KYBERNETIKA — VOLUME *56* (2020), NUMBER 6, PAGES 1090–1110

ON TYPICAL ENCODINGS OF MULTIVARIATE ERGODIC SOURCES

Michal Kupsa

In memory of Fero Matúš

We show that the typical coordinate-wise encoding of multivariate ergodic source into prescribed alphabets has the entropy profile close to the convolution of the entropy profile of the source and the modular polymatroid that is determined by the cardinalities of the output alphabets. We show that the proportion of the exceptional encodings that are not close to the convolution goes to zero doubly exponentially. The result holds for a class of multivariate sources that satisfy asymptotic equipartition property described via the mean fluctuation of the information functions. This class covers asymptotically mean stationary processes with ergodic mean, ergodic processes, irreducible Markov chains with an arbitrary initial distribution. We also proved that typical encodings yield the asymptotic equipartition property for the output variables. These asymptotic results are based on an explicit lower bound of the proportion of encodings that transform a multivariate random variable into a variable with the entropy profile close to the suitable convolution.

Keywords: entropy, entropy rate, multivariate source, ergodic source, a.e.p. property

Classification: 94A24, 94A29

1. INTRODUCTION

In Information theory and namely in Network coding theory, the encoding of multiple possibly correlated sources have been extensively studied in the last two decades (for the main framework, overview and references see [10] and [1]). The sources are usually assumed to be discrete and memory-less, so they are represented by sequences of i.i.d. discrete random variables. The important characteristics of multivariate sources, which we focus on, is its entropy profile (an entropic point corresponding to a given multivariate source, see [5]).

We deal with a problem of how the entropy profile changes when "typical" coordinate-wise encodings into prescribed output alphabets are applied. As was shown in [7, Theorem 3], in an asymptotic case, a typical encoding saves as much of the original information as possible. Namely, the conditional entropy of the encoded variable is naturally bounded from above by the conditional entropy of the source and also by the logarithm

DOI: 10.14736/kyb-2020-6-1090

of the size of the output alphabet. Matúš showed that this bound is asymptotically tight. More literally, a typical coordinate-wise encoding preserves almost all conditional entropy whenever the output alphabet is large enough, i. e., when the logarithm of the alphabet size exceeds the conditional entropy. If the logarithm is not larger, the conditional entropy of the encoded variable is close to the logarithm of the alphabet size. This observation is used to prove the closeness of the entropy region under the convolution with modular polymatroids ([7, Theorem 2]). The role of convolution in the research on the entropy region was then more explored in [8].

In [7], the coordinate-wise encodings of the original multivariate source into prescribed alphabets are applied inductively, coordinate by coordinate, and in between these inductive steps, one has to pass from a random vector to its i.i.d. expansion (see the proof of Theorem 2 in the discussed paper). In particular, each encoding is applied to a different random vector. This procedure can be reinterpreted as simultaneous coordinate encodings used on one fixed i.i.d. expansion of the original entropy vector. But this simple reasoning does not allow to deduce that the entropy profile obtained for a specific encoding is also realized by the most of the coordinate-wise encodings from some natural domain, as it can be concluded in the one-dimensional case.

The first step towards the results for "typical" encodings in the multivariate case was done in [9], where the authors proved that the proportion of encodings of a two-dimensional random vector that realizes a given convolution goes to one doubly exponentially. It is also explained there that the encodings behave well, not only when they are applied on i.i.d. copies of some random vector, but also when we apply them on any vector that is drawn from bi-variate (strictly stationary) ergodic source.

Our work presented in this article extends the control on the entropy-profile of transformed variables for the general multivariate case whenever the original source possesses asymptotic equipartition property (AEP). Our main results are stated in Theorems 2, 3 and 6. In Theorem 2, we introduce an explicit lower bound of the proportion of encodings that transform a multivariate random variable into a variable with the entropy profile close to the suitable convolution. The bound is given in terms of the entropy of the original random variable and works when the mean fluctuation of information functions is small. We use the bound to develop an asymptotic scheme in Theorem 3 that is applied to get the result on typical encodings for ergodic processes (Theorem 6). Last but not least, we control not only the entropy profile of the transformed variables, but we show that they also possess some kind of equipartition property. To describe the equipartition property of a random variable, we introduce a new quantity that measures the non-uniformness in a way that is well preserved via transformations (encodings), conditioning, and i.i.d. expansions, namely the mean fluctuation of the information functions, see Section 2 (and Section 6 for more details).

Let us stress out that the extraction of the critical property, namely the asymptotic equipartition property, which is sufficient assumption in Theorem 3, allows us to extend the previous works on this topics in two significant ways; the source needs to be neither i.i.d., nor stationary. It is satisfactory if the original process is asymptotically mean stationary with ergodic mean, as defined in [4, page 16] (the mentioned result can be found therein as Theorem 4.1 and Section 4.5), e. g., finite-state Markov chains of any order, its functions, block codings of stationary processes, etc. For the same reason, our

results can be extended to the situation when one considers a family of ergodic random fields (corresponding with an action of an amenable group, see [6]) instead of a family of random processes. In Section 5, we introduce an example of a non-stationary Markov chain and its encodings in a binary alphabet that shows the generality of our results.

We generalize the known results also in another important direction. Our results also cover the situation when only some coordinates of the multivariate source are encoded. In particular, the theorem can be used to describe the common entropy profile of the family that consists of the original variables as well as the encoded ones.

2. EQUIPARTITION PROPERTY AND MEAN FLUCTUATION OF THE INFORMATION FUNCTION

Let us recall some basic notions from Information Theory. Let $\mathbb{P}$ be a probability on a finite set $\mathcal{X}$ (not necessarily a subset of real or complex domain). The information function $\mathcal{I}_{\mathbb{P}} : \mathcal{X} \to \mathbb{R}$ is given by the formula $\mathcal{I}_{\mathbb{P}}(x) = -\ln \mathbb{P}(x)$. The entropy $H(\mathbb{P})$ is its expectation, i. e. $H(\mathbb{P}) = \sum \mathbb{P}(x)(-\ln \mathbb{P}(x))$, where we sum over all $x \in \mathcal{X}$ of positive probability. The set of all $x \in \mathcal{X}$ of positive probability is the support of $\mathbb{P}$, denoted by $s(\mathbb{P})$. A discrete finite-valued random variable X, e. g. a measurable map from a probability space $(\Omega, \mathbb{P})$ into a finite set $\mathcal{X}$, induces in a natural way a discrete probability measure $\mathbb{P}_X$ on $\mathcal{X}$, so we can extend immediately the previous notions, the information function, the entropy and the support, as follows:

$$s(X) := s(\mathbb{P}_X),\ \mathcal{I}_X := \mathcal{I}_{\mathbb{P}_X}, H(X) := H(\mathbb{P}_X).$$

We will often use in the text this small abuse of notation when the random variable is written down instead of the induced probability.

Let $\mathbb{X} = (X(n))_{n\in\mathbb{N}}$ be a random process with values in a finite alphabet $\mathcal{A}$. For given n, $X^{(n)} = (X(i))_{i=1}^{n}$ is understood as a random variable with values in $\mathcal{A}^n$. The entropy rate of the process $\mathbb{X}$ is defined by the formula

$$h(\mathbb{X}) = \lim_{n\to\infty} \frac{1}{n} H(X^{(n)}).$$

The asymptotic equipartition property (AEP) claims that the entropy rate is well defined, the limit above exists, and it is equal to the limit of $\frac{1}{n}\mathcal{I}_{X^{(n)}}$ with probability one (it is well known that i.i.d. processes, ergodic processes posses AEP, see [2]). In particular, the AEP claims that $\frac{1}{n}\mathcal{I}_{X^{(n)}}$ is very "flat". In order to describe such behavior in an efficient manner, we introduce the following quantities for a discrete probability $\mathbb{P}$ and a real value $a \in \mathbb{R}$:

$$M(\mathbb{P}, a) = \mathbb{E}_{\mathbb{P}} \left| \mathcal{I}_{\mathbb{P}} - a \right|, \qquad M(\mathbb{P}) = M(\mathbb{P}, H(\mathbb{P})).$$

We call $M(\mathbb{P}, a)$ the mean fluctuation of the information function from a and $M(\mathbb{P})$ simply the mean fluctuation. We again extend these definitions for random variables with finite values in the natural way, $M(X, a) := M(\mathbb{P}_X, a)$ and $M(X) := M(\mathbb{P}_X)$.

3. MULTIVARIATE RANDOM VARIABLES

In order to study the multivariate random variables, we admit that a random variable X has its own structure, namely, that $X = (X_i)_{i=1}^k$, where X_i is a random variable with values in a finite alphabet $\mathcal{A}_i$, $i \leq k$. For a natural number k we denote the set $\{1, 2, \ldots, k\}$ by $\hat{k}$ and the set of all subsets of $\hat{k}$ by $\tilde{k}$. The entropy profile of X is the point $\vec{H}(X) \in \mathbb{R}^{\tilde{k}}$ given by the formula, $\vec{H}(X) = (H(X_I))_{I \in \tilde{k}}$, where X_I denotes the sub-vector $(X_i)_{i \in I}$ ($H(X_\emptyset)$ is defined to be zero). In a consistent manner, we put $M(X_\emptyset) = 0$ and define the maximal mean fluctuation

$$M'(X) = \max_{I \in \tilde{k}} M((X_i)_{i \in I}).$$

We will consider coordinate-wise encodings that encode all or only some of the coordinates. For this generality, $\ell \leq k$ is specified, as well as the family $\mathcal{B} = (\mathcal{B}_i)_{i=1}^\ell$ of finite output alphabets.

A mapping f from $\prod_{i=1}^k \mathcal{A}_i$ to $\prod_{i=1}^\ell \mathcal{B}_i \times \prod_{k=\ell+1}^k \mathcal{A}_i$ is a coordinate-wise encodings of the first ℓ coordinates if it satisfies the formula

$$f(x_1, \ldots, x_k) = (f_1(x_1), f_2(x_2), \ldots, f_\ell(x_\ell), x_{\ell+1}, x_{\ell+2}, \ldots, x_k)$$

for some family of mappings $(f_i)_{i \leq \ell}$, where $f_i : \mathcal{A}_i \to \mathcal{B}_i$, $i \leq \ell$.

Since f is determined by $(f_i)_{i \leq \ell}$, we identify the mapping with the family, i. e. we write $f = (f_i)_{i \leq \ell}$. The set of all these coordinate-wise encodings of the first ℓ coordinates is denoted by $\mathcal{E}_\ell$.

We define also the entropy profile $\vec{H}(\mathcal{B}) \in \mathbb{R}^{\tilde{\ell}}$ of the output alphabets as follows

$$(\vec{H}(\mathcal{B}))_I = \ln |\prod_{i \in I} \mathcal{B}_i|, \qquad I \in \tilde{\ell}.$$

We are interested in the question how the encodings change the entropy profile, i. e. what we can say about $\vec{H}(f(X)) = H(f_I(X_I))_{I \in \tilde{k}}$. It is quite straightforward to show that the profile is coordinate-wise bounded by the convolution $\vec{H}(X) * \vec{H}(\mathcal{B})$. In general, convolution $w = u * v$ of two points $u \in \mathbb{R}^{\tilde{k}}$ and $v \in \mathbb{R}^{\tilde{\ell}}$, $\ell \leq k$, is the point from $\mathbb{R}^{\tilde{k}}$ defined by the formula

$$(u * v)_I = \min_{J \subset I \cap \hat{\ell}} (u_{I \setminus J} + v_J), \qquad I \in \tilde{k}.$$

Proposition 1. Let f be an encoding from $\mathcal{E}_\ell$. Then

$$H(f(X))_I \leq (\vec{H}(X) * \vec{H}(\mathcal{B}))_I, \qquad I \in \tilde{k}.$$

Proof. By the definition of the convolution, it is enough to prove that

$$H(f_I(X_I)) \leq H(X_{I \setminus J}) + (\vec{H}(\mathcal{B}))_J,$$

for all $I \subset \tilde{k}$ and $J \subset I \cap \tilde{\ell}$. But $H(f_I(X_I))$ is bounded from above by the sum of $H(f_{I \setminus J}(X_{I \setminus J}))$ and $H(f_J(X_J))$, where the former entropy is surely bounded by the

entropy of the source $H(X_{I\setminus J})$ and the latter by the logarithm of the cardinality of the output set $\prod_{i\in J}\mathcal{B}_J$. □

In the next theorem, we show much more, namely that for a large part of encodings, the entropy profile $\vec{H}(f(X))$ is not just bounded by the convolution, but it is close to this bound. We also show that the maximal mean fluctuation can be very small at the same moment. The proof of Theorem 2 is postponed to the last section.

Theorem 2. Let $1 \le k$, $1 \ge \varepsilon > 0$, $\ell \le k$, $\delta = \left(\frac{\varepsilon}{121}\right)^{2^{|\ell|}}$, $H > 0$ and $X = (X_i)_{i\le k}$ be a family of discrete random variables such that X_i takes values in $\mathcal{A}_i$, $i \le k$, and

$$H > H(X_{\hat{\ell}}), \qquad H \ge \frac{2\ln 2}{\delta}, \qquad H \ge \frac{M'(X)}{\delta}. \tag{1}$$

The proportion of those encodings $f \in \mathcal{E}_\ell$ that satisfy the conditions

$$M'(f(X)) \le \varepsilon H \qquad \& \qquad \left\| H(f(X)) - \vec{H}(X) * \vec{H}(\mathcal{B}) \right\|_{\max} \le \varepsilon H, \tag{2}$$

is at least

$$1 - |\ell| 2^{k-1} \exp\left(-\frac{\ln 2}{2} e^{\delta H} + (\vec{H}(\mathcal{B}))_{\hat{\ell}} + 2H \right).$$

Let us notice that for a fixed dimension k, the bound for the proportion of the encodings in the theorem goes to one very fast ("doubly exponentially") with respect to H, provided H goes to infinity, and M'/H goes to zero. In the next section, we apply this idea and the theorem in the situation when an ergodic source and an a.m.s. source is encoded.

Another important remark is that we control the behavior of the entropy of the output variable $f(X)$ as well as the quantity $M'(f(X))$. To control both is necessary to use Theorem 2 inductively in the proof of the Theorem 3.

4. ASYMPTOTIC SCHEME

In this section, instead of encodings of one family of random variables, we will consider a sequence of families and their encodings to different alphabets. Our aim is to construct an asymptotic scheme that is presented in Theorem 3.

We fix $\ell \le k$. For given $n \ge 1$, we consider a family of random variables $X^{(n)} = (X_i^{(n)})_{i\le k}$ defined on the same probability space, where $X_i^{(n)}$ takes values in a finite set $\mathcal{A}_i^{(n)}$. Put $\mathcal{A}^{(n)} = \prod_{i\le k}\mathcal{A}_i^{(n)}$. As well as in the previous section, we fix a family of finite sets $\mathcal{B}^{(n)} = (\mathcal{B}_i^{(n)})_{i\le\ell}$ and denote by $\mathcal{E}_\ell^{(n)}$ the set of all mappings from $\mathcal{A}^{(n)}$ to $\mathcal{B}^{(n)}$ of the form

$$f(x_1, \dots, x_k) = (f_1(x_1), f_2(x_2), \dots, f_\ell(x_\ell), x_{\ell+1}, x_{\ell+2}, \dots, x_k),$$

for some $f_i : \mathcal{A}_i^{(n)} \to \mathcal{B}_i^{(n)}$, $i \le \ell$. Let us recall, that we call these mappings coordinate-wise encodings of the first ℓ coordinates.

Theorem 3. Let $\frac{\vec{H}(X^{(n)})}{n}$ converges to a non-zero $h \in \mathbb{R}^{\tilde{k}}$, $\frac{M'(X^{(n)})}{n}$ tends to zero and $\frac{\vec{H}(\mathcal{B}^{(n)})}{n}$ converges to $b \in \mathbb{R}^{\tilde{\ell}}$.

If $1 \geq \varepsilon > 0$, $\delta < \left(\frac{\min(\varepsilon, h_{\tilde{k}})}{121 h_{\tilde{k}}}\right)^{2^{|\ell|}}$, n large enough, then the proportion of those encodings $f \in \mathcal{E}_{\ell}^{(n)}$ that satisfy the conditions

$$\frac{M'(f(X^{(n)}))}{n} \leq \varepsilon \qquad \& \qquad \left\| \frac{\vec{H}(f(X^{(n)}))}{n} - h * b \right\|_{\max} \leq \varepsilon,$$

is at least

$$1 - \exp\left(-e^{\delta n}\right).$$

The proof is postponed to the last section.

We developed the asymptotic scheme in the case when the limits of some numerical characteristics are assumed to exist. Nevertheless, the scheme does not require any structural relation between $X^{(n)}$ and $X^{(n+1)}$. In the next section, we will apply this scheme in the case when $X^{(n)}$ arises as the first n-coordinates of some process $(X(n))_{n\in\mathbb{N}}$ and where $X^{(n+1)}$ contains $X^{(n)}$ as its beginning.

5. ENCODINGS OF ERGODIC PROCESSES AND A.M.S. PROCESSES WITH ERGODIC MEAN

Let $\mathbb{X} = (X(n))_{n\in\mathbb{N}}$ be a multivariate random process with values in a Cantor product of finite sets $\mathcal{A}_i$, $1 \leq i \leq k$. Put $\mathcal{A} = \prod_{i=1}^{k} \mathcal{A}_i$. Hence, each $X(n)$ is a tuple of random variables, $X(n) = (X_i(n))_{i=1}^{k}$. For a subset of coordinates $J \subset \hat{k}$, we define a sub-process $\mathbb{X}_J = (X_J(n))_{n\in\mathbb{N}}$ in the following way: $X_J(n) = (X_j(n))_{j\in J}$. As in the previous sections, $X^{(n)}$ stands for the vector $(X(1), X(2), \ldots, X(n))$, $X_J^{(n)}$ stands for $(X_J(1), X_J(2), \ldots, X_J(n))$.

We define an entropy profile of the multivariate process $\mathbb{X}$ as the vector $\vec{h}(\mathbb{X}) = (h(\mathbb{X}_J))_{J\in\tilde{k}}$, where $h(\mathbb{X}_J)$ is the entropy rate of the process $\mathbb{X}_J$, i. e.

$$h(\mathbb{X}_J) = \lim_{n\to\infty} \frac{1}{n} H(X_J(1), X_J(2), \ldots, X_J(n)), \qquad J \subset \hat{k}.$$

There is a quite large class of processes for which the entropy rates are well defined and $M(X_J^{(n)})/n$ goes to zero for every $J \subset \hat{k}$. In order to explain this class we need to assign a process with the corresponding measure on the output-sequences. Namely, $\mathbb{X} = (X(n))_{n\in\mathbb{N}}$ with values in a finite set A gives rise the measure $\mathbb{X}_*\mathbb{P}$ on $A^{\mathbb{N}}$ that is determined by the equalities

$$\mathbb{X}_*\mathbb{P}([a_1 \ldots a_n]) = \mathbb{P}(X(0) = a_0, \ldots, X(n) = a_n), \qquad n \in \mathbb{N}, a_1, \ldots a_n \in A,$$

where

$$[a_1 \ldots a_n] = \{(x_i)_{i\in\mathbb{N}} \in A^{\mathbb{N}} \mid x_i = a_i, \text{ for } i \leq n\}, \qquad n \in \mathbb{N}, a_1, \ldots a_n \in A.$$

The measure is defined on the σ-field $\mathcal{F}$ generated by the above-mentioned sets that are usually called cylinders.

We define the shift-map T on $A^{\mathbb{N}}$ by the formula $T(x_1x_2...) = (x_2x_3...)$. A probability measure μ on $\mathcal{F}$ is

- *asymptotically mean stationary (a.m.s.)* if $\frac{1}{n}\sum_{i=1}^{n}\mu(T^{-i}F)$ converges, for every $F \in \mathcal{F}$,
- *stationary* if $\mu(T^{-1}F) = \mu(F)$ for every $F \in \mathcal{F}$,
- *ergodic* if it is stationary and $T^{-1}F = F$ and $F \in \mathcal{F}$ implies $\mu(F) = \{0, 1\}$.

We say that a process is a.m.s., stationary or ergodic, if the measure $\mathbb{X}_*\mathbb{P}$ has the corresponding property. If a process is a.m.s., then the formula

$$\mathbb{X}_*^m\mathbb{P} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_*\mathbb{P}(T^{-i}F), \qquad F \in \mathcal{F},$$

defined a probability stationary measure on $\mathcal{F}$ (the upper index "m" stands for "mean"). This measure is called the mean of the process. We will be interested in the a.m.s. processes with ergodic mean. The following theorem is a slightly weaker version of Corollary 4 in [3] translated into our notations and settings.

Proposition 4. (Gray and Kieffer [3]) Let $\mathbb{X}$ be an a.m.s. process with ergodic mean, $\mathbb{Y}$ be a stationary process with the same mean. Then the entropy rate $h(\mathbb{X})$ is well-defined and equal to $h(\mathbb{Y})$. In addition, $\frac{1}{n}M(\mathbb{X}^{(n)})$ goes to zero.

Corollary 5. Let $\mathbb{X}$ be an a.m.s. process with ergodic mean, $\mathbb{Y}$ be a stationary process with the same mean. Then the profile $\vec{h}(\mathbb{X})$ is well-defined and equal to $\vec{h}(\mathbb{Y})$. In addition, $\frac{1}{n}M'(\mathbb{X}_J^{(n)})$ goes to zero for every $J \subset \hat{k}$.

P r o o f . For $J \subset \hat{k}$, the natural projection from $(\prod_{i=1}^{k}A_i)^{\mathbb{N}}$ onto $(\prod_{i\in J}A_i)^{\mathbb{N}}$ intertwines with the shift-map on both spaces, so it is a factor mapping in the category of dynamical systems. In addition, the projection maps $(\mathbb{X})_*\mathbb{P}$ onto $(\mathbb{X}_J)_*\mathbb{P}$ and the a.m.s. property is preserved via the factor mapping. So $(\mathbb{X}_J)_*\mathbb{P}$ is a.m.s. In particular, the entropy rate $h(\mathbb{X}_J)$ is well defined and $\frac{1}{n}M(\mathbb{X}_J^{(n)})$ goes to zero. □

The following theorem is a straightforward consequence of Corollary 5 and Theorem 2.

Theorem 6. Let $\mathbb{X} = (X(n))_{n\in\mathbb{N}}$ be a.m.s. with ergodic mean, h be its entropy-rate profile. If $h_{\hat{k}} > 0$, $1 \geq \varepsilon > 0$, $\delta < \left(\frac{\min(\varepsilon, h_{\hat{k}})}{121h_{\hat{k}}}\right)^{2^{|\ell|}}$ and n large enough, then the proportion of those encodings $f \in \mathcal{E}_\ell^{(n)}$ that satisfy the conditions

$$\frac{M'(f(X^{(n)}))}{n} \leq \varepsilon \qquad \& \qquad \left\|\frac{\vec{H}(f(X^{(n)}))}{n} - h * b\right\|_{\max} \leq \varepsilon,$$

is at least

$$1 - \exp\left(-e^{\delta n}\right).$$

Let us point out that the class of a.m.s. processes with ergodic mean covers all ergodic processes (e. g., i.i.d. processes). It also contains all irreducible (possibly periodic) finite-states Markov chains. Other examples of a.m.s. processes can be found in [3], below Corollary 4.

At the end of the section, we will exhibit the generality of the theory applying previous theorem and corollary to encodings of a non-stationary and non-independent, but Markov process.

Put $k = \ell = 2$, $A_1 = A_2 = \mathbb{Z}_8$. The chain $\mathbb{X}$ is defined as the random walk on $\mathbb{Z}_8 \times \mathbb{Z}_8$ (a screwed chessboard), where we can move only one step in the horizontal direction or a one step in the vertical direction. Since 7 plus 1 is 0, we suppose that 7 and 0 are adjacent values. Namely, the transition probabilities are given by the following formula:

$$p_{(i,j)(i',j')} = \begin{cases} \frac{1}{4}, & \text{if } i = i' \text{ and } |j - j'| \in \{1, n-1\}, \\ \frac{1}{4}, & \text{if } j = j' \text{ and } |i - i'| \in \{1, n-1\}, \\ 0, & \text{otherwise.} \end{cases}$$

Let us fix a deterministic start at the origin, $X(0) = (0, 0)$. By the standard analysis of homogeneous Markov chains with finite states we get that the chain is irreducible, periodic with period two, and non-stationary because the initial distribution is not equal to the stationary one. It is straightforward that the stationary distribution is the uniform distribution. From every states, there are four equiprobable ways out. This leads to the fact that $H(X(n+1)|X(n))$ equals to $2 \ln 2$ (for any initial distribution). If we focus on the first coordinate of the process, one can see that in the next step, we can increase the value by one (modulo n) with probability 1/4, decrease the value by one (modulo n) with probability 1/4 or stay at the same value with probability 1/2. In particular, $H(X_1(n+1)|X_1(n))$ is equal to $\frac{3}{2} \ln 2$. The same is true for the second coordinate. By homogeneity, the entropy rates $\frac{1}{n}H(X^{(n)})$, $\frac{1}{n}H(X_1^{(n)})$ and $\frac{1}{n}H(X_2^{(n)})$ converge to the mentioned conditional entropies $2 \ln 2$, $\frac{3}{2} \ln 2$ and $\frac{3}{2} \ln 2$, respectively. Hence, $\frac{1}{n}\vec{H}(X^{(n)})$ converges to a non-zero $h \in \mathbb{R}^{\tilde{k}}$, where

$$h := (h_\emptyset, h_1, h_2, h_{1,2}) = (0, \frac{3}{2} \ln 2, \frac{3}{2} \ln 2, 2 \ln 2).$$

In addition, let us assume that for encoding of the first n moves in the screwed chessboard, we use 2^n colors for vertical position, as well as, for horizontal position. The alphabet $\mathcal{B}_i^{(n)}$, $i = 1, 2$, can be understood as the set of all binary strings of the length n. In the terms of the entropy,

$$\frac{1}{n}\left(\vec{H}(\mathcal{B}^{(n)})_\emptyset, \vec{H}(\mathcal{B}^{(n)})_1, \vec{H}(\mathcal{B}^{(n)})_2, , \vec{H}(\mathcal{B}^{(n)})_{1,2}\right) = (0, \ln 2, \ln 2, 2 \ln 2).$$

Applying Theorem 6, we can say that a typical pair of encodings $f = (f_1, f_2)$, $f_1 : (\mathbb{Z}_8)^n \to 2^n$ and $f_1 : (\mathbb{Z}_8)^n \to 2^n$, yields the transformations of $X^{(n)}$ with the entropies satisfying:

$$\frac{1}{n}\vec{H}(f(X^{(n)})) \sim (0, \frac{3}{2} \ln 2, \frac{3}{2} \ln 2, 2 \ln 2) * (0, \ln 2, \ln 2, 2 \ln 2) = (0, \ln 2, \ln 2, 2 \ln 2).$$

Let us say, that for the evaluation of $\vec{H}(X_i^{(n)})$, $i = 1, 2$, it was useful, that both processes, $X_i^{(n)}$ and $X_2^{(n)}$ were Markov. In the next example we relax this property.

Let us now consider a slight variation of the previous example, namely the random walk $\mathbb{Y}$ on the standard chess board. So the values 0 and 7 are not adjacent any more. We say that two elements (i, j) and (i', j') from $\mathbb{Z}_8 \times \mathbb{Z}_8$ are adjacent if they are adjacent in one coordinate and equal in the other, i. e. if the sum of differences $|i - i'|$ and $|j - j'|$ equals one. We denote by $V_{i,j}$ the number of pairs from $\mathbb{Z}_8 \times \mathbb{Z}_8$ adjacent to (i, j) and define the transition probabilities as follows,

$$p_{(i,j)(i',j')} = \begin{cases} \frac{1}{V_{i,j}}, & \text{if } (i,j) \text{ and } (i',j') \text{ are adjacent,} \\ 0, & \text{otherwise.} \end{cases}$$

Let us fix a deterministic start at the origin, $Y(0) = (0, 0)$. Again, this Markov chain with finite states is homogeneous, irreducible, periodic with period two and non-stationary because the initial distribution is not equal to the stationary one. In order to find the entropy profile of the Markov chain, it is very handful to replace the process by its stationary version that has the same profile due to Corollary 5. It helps to establish the entropy rate of the whole process. Nevertheless, the process $(Y_1(n))_{n\in\mathbb{N}}$ and $(Y_2(n))_{n\in\mathbb{N}}$ are not Markov, so their entropy rate can be only estimated. Using appropriate formula for the entropy rate of a Markov chain and a lower bound for the entropy rate of a function of a Markov chain, we get that $h(\mathbb{Y})$ is around $1.83 \ln 2$ and $h(\mathbb{Y}_1) = h(\mathbb{Y}_2) > 1.29 \ln 2$. But even very simple observations lead to the facts that all the rates $h(\mathbb{Y})$, $h(\mathbb{Y}_1)$ and $h(\mathbb{Y}_2)$ belong into the interval $(\ln 2, 2 \ln 2)$, what is important to evaluate the convolution below.

If we use an encoding scheme into the same alphabets as in the previous example, we obtain that a typical pair of encodings $f = (f_1, f_2)$, $f_1 : (\mathbb{Z}_8)^n \to 2^n$ and $f_1 : (\mathbb{Z}_8)^n \to 2^n$, yields the transformations of $Y^{(n)}$ with the entropies satisfying:

$$\frac{1}{n}\vec{H}(f(Y^{(n)})) \quad \sim \quad \vec{h}(\mathbb{Y}) * (0, \ln 2, \ln 2, 2 \ln 2) = (0, \ln 2, \ln 2, h(\mathbb{Y})) = (0, 1, 1, 1.83) \ln 2.$$

6. MEAN FLUCTUATION OF THE INFORMATION FUNCTION

As far as we know, the mean fluctuation $M(\mathbb{X})$ of the information function $\mathcal{I}_{\mathbb{X}}$ from its mean $H(\mathbb{X})$ is not established explicitly in the literature. Nevertheless, we found it very efficient and elegant to use it as a quantity that helps to describe asymptotic equipartition property. In this section, we would like to introduce some properties of the mean fluctuation that are interesting in its own.

Let us introduce some similar quantities for a probability measure on a finite set,

$$M^+(\mathbb{P}, a) = \mathbb{E}_{\mathbb{P}} \left(\mathcal{I}_{\mathbb{P}} - a \right)^+, \qquad D(\mathbb{P}) = M(\mathbb{P}, \ln(\#s(\mathbb{P}))),$$
$$M^-(\mathbb{P}, a) = \mathbb{E}_{\mathbb{P}} \left(\mathcal{I}_{\mathbb{P}} - a \right)^-, \qquad D^+(\mathbb{P}) = M^+(\mathbb{P}, \ln(\#s(\mathbb{P}))),$$

where the notation a^+ and a^- stands for positive and negative parts of a number a, respectively. In the similar manner, we define D^-. We can express the entropy and the divergence from the uniform distribution on the support of the measure as follows:

$$H(\mathbb{P}) = M(\mathbb{P}, 0), \qquad D_{KL}(\mathbb{P} || \mathcal{U}(s(\mathbb{P}))) = D(\mathbb{P}) - 2D^+(\mathbb{P}).$$

Let us point out that the above-mentioned Kullback-Leibler divergence from the uniform distribution, as well as the mean fluctuation from $a = \ln(\#s(\mathbb{P}))$, is well defined only if the support $s(\mathbb{P})$ is finite, whereas the mean fluctuation does not need the finiteness (we assumed the finiteness of probability spaces only for simplicity).

For a discrete random variable X, we extend the previous notions in a natural way, $M(X, a) := M(\mathbb{P}_X, a)$, etc.

Let us notice that the term $D^+(X)$ is bounded above by $\frac{\ln e}{e}$ and can be often neglected as a term of small magnitude with respect to $D(X)$. Using the notation $\hat{s}(X)$ for the subset of the support of $\mathbb{P}_X$ that contains very small atoms, i. e. the values whose probability is less than $1/(\#s(X))$, we get the mentioned bound:

$$\begin{aligned} D^+(X) &= \sum_{x\in\hat{s}(X)} \mathbb{P}(x) \ln \frac{1}{\#s(X)\mathbb{P}(x)} \\ &= \mathbb{P}(\hat{s}(X)) \sum_{x\in\hat{s}(X)} \frac{\mathbb{P}(x)}{\mathbb{P}(\hat{s}(X))} \ln \frac{1}{\#s(X)\mathbb{P}(x)} \\ &\le \mathbb{P}(\hat{s}(X)) \ln \sum_{x\in\hat{s}(X)} \frac{1}{\mathbb{P}(\hat{s}(X))\#s(X)} \\ &\le \mathbb{P}(\hat{s}(X)) \ln \frac{1}{\mathbb{P}(\hat{s}(X))} \le \frac{\ln e}{e}. \end{aligned}$$

For a random variable, we define the following relative versions of the mean fluctuations:

$$M_{rel}(X) = \frac{M(X)}{H(X)}, \qquad D_{rel}(X) = \frac{D(X)}{H(X)}.$$

The definition is correct as $H(X) > 0$. Otherwise, M_{rel} and D_{rel} are set to be zero. We call M_{rel} the relative mean fluctuation and D_{rel} the relative index of uniformity. Let us recall that for a positive random variable η the mean fluctuation is bounded as follows:

$$\mathbb{E}\,|\mathbb{E}(\eta) - \eta| = 2\mathbb{E}\,(\mathbb{E}(\eta) - \eta)^+ \le 2\mathbb{E}(\eta).$$

Hence, M_{rel} is bounded by 2, whereas D_{rel} has no reasonable bound. The following lemma shows that D_{rel} dominates M_{rel}.

Lemma 1. If $H(X) > 0$, then

$$M_{rel}(X) \le 2D_{rel}(X).$$

P r o o f . Obviously,

$$M(X, H(X)) < M(X, \ln(\#s(X))) + |H(X) - \ln(\#s(X))| \le 2D(X).$$

□

Since $H(\mathbb{P})$ is the expectation of $\mathcal{I}_{\mathbb{P}}$, we get

$$M^-(\mathbb{P}) = M^+(\mathbb{P}), \quad M(\mathbb{P}) = 2M^-(\mathbb{P}) = 2M^+(\mathbb{P}).$$

Lemma 2. Let $\mathbb{P} = (1-\varepsilon)\mathbb{P}' + \varepsilon\mathbb{P}''$ for three discrete probability measures $\mathbb{P}$, $\mathbb{P}'$ and $\mathbb{P}''$ defined on the same space and $\varepsilon \in (0,1)$. Then

$$M(\mathbb{P}) \leq 2\varepsilon(H(\mathbb{P}'') + 2H(\mathbb{P}')) + 2M(\mathbb{P}') + 10\ln 2.$$

P r o o f . Let us recall that

$$(1-\varepsilon)H(\mathbb{P}') + \varepsilon H(\mathbb{P}'') \leq H(\mathbb{P}) \leq (1-\varepsilon)H(\mathbb{P}') + \varepsilon H(\mathbb{P}'') + 1.$$

Let A denotes the set of all x's such that $(1-\varepsilon)\mathbb{P}'(x) > \varepsilon\mathbb{P}''(x)$. We conclude the proof by the following calculation:

$$\begin{aligned}
\frac{1}{2}M(\mathbb{P}) = M^-(\mathbb{P}) &= \sum_x \mathbb{P}(x)\left(H(\mathbb{P}) + \ln\mathbb{P}(x)\right)^+ \\
&\leq \sum_x \mathbb{P}(x)\left((1-\varepsilon)H(\mathbb{P}') + \varepsilon H(\mathbb{P}'') + \ln 2 + \ln\mathbb{P}(x)\right)^+ \\
&\leq \varepsilon H(\mathbb{P}'') + \ln 2 + \sum_{x\notin A} 2\varepsilon\mathbb{P}''(x)\left(H(\mathbb{P}') + \ln 2\varepsilon\mathbb{P}''(x)\right)^+ \\
&\quad + \sum_{x\in A} 2(1-\varepsilon)\mathbb{P}'(x)\left(H(\mathbb{P}') + \ln 2(1-\varepsilon)\mathbb{P}'(x)\right)^+ \\
&\leq \varepsilon H(\mathbb{P}'') + \ln 2 + 2\varepsilon(H(\mathbb{P}') + \ln 2) + 2\ln 2 + 2M^-(\mathbb{P}').
\end{aligned}$$

□

Lemma 3. Let $\mathbb{P} = (1-\varepsilon)\mathbb{P}' + \varepsilon\mathbb{P}''$ for three discrete probability measures $\mathbb{P}$, $\mathbb{P}'$ and $\mathbb{P}''$ defined on the same space and $\varepsilon \in (0,1)$. Then

$$M(\mathbb{P}) \leq 2\left(\varepsilon H(\mathbb{P}) + \ln 2 + \sum_x \mathbb{P}(x)\left(H(\mathbb{P}) - \mathcal{I}_{\mathbb{P}'}(x)\right)^+\right).$$

P r o o f . Let us recall that

$$(1-\varepsilon)H(\mathbb{P}') + \varepsilon H(\mathbb{P}'') \leq H(\mathbb{P}) \leq (1-\varepsilon)H(\mathbb{P}') + \varepsilon H(\mathbb{P}'') + 1.$$

Let A denotes the set of all x's such that $(1-\varepsilon)\mathbb{P}'(x) > \varepsilon\mathbb{P}''(x)$. We conclude the proof by the following calculation:

$$\begin{aligned}
\frac{1}{2}M(\mathbb{P}) = M^-(\mathbb{P}) &= \sum_x \mathbb{P}(x)\left(H(\mathbb{P}) + \ln\mathbb{P}(x)\right)^+ \\
&= \ln 2 + \sum_{x\notin A} \mathbb{P}(x)\left(H(\mathbb{P}) + \ln\varepsilon\mathbb{P}''(x)\right)^+ \\
&\quad + \sum_{x\in A} \mathbb{P}(x)\left(H(\mathbb{P}) + \ln(1-\varepsilon)\mathbb{P}'(x)\right)^+ \\
&\leq \ln 2 + \varepsilon H(\mathbb{P}) + \sum_{x\in A} \mathbb{P}(x)\left(H(\mathbb{P}) + \ln\mathbb{P}'(x)\right)^+.
\end{aligned}$$

□

We already mentioned that for a variable X with infinite support $s(X)$, the value $D(X)$ and $D_{rel}(X)$ is not well-defined whereas the values $M(X)$ and $M_{rel}(X)$ can take arbitrarily small positive values. In this section, we show that the difference between these two notions remains significant even in the case of finite-valued i.i.d. process.

In the next lemma, we show, that for the ergodic case, M_{rel} goes to zero, whenever the entropy rate is non-null. We generalize the result for the conditional version of quantity M_{rel} in Lemma 5 (for the definition of the conditional M_{rel} see the beginning of Section 7). Lemma 5 reduces to Lemma 4 by choosing $(Y_i)_{i\in\mathbb{N}}$ to be a trivial process with only one possible value.

Lemma 4. Let $\mathbb{X} = (X_i)_{i\in\mathbb{N}}$ be an ergodic stationary process with strictly positive and finite entropy rate h. Then

$$\lim_{n\to\infty} M_{rel}(X_1, X_2, \ldots, X_n) = 0.$$

Lemma 5. Let $\mathbb{X} = (X_i, Y_i)_{i\in\mathbb{N}}$ be an ergodic stationary process with strictly positive (and finite) entropy rate $h(X|Y)$. Then

$$\lim_{n\to\infty} M_{rel}(X^{(n)}|Y^{(n)}) = 0.$$

P r o o f . We use the weaker form of conditional AEP for the ergodic processes. Namely, $\frac{1}{n}\mathcal{I}_{X^{(n)}|Y^{(n)}}$ converges to h in probability. Since $\frac{1}{n}H(X^{(n)}|Y^{(n)})$ goes to $h(X|Y)$ too, the difference

$$\xi_n = \frac{1}{n}\left(\mathcal{I}_{X^{(n)}|Y^{(n)}} - H(X^{(n)}|Y^{(n)})\right)$$

converges to zero in probability. Since $\mathbb{E}\xi_n$ is zero,

$$\mathbb{E}|\xi_n| = 2\mathbb{E}\xi_n^-.$$

But ξ_n^- goes to zero in $\mathcal{L}_1$-norm, because it converges in probability and is bounded by $H(X_1|Y_1)$. It follows that ξ_n goes to zero in $\mathcal{L}_1$-norm, i. e. $\mathbb{E}|\xi_n|$ goes to zero. Thus,

$$\lim_{n\to\infty} M_{rel}(X^{(n)}|Y^{(n)}) = \lim_{n\to\infty} \frac{\mathbb{E}\left|\mathcal{I}_{X^{(n)}|Y^{(n)}} - H(X^{(n)}|Y^{(n)})\right|}{H(X^{(n)}|Y^{(n)})} = \lim_{n\to\infty} \frac{\mathbb{E}|\xi_n|}{H(X^{(n)}|Y^{(n)})/n}$$
$$= \frac{0}{h(X|Y)} = 0.$$

□

The following lemma is a direct consequence of the property of the divergence for the product measures.

Lemma 6. If $= (X_i)_{i\in\mathbb{N}}$ is i.i.d., $s(X_1)$ is finite and $H(X_1) > 0$, then $D_{rel}(X^{(n)})$ converges to $\frac{\ln(\#s(X_1))-H(X_1)}{H(X_1)}$. In particular, the sequence $D_{rel}(X^{(n)})$ converge to zero if and only if every X_i is uniform.

The lemmas show that the control of the uniformity of the distribution of the random variable via the relative mean fluctuation is weaker than that via the relative divergence from the uniform distribution on the set of values.

7. PROOFS OF MAIN THEOREMS

In this section, we prove Theorems 2 and 3. The section is self-contained, and the lemmas from the previous section are not involved. Proposition 7 is proved with several free parameters where the full generality is aimed to serve as a flexible reference in future research. Afterward, many of the parameters are fixed in Corollary 9 to get a more specific result, which is used in the proofs of Theorems 2 and 3.

First, let us introduce conditional counterparts of quantities defined so far. Given two discrete random variables X and Y defined on the same probability space with values in the countable sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, we define the conditional information function by the formula $\mathcal{I}_{X|Y} = \mathcal{I}_{X,Y} - \mathcal{I}_Y$, where all the three functions are considered on the domain $\mathcal{X} \times \mathcal{Y}$.

We can extend the definition of M and M_{rel} to the conditional case as follows:

$$M(X|Y,a) = \mathbb{E}_{\mathbb{P}_{X,Y}} \left| \mathcal{I}_{X|Y} - a \right| \qquad M^+(X|Y,a) = \mathbb{E}_{\mathbb{P}_{X,Y}} \left(\mathcal{I}_{X|Y} - a \right)^+$$
$$M^-(X|Y,a) = \mathbb{E}_{\mathbb{P}_{X,Y}} \left(\mathcal{I}_{X|Y} - a \right)^- .$$

Shorter notation $M(X|Y)$, $M^+(X|Y)$ and $M^-(X|Y)$ is used when $a = H(X|Y)$. In addition, when $H(X|Y) > 0$, then put

$$M_{rel}(X|Y) = \frac{M(X|Y)}{H(X|Y)}.$$

Since the information function satisfies the following chain rule,

$$\mathcal{I}_{X|Y} + \mathcal{I}_Y = \mathcal{I}_{X,Y},$$

we get

$$M(X|Y) \le \mathbb{E}|\,\mathcal{I}_Y - H(Y)| + \mathbb{E}|\,\mathcal{I}_{X,Y} - H(X,Y)| \le M(Y) + M((X,Y)), \tag{3}$$
$$M(X,Y) \le M(Y) + M((X|Y)). \tag{4}$$

Since $H(X|Y)$ is the expectation of $\mathcal{I}_{X|Y}$, we get

$$M^-(X|Y) = M^+(X|Y), \quad M(X|Y) = 2M^-(X|Y) = 2M^+(X|Y). \tag{5}$$

The following lemma is a simplified version of Lemma 6 in [7]. It provides the crucial bound on the probability of the colored atoms that is applied in the next proposition.

Lemma 7. (Matúš [7]) Let $\mathbb{P}$ be a sub-probability measure on a finite set $\mathcal{X}$. For $k \ge 1$, $\varepsilon > 0$, the proportion of those maps (encodings) f from $\mathcal{X}$ into $\hat{k}$ that satisfy

$$\mathbb{P}(f^{-1}(j)) \le \frac{1+\varepsilon}{k} \qquad j \in \hat{k},$$

is at least

$$1 - ke^{-\frac{\varepsilon}{2kq}\ln(1+\varepsilon)},$$

where $q = \max_{x\in\mathcal{X}} \mathbb{P}(x)$.

The next proposition is our core technical result. We introduce it in all its generality and possible flexibility given by free parameters t_1, t_2, δ, r and s. For the purpose of this article, it is enough to reduce the parameters in the way we follow in Corollary 9, where t_1 and t_2 coincide and have the value enough large with respect to the join entropy and mean fluctuation of given variables X and Y, r and s are set to be $1/2$ and δ is a constant possibly large, but definitely significantly smaller than H.

Proposition 7. Let X, Y be random variables with values in finite sets $\mathcal{A}_X$ and $\mathcal{A}_Y$, respectively. Let $\mathcal{B}$ be a finite set, $t_1, t_2, \delta \in \mathbb{R}^+$, $r, s \in (0,1)$. Let $\alpha = \frac{1}{t_1} M(X|Y)$, $\beta = \frac{1}{t_2} M(Y)$, $R = \min(H(X|Y), \ln|\mathcal{B}|)$ and $\gamma = \alpha^{1-r} + \beta^{1-s}$. Then the proportion of those maps (encodings) f from $\mathcal{A}_X$ into $\mathcal{B}$ that satisfy the conditions

$$M(f(X)|Y) \leq 2\gamma R + 2\delta + 4\ln 2,$$

$$|H(f(X)|Y) - R| \leq \gamma R + \delta + 2\ln 2,$$

is at least

$$1 - \exp\left(-\frac{\ln 2}{2} e^{\delta + H(X|Y) - R - \alpha^r t_1} + \ln|\mathcal{B}| + H(Y) + \beta^s t_2\right).$$

Before the rigorous proof, let us explain the route of the proof informally, with some level of simplification. By the assumptions and notations in the proposition, we get that for $\mathbb{P}$-most of $y \in \mathcal{A}_Y$, the conditional probability $\mathbb{P}(x|y)$ is bounded from above by $e^{-H(X|Y)+c}$ for $\mathbb{P}(.|y)$-most of $x \in X$, where c is a constant related to the free parameters r, t_1 and the mean fluctuation $M(X|Y)$. Fix $y_0 \in \mathcal{A}_Y$ with such a property. A random encoding $f : \mathcal{A}_X \to \mathcal{B}$ can be understood as a random grouping of elements from $\mathcal{A}_X$ into $|\mathcal{B}|$ groups. The large deviation principle, introduced as Lemma 7, ensures that for a large majority of encodings (the complement is exponentially small), the sum of conditional probabilities $P(x|y_0)$ over all x from one group does not exceed $e^{-\ln|\mathcal{B}|} + e^{-R+\delta}$. These encodings behave nicely for one given $y_0 \in \mathcal{A}_Y$. We prove that the number of "bad" encodings for a given y_0 is so small, that the most of encodings behave nicely for the most of $y \in \mathcal{A}_Y$ (the complement of all "bad" encodings is still exponentially small). For such an encoding, $e^{-\ln|\mathcal{B}|} + e^{-R+\delta}$ dominates $\mathbb{P}(f^{-1}(z)|y)$ for most of $z \in \mathcal{B}$ and $y \in \mathcal{A}_Y$, so the information function $\mathcal{I}_{f(X)|Y}(z,y)$ is bounded from below by a value close to $R - \delta$ on the most of atoms $(z,y) \in \mathcal{B} \times \mathcal{A}_Y$. Nevertheless, R is also a natural upper bound for the entropy $H(f(X)|Y)$, since the entropy of the output exceeds neither the entropy of the input nor the logarithm of the size of the output alphabet. These two facts then provide the respective bounds on the entropy $H(f(X)|Y)$ and on the mean fluctuation $M(f(X)|Y)$ from the hypothesis of the proposition.

P r o o f . Denote $R = \min(H(X|Y), \ln|\mathcal{B}|)$. Surely, $H(f(X)|Y) \leq R$.

Fix $r, s > 0$ and put

$$B = \{y \mid |\mathcal{I}_Y - H(Y)| < \beta^s t_2\},$$
$$A = \{(x,y) \mid \left|\mathcal{I}_{X|Y} - H(X|Y)\right| < \alpha^r t_1\}.$$

By Markov inequality, we get $\mathbb{P}_{X,Y}(A) > 1 - \alpha^{1-r}$ and $\mathbb{P}_Y(B) > 1 - \beta^{1-s}$. Let $\mathbb{P}'$ be the restriction of $\mathbb{P}_{X,Y}$ on A, i. e. $\mathbb{P}'$ is sub-probability measure defined as follows:

$$\mathbb{P}'(x,y) = \begin{cases} \mathbb{P}_{X,Y}(x,y), & \text{if} \qquad (x,y) \in A, \quad y \in B, \\ 0, & \text{otherwise.} \end{cases}$$

Given $y \in B$, measure $Q_y(x) = \mathbb{P}'(x,y)/\mathbb{P}_Y(y)$ is a sub-probability measure that is bounded by $e^{-H(X|Y)+\alpha^r t_1}$. We apply Lemma 7. Let c_y be the proportion of encodings that satisfy the condition:

$$Q_y(f^{-1}(x')) \le \frac{1}{|\mathcal{B}|} + e^{\delta - R}, \qquad x' \in \mathcal{B}. \tag{6}$$

By Lemma 7,

$$\begin{aligned} 1 - c_y &\le \exp\left(-\frac{e^{\delta - R}|\mathcal{B}|}{2|\mathcal{B}|e^{-H(X|Y)+\alpha^r t_1}} \ln(1 + e^{\delta - R}|\mathcal{B}|) + \ln|\mathcal{B}| \right) \\ &\le \exp\left(-\frac{\ln 2}{2} e^{\delta + H(X|Y) - R - \alpha^r t_1} + \ln|\mathcal{B}| \right). \end{aligned}$$

Let c be the proportion of the encodings that satisfy the above-mentioned conditions for all $y \in B$ simultaneously. Then

$$\begin{aligned} 1 - c &\le (\#B) \exp\left(-\frac{\ln 2}{2} e^{\delta + H(X|Y) - R - \alpha^r t_1} + \ln|\mathcal{B}| \right) \\ &\le \exp\left(-\frac{\ln 2}{2} e^{\delta + H(X|Y) - R - \alpha^r t_1} + \ln|\mathcal{B}| + H(Y) + \beta^s t_2 \right). \end{aligned}$$

For the rest of the proof, let $f : \mathcal{A}_X \to \mathcal{B}$ be an encoding that satisfies condition (6). We denote by $\mathbb{P}_{f(X),Y}$ and $\mathbb{P}'_{f,Id}$ the probability and sub-probability measures on $\mathcal{B} \times \mathcal{A}_Y$ that are the images of the probability $\mathbb{P}_{X,Y}$ and sub-probability $\mathbb{P}'$ under the mapping $(f, Id) : \mathcal{A}_X \times \mathcal{A}_Y \to \mathcal{B} \times \mathcal{A}_Y$, i. e.

$$\mathbb{P}_{f(X),Y}(x',y) = \mathbb{P}_{X,Y}\left(f^{-1}(x') \times \{y\}\right), \qquad \mathbb{P}'_{f,Id}(x',y) = \mathbb{P}'\left(f^{-1}(x') \times \{y\}\right).$$

Put $\mathbb{P}'' = \mathbb{P}_{f(X),Y} - \mathbb{P}'_{f,Id}$, $A' = \{(x',y) \mid \mathbb{P}'_{f,Id}(x',y) > \mathbb{P}''(x',y)\}$. Let us notice, that $(x',y) \in A'$ implies $\mathbb{P}'_{f,Id}(x',y)$ is strictly positive, so $y \in B$.

In the following calculation, we use the fact that $\mathbb{P}_{f(X),Y} \le 2\max(\mathbb{P}'', \mathbb{P}'_{f,Id})$:

$$\begin{aligned}
M^-(f(X)|Y,R) &\le \sum_{(x',y)} \mathbb{P}_{f(X),Y}(x',y)\left(R + \ln\frac{2\max\left(\mathbb{P}'_{f,Id}(x',y), \mathbb{P}''(x',y)\right)}{\mathbb{P}_Y(y)}\right)^+ \\
&\le \ln 2 + \mathbb{P}''(\mathcal{B}\times\mathcal{A}_Y)R \\
&\quad + \sum_{(x',y)\in A', y\in B} \mathbb{P}'_{f,Id}(x',y)\left(R + \ln\frac{\mathbb{P}'_{f,Id}(x',y)}{\mathbb{P}_Y(y)}\right)^+ \\
&\le \ln 2 + (\alpha^{1-r} + \beta^{1-s})R \\
&\quad + \left(R + \ln\left(\frac{1}{|\mathcal{B}|} + e^{\delta - R}\right)\right) \\
&\le \ln 2 + (\alpha^{1-r} + \beta^{1-s})R + (\delta + \ln 2) \le \gamma R + \delta + 2\ln 2.
\end{aligned}$$

In addition,

$$R - H(f(X)|Y) = \mathbb{E}_{\mathbb{P}_{f(X),Y}}\left(R - \mathcal{I}_{f(X)|Y}\right) \le M^-(f(X)|Y,R).$$

Since $H(f(X)|Y) \le R$, $|R - H(f(X)|Y)|$ is bounded by $\gamma R + \delta + 2\ln 2$. Moreover,

$$\begin{aligned}
M(f(X)|Y) &= 2M^-(f(X)|Y) = 2\mathbb{E}_{\mathbb{P}_{f(X),Y}}\left(H(f(X)|Y) - \mathcal{I}_{f(X)|Y}\right)^+ \\
&\le 2\mathbb{E}_{\mathbb{P}_{f(X),Y}}\left(R - \mathcal{I}_{f(X)|Y}\right)^+ \le 2M^-(f(X)|Y,R).
\end{aligned}$$

□

If we choose Y to be a trivial random variable (deterministic one), then we get the following corollary for one random variable.

Corollary 8. Let X be random variables with values in finite set $\mathcal{A}$, $\mathcal{B}$ be a finite set, $t \in \mathbb{R}^+$, $r \in \mathbb{R}$. Let $\alpha = \frac{1}{t}M(X)$, $R = \min(H(X), \ln|\mathcal{B}|)$. Then the proportion of those maps (encodings) f from $\mathcal{A}$ into $\mathcal{B}$ that satisfy the conditions

$$M(f(X)) \le 2\alpha^{1-r}R + 2\delta + 4\ln 2,$$

$$|H(f(X)) - R| \le \alpha^{1-r}R + \delta + 2\ln 2,$$

is at least

$$1 - \exp\left(-\frac{\ln 2}{2}e^{\delta + H(X) - R - \alpha^r t} + \ln|\mathcal{B}|\right).$$

The following corollary introduces the explicit bounds for the change of the joint entropy and the mean error of the joint information function when one variable is encoded into a prescribed alphabet.

Corollary 9. Let X, Y be random variables with values in finite sets $\mathcal{A}_X$ and $\mathcal{A}_Y$, respectively. Let $\mathcal{B}$ be a finite set, $1 \le \varepsilon > 0$ and $H > 0$ such that

$$H \ge H(X,Y), \qquad H \ge \frac{M(X,Y)}{\varepsilon}, \qquad H \ge \frac{M(Y)}{\varepsilon}, \qquad H \ge \frac{4\ln 2}{\varepsilon}. \tag{7}$$

The proportion of those maps (encodings) f from $\mathcal{A}_X$ into $\mathcal{B}$ that satisfy the conditions

$$M(f(X)|Y) \le 10\sqrt{\varepsilon}H,$$

$$|H(f(X)|Y) - R| \le 5\sqrt{\varepsilon}H,$$

is at least

$$1 - \exp\left(-\frac{\ln 2}{2}e^{\varepsilon H} + \ln|\mathcal{B}| + 2H\right),$$

where $R = \min(H(X|Y), \ln|\mathcal{B}|)$.

Proof. Put

$$t_1 = t_2 = H, \qquad r = s = \frac{1}{2}, \qquad \delta = \left(\varepsilon + \sqrt{\varepsilon}\right) H.$$

If we define α, β and γ as in Proposition 7, then the inequalities (7) and subadditivity for M (see (3)) ensures that $\alpha \le 2\varepsilon$, $\beta \le \varepsilon$ and $\gamma \le (\sqrt{2}+1)\sqrt{\varepsilon}$.

By Proposition 7 , the proportion of those maps (encodings) f from $\mathcal{A}_X$ into $\mathcal{B}$ that satisfy the conditions

$$M(f(X)|Y) \le 2(\sqrt{2}+1)\sqrt{\varepsilon}H + 2(\varepsilon+\sqrt{\varepsilon})H + \varepsilon H,$$
$$|H(f(X)|Y) - \min(H(X|Y), \ln|\mathcal{B}|)| \le (\sqrt{2}+1)\sqrt{\varepsilon}H + (\varepsilon+\sqrt{\varepsilon})H + \varepsilon H/2,$$

is at least

$$1 - \exp\left(-\frac{\ln 2}{2}e^{\varepsilon H} + \ln|\mathcal{B}| + H(Y) + \sqrt{\varepsilon}H\right).$$

Since $M(Y) \le \varepsilon H$ and $\varepsilon \le \sqrt{\varepsilon}$, we get

$$2(\sqrt{2}+1)\sqrt{\varepsilon}H + 2(\varepsilon+\sqrt{\varepsilon})H + \varepsilon H \le 10\sqrt{\varepsilon}H$$
$$(\sqrt{2}+1)\sqrt{\varepsilon}H + (\varepsilon+\sqrt{\varepsilon})H + \varepsilon H/2 \le 5\sqrt{\varepsilon}H.$$

In the exponent for the bound of the proportion of the encodings, $H(Y) + \sqrt{\varepsilon}H$ is bounded by $2H$, since $\varepsilon \le 1$. □

Proof. [Proof of Theorem 2]

We will prove the theorem by the induction over ℓ. For all the proof, fix $k, \ell, \varepsilon, \delta, X = (X_i)_{i\le k}, H$ which satisfy the assumptions of the proposition.

Assume that $\ell = 0$. Then the only element f from $\mathcal{E}_\ell$ is identity, $f(X) = X$, $w = H(X)$. The theorem follows immediately.

Let $\ell \geq 1$, $\varepsilon' = \left(\frac{\varepsilon}{121}\right)^2$ and

$$w = \vec{H}(X) * \vec{H}((\mathcal{B}_i)_{i\leq \ell}), \qquad w' = \vec{H}(X) * \vec{H}((\mathcal{B}_i)_{i\leq \ell-1}).$$

The set of all $f' \in \mathcal{E}_{\ell-1}$ that satisfy the conditions

$$M'(f'(X)) \leq \varepsilon' H, \qquad \left\| \vec{H}(f'(X)) - \vec{H}(X) * \vec{H}((\mathcal{B}_i)_{i\leq \ell-1}) \right\|_{\max} \leq \varepsilon' H, \tag{8}$$

is denoted by $\mathcal{G}'$. Let $\pi : \mathcal{E}_\ell \to \mathcal{E}_{\ell-1}$ is the projection defined by the formula $\pi(f) = (f_i)_{i\leq \ell-1} \in \mathcal{E}_\ell$.

Given $f \in \mathcal{E}_{\ell-1}$, $J \subset \hat{k}\backslash\{\ell\}$, $\mathcal{G}_{f,J}$ is the set of all mappings that satisfy the conditions:

$$M(g(X_\ell)|f_J(X_J)) \leq 10\sqrt{\varepsilon'} H \tag{9}$$

$$|H(g(X_\ell)|f_J(X_J)) - \min(H(X_\ell|f_J(X_J)), \ln|\mathcal{B}_\ell|)| \leq 5\sqrt{\varepsilon'} H. \tag{10}$$

The special form of the conditions suits the later application of Corollary 9. Put $\mathcal{G}_f = \bigcap_J \mathcal{G}_{f,J}$, where the intersection goes over all $J \subset \hat{k} \setminus \{\ell\}$,

$$\mathcal{G} = \{f \in \mathcal{E}_\ell \mid \pi(f) \in \mathcal{G}', f_\ell \in \mathcal{G}_{\pi(f)}\}.$$

Let $f \in \mathcal{G}$, $I \subset \hat{k}$. If $\ell \notin I$, then $w_I = w'_I$, f_I equals f'_I and (8) ensures that $M(f_I(X_I))$ and the difference $|H(f_I(X_I)) - w_I|$ are both bounded by εH. If $\ell \in I$, then

$$M(f_I(X_I)) \leq M(f_\ell(X_\ell)|f'_J(X_J)) + M(f'_J(X_J)) \leq 10\sqrt{\varepsilon'} H + \varepsilon' H \leq \varepsilon H,$$

where $J = I \setminus \{\ell\}$. By Proposition , $w_I \geq H(f_I(X_I))$ and

$$\begin{aligned} w_I - H(f_I(X_I)) &= w_I - H(f_J(X_J)) - H(f_\ell(X_\ell)|f_J(X_J)) \\ &= w_I - H(f_J(X_J)) - \min(H(X_\ell|f_J(X_J)), \ln|\mathcal{B}_\ell|) + 5\sqrt{\varepsilon'} H \\ &= w_I - \min(H(f'_I(X_I)), H(f_J(X_J)) + \ln|\mathcal{B}_\ell|) + 5\sqrt{\varepsilon'} H \\ &= w_I - \min(w'_I, w'_J + \ln|\mathcal{B}_\ell|) + \varepsilon' H + 5\sqrt{\varepsilon'} H \\ &\leq w_I - w_I + \varepsilon' H + 5\sqrt{\varepsilon'} H \leq \varepsilon H. \end{aligned}$$

Hence, condition (2) holds for all $f \in \mathcal{G}$.

In order to make statements shorter and notations more readable, we put

$$\ell' = \ell - 1, \qquad D = \exp\left(-\frac{\ln 2}{2} e^{\delta H} + (\vec{H}(\mathcal{B}))_{\hat{\ell}} + 2H\right).$$

Since $\delta = \left(\frac{\varepsilon'}{121}\right)^{2^\ell}$ and $H(X_{\hat{\ell}'}) \leq H(X_{\hat{\ell}})$, the assumption of the theorem remains true when replacing ℓ by ℓ' and ε by ε'. By the inductive assumption and the fact that π is a mapping $|\mathcal{A}_\ell|^{|\mathcal{B}_\ell|}$ to 1, we get

$$c_1 := 1 - \frac{|\pi^{-1}(\mathcal{G}')|}{|\mathcal{E}_\ell|} = 1 - \frac{|\mathcal{G}'|}{|\mathcal{E}_{\ell-1}|} \leq (\ell-1)\, 2^{k-1} D.$$

Let $f' \in \mathcal{G}'$, $J \subset \hat{k} \setminus \{\ell\}$. We will apply Corollary 9, where the variables X and Y are understood as X_ℓ and $f_J(X_J)$, respectively, and ε is replaced by ε'.

First we have to verify the assumptions of the proposition, namely

$$H \geq H(X_\ell, f_J(X_J)), \qquad H \geq \frac{M(X_\ell, f_J(X_J))}{\varepsilon'}, \qquad H \geq \frac{M(f_J(X_J))}{\varepsilon'}$$

and $H \geq \frac{4 \ln 2}{\varepsilon'}$. The first inequality follows from the fact that $H(f_J(X_J)) \leq H(X_J)$ for every $J \in \hat{k}$. The second and the third one are the immediate consequence of the fact that the numerators are bounded $M'(f(X))$. The last one follows from the inequality $\varepsilon' > \varepsilon$.

Applying Corollary 9,

$$c_{f',J} := 1 - \frac{|\,\mathcal{G}_{f',J}|}{|\,\mathcal{B}_\ell|^{|\,\mathcal{A}_\ell|}} \leq \exp\left(-\frac{\ln 2}{2} e^{\delta H} + (\vec{H}(\,\mathcal{B}))_{\hat{\ell}'} + 2H\right) \leq D.$$

Hence,

$$c_{f'} := 1 - \frac{|\,\mathcal{G}_{f'}|}{|\,\mathcal{B}_\ell|^{|\,\mathcal{A}_\ell|}} \leq \sum_{J \subset \hat{k} \setminus \{\ell\}} 1 - \frac{|\,\mathcal{G}_{f',J}|}{|\,\mathcal{B}_\ell|^{|\,\mathcal{A}_\ell|}} \leq 2^{k-1} D.$$

By the definition of $\mathcal{G}$, there is one-to-one correspondence between its elements and pairs (f', g) where $f' \in \mathcal{G}$ and $g \in \mathcal{G}_{f'}$, given by the equality $f_i = f'_i$, $i \leq \ell - 1$, $f_\ell = g$.

Hence

$$\begin{aligned}
1 - \frac{|\,\mathcal{G}|}{|\,\mathcal{E}_\ell|} &= 1 - \frac{\sum_{f' \in \mathcal{G}'} |\,\mathcal{G}_{f'}|}{|\,\mathcal{E}_\ell|} \leq \frac{|\,\mathcal{G}'| \cdot |\,\mathcal{B}_\ell|^{|\,\mathcal{A}_\ell|} - \sum_{f' \in \mathcal{G}'} |\,\mathcal{G}_{f'}|}{|\,\mathcal{E}_\ell|} \\
&\leq \frac{|\,\mathcal{E}_\ell| - |\,\mathcal{G}'| \cdot |\,\mathcal{B}_\ell|^{|\,\mathcal{A}_\ell|} + \sum_{f' \in \mathcal{G}'} |\,\mathcal{B}_\ell|^{|\,\mathcal{A}_\ell|} - |\,\mathcal{G}_{f'}|}{|\,\mathcal{E}_\ell|} \\
&\leq 1 - \frac{|\,\mathcal{G}'|}{|\,\mathcal{E}_{\ell-1}|} + \frac{1}{|\,\mathcal{E}_{\ell-1}|} \sum_{f' \in \mathcal{G}'} \frac{|\,\mathcal{B}_\ell|^{|\,\mathcal{A}_\ell|} - |\,\mathcal{G}_{f'}|}{|\,\mathcal{B}_\ell|^{|\,\mathcal{A}_\ell|}} \\
&\leq c_1 + \frac{1}{|\,\mathcal{E}_{\ell-1}|} \sum_{f' \in \mathcal{G}'} c_{f'} \leq (\ell - 1)\, 2^{k-1} D + \frac{|\,\mathcal{G}'|}{|\,\mathcal{E}_{\ell-1}|} 2^{k-1} D \leq \ell\, 2^{k-1} D.
\end{aligned}$$

□

Before we start to prove Theorem 3, let us recall the following inequalities, that will be used in the proof. For $u, u' \in \mathbb{R}^{\tilde{k}}$ and $v, v' \in \mathbb{R}^{\tilde{\ell}}$,

$$||u * v - u' * v||_{max} \leq ||u - u'||_{max}, \quad ||u * v - u * v'||_{max} \leq ||v - v'||_{max}.$$

It follows from the very general fact; given two vectors of real numbers $(a_i)_{i \leq m}$ and $(b_i)_{i \leq m}$,

$$|\min_{i \leq m} a_i - \min_{i \leq m} b_i| \leq \max_{i \leq m} |a_i - b_i|.$$

P r o o f . [Proof of Theorem 3]

Use the shorter notation:

$$h^{(n)} = \frac{\vec{H}(X^{(n)})}{n}, \qquad g^{(n)} = \frac{\vec{H}(f(X^{(n)}))}{n}, \qquad b^{(n)} = \frac{\vec{H}(\mathcal{B}^{(n)})}{n}.$$

Let $\delta < \left(\frac{\min(\varepsilon, h_{\hat{k}})}{121 h_{\hat{k}}}\right)^{2^{|\ell|}}$. There exist $\varepsilon'' < \varepsilon' < \min(\varepsilon, h_{\hat{k}})$ and $\delta' > \delta$ such that $\delta' = \left(\frac{\varepsilon''}{121 h_{\hat{k}}}\right)^{2^{|\ell|}}$. By the assumptions of the theorem, for n big enough, $H(X^{(n)})$ exceeds both , $(\delta')^{-1} M'(X^{(n)})$ and $(\delta')^{-1} 4 \ln 2$. By Theorem 2, the proportion of the encodings from $\mathcal{E}_\ell^{(n)}$ that satisfy

$$M'(f(X^{(n)})) \le \frac{\varepsilon''}{h_{\hat{k}}} n h_{\hat{k}}^{(n)} \qquad \& \qquad \left\| n g^{(n)} - n h^{(n)} * n b^{(n)} \right\|_{\max} \le \frac{\varepsilon''}{h_{\hat{k}}} n h_{\hat{k}}^{(n)}, \tag{11}$$

is at least

$$1 - \exp\left(-\frac{\ln 2}{2} e^{\delta' n} + n b_{\hat{\ell}}^{(n)} + 2 n h_{\hat{k}}^{(n)} + (k-1)\ln 2 + \ln \ell \right).$$

Since $h^{(n)}$ goes to h and $h_{\hat{k}} > 0$, $\frac{h_{\hat{k}}^{(n)}}{h_{\hat{k}}}$ is smaller than $\frac{\varepsilon'}{\varepsilon''}$ for n large enough. In such a case, condition (11) implies

$$\frac{M'(f(X^{(n)}))}{n} < \varepsilon' \qquad \& \qquad \left\| g^{(n)} - h^{(n)} * b^{(n)} \right\|_{\max} < \varepsilon'. \tag{12}$$

In addition,

$$\begin{aligned} \left\| h * b - h^{(n)} * b^{(n)} \right\|_{\max} &\le \left\| h * b - h^{(n)} * b \right\|_{\max} + \left\| h^{(n)} * b - h^{(n)} * b^{(n)} \right\|_{\max} \\ &\le \left\| h - h^{(n)} \right\|_{\max} + \left\| b - b^{(n)} \right\|_{\max}. \end{aligned}$$

The last two terms goes to zero. Thus, for n large enough, their sum is bounded by $\varepsilon - \varepsilon'$ and condition (11) implies

$$\left\| g^{(n)} - h * b \right\|_{\max} < \varepsilon.$$

It remains to prove, that the lower bound for the proportion of the encodings satisfying (11) is larger than $1 - \exp\left(-e^{\delta n}\right)$. But in the exponent of the exponential term in the bound, there is only one exponential term $-\frac{\ln 2}{2} e^{\delta' n}$, the others are at most linear. Since $\delta < \delta'$, the term $-e^{\delta n}$ is eventually bigger than all the exponent term from the bound and

$$1 - \exp\left(-\frac{\ln 2}{2} e^{\delta' n} + n b_{\hat{\ell}}^{(n)} + 2 n h_{\hat{k}}^{(n)} + (k-1)\ln 2 + \ln \ell \right) > 1 - \exp\left(-e^{\delta n}\right).$$

□

ACKNOWLEDGMENTS

We are greatly indebted to Prof. Laszlo Csirmaz for useful discussions and comments that helped to find the final shape of the presented results. We would also like to thank anonymous reviewers for their comments that improved the quality and exposition of this paper.

(Received June 6, 2019)

REFERENCES

[1] R. Bassoli, H. Marques, J. Rodriguez, K. W. Shum, and R. Tafazolli: Network coding theory: A survey. IEEE Commun. Surveys Tutor. *15* (2013), 1950–1978. DOI:10.1109/surv.2013.013013.00104

[2] T. M. Cover and J. A. Thomas: Elements of Information Theory. John Wiley and Sons, 2012. DOI:10.1002/0471200611

[3] R. M. Gray and J. C. Kieffer: Asymptotically mean stationary measures. Ann. Probab. *8* (1980), 962–973. DOI:10.1214/aop/1176994624

[4] R. M. Gray: Entropy and Information Theory. Springer Science and Business Media, 2011.

[5] T. Kaced: Partage de secret et théorie algorithmique de l'information. PhD. Thesis, Université Montpellier 2, 2012.

[6] J. C. Kieffer: A generalized Shannon-McMillan theorem for the action of an amenable group on a probability space. Ann. Probab. *3* (1975), 1031–1037. DOI:10.1214/aop/1176996230

[7] F. Matúš: Two constructions on limits of entropy functions. IEEE Trans. Inform. Theory *53* (2007), 320–330. DOI:10.1109/tit.2006.887090

[8] F. Matúš and L. Csirmaz: Entropy region and convolution. IEEE Trans. Inform. Theory *62* (2016), 6007–6018. DOI:10.1109/tit.2016.2601598

[9] F. Matúš and M. Kupsa: On colorings of bivariate random sequences. In: Proc. IEEE International Symposium on Information Theory 2010, pp. 1272–1275. DOI:10.1109/isit.2010.5513700

[10] R. W. Yeung: Information Theory and Network Coding. Springer Science and Business Media, 2008. DOI:10.1007/978-0-387-79234-7_1

Michal Kupsa, Institute of Information Theory and Automation, The Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: kupsa@utia.cas.cz