by $P\times W^*$ and where the last summation over $\mathcal{E}\cup\mathcal{S}^c$ is bounded in terms of the probability of $\mathcal{E}\cup\mathcal{S}^c$ by using the fact that $W^*(y^n|x^m)\geq 1/|T_{x^m}|\geq|\mathcal{A}|^{-m}$ whenever $W^*(y^n|x^m)>0$. Since $\log m=o(n)$, to complete the analysis for the case $H<R$, it suffices to show that the probability of $\mathcal{E}\cup\mathcal{S}^c$ w.r.t. $P\times W^*$ is exponentially small in this case. The following lemma, whose proof appears in Appendix C, establishes this fact.

Lemma 2 *Let $P\in\mathcal{P}$ be a given source for which $H<R$, and assume that $\log m=o(n)$. Then, for all ϵ in the range $0<\epsilon<(R-H)/3$,*

$$\limsup_{n\to\infty}\frac{1}{n}\log\Pr\{\mathcal{E}\cup\mathcal{S}^c\}<0$$

where $\Pr\{\cdot\}$ is defined w.r.t. $P\times W^*$.

Notice that if not only $R>H$, but, moreover, $R>HC$, where $C=\limsup_{n\to\infty}(m/n)$, then the set of sequences x^m such that $n(R-\epsilon)>|T_{x^m}|$ is typical provided that $\epsilon<(R/C)-H$, since the set $\mathcal{D}_\delta$, with $\delta<(R/C)-H-\epsilon$, is a large deviations event (Lemma A.1). Thus, by Lemma 1, (34) implies that any permutation of T_{x^m} can be used.

Consider next the case $H>R$. With a slight abuse of notation, let us redefine now $\mathcal{E}$ as $\{x^m:\log|T_{x^n}|\geq nR\}$, without the Cartesian product with $\mathcal{A}^n$ as before. Let $\mathcal{Y}_0(x^m)$ denote the set of $2^k=2^{nR}$ output sequences that can be obtained from the second line of the right hand side of eq. (32) as u^k exhausts $\mathcal{B}^k$. Now,

$$\begin{aligned}
I(X^m;Y^n) &= \sum_{x^m,y^n}P(x^m)W(y^n|x^m)\log\frac{W(y^n|x^m)}{P(y^n)}\\
&= \sum_{x^m\in\mathcal{E}}P(x^m)\sum_{y^n\in\mathcal{Y}_0(x^m)}W(y^n|x^m)\log\frac{W(y^n|x^m)}{P(y^n)}+\\
&\quad\sum_{x^m\in\mathcal{E}^c}P(x^m)\sum_{y^n}W(y^n|x^m)\log\frac{W(y^n|x^m)}{P(y^n)}\\
&= \sum_{x^m\in\mathcal{E}}P(x^m)\sum_{y^n\in\mathcal{Y}_0(x^m)}2^{-nR}\log\frac{2^{-nR}}{P(x^n)}+\\
&\quad\sum_{x^m\in\mathcal{E}^c}P(x^m)\sum_{y^n}W(y^n|x^m)\log\frac{W(y^n|x^m)}{P(y^n)}
\end{aligned}\tag{35}$$

where the last equality follows from the fact that $\mathcal{Y}_0(x^m)$ is a subset of T_{x^n} and therefore all the sequences in the set have probability $P(x^n)$. Letting $P_{\min}$ denote the minimum non-zero symbol probability assigned by P,[4] we have $P(y^n) \geq P_{\min}^n$. Since $W(y^n|x^m) \leq 1$, (35) implies

$$\begin{aligned} I(X^m;Y^n) &\leq nH - nR \cdot \Pr\{\mathcal{E}\} + n\left(\log \frac{1}{P_{\min}}\right) \cdot \Pr\{\mathcal{E}^c\} \\ &= I_{\min}(X^m;Y^n) + n(R + \log \frac{1}{P_{\min}}) \cdot \Pr\{\mathcal{E}^c\}\,. \end{aligned} \tag{36}$$

Clearly, the probability of $\mathcal{E}^c$ is the same as the probability of the set $\mathcal{D}^c$ defined in Subsection 3.2, which is shown in Appendix A to vanish exponentially rapidly. Thus, the upper bound in the case $H > R$ also approaches $I_{\min}(X^m;Y^n)$ exponentially rapidly. This completes the proof of part (b).

To prove part (c), it suffices to notice that under Assumption A2, if m/n is bounded, $I_{\min}(X^m;Y^n)$ is bounded away from zero, and thus dominates over ζ_n. □

5 Extension to More General Information Measures

Our results can be extended in several directions. Two of these directions will be outlined informally in this section: The first one extends the class $\mathcal{P}$ to sources with memory; the second one extends the dependency measure to a more general class of information measures, which includes the Shannon mutual information as a special case.

5.1 Sources with Memory

For simplicity, we will assume that $\mathcal{P}$ is the *entire* class of finite-state (FS) sources with a given next-state function; extensions to parametric subfamilies are possible depending on the validity of assumptions A1 and A2. Our model is defined by a FS machine with a finite set of states S, driven by a deterministic next-state function $s_{t+1} = f(s_t, x_t)$, where $s_t \in S$, $1 \leq t < n$, and s_1 is a *fixed* initial state. The model is parametrized by conditional probabilities $p(x|s)$, $x \in \mathcal{A}$, $s \in S$. The probability of a sequence x^m is given by

$$P(x^m) = \prod_{t=1}^{n} p(x_i|s_i)\,.$$

[4] Without loss of generality, we can assume that all symbol probabilities are positive. For the results to hold *uniformly* over Ω, the parameter space should be such that the probabilities are bounded away from zero.

Notice that, under this definition, in which the initial state is fixed, the sources in the class will be, in general, nonstationary. Clearly, the type class of a sequence is given by its FS-type, i.e., the number of transitions between each pair of states (starting from the fixed initial state), and the class satisfies Assumption A1. Also, from Stirling's formula applied to the size of a FS-type (or using a local limit theorem for sources with memory [2]), the asymptotic behavior of $\boldsymbol{E}\log|T_{X^m}|$ is still as in (A.5), with nH replaced with $H(X^n|s_1)$ (the entropy of X^n when the source is started at the initial state s_1), and where $d = d'$ is given by the number of free parameters $|S|(|\mathcal{A}| - 1)$. Here, the fact that the expected divergence between the distributions of X^n with the MLE and the true parameters (both conditioned on s_1) tends to $\frac{d}{2}\log e$ also for sources with memory, is proved in [1].

First, we notice that all the steps that lead to the main result in Section 3 hold *verbatim*, with nH replaced with $H(X^n|s_1)$, as the i.i.d. assumption is not used elsewhere. Moreover, since $n^{-1}H(X^n|s_1)$ converges to the *entropy rate* H of the FS source, the asymptotic behavior of the lower bound on the mutual information is unchanged, and H is still the critical threshold for the random bit-rate R needed for the per-symbol mutual information to vanish.

The situation is slightly more involved when $n < m$, due to the fact that the final state s_{t+1} to which the FS source is driven by a feasible y^n, may not agree with s_1, and therefore the initial conditions for z^r (defined such that $y^n z^r \in T_{x^m}$) are different. It should be noted, however, that if $\tilde{y}^n \in T_{y^n}$, the final states corresponding to y^n and $\tilde{y}^n$ must coincide, since the initial states do, and the number of transitions between any pair of states is the same for both sequences, with the initial and final states being the only ones for which the nodes in the associated graph may have an outgoing degree that differs from the incoming degree. As a result, a key property in the memoryless case, namely that the feasibility of y^n w.r.t. x^m depends only on the type classes of these sequences, is preserved. Notice, however, that in order to generalize our results, we need to change the definition of the channel $W^*(y^n|x^m)$ in (17) to reflect the difference in the initial state of z^r: Specifically, the type class in the numerator assumes that the *initial state* is given by the *final state* prescribed by y^n (or any other sequence in the type class of y^n). This change in the definition of T_{z^r} throughout the analysis guarantees that the results will remain valid.

5.2 More General Information Measures

In [18] and [19], Ziv and Zakai proposed a generalized functional that satisfies a data processing theorem, in the context of deriving tighter joint source–channel distortion bounds for short block codes. If $P(u,v) = P(u)W(v|u)$ denotes the joint distribution of a pair of random variables (or random vectors) (U,V), and Q denotes the marginal of V, this functional is defined as

$$I^S(U;V) \stackrel{\triangle}{=} \boldsymbol{E}\left\{S\left(\frac{P(U)Q(V)}{P(U,V)}\right)\right\} = \boldsymbol{E}\left\{S\left(\frac{Q(V)}{W(V|U)}\right)\right\} \tag{37}$$

where the expectation is w.r.t. $P \times W$ and $S : \mathbb{R}^+ \to \mathbb{R}$ is a monotonically nonincreasing, convex $(\cup)$ function, with $S(1) = 0$ and $0 \cdot S(1/0) \stackrel{\triangle}{=} \lim_{t\to 0} tS(1/t) = 0$. Of course, $S(t) = -\log t$ corresponds to the Shannon mutual information. It is easy to see that I^S is a 'reasonable' measure of statistical dependence in the sense that it takes its minimum value, zero, if U and V are statistically independent, and its maximum value, $\boldsymbol{E}S(P(U))$, if there is a one-to-one correspondence between U and V. As is shown in [18] and [19], I^S satisfies the data processing theorem: If $U \to V \to Z$ is a Markov chain, then

$$I^S(U;V) \geq I^S(U;Z) \leq I^S(V;Z).$$

We now demonstrate that our earlier results can be extended to account for $I^S(X^m;Y^n)$ as the objective function to be minimized in Condition C3. It turns out that the same simulation schemes that we discussed earlier, essentially minimize $I^S(X^m;Y^n)$ for a general S.

Let us start, once again, with the case $m = n$. We first derive a lower bound to $I^S(X^n;Y^m)$. As explained in the Introduction, for a given value of $k = nR$, $W(y^n|x^n)$ is always of the form $W(y^n|x^n) = l(y^n|x^n)2^{-nR}$, where $\{l(y^n|x^n)\}$ are non-negative integers whose sum over $\{y^n\}$ is 2^{nR} for every x^n. Let $\mathcal{Y}(x^n) \stackrel{\triangle}{=} \{y^n : W(y^n|x^n) > 0\}$. As we have shown in the proof of Theorem 1, part (a), under Assumption A1, to meet Condition C2, y^n must always be of the same type class as x^n. Thus, $\mathcal{Y}(x^n)$ must be a subset of T_{x^n}. Since $|\mathcal{Y}(x^n)|$ cannot exceed 2^{nR}, it follows then that $|\mathcal{Y}(x^n)| \leq \min\{2^{nR}, |T_{x^n}|\}$, and so:

$$\begin{aligned} I^S(X^n;Y^n) &\stackrel{\text{(a)}}{=} \sum_{x^n} P(x^n) \sum_{y^n \in \mathcal{Y}(x^n)} W(y^n|x^n) S\left(\frac{P(y^n)}{W(y^n|x^n)}\right) \\ &\stackrel{\text{(b)}}{\geq} \sum_{x^n} P(x^n) S\left(\sum_{y^n \in \mathcal{Y}(x^n)} W(y^n|x^n) \cdot \frac{P(y^n)}{W(y^n|x^n)}\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{x^n} P(x^n) S\left(\sum_{y^n \in \mathcal{Y}(x^n)} P(y^n)\right) \\
&\overset{(c)}{=} \sum_{x^n} P(x^n) S\left(|\mathcal{Y}(x^n)| \cdot P(x^n)\right) \\
&\overset{(d)}{\geq} \sum_{x^n} P(x^n) S\left(P(x^n) \cdot \min\{2^{nR}, |T_{x^n}|\}\right) \\
&= \boldsymbol{E} S\left(P(X^n) \cdot \min\{2^{nR}, |T_{X^n}|\}\right) \qquad (38)
\end{aligned}
$$

where (a) follows from the assumption that $0 \cdot S(1/0) = 0$, (b) – from Jensen's inequality, (c) – from the fact that $\mathcal{Y}(x^n) \subseteq T_{x^n}$ (hence $P(y^n) = P(x^n)$ for all $y^n \in \mathcal{Y}(x^n)$), and (d) – from the aforementioned upper bound to $|\mathcal{Y}(x^n)|$ and the monotonicity of S. It is straightforward to see that this lower bound is essentially achieved using the same scheme as in Theorem 1, part (b), because for that scheme, $W(y^n|x^n)$ is essentially uniformly distributed within $\mathcal{Y}(x^n)$ whose size is exactly $\min\{2^{nR}, |T_{x^n}|\}$.

For $m > n$ and $R = \infty$, the data processing theorem w.r.t. I^S gives (as in the proof of Theorem 2, part (a)), the lower bound

$$
I^S(X^m; Y^n) \geq \boldsymbol{E}_{P \times W^*} S\left(\frac{P(Y^n)}{W^*(Y^n|X^m)}\right) \qquad (39)
$$

where $\boldsymbol{E}_{P \times W^*}$ denotes expectation w.r.t. $P \times W^*$ and W^* is the same channel as in Section 4, which attains the bound. This extension follows from a similar chain of equalities as in eq. (25), where the function $-\log(\cdot)$ is replaced by the general function $S(\cdot)$.

Finally, consider the case $n < m$ and $R < \infty$. Redefining $\mathcal{Y}(x^m)$ as the set of $\{y^n\}$ with strictly positive probabilities given x^m, and $\mathcal{Y}_0(x^m, T_{y^n}) \stackrel{\triangle}{=} \mathcal{Y}(x^m) \cap T_{y^n}$, we have the following:

$$
\begin{aligned}
I^S(X^m; Y^n) &= \sum_{x^m} P(x^m) \sum_{y^n \in \mathcal{Y}(x^m)} W(y^n|x^m) S\left(\frac{P(y^n)}{W(y^n|x^m)}\right) \\
&\overset{(a)}{=} \sum_{x^m} P(x^m) \sum_{\{T_{y^n}\}} \sum_{\tilde{y}^n \in \mathcal{Y}_0(x^m, T_{y^n})} W(\tilde{y}^n|x^m) S\left(\frac{P(y^n)}{W(\tilde{y}^n|x^m)}\right) \\
&\overset{(b)}{\geq} \sum_{x^m} P(x^m) \sum_{\{T_{y^n}\}} \sum_{\tilde{y}^n \in \mathcal{Y}_0(x^m, T_{y^n})} W(\tilde{y}^n|x^m) S\left(\frac{P(y^n)}{2^{-nR}}\right) \\
&= \boldsymbol{E} S(2^{nR} P(Y^n)) \\
&= \boldsymbol{E} S(2^{nR} P(X^n)) \qquad (40)
\end{aligned}
$$

where (a) follows from the facts that $P(\tilde{y}^n) = P(y^n)$ for all $\tilde{y}^n \in \mathcal{Y}_0(x^m, T_{y^n}) \subseteq T_{y^n}$ and

(b) follows from the assumption that S is monotonically nonincreasing and the fact that $W(y^n|x^m) \geq 2^{-nR}$ whenever $W(y^n|x^m) > 0$. This bound (for $R < H$) and the previous bound of (39) (for $R > H$) can essentially be attained jointly, by the same scheme as described in the proof of Theorem 3, provided that S satisfies some additional regularity conditions that account, among other things, for insensitivity to small differences between W and W^* (cf. Lemma 1). We will not get into the technical details of this case any further.

Acknowledgment

We would like to thank Gadiel Seroussi for useful suggestions.

Appendix A

In this appendix, we examine the asymptotic behavior of the lower bound for $n = m$. To this end, we use an asymptotic expansion of the expression $\boldsymbol{E} \log |T_{X^n}|$, based on a local limit theorem for *lattice random vectors* [3, Chapter 5] (see [9, Chapter XV.5] for the scalar case) under Assumption A2 on the exponential family $\mathcal{P}$, stated in Subsection 3.2. We start with a discussion of Property P1, which links our problem to lattice distributions, and, in particular, the relation between the parameter dimension d and the relevant lattice rank d'. We also discuss Property P2 (Assumption A2 states standard regularity conditions).

First, we observe that Property P1 holds for the exponential families (2), as a result of $n\boldsymbol{\tau}(X^n)$ being a one-to-one function of T_{X^n} over $\mathbb{R}^d$, which can be written as a sum of n i.i.d. random variables $\tau(X_i)$. The idea is that each random variable $\tau(X_i)$ can be transformed into a random d'-vector of integers as follows. Consider the (at most $|\mathcal{A}|$) real values that the j-th coordinate of $\tau(X_i)$ can take, $j = 1, 2, \ldots, d$. For each j, we can always find a positive integer r_j, and r_j pairwise incommensurable real numbers $\beta[j]_h$, $h = 1, 2, \ldots, r_j$, that yield a (unique) decomposition of any value of the j-th coordinate, denoted $\tau[j]$, in the form

$$\tau[j] = \beta[j]_0 + \sum_{h=1}^{r_j} \beta[j]_h L[j]_h \tag{A.1}$$

where $\beta[j]_0$ is a fixed real number and $L[j]_h$, $h = 1, 2, \ldots, r_j$, are integers that depend on $\tau[j]$. The uniqueness of the decomposition clearly follows from the irrationality of the ratios $\beta[j]_{h_1}/\beta[j]_{h_2}$, $1 \leq h_1, h_2 \leq r_j$, $h_1 \neq h_2$. Now, to each (real) d-vector τ we associate the unique (integer) d'-vector with coordinates $L[j]_h$, $1 \leq j \leq d$, $1 \leq h \leq r_j$, $d' \stackrel{\triangle}{=} \sum_{j=1}^{d} r_j$,

obtained by decomposing each coordinate of τ according to (A.1). Clearly, this process defines a random variable L_i which is a one-to-one function of $\tau(X_i)$, $i = 1, 2, \ldots, n$, and, accordingly, the random variable $\boldsymbol{L} \stackrel{\triangle}{=} \sum_{i=1}^{n} L_i$ is a one-to-one function of T_{x^n}. The random variables L_i are clearly i.i.d. In addition, we choose the real numbers $\beta[j]_h$ to be the *spans* of the distribution of the j-th component, that is, to have maximum magnitude among those real numbers yielding a decomposition of the form (A.1). Thus, for each coordinate j, h, the values taken with positive probability by $L[j]_h$ span $\mathbb{Z}$, as they are mutually prime. We can therefore assume, without loss of generality, that d' is the rank of the lattice random vectors L_i, in the sense that for some d'-vector v such that $\Pr\{L_i = v\} > 0$, $u - v$ spans $\mathbb{Z}^{d'}$ as u ranges over all d'-vectors with positive probability (otherwise, all the vectors $u - v$ would lie in a hyperplane and the dimension d' could be reduced by standard transformations).

Example. Consider again the example presented in Section 2, of a ternary source with a scalar parameter ($d = 1$) and symbol probabilities taking the form $C(\theta) \cdot e^{\theta \beta_h}$, $h = 0, 1, 2$, where the ratio $(\beta_1 - \beta_0)/(\beta_2 - \beta_0)$ is irrational. Here, $\tau(X)$ (in this case, a scalar, so that the index j is omitted) can take the values β_0, β_1, and β_2. Therefore, for each value of τ, there exists a unique pair of integers (L_1, L_2) such that $\tau = \beta_0 + (\beta_1 - \beta_0)L_1 + (\beta_2 - \beta_0)L_2$ (specifically, the pairs $(0, 0)$, $(1, 0)$, and $(0, 1)$, for τ taking the values β_0, β_1, and β_2, respectively), and $d' = r_1 = 2$. As a result, the lattice random vectors $L(X_i)$ take the values $(0, 0)$, $(1, 0)$, and $(0, 1)$, with positive probability. Notice that we have chosen the maximal possible values for the incommensurable real constants in the representation, and the vectors $(1, 0)$ and $(0, 1)$ indeed span $\mathbb{Z}^2$.

As for Property P2, it clearly follows from the construction of $\boldsymbol{L}(x^n)$, and the fact that, for every given sequence x^n, the exponential families satisfy, under Assumption A2,

$$\boldsymbol{E}_{\hat{\theta}(x^n)} \boldsymbol{\tau}(X^n) = \boldsymbol{\tau}(x^n) \,.$$

In some cases, it is convenient to use the MLE $\hat{\theta}(X^n)$ as the (real) random vector from which $\boldsymbol{L}(X^n)$ is derived. Clearly, Property P2 will follow from the unbiasedness of the MLE. Notice that since the mapping must be one-to-one, this is only possible when, as in the case of exponential families, two sequences belong to the same type class if and only if they yield the same MLE.

Next, to evaluate $\log |T_{x^n}|$ for a given sequence x^n, we observe that since, by Property P1,

$\boldsymbol{L}(X^n)$ is a one-to-one function of T_{X^n}, for any parameter value θ we have

$$|T_{x^n}| = \frac{P_\theta\{\boldsymbol{L}(X^n) = \boldsymbol{L}(x^n)\}}{P_\theta(x^n)}\,.$$

In particular, if the MLE of θ for x^n lies in Ω, we can choose $\theta = \hat{\theta}(x^n)$ and use Property P2 to obtain

$$|T_{x^n}| = \frac{P_{\hat{\theta}(x^n)}\{\boldsymbol{L}(X^n) = \boldsymbol{E}_{\hat{\theta}(x^n)}[\boldsymbol{L}(X^n)]\}}{P_{\hat{\theta}(x^n)}(x^n)}\,. \tag{A.2}$$

The numerator in (A.2) is the probability that an integer-valued random d'-vector, which by Property P1 is the sum of n i.i.d. random vectors, equals its expected value (which is also in $\mathbb{Z}^{d'}$). This probability can be evaluated using the local limit theorem of [4, Corollary I, part II], to obtain[5]

$$P_{\hat{\theta}(x^n)}\{\boldsymbol{L}(X^n) = \boldsymbol{E}_{\hat{\theta}(x^n)}[\boldsymbol{L}(X^n)]\} = \frac{1+o(1)}{(2\pi n)^{d'/2}\sqrt{\det M[\hat{\theta}(x^n)]}}$$

where $M[\hat{\theta}(x^n)]$ denotes the covariance matrix of the random variables L_i evaluated at the MLE of θ for x^n. Thus, (A.2) takes the form

$$\log|T_{x^n}| = -\log P_{\hat{\theta}(x^n)}(x^n) - \frac{d'}{2}\log(2\pi n) - \frac{1}{2}\log\det M[\hat{\theta}(x^n)] + o(1)\,. \tag{A.3}$$

Now, to compute $\boldsymbol{E}_\theta \log|T_{X^n}|$ for a given parameter $\theta \in \Omega$, we take a ball around θ that lies inside Ω, and for those (typical) sequences x^n for which the MLE lies inside the ball we use (A.3), whereas for the non-typical sequences we use the trivial upper bound $n\log|\mathcal{A}|$ on the left hand side. After standard manipulations involving typicality arguments and the almost sure convergence of $\hat{\theta}(x^n)$ to θ, we obtain

$$\boldsymbol{E}_\theta \log|T_{X^n}| = -\boldsymbol{E}_\theta \log P_{\hat{\theta}(X^n)}(X^n) - \frac{d'}{2}\log(2\pi n) - \frac{1}{2}\log\det M(\theta) + o(1)\,. \tag{A.4}$$

Notice that for the case in which $\mathcal{P}$ is the entire DMS class with positive symbol probabilities, the asymptotic expansion (A.4) can be obtained by applying Stirling's formula to the type size

$$\log|T_{x^n}| = \log\left[\frac{n!}{\prod_{a\in\mathcal{A}}(n_{x^n}(a))!}\right]$$

where $n_{x^n}(a)$ denotes the number of occurrences of a symbol a in x^n, and taking the expectation. Here, $d' = d$ and $\det M(\theta) = \prod_{a\in\mathcal{A}} P_\theta(a)$.

[5]The local limit theorem of [4] was chosen due to the simplicity of the required assumptions, but earlier similar results exist in the literature. In particular, the local limit theorem for lattice random vectors of [2] does not require $\boldsymbol{L}$ to be a sum of i.i.d. random vectors, and can therefore be used when $\mathcal{P}$ is not a family of DMS's, e.g., in the Markov case.

Finally, using the fact that $n\boldsymbol{E}_\theta D(P_{\hat{\theta}(X^n)}\|P_\theta)$ tends to $\frac{d}{2}\log e$, where $D(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence between two probability mass functions on $\mathcal{A}$ (see [5] and ref. [19, Proposition 5.2] therein), we conclude that

$$\boldsymbol{E}_\theta \log|T_{X^n}| = nH - \frac{d'}{2}\log(2\pi n) - \frac{d}{2}\log e - \frac{1}{2}\log[\det M(\theta)] + o(1)\,. \tag{A.5}$$

Next, we use the asymptotic expansion (A.5) to evaluate the lower bound of Theorem 1. With $\mathcal{D} \stackrel{\triangle}{=} \{x^n : \log|T_{x^n}| \geq nR\}$, we have the decomposition

$$\boldsymbol{E}\min\{nR, \log|T_{X^n}|\} = nR\cdot\sum_{x^n\in\mathcal{D}} P(x^n) + \sum_{x^n\in\mathcal{D}^c} P(x^n)\log|T_{x^n}|\,. \tag{A.6}$$

Thus,

$$\boldsymbol{E}\log|T_{X^n}| \geq \boldsymbol{E}\min\{nR, \log|T_{X^n}|\} \geq \boldsymbol{E}\log|T_{X^n}| - n\log(|\mathcal{A}|)\cdot\Pr\{\mathcal{D}\}$$

where the second inequality follows from omitting the first term in the decomposition (A.6), and using the trivial bound $\log|T_{x^n}| \leq |\mathcal{A}|^n$ for $x^n \in \mathcal{D}$. Now, for $R > H$, we will show that $\Pr\{\mathcal{D}\}$ decays exponentially fast with n, and therefore, by (A.5), the asymptotic value of the lower bound of Theorem 1 is $\frac{d'}{2}\log(2\pi n) + \frac{d}{2}\log e + \frac{1}{2}\log[\det M(\theta)]$. Similarly, omitting the second term in (A.6),

$$nR \geq \boldsymbol{E}\min\{nR, \log|T_{X^n}|\} \geq nR(1-\Pr\{\mathcal{D}^c\})\,. \tag{A.7}$$

Considering now the case $R < H$, we will show that $\Pr\{\mathcal{D}^c\}$ decays exponentially fast with n, and therefore the lower bound of Theorem 1 behaves like $n(H-R)$ within an exponentially vanishing term.

To show that the probability of the set $\mathcal{D}$ behaves as claimed, it suffices to prove the following lemma, which generalizes analogous results for elementary type classes [8].

Lemma A.1 *For any $\delta > 0$, let $\mathcal{D}_\delta \stackrel{\triangle}{=} \{x^n : \ |\log|T_{x^n}| - nH| > n\delta\}$. Then,*

$$\limsup_{n\to\infty}\frac{1}{n}\log\Pr\{\mathcal{D}_\delta\} < 0\,.$$

Proof. By the definition of a class type,

$$0 \geq \log|T_{x^n}| - \log\frac{1}{P(x^n)} = \log P(T_{x^n}) \tag{A.8}$$

and

$$\Pr\{P(T_{X^n}) < 2^{-n\delta/2}\} \leq N(\mathcal{P}, n)2^{-n\delta/2}$$

where $N(\mathcal{P}, n)$ grows *polynomially* fast with n (as the number of type classes is no larger than the number of elementary type classes). Therefore, with probability that approaches one exponentially fast, the right hand side of (A.8) is lower-bounded by $-n\delta/2$. In addition, by the Asymptotic Equipartition Property (AEP), $-\log P(x^n)$ is within $-n\delta/2$ of nH with probability that approaches one exponentially fast, implying the exponentially vanishing probability of $\mathcal{D}_\delta$. □

Appendix B

Proof of Lemma 1. For a given $\epsilon > 0$, we shall estimate the relative number of 'bad' members of $\mathcal{F}$ which violate eq. (31) for some $(x^m, y^n) \in \mathcal{S}$. We show that the number of such 'bad' schemes is a doubly exponentially small fraction of the total size of the family, $|\mathcal{F}| = \prod_{T \in \mathcal{T}^m} |T|!$.

For a given x^m, let us denote by y^m, the permuted m-sequence defined by eq. (30), but without applying the truncation operator $[\cdot]_1^n$, that is,

$$y^m = J_m^{-1}\left(J_m(x^m) \oplus \sum_{i=1}^{k} 2^{i-1} u_i\right). \tag{B.1}$$

First, observe that given x^m, there is a set of $2^{nR'(x^m)} \stackrel{\triangle}{=} 2^{nR} \bmod |T_{x^m}|$ different sequences $\{y^m\}$, which we shall denote by $\mathcal{Y}(x^m)$, that are obtained from (B.1) one more time than the other sequences in T_{x^m} as u^k exhausts $\mathcal{B}^k$. Also, since $y^n \tilde{z}^r \in T_{x^m}$ for every $\tilde{z}^r \in T_{z^r}$, there are $|T_{z^r}|$ sequences in T_{x^m} that start with y^n as a prefix. Therefore, it is easy to see that the probability of y^n given x^m, $W(y^n|x^m)$, takes the form

$$W(y^n|x^m) = (1 - 2^{-n(R-R'(x^m))}) W^*(y^n|x^m) + 2^{-nR} |\mathcal{Y}(x^m, y^n)| \tag{B.2}$$

where $|\mathcal{Y}(x^m, y^n)|$ denotes the number of members of $\mathcal{Y}(x^m)$ whose first n coordinates coincide with y^n. So the first question to be addressed is the following: How many 'bad' permutations of the members of T_{x^m} are there such that more than $L_0 \stackrel{\triangle}{=} W^*(y^n|x^m)[2^{nR'(x^m)} + 2^{n(R-\epsilon)}]$ members of $\mathcal{Y}(x^m)$ start with y^n as a prefix, and thus, by (B.2), do not satisfy the second inequality in (31)? Similarly, for how many permutations, less than $L_1 \stackrel{\triangle}{=} W^*(y^n|x^m)[2^{nR'(x^m)} - 2^{n(R-\epsilon)}]$ members of $\mathcal{Y}(x^m)$ begin with y^n, and thus do not satisfy the first inequality in (31)?

Notice that if $2^{n(R-\epsilon)} > |T_{x^m}|$, then $L_0 > |T_{z^r}|$ and $L_1 < 0$, so that the answer to the above questions is trivial in this case: all permutations satisfy eq. (31). Thus, we shall

upper-bound the number of 'bad' permutations for the first question under the assumption $L_0 \leq |T_{z^r}|$ (implying $2^{n(R-\epsilon)} \leq |T_{x^m}|$). For the second one (under the analogous assumption $L_1 \geq 0$), the technique is very similar, and the analysis will be omitted. Assume that the given x^m is in a certain fixed position in its type class, say, $J(x^m) = 0$. The number of permutations μ of the remaining sequences, such that at least L_0 members of $\mathcal{Y}(x^m)$ begin with y^n, is given by

$$\mu = (|T_{x^m}| - 2^{nR'(x^m)})! \sum_{\ell \geq L_0} \left(\begin{array}{c} 2^{nR'(x^m)} \\ \ell \end{array} \right) \prod_{i=0}^{\ell-1} (|T_{z^r}| - i) \prod_{j=0}^{2^{nR'(x^m)}-\ell-1} (|T_{x^m}| - |T_{z^r}| - j) \quad \text{(B.3)}$$

where each summand corresponds to all combinations of $2^{nR'(x^m)}$ sequences (that form $\mathcal{Y}(x^m)$) such that exactly ℓ members of them are prefixed by y^n, and the factor in front of the summation is the number of permutations of the members of $T_{x^m} \cap [\mathcal{Y}(x^m)]^c$. Equivalently, μ can be rewritten as follows:

$$\begin{aligned} \mu &= (|T_{x^m}| - 2^{nR'(x^m)})! \sum_{\ell \geq L_0} \frac{(2^{nR'(x^m)})!}{\ell!(2^{nR'(x^m)} - \ell)!} \cdot \frac{|T_{z^r}|!}{(|T_{z^r}| - \ell)!} \cdot \frac{(|T_{x^m}| - |T_{z^r}|)!}{(|T_{x^m}| - |T_{z^r}| - (2^{nR'(x^m)} - \ell))!} \\ &= |T_{x^m}|! \cdot \frac{\sum_{\ell \geq L_0} \left(\begin{array}{c} |T_{z^r}| \\ \ell \end{array} \right) \cdot \left(\begin{array}{c} |T_{x^m}| - |T_{z^r}| \\ 2^{nR'(x^m)} - \ell \end{array} \right)}{\left(\begin{array}{c} |T_{x^m}| \\ 2^{nR'(x^m)} \end{array} \right)}. \end{aligned} \quad \text{(B.4)}$$

Since the first factor of the last expression, $|T_{x^m}|!$, is the total number of permutations of the members of T_{x^m}, the second factor, is the fraction of permutations for which $W(y^n|x^m) \geq (1 + 2^{-n\epsilon})W^*(y^n|x^m)$. 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 right–most side of (B.4). The numerator is upper-bounded using the fact that for any two nonnegative integers N and K ($K \leq N$):

$$\left(\begin{array}{c} N \\ K \end{array} \right) \leq 2^{Nh(K/N)}$$

where $h(t) \stackrel{\triangle}{=} -t \log t - (1-t)\log(1-t)$, $t \in [0,1]$. Specifically,

$$\begin{aligned} & \sum_{\ell \geq L_0} \left(\begin{array}{c} |T_{z^r}| \\ \ell \end{array} \right) \cdot \left(\begin{array}{c} |T_{x^m}| - |T_{z^r}| \\ 2^{nR'(x^m)} - \ell \end{array} \right) \\ \leq & \sum_{\ell \geq L_0} \exp_2 \left\{ |T_{z^r}| \cdot h\left(\frac{\ell}{|T_{z^r}|} \right) \right\} \cdot \exp_2 \left\{ (|T_{x^m}| - |T_{z^r}|) \cdot h\left(\frac{2^{nR'(x^m)} - \ell}{|T_{x^m}| - |T_{z^r}|} \right) \right\} \\ \leq & \; 2^{nR'(x^m)} \max_{\ell \geq L_0} \exp_2 \left\{ |T_{z^r}| \cdot h\left(\frac{\ell}{|T_{z^r}|} \right) + (|T_{x^m}| - |T_{z^r}|) \cdot h\left(\frac{2^{nR'(x^m)} - \ell}{|T_{x^m}| - |T_{z^r}|} \right) \right\} \\ \leq & \; 2^{nR'(x^m)} \cdot 2^{|T_{x^m}| \cdot F} \end{aligned} \quad \text{(B.5)}$$

where $F = \max\{qh(\alpha) + (1-q)h(\beta)\}$, $q \stackrel{\triangle}{=} |T_{z^r}|/|T_{x^m}| = W^*(y^n|x^m)$, the maximum being over all pairs (α, β) for which $\alpha \geq (1+2^{-n\nu})\gamma$ and $q\alpha+(1-q)\beta = \gamma$, with $\gamma \stackrel{\triangle}{=} 2^{nR'(x^m)}/|T_{x^m}|$ and $\nu \stackrel{\triangle}{=} R'(x^m) - R + \epsilon$. It is easy to show that the function $qh(\alpha)+(1-q)h((\gamma-q\alpha)/(1-q))$ is monotonically decreasing in α for $\alpha \geq \gamma$. Thus, the maximum defining F is attained for $\alpha = \alpha_0 \stackrel{\triangle}{=} (1+2^{-n\nu})\gamma \leq 1$, where the inequality follows from the assumption $L_0 \leq |T_{z^r}|$. As a result, the numerator of the expression at hand is upper-bounded by

$$2^{nR'(x^m)} \cdot \exp_2\left\{|T_{x^m}| \cdot [qh(\alpha_0) + (1-q)h(\beta_0)]\right\}$$

where $\beta_0 \stackrel{\triangle}{=} (\gamma - q\alpha_0)/(1-q)$. The denominator, on the other hand, is lower-bounded [8] by

$$\left(\begin{array}{c} |T_{x^m}| \\ 2^{nR'(x^m)} \end{array} \right) \geq \frac{1}{|T_{x^m}|+1} \cdot \exp_2\{|T_{x^m}| \cdot h(\gamma)\}\,. \tag{B.6}$$

On substituting the upper bound on the numerator and the lower bound on the denominator into eq. (B.4), the exponent of the denominator is subtracted from that of the numerator and we obtain:

$$\begin{aligned} qh(\alpha_0) + (1-q)h(\beta_0) - h(\gamma) &= -qD(\alpha_0\|\gamma) - (1-q)D(\beta_0\|\gamma) \\ &\leq -qD(\alpha_0\|\gamma) \end{aligned} \tag{B.7}$$

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

$$\mu \leq (|T_{x^m}|+1)! \cdot 2^{nR'(x^m)} \exp_2\left\{-|T_{z^r}| \cdot D((1+2^{-n\nu})\gamma\|\gamma)\right\}\,. \tag{B.8}$$

To further upper bound μ, we next derive a lower bound on $|T_{z^r}| \cdot D((1+2^{-n\nu})\gamma\|\gamma)$. Using the fact that

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

we have the following lower bound on the divergence:

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

$$\begin{aligned} & \frac{1}{\ln 2}[1-(1+2^{-n\nu})\gamma]\cdot\frac{\gamma 2^{-n\nu}}{1-(1+2^{-n\nu})\gamma} \\ = \;& \frac{\gamma}{\ln 2}[(1+2^{-n\nu})\ln(1+2^{-n\nu})-2^{-n\nu}] \\ \geq \;& \frac{\gamma 2^{-2n\nu}}{3\ln 2} \quad \forall \text{ large } n \end{aligned} \tag{B.9}$$

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

$$\begin{aligned} |T_{z^r}|\cdot D((1+2^{-n\nu})\gamma\|\gamma) &\geq \frac{1}{3\ln 2}\cdot 2^{\log|T_{z^r}|-\log|T_{x^m}|-nR'(x^m)+2nR-2n\epsilon} \\ &\geq \frac{1}{3\ln 2}\cdot 2^{\log|T_{z^r}|-\log|T_{x^m}|+nR-2n\epsilon} \geq \frac{2^{n\epsilon}}{3\ln 2} \end{aligned} \tag{B.10}$$

where the second inequality follows from $R'(x^m) < R$, and the last one from the assumption that $(x^m, y^n) \in \mathcal{S}$.

We conclude that for a given $(x^m, y^n) \in \mathcal{S}$, and a given location $J(x^m)$ in the list of T_{x^m}, the number of permutations of the remaining sequences in T_{x^m} for which L_0 or more members of $\mathcal{Y}(x^m)$ begin with y^n as a prefix, is upper-bounded by

$$\mu \leq (|T_{x^m}|+1)!\cdot 2^{nR}\cdot\exp_2\left\{-\frac{2^{n\epsilon}}{3\ln 2}\right\}$$

for large n, where we have used, again, the inequality $R'(x^m) < R$. Multiplying this bound by the $|T_{x^m}|$ possible locations of x^m in T_{x^m} and by the cardinality of $\mathcal{S}$, which are both bounded by exponential functions of m, we deduce that the total number of permutations that have this property for *some* location of x^m in the list and for *some* pair $(x^m, y^n) \in \mathcal{S}$ is still a vanishing fraction of the total number of permutations, $|T_{x^m}|!$ (as we assume $\log m = o(n)$). This conclusion remains unchanged even after taking into account also the permutations for which $\ell \leq L_1$, whose number is also bounded (similarly) by a doubly exponentially small fraction of $|T_{x^m}|!$. Thus, not only does a good channel W (corresponding to a scheme $\phi \in \mathcal{F}$) exist in the sense of satisfying Lemma 1, but actually the vast majority of channels satisfy Lemma 1. □

Appendix C

Proof of Lemma 2. Proceeding as with the set $\mathcal{D}$ in Appendix A, $\Pr\{\mathcal{E}\}$ clearly vanishes exponentially fast with n for $H < R$. Thus, by the union bound, it suffices to show that $\Pr\{(x^m, y^n) : nR < \log|T_{x^m}| - \log|T_{z^r}| + 3n\epsilon\}$ also vanishes exponentially fast with n. Since

Z^r is drawn according to P, we have

$$\log|T_{x^m}| - \log|T_{z^r}| \leq \log\frac{1}{P(x^m)} - \log\frac{P(T_{z^r})}{P(z^r)} = \log\frac{1}{P(y^n)} - \log P(T_{z^r})$$

where the last term on the right hand side satisfies, with $\delta = (R - H - 3\epsilon)/2 > 0$,

$$\Pr\{P(T_{Z^r}) < 2^{-n\delta}\} < N(\mathcal{P}, r)2^{-n\delta}\,.$$

Since $N(\mathcal{P}, r)$ grows polynomially fast with r and $\log r = o(n)$, the result follows from noting that, by the AEP, $-\log P(y^n)$ is within δ of nH with probability that approaches one exponentially fast with n. □

References

[1] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 2104–2109, September 1999.

[2] E. A. Bender and L. B. Richmond, "Central and local limit theorems applied to asymptotic enumeration II: Multivariate generating functions," *Journal of Combinatorial Theory*, ser. A 34, pp. 255-265, 1983.

[3] R. N. Bhattacharya and R. R. Rao, *Normal Approximation and Asymptotic Expansions*. New York: John Wiley, 1976.

[4] A. A. Borovkov and A. A. Mogulskii, "Integro-local limit theorems including large deviations for sums of random vectors," *Theory Probab. Appl.*, vol. 43, pp. 1-12, 1999.

[5] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453-471, May 1990.

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

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley, 1991.

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

[9] W. Feller, *An Introduction to Probability Theory and its Applications.* New York: John Wiley, 1971.

[10] R. G. Gallager, "Variations on a theme by Huffman," *IEEE Trans. Inform. Theory*, vol. 24, pp. 668–674, November 1978.

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

[12] 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.

[13] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Transactions on Information Theory*, vol. IT–30, no. 4, pp. 629–636, July 1984.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications," in: *Information Theory New Trends and Open Problems*, edited by G. Longo, Springer-Verlag, 1975, pp. 87–123.

[19] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Trans. Inform. Theory*, vol. IT–19, no. 3, pp. 275–283, May 1973.