Rényi Entropy, Guesswork Moments and Large Deviations

C-E Pfister and WG Sullivan

Abstract

For a large class of stationary probability measures on $\mathtt{A}^{\mathbf{N}}$, where $\mathtt{A}$ is a finite alphabet, we compute the specific Rényi entropy of order α and the specific guesswork moments of order $\beta > -1$. We show that the specific guesswork moment of order β equals the specific Rényi entropy of order $\alpha = 1/(1+\beta)$ multiplied by β. The method is based on energy–entropy estimates suggested by statistical physics. The technique also yields a simple proof of the large deviation principle for the empirical measure on the space of an irreducible sofic shift with reference probability measure ν, which is stationary and satisfies a rate condition on the probability of allowed words.

Keywords

Guesswork, large deviation, Renyi entropy, sofic shift.

I. Introduction

Assume that a certain system allows n possible options and that an option is chosen from the probability distribution specified by $p_1, \dots, p_n$, where $p_1 \geq \cdots \geq p_n \geq 0$, $\sum_i p_i = 1$. If one tries to guess the chosen option using the strategy of guessing in decreasing order of probability, the expected number of guesses required to determine the chosen option is

$$\mathcal{G}(\{p_i\}) := \sum_{i=1}^{n} i\, p_i. \tag{1}$$

In practice one rarely knows $\{p_i\}$ exactly, but the above expression gives a lower bound for the expected number of guesses required by an exhaustive search to find the chosen value. We write $\{p_i\}$ as a simple probability distribution; often the guesses involve additional data so that $\{p_i\}$ is conditional distribution. This type of search arises in several contexts. One case is that of a distributed data base. A central server can repeatedly query secondary servers until the desired record is found. The order in which the secondaries are polled may be conditioned on the nature of the request. As the number of servers grows, the average number of machines queried per record requested may increase significantly.

A more familiar context is that of guessing passwords. Here the options are the possible passwords. Hackers have complied lists of frequently chosen passwords to facilitate access to computer accounts of careless users. As a countermeasure, the system may disallow passwords which appear to be vulnerable. Alternatively, the system may assign passwords using algorithms intended to provide more secure choices.

Essentially the same problem arises with ciphers used to impede unauthorized access. Wireless ethernets, digital television and DVDs have schemes which make access by hackers marginally more difficult. Public key encryption also makes use of ciphers. In order for a public key system to be secure, attack by exhaustive search must be beyond current hardware and software capabilities. However, calculations with public keys are time consuming. In practice public key encryption is used to exchange keys for cipher methods requiring much less computing power.

One can consider (1) in the context of guessing a cipher key. The distribution $\{p_i\}$ will depend on the program which selects keys. Typically, such programs combine a pseudo-random number generator algorithm with external information, which is presumed to be random. There may also be checking to eliminate keys which are cryptographically weak. In the past, data such as the system time and process numbers have been used to provide "random" input. However, such data may be readily available to an intruder: the resulting $\{p_i\}$ may have a much smaller value for $\mathcal{G}(\{p_i\})$.

Arikan [1] has applied guesswork techniques to sequential decoding. Here we outline a simpler channel decoding model. Consider the data sequence $u_1, u_2, \dots$ in which each u_i is a character from a finite alphabet, perhaps a binary digit or a byte. We take a block of characters $u_1, \dots, u_k$ and append verification characters $u_{k+1}, \dots, u_n$. A simple case is a parity bit appended to seven data bits, which allows single error detection. Here we assume that the verification data contains more sophisticated data digests, e.g., md5 (see [11]). The input and verification data may be distributed throughout $u_1, \dots, u_n$. For transmission (or storage) the data will be coded to $v_1, \dots, v_m$ in such a way that $u_1, \dots, u_n$ can be recovered exactly from $v_1, \dots, v_m$. An example in common practice is the storage of blocks of bytes in sectors of magnetic disks using run length limiting coding; each sector contains check data to allow for error detection. Let us assume that the transmitted data $v_1, \dots, v_m$ is received as $v'_1, \dots, v'_m$, where $v_k \neq v'_k$ for some k. If the verification data

C-E Pfister is with the IACS, EPF-L, CH-1015 Lausanne, Switzerland.

WG Sullivan is with the Communications Network Research Institute, Dublin Institute of Technology, Ireland, on leave of absence from University College, Dublin.

appended to the original $u_1, \ldots, u_k$ is sufficiently detailed, we may be able to recover the original message by guessing. Experience could be used as a guide to the likely nature of how $v_1, \ldots, v_m$ might have been changed. For example, if only one v_k is changed, the number of cases to check is at most m times the number of characters allowed for each v_k.

To place the channel decoding problem in the context of guesswork moments, let us assume that given the received data $v'_1, \ldots, v'_m$, one repeatedly guesses $v''_1, \ldots, v''_m$ until values are obtained so that the corresponding $u'_1, \ldots, u'_n$ satisfy the verification conditions. If $v_1, \ldots, v_m$ is run length limited coded, the guesses are naturally restricted to the space of this code. Next one needs an estimate of the distribution of $v_1, \ldots, v_m$ conditioned on observing $v'_1, \ldots, v'_m$. Then (1) gives a lower bound for the expected number of guesses required, conditioned on $v'_1, \ldots, v'_m$. The bound for the the expected number of guesses is obtained by averaging over the distribution of $v'_1, \ldots, v'_m$. For distributions of the type considered here, this expected number of guesses grows exponentially in m. From knowledge of limits of the form (3), one can estimate how large m can be, given the available computing power.

In this note we consider the asymptotic behavior of

$$\mathcal{G}^{\beta}(\{p_i\}) := \sum_{i=1}^{m_n} i^{\beta} p_i \tag{2}$$

as $n \to \infty$ in the context of keys of the form $(u_1, \ldots, u_n)$ arising from a stationary distribution on a finite alphabet. The probabilities of the m_n possible keys in decreasing order are $p_1, \ldots, p_{m_n}$. The case in which the $\{u_i\}$ are independent and identically distributed is treated in Arikan [1]. This is extended to the case of an ergodic Markov chain in [7]. However, the distributions which arise in guesswork problems may have greater complexity. The distribution of $\{u_i\}$ may be supported by a shift space which is a proper subset of the full shift. An important case is that of sofic shifts, which includes subshifts of finite type such as run length limited codes. Sofic shifts can be generated by finite state machines. See [6] for more details about shift spaces.

Let $\mathtt{A} = \{0, \ldots, r-1\}$ be a finite alphabet with r characters. Let ν be a stationary probability measure on $\mathtt{A}^{\mathbf{N}}$, with $\mathbf{N} := \{1, 2, \ldots\}$. Corresponding to (2) we define the β^{th} *specific guesswork moment*

$$\lim_n \frac{1}{n} \log \sum_{\mathrm{w} \in \mathtt{A}^n} \nu([\mathrm{w}]_1^n) \operatorname{rank}^{\beta}(\nu([\mathrm{w}]_1^n)), \tag{3}$$

where $\operatorname{rank}(\nu([\mathrm{w}]_1^n))$ takes on the values $1, \ldots, r^n$ and $\operatorname{rank}(\nu([\mathrm{w}]_1^n))$ is nonincreasing. Here $[\mathrm{w}]_1^n$ denotes the points of $\mathtt{A}^{\mathbf{N}}$ whose first n coordinates coincide with those of w. We use log to denote the logarithm. In information theory the common practice is to use $\log_2$, but $\log_b$ for any real $b > 1$ suffices. Whenever exp is used below, it refers to the same base as used with log: if $\log := \log_2$ is used, then $\exp(x) := 2^x$.

A related quantity is the *specific Rényi entropy of order $\alpha \neq 1$,*

$$\lim_n \frac{1}{n(1-\alpha)} \log \sum_{\mathrm{w} \in \mathtt{A}^n} \left(\nu([\mathrm{w}]_1^n)\right)^{\alpha}. \tag{4}$$

For many problems in information theory, e.g., noiseless data compression, the significant asymptotic limit is the specific Shannon entropy. In this paper we show that, under appropriate hypotheses, the β^{th} specific guesswork moment is β times the specific Rényi entropy of order $\alpha = 1/(1+\beta)$.

Here the case in which ν is a stationary probability measure on the shift space $\Sigma \subset \mathtt{A}^{\mathbf{N}}$ is considered. Our approach is based on ideas related to simplification of the treatment of large deviations of the empirical measure on shift spaces. The standard approach is to place various mixing conditions on the stationary distribution ν (see [2]). Instead of mixing conditions on ν, we assume a limit property for the marginal distributions $\{\nu([\mathrm{w}]_1^n)\}$ and an entropy property for the shift space Σ^{ν} defined by (6) .

For a given stationary probability measure ν of the type considered here, there are considerable theoretical similarities between specific entropy and specific guesswork moments. In practice there is a marked difference. The specific entropy specifies the logarithmic growth rate of the number of "words" of length n whose statistics are "typical" for ν; but there is no need to work with all such words. Exhaustive guessing involves calculations with each word, so practical values of n are relatively small. Also, a very significant contribution to (3) comes from words whose probability is very low. In practice it is useful to allow truncation of the guessing procedure when the probability of the untried words becomes sufficiently small. This can reduce significantly the expected number of guesses, but it allows the possibility of failure.

The structure of the remainder of this note is as follows. We first give a precise mathematical formulation and state our results. Then we give two related examples of channel decoding by guesswork, including numerical calculations to show the effect of truncation. The final section gives proofs of the theoretical results.

II. Definitions, notation and main results

A. Notation

The cardinality of any finite set C is denoted by $|C|$. Let $\mathtt{A} := \{0, 1, \ldots, r-1\}$ (with the discrete topology) and $\Omega := \mathtt{A}^{\mathbf{N}}$ be the product space. $X_n : \Omega \to \mathtt{A}^n$ indicates the mapping $\omega \to (\omega_1, \ldots, \omega_n)$. Note $[\mathrm{w}]_1^n = X_n^{-1}\{\mathrm{w}\}$. $\mathcal{M}$ denotes the space of Borel probability measures on Ω.

We use σ to denote the probability measure on Ω which is the product of the equiprobable distribution on each factor $\mathtt{A}$. Given $n \in \mathbf{N}$, the restriction of $\rho \in \mathcal{M}$ to $\mathtt{A}^n$ is denoted ρ_n, so that for $C \subset \mathtt{A}^n$,

$$\frac{1}{n}\log \sigma_n(C) = \frac{\log|C|}{n} - \log r. \tag{5}$$

S: $\Omega \to \Omega$ is the shift operator, $(S\omega)_j := \omega_{j+1}$, for each $j \in \mathbf{N}$. The action of S on the function f is given by $(Sf)(\omega) := f \circ S(\omega)$. $\mathcal{M}_S \subset \mathcal{M}$ denotes shift-invariant probability measures. For $\nu \in \mathcal{M}_S$ we define the shift space $\Sigma^\nu \subset \Omega$ by

$$\Sigma_n^\nu := \{\mathrm{w} \in \mathtt{A}^n \,:\, \nu([\mathrm{w}]_1^n) > 0\}, \quad \Sigma^\nu := \bigcap_n X_n^{-1}(\Sigma_n^\nu). \tag{6}$$

Thus $\omega \in \Sigma^\nu$ if and only if $\nu([\omega_1, \ldots, \omega_n]_1^n) > 0$ for all n.

Let f: $\Omega \to \mathbf{R}$. $\|f\| := \sup_\omega |f(\omega)|$. $\mathcal{F}_n$ denotes the σ-algebra generated by X_n. We write $f \in \mathcal{F}_n$ to mean the function f is $\mathcal{F}_n$ measurable. f is *local* if there exists $n \in \mathbf{N}$ so that $f \in \mathcal{F}_n$. We use the weak topology on $\mathcal{M}$ (see [2]). There exists a sequence $\{f_j\}$ of local functions which determines this topology. In fact one could take $\{f_j\}$ to be the set of indicator functions of each word in $\mathtt{A}^n$ for each $n \in \mathbf{N}$. The integral of f with respect to ρ is denoted by $\langle\, f, \rho\,\rangle$. S acts on measures by $\langle\, f, S\rho\,\rangle := \langle\, Sf, \rho\,\rangle$.

We use $T_n(\omega)$ to denotes the *empirical measure,*

$$T_n(\omega) := \frac{1}{n}\left(\delta_\omega + \delta_{S(\omega)} + \cdots + \delta_{S^{n-1}(\omega)}\right), \tag{7}$$

where $\delta_{S^j\omega}$ denotes the measure concentrated on the point $S^j(\omega) = (\omega_{j+1}, \omega_{j+2}, \ldots)$. This is convenient notation, but its conceals the real significance. Measures concentrated at single points of the space are unimportant. However, the sequence $\{T_n(\omega) : n \in \mathbf{N}\}$ gives easy access to the statistics of $\omega = (\omega_1, \omega_2, \ldots)$. For $a \in \mathtt{A}$ define

$$f_a(\omega) := \begin{cases} 1 & \text{if } \omega_1 = a; \\ 0 & \text{otherwise.} \end{cases}$$

Then $\langle\, f_a, T_n(\omega)\,\rangle$ gives the fraction of times the character a appears in the first n entries of ω. By using functions of ω which depend only on ω_1, ω_2, we can use $\{T_n(\omega)\}$ to obtain pair probability distributions; functions which depend on $\omega_1, \ldots, \omega_m$ yield distributions on $\mathtt{A}^m$. For these to be useful we need n to be much larger than m. A complication is the fact that the probability distributions on $\mathtt{A}^m$ resulting from $T_n(\omega)$ depend on $\omega_{n+1}, \ldots, \omega_{n+m-1}$. In the analysis below we have to prove that, in the limits we consider, the dependence of $T_n(\omega)$ on ω_k for $k > n$ is negligible.

The *Shannon entropy* of $\rho \in \mathcal{M}_S$ is

$$h_{\mathrm{Sh}}(\rho) := \lim_{n\to\infty} \frac{1}{n} H_n(\rho) \tag{8}$$

with

$$H_n(\rho) := -\sum_{\mathrm{w}\in\mathtt{A}^n} \rho([\mathrm{w}]_1^n)\log\rho([\mathrm{w}]_1^n)\,. \tag{9}$$

We have $h_{\mathrm{Sh}}(\sigma) = \log r$, since $|\mathtt{A}| = r$.

The *specific I-divergence* of $\rho \in \mathcal{M}_S$ with respect to $\nu \in \mathcal{M}_S$ is given by

$$h(\rho\,|\,\nu) := \lim_{n\to\infty}\frac{1}{n}\sum_{\mathrm{w}\in\mathtt{A}^n} \rho([\mathrm{w}]_1^n)\log\frac{\rho([\mathrm{w}]_1^n)}{\nu([\mathrm{w}]_1^n)}, \tag{10}$$

when the limit exists. For all $\rho \in \mathcal{M}_S$, $h(\rho\,|\,\sigma)$ exists and equals $-h_{\mathrm{Sh}}(\rho) + \log r$.

B. Main results

Let $\nu \in \mathcal{M}_S$ be a given stationary probability measure. Σ^ν is the shift space defined by (6). $\mathcal{M}^\nu$ denotes the set of Borel probability measures on Σ^ν, and $\mathcal{M}_S^\nu$ the shift-invariant probability measures on Σ^ν. Our main hypotheses are formulated for convenience as follows.

H1 For any neighbourhood U of $\rho \in \mathcal{M}_S^\nu$, and for any $\varepsilon > 0$, there exists an ergodic $\rho' \in U \cap \mathcal{M}_S^\nu$ such that $h_{\mathrm{Sh}}(\rho') \geq h_{\mathrm{Sh}}(\rho) - \varepsilon$.

H2 The given reference probability measure ν is shift-invariant. There exists a continuous nonnegative function $e_\nu\colon \Omega \to \mathbf{R}$ satisfying

$$\lim_n \sup_{\mathrm{w}\in\Sigma_n^\nu} \frac{1}{n} |\log \nu([\mathrm{w}]_1^n) + e_\nu(\omega)| = 0. \tag{11}$$

Several authors have employed conditions similar to H1 (see [3], [8]). These works often include a "specification" hypothesis, which implies strong mixing, so periodic subshifts of finite type are not covered. Note that H1 depends only on the allowed words of the shift space Σ^ν. Spaces which satisfy H2 of [9] satisfy H1. In particular, the space of any irreducible (possibly periodic) sofic shift satisfies H1. An example of a space which does *not* satisfy H1 is the space $\Sigma_{e,o}$ over the alphabet $\{0,1,2,3\}$ in which $\omega \in \Sigma_{e,o}$ if and only if all $\{\omega_i\}$ are even or all $\{\omega_i\}$ are odd.

Hypothesis H2 corresponds to the probability $\nu([\mathrm{w}]_1^n)$ being determined (in an approximate sense) by the character combinations in w. The Parry measure ν of an irreducible sofic shift satisfies H2 with the function e_ν equal to the constant $h_{\mathrm{Sh}}(\nu)$ (see [9]). When different weights are assigned to the characters of the alphabet, one obtains an invariant measure ν' with $\Sigma^\nu = \Sigma^{\nu'}$, and $e_{\nu'}(\omega)$ is a function of ω_1. The continuity requirement for e_ν allows for more complex weightings.

The basic ideas in the proofs below are quite simple, but the details can be confusing. Using assumption H2 we use e_ν to split the measures into K parts and then use T_n to split the words of length n, Σ_n^ν, into K corresponding parts. Actually we first approximate e_ν by f_δ, but this is just a technical detail. Two words in the same part have nearly equal ν probability. The appearance of $(1/n)\log$ in the asymptotic limits means that we only need to consider a single term which maximizes a combination of the entropy of a part and the probability of words in that part. This approach suffices for the large deviation results and specific Rényi entropy. Some elementary inequalities are needed with the guesswork asymptotics. In expressions involving ranks, we show that the same result obtains if parts other than that where the maximum occurs are ignored.

First we consider the large deviation problem, using well-known result that the sequence $\{\sigma \circ T_n^{-1} : n \in \mathbf{N}\}$ satisfies a large deviation principle with rate function $h(\cdot\,|\,\sigma)$.

Theorem II.1: For each closed subset $F \subset \mathcal{M}$

$$\limsup_n \frac{1}{n} \log \sigma(\{\omega \in \Omega :\, T_n(\omega) \in F\}) \le \sup_{\rho \in F \cap \mathcal{M}_S} h_{\mathrm{Sh}}(\rho) - \log r. \tag{12}$$

For each open subset $G \subset \mathcal{M}$

$$\liminf_n \frac{1}{n} \log \sigma(\{\omega \in \Omega :\, T_n(\omega) \in G\}) \ge \sup_{\rho \in G \cap \mathcal{M}_S} h_{\mathrm{Sh}}(\rho) - \log r. \tag{13}$$

Basic to the discussion below are sets of the form

$$\{\omega \in \Sigma_\nu\colon T_n(\omega) \in B\}, \tag{14}$$

where $B \subset \mathcal{M}$. We treat such sets by considering the corresponding words of length n, $\mathrm{w} = X_n(\omega)$. One difficulty is that there can be ω, $\omega' \in \Sigma^\nu$ so that $X_n(\omega) = X_n(\omega')$, $T_n(\omega) \in B$ but $T_n(\omega') \notin B$. To handle this situation we define two sets of words of length n associated with (14):

$$\widehat{\Gamma}_{n,B} := \{\mathrm{w} \in \Sigma_n^\nu \,:\, \exists\, \omega \in \Sigma^\nu, X_n(\omega) = \mathrm{w} \text{ and } T_n(\omega) \in B\} \tag{15}$$

and

$$\widetilde{\Gamma}_{n,B} := \{\mathrm{w} \in \Sigma_n^\nu \,:\, \forall\, \omega \in \Sigma^\nu, X_n(\omega) = \mathrm{w} \Rightarrow T_n(\omega) \in B\}. \tag{16}$$

Note $\widetilde{\Gamma}_{n,B} \subset \Sigma_n^\nu \cap X_n(T_n^{-1}B) = \widehat{\Gamma}_{n,B}$ and

$$|\widetilde{\Gamma}_{n,B}| \le r^n \sigma(T_n^{-1}B) \le |\widehat{\Gamma}_{n,B}|. \tag{17}$$

Consider the word $\mathrm{w} \in \Sigma_n^\nu$ and all possible infinite extensions: $[\mathrm{w}]_1^n \cap \Sigma_n^\nu$. If $\omega \in [\mathrm{w}]_1^n \cap \Sigma^\nu$ satisfies $T_n(\omega) \in B$, then $\mathrm{w} \in \widehat{\Gamma}_{n,B}$; if for *all* $\omega \in [\mathrm{w}]_1^n \cap \Sigma^\nu$, $T_n(\omega) \in B$, then $\mathrm{w} \in \widetilde{\Gamma}_{n,B}$.

Here is an example in which $\widetilde{\Gamma}_{n,B}$ and $\widehat{\Gamma}_{n,B}$ differ. Let $\mathtt{A} := \{0,1\}$. Let $f^*\colon \Omega \to \mathbf{R}$ be given by $f^*(\omega) = \omega_1\omega_2$. Then $\langle\, f^*, T_n(\omega)\,\rangle$ is equal to $1/n$ times the number of adjacent pairs of $1's$ in $\omega_1, \ldots, \omega_{n+1}$. Define

$$B^* := \{\rho \in \mathcal{M}\colon \langle\, f^*, \rho\,\rangle \ge 0.5\}.$$

Let $m \geq 1$ and $n := 2m$. Let w^* be the word consisting of m 0's followed by m 1's. If $X_n(\omega) = \mathrm{w}^*$, then $T_n(\omega) \in B^*$ if and only if $\omega_{n+1} = 1$, so $\mathrm{w}^* \in \widehat{\Gamma}_{n,B^*} \setminus \widetilde{\Gamma}_{n,B^*}$. However, lemma IV.1 shows that, if $F \subset G$, where F is closed and G is open, then for sufficiently large n, $\widehat{\Gamma}_{n,F} \subset \widetilde{\Gamma}_{n,G}$.

Theorem II.2: For each closed subset $F \subset \mathcal{M}^\nu$

$$\limsup_n \frac{1}{n} \log |\widehat{\Gamma}_{n,F}| \leq \sup_{\rho \in F \cap \mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho). \tag{18}$$

If Σ^ν satisfies hypothesis H1, then for each open subset $G \subset \mathcal{M}^\nu$

$$\liminf_n \frac{1}{n} \log |\widetilde{\Gamma}_{n,G}| \geq \sup_{\rho \in G \cap \mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho). \tag{19}$$

Theorem II.1 is formulated for the full shift space Ω. For the slightly stronger theorem II.2, the upper bound will be shown to follow from that for Ω; the lower bound does not obtain in general, but is valid under H1.

The next results relate to the large deviation principle for the empirical measure relative to ν.

Proposition II.1: Let $\nu \in \mathcal{M}_S$ satisfy H2. Then for each $\rho \in \mathcal{M}_S^\nu$,

$$h(\rho \,|\, \nu) := \lim_n \frac{1}{n} \sum_{\mathrm{w} \in \Sigma_n^\nu} \rho([\mathrm{w}]_1^n) \log \frac{\rho([\mathrm{w}]_1^n)}{\nu([\mathrm{w}]_1^n)} \tag{20}$$

exists and equals $\langle e_\nu, \rho \rangle - h_{\mathrm{Sh}}(\rho)$.

Proposition II.2: Let $\nu \in \mathcal{M}_S$ satisfy H2. Let F be a closed subset of $\mathcal{M}^\nu$. Then

$$\limsup_n \frac{1}{n} \log \nu(\{\omega \,:\, T_n(\omega) \in F\}) \leq \sup_{\rho \in F \cap \mathcal{M}_S^\nu} -h(\rho \,|\, \nu). \tag{21}$$

Proposition II.3: Let $\nu \in \mathcal{M}_S$ satisfy H1 and H2. Let G be an open subset of $\mathcal{M}^\nu$. Then

$$\liminf_n \frac{1}{n} \log \nu(\{\omega \,:\, T_n(\omega) \in G\}) \geq \sup_{\rho \in G \cap \mathcal{M}_S^\nu} -h(\rho \,|\, \nu). \tag{22}$$

Since $\mathcal{M}^\nu$ is compact, the above propositions show that $\{\nu \circ T_n^{-1}\}$ satisfies a large deviation principle with rate function $h(\cdot \,|\, \nu)$, when H1 and H2 obtain. The same techniques allow us to compute specific Rényi entropies and guessswork moments. Note, however, that a set which supports the β^{th} specific guesswork moment of ν can be distinct from one which supports ν. For example in the Bernoulli $\frac{1}{3}$ case with $\mathtt{A} = \{0, 1\}$, for large n the measure is concentrated on ω which have $\sum_1^n \omega_i \approx n/3$, while the dominant term in the $\beta = 1$ guesswork moment comes from ω which have $\sum_1^n \omega_i \approx n\sqrt{1/3}/(\sqrt{1/3} + \sqrt{2/3})$.

Theorem II.3: Let $\nu \in \mathcal{M}_S$ satisfy H1 and H2. If $\alpha \geq 0$, $\alpha \neq 1$, then

$$\lim_n \frac{1}{n(1-\alpha)} \log \sum_{\mathrm{w} \in \Sigma_n^\nu} \big(\nu([\mathrm{w}]_1^n)\big)^\alpha$$

$$= \frac{1}{(1-\alpha)} \sup_{\rho \in \mathcal{M}_S^\nu} \Big[h_{\mathrm{Sh}}(\rho) - \alpha \langle e_\nu, \rho \rangle\Big]. \tag{23}$$

Theorem II.4: Let $\nu \in \mathcal{M}_S$ satisfy H1 and H2. If $\beta > -1$, then

$$\lim_n \frac{1}{n} \log \sum_{\mathrm{w} \in \Sigma_n^\nu} \nu([\mathrm{w}]_1^n)\, \mathrm{rank}^\beta(\nu([\mathrm{w}]_1^n))$$

$$= (1+\beta) \sup_{\rho \in \mathcal{M}_S^\nu} \Big[h_{\mathrm{Sh}}(\rho) - \frac{\langle e_\nu, \rho \rangle}{1+\beta}\Big]. \tag{24}$$

III. Application to decoding

Proofs of the above results are given in a section below. First we discuss the implications and give an application. Let $\nu' \in \mathcal{M}_S^\nu$ and let F be a closed neighbourhood of ν' such that

$$\sup_{\rho \in F \cap \mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho) \approx h_{\mathrm{Sh}}(\nu').$$

If $T_n(\omega) \in F$, then the sample statistics of ω approximate the distribution of ν'. The nature of the approximation depends on the choice of neighbourhood F. Theorem II.2 means that the number of words of length n with these

m	$1-\mathcal{P}^{s_m}$	$\mathcal{G}^{s_m}$	$0.5\mathcal{R}^{s_m}$
1	2.64×10^{-1}	$1.85\times 10^{+2}$	$1.96\times 10^{+2}$
2	8.02×10^{-2}	$4.63\times 10^{+4}$	$5.22\times 10^{+4}$
3	1.89×10^{-2}	$5.17\times 10^{+6}$	$6.17\times 10^{+6}$
4	3.64×10^{-3}	$3.24\times 10^{+8}$	$4.11\times 10^{+8}$
5	5.88×10^{-4}	1.30×10^{10}	1.75×10^{10}
6	8.20×10^{-5}	3.63×10^{11}	5.20×10^{11}
42	$< 1\times 10^{-9}$	1.15×10^{25}	2.01×10^{26}

TABLE I
Probability and expected guesses, $n = 1000$, $\theta = 0.001$

sample statistics, $|\widetilde{\Gamma}_{n,F}|$, is approximately $\exp(nh_{\mathrm{Sh}}(\nu'))$. Propositions II.2 and II.3 are similar, but the conclusion is that $\nu_n(\widetilde{\Gamma}_{n,F})$ is approximated by $\exp -nh(\nu' \,|\, \nu)$.

The guesswork moment of most significance is that for $\beta = 1$. Let $\mathtt{G}$ denote the limit in (24) with $\beta = 1$. Note that (23) with $\alpha = 1/2$ also yields $\mathtt{G}$. The expected number of guesses required to determine a word of length n chosen according to the distribution ν is approximately $\exp n\mathtt{G}$.

One aspect of expectations of the form (1) is that a large contribution can be due to events with very low probability. This suggests truncating the procedure when the probability of untried words reaches a certain level. Truncation can reduce the expected number of guesses dramatically, but this means the procedure will occasionally fail.

We now discuss two related examples of decoding by repeated guessing. We give an upper bound for the guesswork expectation which allows for truncation. Also, the ranking function $2i-1$ is technically convenient. For the probability distribution $p_1 \geq p_2 \geq \cdots \geq p_n$ we define for $1 \leq m \leq n$,

$$\mathcal{G}^m := \sum_{i=1}^{m} i\, p_i; \quad \widehat{\mathcal{G}}^m := \sum_{i=1}^{m} (2i-1)p_i; \tag{25}$$

$$\mathcal{P}^m := \sum_{i=1}^{m} p_i; \quad \mathcal{R}^m := \left(\sum_{i=1}^{m} p_i^{1/2}\right)^2 . \tag{26}$$

Note $\mathcal{G}^m = 0.5(\widehat{\mathcal{G}}^m + \mathcal{P}^m)$. By expanding (26) and noting that there are exactly $2i-1$ pairs $p_j^{1/2}p_k^{1/2}$ with $i = \max\{j,k\}$, we deduce

$$\widehat{\mathcal{G}}^m \leq \mathcal{R}^m, \quad \mathcal{G}^m \leq 0.5(\mathcal{R}^m + \mathcal{P}^m), \tag{27}$$

with equality if an only if $p_1 = \cdots = p_m$.

Now we apply guessing to channel decoding. First we assume a source of words $u_1, \ldots, u_n$, which have k data bits with $n-k$ verification bits embedded. We assume that the data bits are independent and identically distributed with equal probability for $0, 1$. The transmitted $u_1, \ldots, u_n$ is received as the binary word $v_1, \ldots, v_n$. We assume a binary symmetric channel: the probability that $u_i \neq v_i$ is $\theta \ll 1/2$ and independent of the remaining u_j, v_j. Under the simplifying assumption that the verification data has negligible effect on conditional distributions, it follows that the events $u_i \neq v_i$, $i = 1, \ldots, n$, conditioned on receiving $v_1, \ldots, v_n$, are independent with probability θ. Then $\mathcal{R}^{2^n} =$

$$\left(\sum_{0}^{n} \binom{n}{k} \theta^{k/2}(1-\theta)^{(n-k)/2}\right)^2 = \left(1 + 2\sqrt{\theta(1-\theta)}\,\right)^n ,$$

which, in this case, coincides with the asymptotic estimate $\exp n\mathtt{G}$ mentioned above. Define

$$s_m := \sum_{k=0}^{m} \binom{n}{k}, \quad \mathcal{P}^{s_m} := \sum_{k=0}^{m} \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

We consider the specific example with $n = 1000$, $\theta = 0.001$, Table I shows the residual probability and expected number of guesses for truncated guessing allowing for up to m changed bits Here the Rényi entropy upper bound is $0.5\,\mathcal{R}^{2^{1000}} = 2.09 \times 10^{26}$, while the actual total expected number of guesses, 1.15×10^{25}, is essentially achieved when terms for $m \leq 42$ are included.

We consider a second example in the same context: a stationary binary input $u_1, \ldots, u_n$ producing the binary output $v_1, \ldots, v_n$. The basic aim is to compute or bound $\mathcal{G}^m$ from the conditional distribution of $u_1, \ldots, u_n$ given $v_1, \ldots, v_n$.

j	k	$1-\mathcal{P}$	$\mathcal{G}$	$0.5\mathcal{R}$
0	15	6.96×10^{-2}	1.65×10^{3}	1.65×10^{3}
1	18	6.87×10^{-3}	1.01×10^{5}	1.39×10^{5}
1	21	9.11×10^{-4}	1.66×10^{5}	2.30×10^{5}
2	24	4.13×10^{-5}	1.46×10^{6}	2.74×10^{6}
2	26	7.52×10^{-6}	1.57×10^{6}	2.97×10^{6}
3	28	5.94×10^{-7}	5.40×10^{6}	1.36×10^{7}
10	36	$< 1 \times 10^{-9}$	2.36×10^{7}	1.76×10^{8}

TABLE II

Probability and expected guesses, $n = 2000$, $\theta = 0.005$

The difficulty is that this conditional distribution may not have a simple mathematical form. A case which does have a relatively simple form is as follows. The distribution of $u_1, \dots, u_n$ is the double even shift: each "run" of 1's or 0's must have even length, and all allowed words of length n have equal probability. The distribution of $v_1, \dots, v_n$ is obtained from $u_1, \dots, u_n$ by requiring that the events $u_i \neq v_i$ are independent with probability θ. We take n to be even. For a long sequence $\{u_i\}$, by shifting the index by 1 if necessary, we can arrange for $u_{k+1} \neq u_k$ only when k is even. We consider $v_1, \dots, v_n$ as $n/2$ adjacent o-pairs: (v_k, v_{k+1}) with k *odd.* We say that a *defect* occurs at k if k is odd and $v_{k+1} \neq v_k$, which means that exactly one of $u_k \neq v_k$ and $u_{k+1} \neq v_{k+1}$ obtains. The probability that there are exactly k defects and exactly j o-pairs both of which have been changed is

$$2^k \theta^k (1-\theta)^k \binom{n/2}{k} \binom{n/2-k}{j} \theta^{2j} (1-\theta)^{n-2k-2j}.$$

Given $v_1, \dots, v_n$ with k defects, there are 2^k choices for which member of each of these pairs which has changed; this is combined with the probabilities corresponding to o-pairs both of which have changed. One can show that $\widehat{\mathcal{G}}$ in this case is 2^k times $\widehat{\mathcal{G}}$ for the probability distribution of the non-defect o-pairs (the corresponding result also holds for $\mathcal{R}$). Thus we compute $\widehat{\mathcal{G}}$ for the distribution

$$\binom{n/2-k}{j} \theta^{2j} (1-\theta)^{n-2j} (\theta^2 + (1-\theta)^2)^{-n/2+k},$$

$j = 0, \dots n-k$, which we denote $\widehat{\mathcal{G}}^{(n,k)}$. It is easy to compute the corresponding Rényi entropy:

$$\widehat{\mathcal{R}}^{(n,k)} = (\theta^2 + (1-\theta)^2)^{-n/2+k}.$$

The expected $\widehat{\mathcal{G}}$ number of guesses is

$$\widehat{\mathcal{G}}^* := \sum_{k=0}^{n/2} 2^k \widehat{\mathcal{G}}^{(n,k)} \binom{n/2}{k} 2^k \theta^k (1-\theta)^k (\theta^2 + (1-\theta)^2)^{n/2-k}.$$

Since $\widehat{\mathcal{G}}^{(n,k)} \leq \widehat{\mathcal{R}}^{(n,k)}$, we obtain

$$\widehat{\mathcal{G}}^* \leq (1 + 4\theta(1-\theta))^{n/2}.$$

Table II shows the residual probability and expected number of guesses for truncated guessing allowing for up to j changed o-pairs and k defects.

IV. Proofs

For a proof of theorem II.1 see [2], section 6.2. We prove theorem II.2 using theorem II.1, noting that Σ^ν is a closed subset of Ω and $\mathcal{M}^\nu$ is a closed subset of $\mathcal{M}$.

Lemma IV.1: Let $F \subset G \subset \mathcal{M}$ with F closed and G open. Then there exists $n' \in \mathbf{N}$ such that for all $n \geq n'$, $\widehat{\Gamma}_{n,F} \subset \widetilde{\Gamma}_{n,G}$.

Proof: If there were no such n', then we could find a sequence $\{(\omega_{n_k}, \omega'_{n_k})\}$ with $X_{n_k}(\omega_{n_k}) = X_{n_k}(\omega'_{n_k})$ such that

$$T_{n_k}(\omega_{n_k}) \in F, \; T_{n_k}(\omega'_{n_k}) \notin G, \tag{28}$$

$$\lim_{n_k} T_{n_k}(\omega_{n_k}) = \rho^* \in F, \; \lim_{n_k} T_{n_k}(\omega'_{n_k}) = \rho' \in \mathcal{M} \setminus G. \tag{29}$$

Now if $f \in \mathcal{F}_m$ and $X_n(\omega) = X_n(\omega')$, then

$$|\langle f, T_n(\omega) \rangle - \langle f, T_n(\omega') \rangle| \leq 2\|f\| \frac{m-1}{n}. \tag{30}$$

Hence we have $\langle f, \rho^* \rangle = \langle f, \rho' \rangle$. Since this holds for all local f, we have $\rho^* = \rho'$, which contradicts (29). ■

Lemma IV.2: Let G be an open set in $\mathcal{M}^\nu$. Let $\rho \in G$ be an ergodic probability measure on Σ^ν. Then

$$\liminf_n \frac{1}{n} \log |\widetilde{\Gamma}_{n,G}| \geq h_{\mathrm{Sh}}(\rho). \tag{31}$$

Proof: Let $\{f_j\}$ be a sequence of local functions which determines the topology of $\mathcal{M}$. The open set $G \subset \mathcal{M}^\nu$ can be expressed as $G = G' \cap \mathcal{M}^\nu$, where G' is open in $\mathcal{M}$. There exist $\{\varepsilon_1 > 0, \ldots, \varepsilon_m > 0\}$ so that

$$N := \{\rho' \in \mathcal{M} \,:\, |\langle f_i, \rho' - \rho \rangle| \leq \varepsilon_i,\ i = 1, \ldots, m\} \subset G'. \tag{32}$$

By lemma IV.1 for all sufficiently large n

$$\widehat{\Gamma}_{n,N} \subset \widetilde{\Gamma}_{n,G'} = \widetilde{\Gamma}_{n,G}. \tag{33}$$

Since ρ is assumed to be ergodic, there exists a Borel set $B \subset \Sigma^\nu$ so that $\rho(B) = 1$ and

$$\omega \in B \Longrightarrow \lim_n \langle f_i, T_n(\omega) \rangle = \langle f_i, \rho \rangle,\ i = 1, \ldots, m. \tag{34}$$

It follows that for each $\omega \in B$ there exists n_ω so that $n > n_\omega \Rightarrow T_n(\omega) \in N$; hence

$$\lim_n \rho_n(\widetilde{\Gamma}_{n,G}) = 1. \tag{35}$$

Then (31) follows from lemma 2.1 of [4]. ■

Proof of (19): Hypothesis H1 implies that for open $G \subset \mathcal{M}^\nu$

$$\sup_{\rho \in G \cap \mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho) = \sup_{\text{ergodic } \rho \in G \cap \mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho); \tag{36}$$

the rest follows from the above lemma. ■

Proof of (18): We consider $F = F \cap \mathcal{M}^\nu$ as a subset of $\mathcal{M}$. For each $\rho \in F \setminus \mathcal{M}_S$, we can find a closed neighbourhood N_ρ with $N_\rho \cap \mathcal{M}_S = \emptyset$. Given $\varepsilon > 0$ and $\rho \in F \cap \mathcal{M}_S$, we can find a closed neighbourhood N_ρ of ρ of the form (32) so that

$$\sup_{\rho' \in N_\rho \cap \mathcal{M}_S} h_{\mathrm{Sh}}(\rho') \leq h_{\mathrm{Sh}}(\rho) + \varepsilon. \tag{37}$$

Here we use the upper semicontinuity of $h_{\mathrm{Sh}}(\,\cdot\,)$ which follows from the lower semicontinuity of $h(\,\cdot\,|\,\sigma)$ (see [2]). By compactness, there exist $\rho_1, \ldots, \rho_m$ so that the interiors $\{N_{\rho_i}^\circ\}$ cover F. Lemma IV.1 shows that there exists n' so that for $n \geq n'$, $T_n(\omega') \in \bigcup_i N_{\rho_i}^\circ$ whenever both $T_n(\omega) \in F$ and $X_n(\omega') = X_n(\omega)$. We apply (12) to deduce

$$\limsup_n \frac{1}{n} \log |\widetilde{\Gamma}_{n,\bigcup_i N_{\rho_i}^\circ}| \leq$$

$$\log r + \limsup_n \frac{1}{n} \log \sigma(\{\omega \,:\, T_n(\omega) \in \bigcup_i N_{\rho_i}\}) \tag{38}$$

$$\limsup_n \frac{1}{n} \log |\widehat{\Gamma}_{n,F}| \leq \sup_{\rho \in \mathcal{M}_S \cap \bigcup N_{\rho_i}} h_{\mathrm{Sh}}(\rho). \tag{39}$$

From (37) and the fact that $\varepsilon > 0$ is arbitrary we deduce (18). ■

An elementary consequence of the product topology on $\Omega = \mathtt{A}^{\mathbf{N}}$ and compactness of $\mathtt{A}$ is the following.

Lemma IV.3: Let $\nu \in \mathcal{M}_S$ be a probability measure verifying hypothesis H2. Then for each $\delta > 0$ there exist $m_\delta, N_\delta \in \mathbf{N}$ and f_δ, which is $\mathcal{F}_{m_\delta}$ measurable, so that $\forall\, n \geq N_\delta$, $\forall\, \omega \in \Sigma^\nu$, $|e_\nu(\omega) - f_\delta(\omega)| \leq \delta/2$ and

$$|\langle f_\delta, T_n(\omega) \rangle + \frac{1}{n} \log \nu([X_n(\omega)]_1^n)| < \delta\,.$$

Corollary IV.1: Let $\nu \in \mathcal{M}_S$ be a probability measure verifying hypothesis H2. For $\rho \in \mathcal{M}_S^\nu$ we have

$$\lim_{n \to \infty} -\frac{1}{n} \sum_{\mathrm{w} \in \Sigma_n^\nu} \rho([\mathrm{w}]_1^n) \log \nu([\mathrm{w}]_1^n) = \langle e_\nu, \rho \rangle\,. \tag{40}$$

The above corollary and the well-known limit (8) prove proposition II.1.

Lemma IV.4: For $\delta > 0$ and f_δ, m_δ, N_δ as in lemma IV.3, there exist an integer K_δ, numbers $0 \leq a_0 < \cdots < a_{K_\delta}$ with $a_j - a_{j-1} < \delta$, $j = 1, \ldots, K_\delta$ and sets $\{G_j^\delta \subset F_j^\delta \subset \mathcal{M}^\nu \,:\, j = 0, \ldots, K_\delta\}$ so that each G_j^δ is open and each F_j^δ is closed and

$$\bigcup_{j=0}^{N_\delta} G_j^\delta = \mathcal{M}^\nu, \tag{41}$$

$$\rho \in G_j^\delta \Rightarrow |\langle\, f_\delta, \rho\,\rangle - a_j| < \delta, \tag{42}$$

$$\rho \in F_j^\delta \Rightarrow |\langle\, f_\delta, \rho\,\rangle - a_j| \leq \delta. \tag{43}$$

Proof: Define K_δ, a_j, G_j^δ and F_j^δ by

$$K_\delta := \left\lceil \frac{1 + \|f_\delta\|}{\delta} \right\rceil, \quad a_j := \frac{j}{K_\delta}\|f_\delta\|, \tag{44}$$

$$G_j^\delta := \{\rho \in \mathcal{M}^\nu \,:\, |\langle\, f_\delta, \rho\,\rangle - a_j| < \delta\}, \tag{45}$$

$$F_j^\delta := \{\rho \in \mathcal{M}^\nu \,:\, |\langle\, f_\delta, \rho\,\rangle - a_j| \leq \delta\}. \tag{46}$$

■

Proof of proposition II.2: Using the notation of (15) from (18) we have

$$\limsup_n \frac{1}{n} \log |\widehat{\Gamma}_{n, F\cap F_j^\delta}| \leq \sup_{\rho \in F_j^\delta \cap F \cap \mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho). \tag{47}$$

Also

$$\frac{1}{n} \log \nu(\{\omega \,:\, T_n(\omega) \in F \cap F_j^\delta\}) \leq$$

$$\frac{1}{n} \log |\widehat{\Gamma}_{n, F\cap F_j^\delta}| + \max_{\mathrm{w} \in \widehat{\Gamma}_{n, F\cap F_j^\delta}} \frac{1}{n} \log \nu([\mathrm{w}]_1^n). \tag{48}$$

From lemma IV.3 we deduce that for $n \geq N_\delta$, $\rho \in F_j^\delta \cap \mathcal{M}_S^\nu$ and $\mathrm{w} \in \widehat{\Gamma}_{n, F_j^\delta}$,

$$\frac{1}{n} \log \nu([\mathrm{w}]_1^n) \leq -a_j + 2\delta, \ \langle\, e_\nu, \rho\,\rangle \leq a_j + 2\delta. \tag{49}$$

From proposition II.1 we deduce

$$\sup_{\rho \in F \cap F_j^\delta \cap \mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho) \leq \sup_{\rho \in F \cap F_j^\delta \cap \mathcal{M}_S^\nu} -h(\rho \,|\, \nu) + a_j + 2\delta. \tag{50}$$

Now

$$\limsup_n \frac{1}{n} \log \nu(\{\omega \,:\, T_n(\omega) \in F\}) \tag{51}$$

equals the maximum over the corresponding $\limsup$'s with F replaced by $F \cap F_j^\delta$. We then have

$$\limsup_n \frac{1}{n} \log \nu(\{\omega \,:\, T_n(\omega) \in F\}) \leq$$

$$\max_{0 \leq j \leq K_\delta} \sup_{\rho \in F \cap F_j^\delta \cap \mathcal{M}_S^\nu} -h(\rho \,|\, \nu) + 4\delta. \tag{52}$$

As $\delta > 0$ is arbitrary, the proposition follows. ■

Proof of proposition II.3: Let $G \subset \mathcal{M}^\nu$ be open and let $\rho \in G \cap \mathcal{M}_S^\nu$. Then there exists j so that $\rho \in G_j^\delta$. From (19) given $\varepsilon > 0$, for all sufficiently large n

$$\frac{1}{n} \log |\widetilde{\Gamma}_{n, G\cap G_j^\delta}| \geq h_{\mathrm{Sh}}(\rho) - \varepsilon. \tag{53}$$

By an argument similar to that in the proof of proposition II.2,

$$-h(\rho \,|\, \nu) = h_{\mathrm{Sh}}(\rho) - \langle\, e_\nu, \rho\,\rangle \leq h_{\mathrm{Sh}}(\rho) - a_j + 2\delta. \tag{54}$$

Also for $\mathrm{w} \in \widetilde{\Gamma}_{n,G\cap G_j^\delta}$,

$$\frac{1}{n}\log \nu([\mathrm{w}]_1^n) \geq -a_j - 2\delta. \tag{55}$$

Then

$$\frac{1}{n}\log \nu(\widetilde{\Gamma}_{n,G\cap G_j^\delta}) \geq h_{\mathrm{Sh}}(\rho) - \varepsilon - a_j - 2\delta \geq -h(\rho\,|\,\nu) - \varepsilon - 4\delta. \tag{56}$$

Since ε and δ are arbitrary and $\nu(T_n^{-1}G) \geq \nu(\widetilde{\Gamma}_{n,G\cap G_j^\delta})$, the proposition follows. ■

Proof of theorem II.3: We use the covers $\{G_j^\delta\}$ and $\{F_j^\delta\}$ introduced in lemma IV.4 and the notation (15), (16). For $\alpha \geq 0$, $\rho \in F_j^\delta \cap \mathcal{M}_S^\nu$ and $n \geq N_\delta$, arguing as above we deduce

$$\mathrm{w} \in \widehat{\Gamma}_{n,F_j^\delta} \Rightarrow \frac{1}{n}\log\left(\nu([\mathrm{w}]_1^n)\right)^\alpha \leq \alpha(-\langle\, e_\nu, \rho\,\rangle + 4\delta) \tag{57}$$

and

$$\limsup_n \frac{1}{n}\log|\widehat{\Gamma}_{n,F_j^\delta}| \leq \sup_{\rho\in F_j^\delta\cap\mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho), \tag{58}$$

so

$$\begin{aligned}&\limsup_n \frac{1}{n}\log \sum_{\mathrm{w}\in\widehat{\Gamma}_{n,F_j^\delta}} \left(\nu([\mathrm{w}]_1^n)\right)^\alpha \\ &\leq \sup_{\rho\in F_j^\delta\cap\mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho) - \alpha\langle\, e_\nu, \rho\,\rangle + 4\alpha\delta.\end{aligned} \tag{59}$$

The lower bound is similar.

$$\begin{aligned}&\liminf_n \frac{1}{n}\log \sum_{\mathrm{w}\in\widetilde{\Gamma}_{n,G_j^\delta}} \left(\nu([\mathrm{w}]_1^n)\right)^\alpha \\ &\geq \sup_{\rho\in G_j^\delta\cap\mathcal{M}_S^\nu} h_{\mathrm{Sh}}(\rho) - \alpha\langle\, e_\nu, \rho\,\rangle - 4\alpha\delta.\end{aligned} \tag{60}$$

The theorem follows by noting that $\bigcup\widetilde{\Gamma}_{n,G_j^\delta} \subset \Sigma_n^\nu = \bigcup\widehat{\Gamma}_{n,F_j^\delta}$, that the lim inf and lim sup of the total sum equals the maximum over the $K_\delta + 1$ sets of each cover and that $\delta > 0$ is arbitrary. ■

Proof of theorem II.4: This proof is similar to that just completed, but treatment of the rank function requires some care. We use $\{\widehat{\Gamma}_{n,F_j^\delta}\}$, which splits Σ_n^ν into $K_\delta + 1$ overlapping parts. For $j = 0, \ldots, K_\delta$, we set

$$h_j := |\widehat{\Gamma}_{n,F_j^\delta}|,\ g_0 := 0,\ g_{j+1} := g_j + h_j, \tag{61}$$

select the ranking functions $\{\mathrm{rnk}_j\}$ so that

$$\mathrm{rnk}_j : \widehat{\Gamma}_{n,F_j^\delta} \to \{g_j + 1, \ldots, g_j + h_j\} \tag{62}$$

and define the injection $\mathrm{rnk} : \Sigma_n^\nu \to \{1, \ldots, g_{K_\delta+1}\}$,

$$\mathrm{rnk}(\mathrm{w}) := \min_j\{\mathrm{rnk}_j(\mathrm{w})\colon \mathrm{w} \in \widehat{\Gamma}_{n,F_j^\delta}\}. \tag{63}$$

The properties (see (3)) of $\mathrm{rank}(\,\cdot\,)$ imply

$$\begin{aligned}&\sum_{\mathrm{w}\in\Sigma_n^\nu} \nu([\mathrm{w}]_1^n)\,\mathrm{rank}^\beta(\nu([\mathrm{w}]_1^n)) \begin{cases}\leq & (\text{if } \beta \geq 0)\\ \geq & (\text{if } \beta < 0)\end{cases} \\ &\sum_{\mathrm{w}\in\Sigma_n^\nu} \nu([\mathrm{w}]_1^n)\,\mathrm{rnk}^\beta(\mathrm{w}).\end{aligned} \tag{64}$$

For the $\beta \geq 0$ case we then have

$$\sum_{\mathrm{w}\in\widehat{\Gamma}_{n,F_j^\delta}} \nu([\mathrm{w}]_1^n)\,\mathrm{rank}^\beta(\nu([\mathrm{w}]_1^n)) \leq$$

$$\sum_{k=1}^{h_j}(g_j+k)^\beta \max_{\mathrm{w}\in\widehat{\Gamma}_{n,F_j^\delta}} \nu([\mathrm{w}]_1^n). \tag{65}$$

Using (49) and the bound

$$(g+h)^\beta h \geq \sum_{k=g+1}^{g+h} k^\beta \geq \int_0^h x^\beta\,dx = \frac{h^{1+\beta}}{1+\beta}, \tag{66}$$

we deduce

$$\limsup_n \frac{1}{n}\log\sum_{j=0}^{K_\delta}\sum_{k=1}^{h_j}(g_j+k)^\beta \max_{\mathrm{w}\in\widehat{\Gamma}_{n,F_j^\delta}} \nu([\mathrm{w}]_1^n) \leq$$

$$\max_{j=0,\ldots,K_\delta}\left[\limsup_n \frac{1}{n}(\log h_j + \beta\log(g_{j+1})\,) - a_j\right] + 2\delta. \tag{67}$$

Define H_j

$$H_j := \limsup_n \frac{1}{n}\log h_j \tag{68}$$

and then choose j^* so that

$$(1+\beta)H_j - a_j \leq (1+\beta)H_{j^*} - a_{j^*},\ j=0,\ldots,N_\delta. \tag{69}$$

Since $g_{j+1} = \sum_0^j h_j$, we have

$$\max_{j=0,\ldots,K_\delta}\ \limsup_n \frac{1}{n}(\log h_j + \beta\log(g_{j+1})\,) - a_j \leq$$

$$\max_{j=0,\ldots,K_\delta}\left[H_j + \max_{k\leq j}\beta H_k - a_j\right] = (1+\beta)H_{j^*} - a_{j^*}, \tag{70}$$

because if $k \leq j$, $-a_j \leq -a_k$, so that if $H_k \geq H_j$, we have

$$H_j + \beta H_k - a_j \leq (1+\beta)H_k - a_k \leq (1+\beta)H_{j*} - a_{j*}. \tag{71}$$

Using the same techniques as in the previous proof, we deduce

$$\limsup_n \frac{1}{n}\log\sum_{\mathrm{w}\in\Sigma_n^\nu} \nu([\mathrm{w}]_1^n)\,\mathrm{rank}^\beta(\nu([\mathrm{w}]_1^n)) \leq$$

$$\sup_{\rho\in\mathcal{M}_S^\nu}(1+\beta)h_{\mathrm{Sh}}(\rho) - \langle\, e_\nu, \rho\,\rangle + 4\delta. \tag{72}$$

Next we use the second inequality in (66) to deduce

$$\sum_{\mathrm{w}\in\widehat{\Gamma}_{n,F_j^\delta}} \nu([\mathrm{w}]_1^n)\,\mathrm{rank}^\beta(\nu([\mathrm{w}]_1^n)) \geq$$

$$\frac{h_j^{1+\beta}}{1+\beta}\min_{\mathrm{w}\in\widehat{\Gamma}_{n,F_j^\delta}} \nu([\mathrm{w}]_1^n) \tag{73}$$

for each j.

Since $\Gamma_{n,G_j^\delta} \subset \widehat{\Gamma}_{n,F_j^\delta}$, we have

$$\liminf_n \frac{1}{n}\log\sum_{\mathrm{w}\in\Sigma_n^\nu} \nu([\mathrm{w}]_1^n)\,\mathrm{rank}^\beta(\nu([\mathrm{w}]_1^n)) \geq$$

$$\sup_{\rho\in\mathcal{M}_S^\nu}(1+\beta)h_{\mathrm{Sh}}(\rho) - \langle\, e_\nu, \rho\,\rangle - 4\delta. \tag{74}$$

This covers the $\beta \geq 0$ case. The $-1 < \beta < 0$ is similar, using

$$(g+h)^\beta\, h \leq \sum_{k=g+1}^{g+h} k^\beta \leq \int_0^h x^\beta\,dx = \frac{h^{1+\beta}}{1+\beta}. \tag{75}$$

From this it is not difficult to deduce (72). For the lower bound, when $-1 < \beta < 0$, we note that

$$\sum_{k=g_j+1}^{g_j+h_j} k^{\beta} \geq (g_j + h_j)^{\beta}\, h_j. \tag{76}$$

Notice that, although the rank function we use may not correspond strictly to the actual ordering, we still have a lower bound. We have

$$\sum_{\mathrm{w}\in\Sigma_n^{\nu}} \nu([\mathrm{w}]_1^n)\,\mathrm{rank}^{\beta}(\nu([\mathrm{w}]_1^n)) \geq h_j \min_{\mathrm{w}\in\widehat{\Gamma}_{n,F_j^{\delta}}} \nu([\mathrm{w}]_1^n)\Big(\sum_{k=0}^{j} h_k\Big)^{\beta}. \tag{77}$$

We redefine H_j

$$H_j := \liminf_n \frac{1}{n}\log h_j, \tag{78}$$

and then choose j^* so that (69) obtains. Then

$$\liminf_n \frac{1}{n}\log \sum_{\mathrm{w}\in\Sigma_n^{\nu}} \nu([\mathrm{w}]_1^n)\,\mathrm{rank}^{\beta}(\nu([\mathrm{w}]_1^n)) \geq$$

$$H_{j^*} + \beta \max_{k\leq j^*} H_k - a_{j^*} - 2\delta. \tag{79}$$

Now $k < j^* \Rightarrow a_{j^*} - a_k > 0$, so

$$(1+\beta)H_k \leq (1+\beta)H_{j^*} - (a_{j^*} - a_k) \Rightarrow H_k < H_{j^*}. \tag{80}$$

This means

$$\liminf_n \frac{1}{n}\log \sum_{\mathrm{w}\in\Sigma_n^{\nu}} \nu([\mathrm{w}]_1^n)\,\mathrm{rank}^{\beta}(\nu([\mathrm{w}]_1^n))$$

$$\geq \max_{j=0,\ldots,K_{\delta}} [(1+\beta)H_j - a_j] - 2\delta. \tag{81}$$

Since $\widetilde{\Gamma}_{n,G_j^{\delta}} \subset \widehat{\Gamma}_{n,F_j^{\delta}}$,

$$\max_{j=0,\ldots,K_{\delta}} [(1+\beta)H_j - a_j] - 2\delta \geq$$

$$\sup_{\rho\in\mathcal{M}_S^{\nu}} (1+\beta)h_{\mathrm{Sh}}(\rho) - \langle\, e_{\nu}, \rho\,\rangle - 4\delta, \tag{82}$$

so we have (74). ■

References

[1] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Transactions on Information Theory*, vol. 34, pp. 55-63, January 1988.

[2] A. Dembo and O. Zeitouni, *Large deviation techniques and applications*, Springer, New York 1998.

[3] A. Eizenberg, Y. Kifer and B. Weiss, "Large Deviations for $\mathbf{Z}^d$-actions," *Commun. Math. Phys.* **164** 433-454, 1994.

[4] J.T. Lewis, C.-E. Pfister, R. Russell, W.G. Sullivan, "Reconstruction sequences and equipartition measures: an examination of the asymptotic equipartition property," *IEEE Information Theory* **43**, 1935-1947, 1997.

[5] J.T. Lewis, C.-E. Pfister and W.G. Sullivan, "Entropy, Concentration of Probability and Conditional Limit Theorems," *Markov Processes and Related Fields* **1**, 319–386 1995.

[6] Douglas Lind and Brian Marcus, *Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge CB2 1RP UK, 1995.

[7] D. Malone and W.G. Sullivan, "Guesswork and Entropy," *IEEE Transactions on Information Theory*, vol. 50, 525-526, March 2004.

[8] B. Marcus, "A note on periodic points for ergodic toral automorphisms," *Mh. Math.* **89**, 121-129, 1980.

[9] C.-E. Pfister and W.G. Sullivan, "Billingsley dimension on shift spaces," *Nonlinearity* **16**, 661-682, 2003.

[10] A. Rényi, *Probability theory*, North-Holland, Amsterdam, 1970.

[11] R. Rivest, "The MD5 Message-Digest Algorithm," ftp://ftp.rfc-editor.org/in-notes/rfc1321.txt, 1992.

Charles Ed. Pfister received the Ph.D. degree from the Swiss Federal Institute of Technology in Zürich, Switzerland. He has held research positions at the Zentrum für interdisziplinäre Forschung in Bielefeld, Germany; at CNRS in Marseille, France; and at Rutgers University in New Brunswick, N.J. USA. Since 1979, he has a research and teaching position at the Swiss Federal Institute of Technology in Lausanne, Switzerland. He has worked in statistical mechanics, quantum theory, and applied probability, in particular on Gibbs random fields. His current research interests are in application of large deviations theory to information theory and ergodic theory.

Wayne G. Sullivan received a D.Phil. in Mathematics from Oxford University in 1968 after receiving an undergraduate degree in Chemistry from Georgia Institute of Technology. On the completion of his D.Phil., he returned to the Georgia Institute of Technology as an Assistant Professor in Mathematics. In 1973 he moved to Dublin, initially to the Dublin Institute for Advanced Studies and then to the Department of Mathematics, University College Dublin. He is currently on leave of absence from UCD, working as Senior Researcher at the Communications Network Research Institute of the Dublin Institute of Technology. His current interest is the application of large deviation techniques to problems in information theory, symbolic dynamics and networks.