Ruminations on Information Theory and Statistical Mechanics. Winter 2024 Toulouse Lectures.

Vojkan Jakšić

Department of Mathematics and Statistics
McGill University
805 Sherbrooke Street West
Montreal, QC, H3A 2K6, Canada

January 18, 2024

Contents

1 Prologue

Throughout this section $\mathcal{A}$ denotes a finite set. In information theory, the elements of $\mathcal{A}$ are signals from an underlying information source. In statistical mechanics, $\mathcal{A}$ is the set of configurations of the physical system under consideration. To uniformise the terminology, we will refer to $\mathcal{A}$ as the alphabet of the system. The set of probability measures on $\mathcal{A}$ is denoted by $\mathcal{P}(\mathcal{A})$. In information theory the source statistics is described by elements of $\mathcal{P}(\mathcal{A})$. In statistical mechanics the Gibbs ensemble statistics of the system is described by elements of $\mathcal{P}(\mathcal{A})$.

$P \in \mathcal{P}(\mathcal{A})$ is called faithful if $P(a) > 0$ for all $a \in \mathcal{A}$. The chaotic (or uniform) probability measure on $\mathcal{A}$ is $P_{\mathrm{ch}}(a) = 1/|\mathcal{A}|$, where $|\mathcal{A}|$ is the number of elements of $\mathcal{A}$.

Throughout the notes we adopt the convention $0 \log 0 = 0$, and $0^0 = 1$

1.1 Boltzmann's entropy

Write $|\mathcal{A}| = L$ and enumerate the elements of $\mathcal{A}$ as $\mathcal{A} = \{a_1, \cdots, a_L\}$. The elements of $\mathcal{A}$ correspond to physical configurations of a "gas molecule". An example is $\mathcal{A} = \{0, 1\}$ (see **(b)** below), where 0 corresponds to the configuration where the "gas molecule" is absent, and 1 corresponds to the configuration where the "gas molecule" is present. The elements of $\mathcal{A}^N$ correspond to physical configurations of a "gas of N molecules". Each "microstate" $\omega = (\omega_1, \cdots, \omega_N) \in \mathcal{A}^N$ is identified with the word $\omega = \omega_1 \cdots \omega_N$ of length N with letters ω_j from alphabet $\mathcal{A}$. Let $k_{a_1}, \cdots, k_{a_L}$ be non-negative integers such that $k_{a_1} + \cdots + k_{a_L} = N$. Then

$$T_N(k_{a_1}, \cdots, k_{a_L}) = \frac{N!}{k_{a_1}! \cdots k_{a_L}!} \tag{1.1}$$

is the number of words in $\mathcal{A}^N$ in which the letter a_j appears k_{a_j} times. We write k_j for k_{a_j}.

Suppose that for $1 \leq j \leq L$ natural numbers $k_j = k_j(N)$ are chosen so that for some $p_j \geq 0$

$$\lim_{N \to \infty} \frac{k_j(N)}{N} = p_j. \tag{1.2}$$

Obviously, $\sum_{j=1}^{L} p_j = 1$.

Proposition 1.1

$$\lim_{N \to \infty} \frac{1}{N} \log_{\mathrm{e}} T_N(k_1(N), \cdots, k_L(N)) = -\sum_{j=1}^{L} p_j \log_{\mathrm{e}} p_j.$$

Proof. The Stirling's approximation

$$\sqrt{2\pi} n^{n+\frac{1}{2}} \mathrm{e}^{-n} \le n! \le \mathrm{e}\, n^{n+\frac{1}{2}} \mathrm{e}^{-n}. \tag{1.3}$$

gives that

$$\lim_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} \frac{N^N}{N!} = 1 \qquad \text{and} \qquad \lim_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} \frac{k_j(N)^{k_j(N)}}{k_j(N)!} = p_j.$$

Since $\sum_{j=1}^{L} p_j = 1$, it follows that

$$\lim_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} T_N(k_1(N), \cdots, k_L(N)) = \lim_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} \frac{N^N}{k_1(N)^{k_1(N)} \cdots k_L(N)^{k_L(N)}},$$

assuming that the limit on the right-hand side exists. Since

$$\begin{aligned} \frac{1}{N} \log_{\mathrm{e}} \frac{N^N}{k_1(N)^{k_1(N)} \cdots k_L(N)^{k_L(N)}} &= \sum_{j=1}^{L} \frac{1}{N} \log_{\mathrm{e}} \frac{N^{k_j(N)}}{k_j(N)^{k_j(N)}} \\ &= \sum_{j=1}^{L} \frac{k_j(N)}{N} \log_{\mathrm{e}} \left[\frac{k_j(N)}{N} \right]^{-1}, \end{aligned}$$

(1.2) gives

$$\lim_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} \frac{N^N}{k_1(N)^{k_1(N)} \cdots k_L(N)^{k_L(N)}} = -\sum_{j=1}^{L} p_j \log_{\mathrm{e}} p_j,$$

and the result follows. □

For $P \in \mathcal{P}(\mathcal{A})$ we set

$$S_B(P) := -\sum_{a\in\mathcal{A}} P(a) \log_{\mathrm{e}} P(a)$$

and call $S_B(P)$ the Boltzmann entropy of P.

We now go a bit further following the same line of thought. Let $P \in \mathcal{P}(\mathcal{A})$ be faithful. Denote by $k_j(\omega)$ the number of times the letter a_j appears in $\omega = \omega_1 \cdots \omega_N$. For small $\epsilon > 0$ set

$$T_N(\epsilon) = \left\{ \omega \in \mathcal{A}^N \mid P(a_j) - \epsilon \le \frac{k_j(\omega)}{N} \le P(a_j) + \epsilon \text{ for } 1 \le j \le L \right\}.$$

Then

Proposition 1.2

$$\lim_{\epsilon\downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log_e |T_N(\epsilon)| = \lim_{\epsilon\downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log_e |T_N(\epsilon)| = S_B(P), \tag{1.4}$$

where $|T_N(\epsilon)|$ denotes the number of the elements of the set $T_N(\epsilon)$.

Proof. Let $k_1(N), \cdots, k_L(N)$ be non-negative integers such that

$$k_1(N) + \cdots + k_L(N) = N \tag{1.5}$$

and

$$\lim_{N\to\infty} \frac{k_j(N)}{N} = P(a_j).$$

Then, for any $\epsilon > 0$ we have that for N large enough,

$$|T_N(\epsilon)| \geq T_N(k_1(N), \cdots, k_L(N)).$$

This gives

$$\lim_{\epsilon\downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log_e |T_N(\epsilon)| \geq \lim_{N\to\infty} \frac{1}{N} \log_e T_N(k_1(N), \cdots, k_L(N)) = S_B(P).$$

We now turn to the upper bound. A sequence $(k_1(N), \cdots, k_L(N))$ of non-negative integers satisfying (1.5) is called (N, ϵ)-admissible if for $1 \leq j \leq L$,

$$P(a_j) - \epsilon \leq \frac{k_j(N)}{N} \leq P(a_j) + \epsilon.$$

Denote by $A(N, \epsilon))$ the set of all N-admissible sequences. Obviously, $|A(N, \epsilon)| \leq N^L$. The proof of Proposition 1.1 gives that

$$\begin{aligned} \limsup_{N\to\infty} \frac{1}{N} \sup_{(k_1(N),\cdots,k_L(N))\in A(N,\epsilon)} \log_e T_N(k_1(N), \cdots, k_L(N)) \\ \leq L\epsilon + \sum_{a\in\mathcal{A}} (P(a) + \epsilon) \log_e \frac{1}{P(a) - \epsilon}. \end{aligned} \tag{1.6}$$

We have

$$\begin{aligned} |T_N(\epsilon)| &= \sum_{(k_1(N),\cdots,k_L(N))\in A(N,\epsilon)} T_N(k_1(N), \cdots, k_L(N)) \\ &\leq N^L \sup_{(k_1(N),\cdots,k_L(N))\in A(N,\epsilon)} T_N(k_1(N), \cdots, k_L(N)). \end{aligned}$$

This inequality and (1.2) give that

$$\lim_{\epsilon\downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log_e |T_N(\epsilon)| \leq S_B(P).$$

□

Exercise 1. Write detailed proof of (1.6).

The conclusion of Proposition 1.2 can be reformulated as follows. The variational metric d_{var} on $\mathcal{P}(\mathcal{A})$ is given by

$$d_{\text{var}}(P, Q) = \sum_{a\in\mathcal{A}} |P(a) - Q(a)|.$$

To a microstate $\omega = \omega_1 \cdots \omega_N$ we associate the empirical probability measure $P_\omega \in \mathcal{P}(\mathcal{A})$ by

$$P_\omega(a_j) = \frac{k_j(\omega)}{N}. \tag{1.7}$$

Then (1.4) is equivalent to

$$\begin{aligned} &\lim_{\epsilon \downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} \left|\{\omega \in \mathcal{A}^N \,|\, d_{\mathrm{var}}(P, P_\omega) \le \epsilon\}\right| \\ &= \lim_{\epsilon \downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} \left|\{\omega \in \mathcal{A}^N \,|\, d_{\mathrm{var}}(P, P_\omega) \le \epsilon\}\right| = S_B(P). \end{aligned} \tag{1.8}$$

Exercise 2. Prove that (1.4) $\Leftrightarrow$ (1.8).

The above discussion leads to the following points that will be all discussed latter in the notes.

(a) $S_B(P_{\mathrm{ch}}) = \log_{\mathrm{e}} |\mathcal{A}|$. For any $P \in \mathcal{P}(\mathcal{A})$,

$$S_B(P) \le S_B(P_{\mathrm{ch}})$$

with equality iff $P = P_{\mathrm{ch}}$. This result follows from the concavity of the logarithm and the Jensen inequality.

(b) Let $\mathcal{A} = \{0, 1\}$ and suppose that 0 corresponds to the configuration where the "gas molecule" is absent, and 1 corresponds to the configuration where the "gas molecule" is present. Let $0 < \rho < 1$. Then $k_1(\omega)$ corresponds to the number of "gas molecules" present in the "microstate" $\omega = \omega_1 \cdots \omega_N$. It follows from Proposition 1.6 that

$$\begin{aligned} &\lim_{\epsilon \downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} \left|\left\{\omega \in \mathcal{A}^N \,|\, \rho - \epsilon \le \frac{k_1(\omega)}{N} \le \rho + \epsilon\right\}\right| \\ &= \lim_{\epsilon \downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} \left|\left\{\omega \in \mathcal{A}^N \,|\, \rho - \epsilon \le \frac{k_1(\omega)}{N} \le \rho + \epsilon\right\}\right| \\ &= -\rho \log_{\mathrm{e}} \rho - (1-\rho) \log_{\mathrm{e}} (1-\rho). \end{aligned} \tag{1.9}$$

The number ρ is the "macrostate" of our "ideal gas" associated to its density per unit volume. The Boltzmann entropy of the "macrostate" ρ is given by (1.9),

$$S_B(\rho) := -\rho \log_{\mathrm{e}} \rho - (1-\rho) \log_{\mathrm{e}} (1-\rho).$$

Obviously, $S_B(\rho) = S_B(P)$ where P is the probability measure on $\mathcal{A} = \{0, 1\}$ given by $P(0) = 1 - \rho$, $P(1) = \rho$.

(c) Returning to the general finite $\mathcal{A}$, let $H : \mathcal{A} \to \mathbb{R}$ be a function.[1] The value $H(a_j) = e_j$ is interpreted as the energy of the configuration a_j. The energy of microstate $\omega = \omega_1 \cdots \omega_N$ is

$$H_N(\omega) := H(\omega_1) + \cdots + H(\omega_N).$$

Note that

$$H_N(\omega) = \sum_{j=1}^{L} e_j k_j(\omega).$$

Set

$$\underline{e} = \min_j e_j, \qquad \overline{e} = \max_j e_j.$$

Obviously, $H_N(\omega)/N \in [\underline{e}, \overline{e}]$. We will prove latter in the notes that Proposition 1.2 gives that for any $e \in [\underline{e}, \overline{e}]$,

$$\begin{aligned} &\lim_{\epsilon \downarrow 0} \limsup_{N \to \infty} \frac{1}{N} \log_{\mathrm{e}} \left| \left\{ \omega \in \mathcal{A}^N \,|\, e - \epsilon \le \frac{H_N(\omega)}{N} \le e + \epsilon \right\} \right| \\ &= \lim_{\epsilon \downarrow 0} \liminf_{N \to \infty} \frac{1}{N} \log_{\mathrm{e}} \left| \left\{ \omega \in \mathcal{A}^N \,|\, \mathrm{e} - \epsilon \le \frac{H_N(\omega)}{N} \le e + \epsilon \right\} \right|. \end{aligned} \tag{1.10}$$

The number e is interpreted as the "macrostate" of our "ideal gas" associated to its "energy" per unit volume, and

$$\left| \left\{ \omega \in \mathcal{A}^N \,|\, e - \epsilon \le \frac{H_N(\omega)}{N} \le e + \epsilon \right\} \right|$$

is the number of "microstates" of the N-molecules ideal gas within ϵ-tolerance corresponding to e. The common value of the limits (1.10) is denoted by $S_B(e)$ and is called the Boltzmann entropy of the the macrostate e. Set

$$\mathcal{P}_{H,e} = \left\{ P \in \mathcal{P}(\mathcal{A}) \,|\, \int_{\mathcal{A}} H \mathrm{d}P = e \right\} \tag{1.11}$$

One further shows that

$$S_B(e) = \sup_{P \in \mathcal{P}_{H,e}} S_B(P), \tag{1.12}$$

and that there exists unique $P_e \in \mathcal{P}_{H,e}$ such that

$$S_B(e) = S(P_e). \tag{1.13}$$

To elaborate further connection with physics, we now describe P_e.

In the boundary cases $e = \overline{e}$ and $e = \underline{e}$, the identification of P_e follows from the fact that

$$S_B(\overline{e}) = -\log_{\mathrm{e}} |\mathcal{A}_{\overline{e}}|. \qquad S_B(\underline{e}) = -\log_{\mathrm{e}} |\mathcal{A}_{\underline{e}}|,$$

[1] To avoid trivialities we will assume that H is not a constant function.

where

$$\mathcal{A}_{\overline{e}} = \{a \,|\, H(a) = \overline{e}\}, \qquad \mathcal{A}_{\underline{e}} = \{a \,|\, H(a) = \underline{e}\}.$$

It follows that

$$P_{\overline{e}}(a) = \begin{cases} 1/|\mathcal{A}_{\overline{e}}| & \text{if } a \in \mathcal{A}_{\overline{e}} \\ 0 & \text{otherwise} \end{cases}, \qquad P_{\underline{e}}(a) = \begin{cases} 1/|\mathcal{A}_{\underline{e}}| & \text{if } a \in \mathcal{A}_{\underline{e}} \\ 0 & \text{otherwise} \end{cases}.$$

We now turn to the case $e = (\underline{e}, \overline{e})$.

For $\beta \in (-\infty, \infty)$ set

$$P_\beta^G(a) = \frac{\mathrm{e}^{-\beta H(a)}}{\sum_b \mathrm{e}^{-\beta H(b)}}.$$

We will refer to P_β^G as the Gibbs canonical ensemble at the inverse temperature β and to

$$\mathsf{P}(\beta) = \log_{\mathrm{e}} \left[\sum_{a \in \mathcal{A}} \mathrm{e}^{-\beta H(a)} \right]$$

as the pressure. The identification of β as the inverse temperature of our ideal gas will follow from the discussion below. Denote for a moment $\langle F \rangle_\beta = \int_{\mathcal{A}} F \mathrm{d}P_\beta^G$. The function $\beta \mapsto \langle \mapsto \langle H \rangle_\beta$ is obviously real analytic and [2]

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \langle H \rangle_\beta = -\langle (H - \langle H \rangle_\beta)^2 \rangle_\beta < 0.$$

It follows that the function $\beta \mapsto \langle H \rangle_\beta$ is strictly decreasing, and we denote by $e \mapsto \beta(e)$ its inverse. Obviously, $\beta(e) \in (-\infty, \infty)$ is uniquely specified by

$$e = \int_{\mathcal{A}} H \mathrm{d}P_{\beta(e)}^G, \tag{1.14}$$

and the map $e \mapsto \beta(e)$ is real-analytic and strictly decreasing. The Gibbs variational principle, see Theorem 2.1 (6) and the discussion after this theorem, gives that

$$P_e = P_{\beta(e)}^G.$$

Note that $\lim_{e \uparrow \overline{e}} P_e = P_{\overline{e}}$, $\lim_{e \downarrow \overline{e}} P_e = P_{\overline{e}}$ The relation

$$S_B(e) = S(P_{\beta(e)}) \tag{1.15}$$

gives that for $e \in (\underline{e}, \overline{e})$,

$$\begin{aligned} S_B(e) &= e\beta(e) + \mathsf{P}(\beta(e)), \\ \frac{\mathrm{d}S_B(e)}{\mathrm{d}e} &= \beta(e). \end{aligned} \tag{1.16}$$

[2]Recall our standing assumption that H is not a constant function.

This leads to the identification of β with the inverse temperature. (1.16) are the fundamental thermodynamical relation between energy, entropy, temperature, and pressure.

The second relation in (1.16) gives that the function $e \mapsto S_B(e)$ is strictly concave. Note that

$$\beta(e_0) = 0 \qquad \Longleftrightarrow \qquad e_0 = \frac{1}{N}\sum_{a\in\mathcal{A}} H(a),$$

and that the function $e \mapsto S_B(e)$ is strictly increasing on $(\underline{e}, e_0]$ and strictly decreasing on $[e_0, \overline{e})$. The second law of thermodynamics postulates that entropy increases with energy and hence selects the energy interval $(\underline{e}, e_0]$ and positive values of β as physically relevant.

1.2 Shannon's entropy

Suppose that $P \in \mathcal{P}(\mathcal{A})$ is faithful. The entropy function of P is the map $S_P : \mathcal{A} \to \mathbb{R}$ defined by

$$S_P(a) = -\log_{\mathrm{e}} P(a).$$

Obviously,

$$\int_{\mathcal{A}} S_P \mathrm{d}P = \sum_{a\in\mathcal{A}} S_P(a)P(a) = S_B(P).$$

We denote by $P_N = P \times \cdots \times P$ the product probability measure on $\mathcal{A}^N$. For a given $\epsilon > 0$ let

$$\begin{aligned} T_N(\epsilon) &= \left\{\omega = \omega_1\cdots\omega_N \in \mathcal{A}^N \mid \left|\frac{S_P(\omega_1) + \cdots S_P(\omega_N)}{N} - S_B(P)\right| < \epsilon\right\} \\ &= \left\{\omega \in \mathcal{A}^N \mid \left|-\frac{\log_{\mathrm{e}} P_N(\omega)}{N} - S_B(P)\right| < \epsilon\right\} \\ &= \left\{\omega \in \mathcal{A}^N \mid \mathrm{e}^{-N(S_B(P)+\epsilon)} < P_N(\omega) < \mathrm{e}^{-N(S_B(P)-\epsilon)}\right\}. \end{aligned}$$

The Law of Large Numbers (LLN) gives

$$\lim_{N\to\infty} P_N(T_N(\epsilon)) = 1.$$

We also have the following obvious bounds on the cardinality of $T_N(\epsilon)$:

$$P_N(T_N(\epsilon))\mathrm{e}^{N(S_B(P)-\epsilon)} < |T_N(\epsilon)| < \mathrm{e}^{N(S_B(P)+\epsilon)} \tag{1.17}$$

Since $|\mathcal{A}^N| = |\mathcal{A}|^N = \mathrm{e}^{N\log_{\mathrm{e}} S_B(P_{\mathrm{ch}})}$, (1.18) can be written as

$$P_N(T_N(\epsilon))\mathrm{e}^{N(S_B(P)-S_B(P_{\mathrm{ch}})-\epsilon)} < \frac{|T_N(\epsilon)|}{|\mathcal{A}|^N} < \mathrm{e}^{N(S_B(P)-S_B(P_{\mathrm{ch}})+\epsilon)}, \tag{1.18}$$

which in particular gives

$$S_B(P) - S_B(P_{\mathrm{ch}}) - \epsilon \le \liminf_{N\to\infty}\frac{1}{N}\log\frac{|T_N(\epsilon)|}{|\mathcal{A}|^N} \le \limsup_{N\to\infty}\frac{1}{N}\log\frac{|T_N(\epsilon)|}{|\mathcal{A}|^N} \le S_B(P) - S_B(P_{\mathrm{ch}}) + \epsilon.$$

It follows that if $P \neq P_{\mathrm{ch}}$, then, as $N \to \infty$, the measure P_N is "concentrated" and "equipartioned" on the set $T_N(\epsilon)$ whose size is "exponentially small" with respect to the size of $\mathcal{A}^N$.

Let $\gamma \in]0,1[$ be fixed. The (N,γ) covering exponent is defined by

$$c_N(\gamma) = \min \left\{ |A| \,|\, A \subset \mathcal{A}^N,\, P_N(A) \geq \gamma \right\}. \tag{1.19}$$

One can find $c_N(\gamma)$ according to the following algorithm:

(a) List the words $\omega = \omega_1 \cdots \omega_N$ in order of decreasing probabilities.

(b) Count the listed words until the first time the total probability is $\geq \gamma$.

Proposition 1.3 *For all $\gamma \in]0,1[$,*

$$\lim_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} c_N(\gamma) = S_B(P).$$

Proof. Fix $\epsilon > 0$ and recall the definition of $T_N(\epsilon)$. For N large enough, $P_N(T_N(\epsilon)) \geq \gamma$, and so for such N's,

$$c_N(\gamma) \leq |T_N(\epsilon)| \leq \mathrm{e}^{N(S_B(P)+\epsilon)}.$$

It follows that

$$\limsup_{N\to\infty} \frac{1}{N} \log c_N(\gamma) \leq S_B(P).$$

To prove the lower bound, let $A_{N,\gamma}$ be a set for which the minimum in (1.19) is achieved. Let $\epsilon > 0$. Note that

$$\liminf_{N\to\infty} P_N(T_N(\epsilon) \cap A_{N,\gamma}) \geq \gamma. \tag{1.20}$$

Since $P_N(\omega) \leq \mathrm{e}^{-N(S_B(P)-\epsilon)}$ for $\omega \in T_N(\epsilon)$,

$$P_N(T_N(\epsilon) \cap A_{N,\gamma}) = \sum_{\omega \in T_N(\epsilon)\cap A_{N,\gamma}} P_N(\omega) \leq \mathrm{e}^{-N(S_B(P)-\epsilon)} |T_N(\epsilon) \cap A_{N,\gamma}|.$$

Hence,

$$|A_{N,\gamma}| \geq \mathrm{e}^{N(S_B(P)-\epsilon)} P_N(T_N(\epsilon) \cap A_{N,\gamma}),$$

and it follows from (1.20) that

$$\liminf_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} c_N(\gamma) \geq S_B(P) - \epsilon.$$

Since $\epsilon > 0$ is arbitrary,

$$\liminf_{N\to\infty} \frac{1}{N} \log_{\mathrm{e}} c_N(\gamma) \geq S_B(P),$$

and the proposition is proven. $\square$

We now turn to the result known as the Shannon's source coding theorem. Given a pair of positive integers N, M, the *encoder* is a map

$$F_N : \mathcal{A}^N \to \{0,1\}^M.$$

The *decoder* is a map

$$G_N : \{0,1\}^M \to \mathcal{A}^N.$$

The error probability of the coding pair (F_N, G_N) is

$$P_N \left\{ G_N \circ F_N(\omega) \neq \omega \right\}.$$

If this probability is less than some prescribed $1 > \epsilon > 0$, we shall say that the coding pair is ϵ-good. Note that to any ϵ-good coding pair one can associate the set

$$A = \{\omega \,|\, G_N \circ F_N(\omega) = \omega\}$$

which satisfies

$$P_N(A) \geq 1 - \epsilon, \qquad |A| \leq 2^M. \tag{1.21}$$

On the other hand, if $A \subset \mathcal{A}^N$ satisfies (1.21), we can associate to it an ϵ-good pair (F_N, G_N) by setting F_N to be one-one on A (and arbitrary otherwise), and $G_N = F_N^{-1}$ on $F_N(A)$ (and arbitrary otherwise).

In the source coding we wish to find M that minimizes the compression coefficients M/N subject to an allowed ϵ-error probability. Clearly, the optimal M is

$$M_N = \lfloor \log_2 \min \left\{ |A| \,|\, A \subset \mathcal{A}^N, \, P_N(A) \geq 1 - \epsilon \right\} \rfloor.$$

Shannon's source coding theorem now follows from Proposition 1.3: the limiting optimal compression coefficient is

$$\lim_{N \to \infty} \frac{M_N}{N} = \frac{1}{\log 2} S_B(P) = -\sum_{a \in \mathcal{A}} P(a) \log_2 P(a).$$

The Shannon entropy of $P \in \mathcal{P}(\mathcal{A})$ is defined by

$$S_{\mathrm{Sh}}(P) = -\sum_{a \in \mathcal{A}} P(a) \log_2 P(a).$$

1.3 Notes

2 Entropies on finite sets

2.1 Notation

We continue with finite alphabet $\mathcal{A}$. We equip $\mathcal{P}(\mathcal{A})$ with variational metric d_{var}. The support of $P \in \mathcal{P}(\mathcal{A})$ is the set $\mathrm{supp}P = \{a : P(a) > 0\}$. P is called pure if for some $\mathrm{supp}P = \{a\}$ for some a, that is, if $P(a) = 1$ for some a. P is absolutely continuous with respect to Q, denoted $P \ll Q$, if $\mathrm{supp}P \subset \mathrm{supp}Q$, or equivalently, if $Q(a) = 0 \Rightarrow P(a) = 0$.

If $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ and $P \in \mathcal{P}(\mathcal{A}_1 \times \mathcal{A}_2)$, the marginals of P are

$$P_1(a) = \sum_{b \in \mathcal{A}_2} P(a,b), \qquad P_2(b) = \sum_{a \in \mathcal{A}_1} P(a,b).$$

Obviously, $P_1 \in \mathcal{P}(\mathcal{A}_1)$, $P_2 \in \mathcal{P}(\mathcal{A}_2)$.

Continuing with the product case $\mathcal{A}_1 \times \mathcal{A}_2$, a $|\mathcal{A}_1| \times |\mathcal{A}_2|$ matrix $[M(a,b)]_{a\mathcal{A}_1, b\in\mathcal{A}_2}$ with non-negative entries is called stochastic if for all $a \in \mathcal{A}_1$,

$$\sum_{b \in \mathcal{A}_2} M(a,b) = 1.$$

A stochastic matrix induces a map[3] $M : \mathcal{P}(\mathcal{A}_1) \to \mathcal{P}(\mathcal{A}_2)$ by

$$(MP)(b) = \sum_{a \in \mathcal{A}_1} P(a)M(a,b).$$

In the sequel $\log$ denotes the logarithm function with an unspecified but fixed base $b > 1$. In information theory the common choice is $b = 2$. In statistical mechanics one takes $b = \mathrm{e}$.

$\mathcal{P}_n$ denotes the set of all probability vectors $(p_1, \cdots, p_n)$.[4]

2.2 Entropies

The Boltzmann-Gibbs-Shannon (BGS) entropy of $P \in \mathcal{P}(\mathcal{A})$ is

$$S(P) = -\sum_{a \in \mathcal{A}} P(a) \log P(a).$$

In what follows we will often simply refer to $S(P)$ as the entropy of P.

The cross entropy of a pair $(P, Q) \in \mathcal{P}(\mathcal{A}) \times \mathcal{P}(\mathcal{A})$ is

$$S_{\text{cross}}(P|Q) = -\sum_{a \in \mathcal{A}} P(a) \log Q(a).$$

The cross entropy is finite if $P \ll Q$, otherwise it takes value ∞.

The relative entropy of a pair (P, Q) is

$$\begin{aligned} S(P|Q) &= S_{\text{cross}}(P|Q) - S(P) \\ &= \sum_{a \in \mathcal{A}} P(a)(\log P(a) - \log Q(a)). \end{aligned}$$

[3]Called a stochastic transformation.
[4]$p_k \geq 0$, $\sum p_k = 1$.

The α-Renyi entropy of P, $\alpha \in \mathbb{R}$, is

$$S_\alpha(P) = \log \left(\sum_{a \in \mathrm{supp} P} P(a)^\alpha \right).$$

The α-relative Renyi entropy, $\alpha \in \mathbb{R}$, of a pair (P, Q) with $\mathrm{supp} P = \mathrm{supp} Q$ is

$$S_\alpha(P|Q) = \log \left(\sum_{a \in \mathrm{supp} P} P(a)^\alpha Q(a)^{1-\alpha} \right).$$

The above definitions of Renyi's entropies are somewhat uncommon due to missing normalizations and the fact that they are defined for all $\alpha \in \mathbb{R}$. We will comment in more details on those points latter in the notes.

The basic relations between the entropies are

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} S_\alpha(P)\big|_{\alpha=1} = -S(P),$$

$S_\alpha(P|Q) = S_{1-\alpha}(Q|P)$,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} S_\alpha(P|Q)\big|_{\alpha=0} = -S(Q|P), \qquad \frac{\mathrm{d}}{\mathrm{d}\alpha} S_\alpha(P|Q)\big|_{\alpha=1} = S(P|Q),$$

$$S(P|P_{\mathrm{ch}}) = \log |\mathcal{A}| - S(P), \qquad S_\alpha(P|P_{\mathrm{ch}}) = S_\alpha(P) - (1-\alpha) \log |\mathcal{A}|.$$

Obviously, $S_0(P) = \log |\mathrm{supp} P|$, $S_1(P) = 0$, and $S_0(P|Q) = S_1(P|Q) = 0$. $S_0(P)$ is sometimes called the Hartley entropy and is denoted by $S_H(P)$.

2.3 Proprerties of BGS entropy

Theorem 2.1 (1) *$S(P) \geq 0$ and $S(P) = 0$ iff P is pure.*

(2) *$S(P) \leq \log |\mathcal{A}|$ and $S(P) = \log |\mathcal{A}|$ iff $P = P_{\mathrm{ch}}$.*

(3) *The map $\mathcal{P}(\Omega) \ni P \mapsto S(P)$ is continuous and strictly concave.*

(4) *The entropy map is "almost convex" in the following sense: For any probability vector $(p_1, \cdots, p_n)$ with $p_k > 0$,*

$$S(p_1 P_1 + \cdots + p_n P_n) \leq p_1 S(P_1) + \cdots + p_n S(P_n) + S(p_1, \cdots, p_n),$$

with equality iff $\mathrm{supp}\, P_k \cap \mathrm{supp}\, P_j = \emptyset$ for $k \neq j$.

(5) *The entropy is strictly subadditive: if $P \in \mathcal{P}(\mathcal{A}_1 \times \mathcal{A}_2)$, then*

$$S(P) \le S(P_1) + S(P_2)$$

with equality iff $P = P_1 \times P_2$.

(6)

$$S(P) = \inf_{X:\mathcal{A}\to\mathbb{R}} \left(\log\left(\sum_{a\in\mathcal{A}} \mathrm{e}^{X(a)} \right) - \int_{\mathcal{A}} X \mathrm{d}P \right).$$

The infimum is achieved if P is faithful and $X(a) = -\log P(a) + \mathrm{const}$.

(7) *For any $X : \mathcal{A} \to \mathbb{R}$,*

$$\log\left(\sum_{a\in\mathcal{A}} \mathrm{e}^{X(a)} \right) = \max_{P\in\mathcal{P}(\mathcal{A})} \left(\int_{\mathcal{A}} X \mathrm{d}P + S(P) \right).$$

The maximizer is unique and is given by

$$P(a) = \frac{\mathrm{e}^{X(a)}}{\sum_{b\in\mathcal{A}} \mathrm{e}^{X(b)}}. \tag{2.1}$$

Parts (6) and (7) are known as the Gibbs variational principle. Going back to (1.11)-(1.12), Part (3) gives that there exists unique $P_e \in \mathcal{P}_{H,e}$ such that

$$\sup_{P\in\mathcal{P}_{H,e}} S_B(P_e).$$

That $P_e = P^G_{\beta(e)}$ and another uniqueness argument follow from the Gibbs variational principle (7): for any $P \in \mathcal{P}_{H,e}$,

$$\begin{aligned} S_B(P) &= S_P(P) - \beta(e) \int_{\mathcal{A}} H \mathrm{d}P + \beta(e)e \le \max_{Q\in\mathcal{P}(\mathcal{A})} \left(S_P(P) - \beta(e) \int_{\mathcal{A}} H \mathrm{d}P \right) + \beta(e)e \\ &= \mathsf{P}(\beta(e)) + \beta(e)e = S_B(P^G_{\beta(e)}) \end{aligned}$$

with equality iff $P = P^G_{\beta(e)}$.

It is a fundamental fact that either "almost convexity" or strict subadditivity uniquely characterize Boltzmann-Gibbs-Shannon entropy up to a choice of the base of logarithm. We proceed to describe this aspect of entropy.

Set $\mathcal{P} = \cup_{\mathcal{A}} \mathcal{P}(\mathcal{A})$ and consider functions $\mathfrak{S} : \mathcal{P} \to \mathbb{R}$ that satisfy properties that correspond intuitively to those of *entropy* as a measure of *randomness* of probability measures. We wish to show that those intuitive natural demands uniquely specify $\mathfrak{S}$ up to a choice of units (base of logarithm) and that and that for some choice of this base and for all $P \in \mathcal{P}$, $\mathfrak{S}(P) = S(P)$.

We describe first three basic properties that any candidate for $\mathfrak{S}$ should satisfy. The first is the positivity and non-triviality requirement: $\mathfrak{S}(P) \ge 0$ and this inequality is strict for at least one $P \in \mathcal{P}$. The second is that if $|\mathcal{A}_1| = |\mathcal{A}_2|$ and $\theta : \mathcal{A}_1 \to \mathcal{A}_2$ is a bijection, then for any $P \in \mathcal{P}(\mathcal{A}_1)$, $\mathfrak{S}(P) = \mathfrak{S}(P \circ \theta)$. In other words, the entropy of P should not depend on the labeling of the letters.

In the rest of this section we assume that the above three properties hold.

If $\mathcal{A}_1, \mathcal{A}_2$ are two disjoint sets, we denote by $\mathcal{A}_1 \oplus \mathcal{A}_2$ their union (the symbol $\oplus$ is used to emphasize the fact that the sets are disjoint). If $\mu_j : \mathcal{A}_j \to \mathbb{R}$, $j = 1, 2$, then $\mu := \mu_1 \oplus \mu_2 : \mathcal{A}_1 \oplus \mathcal{A}_1 \to \mathbb{R}$ is defined by $\mu(a) = \mu_1(a)$ if $a \in \mathcal{A}_1$ and $\mu(b) = \mu_2(b)$ if $b \in \mathcal{A}_2$.

The axiomatic characterization of entropy based on Theorem 2.1 (4) is:

Theorem 2.2 *Let $\mathfrak{S} : \mathcal{P} \to [0, \infty[$ be a function such that:*

(a) *$\mathfrak{S}$ is continuous on $\mathcal{P}_2$.*

(b) *For any finite collection of disjoint sets $\mathcal{A}_j$, $j = 1, \cdots, n$,*

$$\mathfrak{S}\left(\bigoplus_{k=1}^{n} p_k P_k\right) = \sum_{k=1}^{n} p_k \mathfrak{S}(P_k) + \mathfrak{S}(p_1, \cdots, p_n). \tag{2.2}$$

where $P_j \in \mathcal{A}_j$ and $(p_1, \cdots, p_n) \in \mathcal{P}_n$.

Then for some base of the logarithm and all $P \in \mathcal{P}$,

$$\mathfrak{S}(P) = S(P). \tag{2.3}$$

Remark 2.1 *If the positivity is dropped, then the proof gives for some base of the logarithm either $\mathfrak{S}(P) = S(P)$ for all P, or $\mathfrak{S}(P) = -S(P)$ for all P.*

Remark 2.2 *The property (2.2) is sometimes called the chain rule for entropy. It can be verbalized as follows: if the initial choices $(1, \cdots, n)$, realized with probabilities $(p_1, \cdots, p_n)$, are split into sub-choices described by probability spaces $(\mathcal{A}_k, P_k)$, $k = 1, \cdots, n$, then the new entropy is the sum of the initial entropy and the entropies of sub-choices weighted by their probabilities.*

The axiomatic characterization of the entropy based on strict subadditivity is:

Theorem 2.3 *Let $\mathfrak{S} : \mathcal{P} \to \mathbb{R}$ be a strictly sub-additive map, namely if $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ and $P \in \mathcal{P}(\mathcal{A}_1 \times \mathcal{A}_2)$, then*

$$\mathfrak{S}(P) \le \mathfrak{S}(P_1) + \mathfrak{S}(P_2)$$

with equality iff $P = P_1 \otimes P_2$. Then for some base of the logarithm and all $P \in \mathcal{P}$,

$$\mathfrak{S}(P) = S(P) \tag{2.4}$$

Remark 2.3 *The strict subadditivity assumption ensures that $\mathfrak{S}$ is positive and non-trivial.*

2.4 Properties of relative entropy

Theorem 2.4 (1) *$S(P|Q) \geq 0$ with equality iff $P = Q$.*

(2)

$$S(P|Q) \geq \frac{1}{2} d_{\mathrm{var}}(P, Q)^2$$

with the equality iff $P = Q$.

(3) *The map*

$$(P, Q) \mapsto S(P|Q)$$

is lower-semicontinuous. This restriction of this map to the convex set $\{(P, Q) \,|\, P \ll Q\}$ is continuous.

(4) *The relative entropy is jointly convex: for $\lambda \in]0, 1[$ and $P_1, P_2, Q_1, Q_2 \in \mathcal{P}(\mathcal{A})$,*

$$S(\lambda P_1 + (1-\lambda)P_2 | \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda S(P_1|Q_1) + (1-\lambda)S(P_2|Q_2). \tag{2.5}$$

(5) *Part (4) has the following generalization. Let $P_1, \cdots, P_n, Q_1, \cdots, Q_n \in \mathcal{P}(\Omega)$ and $p = (p_1, \cdots, p_n), q = (q_1, \cdots, q_n) \in \mathcal{P}_n$. Then*

$$S(p_1 P_1 + \cdots + p_n P_n | q_1 Q_1 + \cdots + q_n Q_n) \leq p_1 S(P_1|Q_1) + \cdots + p_n S(P_n|Q_n) + S(p|q). \tag{2.6}$$

If the r.h.s. in (2.6) is finite, then the equality holds iff for all j, k such that $q_j > 0, q_k > 0$,

$$\frac{p_j P_j(\omega)}{q_j Q_j(\omega)} = \frac{p_k P_k(\omega)}{q_k Q_k(\omega)}$$

holds for all $\omega \in \mathrm{supp}\, Q_k \cap \mathrm{supp}\, Q_j$.

(6) *Relative entropy is stochastically monotone, that is, for any stochastic transformation $M : \mathcal{P}(\mathcal{A}_1) \to \mathcal{P}(\mathcal{A}_2)$,*

$$S(M(P)|M(Q) \leq S(P|Q).$$

(7)

$$S(P|Q) = \sup_{X:\mathcal{A}\to\mathbb{R}} \left(\int_{\mathcal{A}} X \mathrm{d}P - \log \int_{\mathrm{supp}P} \mathrm{e}^X \mathrm{d}Q \right). \tag{2.7}$$

If $S(P|Q) < \infty$, then the supremum is achieved, and each maximizer is equal to $\log \frac{P(a)}{Q(a} + \mathrm{const}$ for $a \in \mathrm{supp}P$ and is arbitrary otherwise

(8) *For $X : \mathcal{A} \to \mathbb{R}$ and $Q \in \mathcal{P}(\mathcal{A})$,*

$$\log \int_{\mathcal{A}} \mathrm{e}^X \mathrm{d}Q = \max_{P \in \mathcal{P}(\mathcal{A})} \left(\int_{\mathcal{A}} X \mathrm{d}P - S(P|Q) \right).$$

The maximizer is unique and is given by

$$P_{X,Q}(a) = \frac{\mathrm{e}^{X(a)} Q(a)}{\sum_{b \in \mathcal{A}} \mathrm{e}^{X(b)} Q(b)}.$$

The Gibbs variational principle part of Theorem 2.1 follows from (7) and (8) by setting $Q = P_{\rm ch}$. It is important to note that Theorem 2.1 (7) follows from the most basic property of relative entropy stated in (1) above. Indeed, denoting by $P_{\max}$ the probability measure (2.1), we have

$$S(P|P_{\max}) = -S(P) - \int_{\mathcal{A}} X \mathrm{d}P + \log\left(\sum_a \mathrm{e}^{X(a)}\right) \geq 0,$$

with equality iff $P = P_{\max}$.

We now turn to the *Boltzmann-Sanov Large Deviation Principle* that generalizes and sheds light on the results of Section 1.1. Recall the definition (1.7) of empirical probability measures. We fix faithful $P \in \mathcal{P}(\mathcal{A})$. By the LLN, for any $\epsilon > 0$,

$$\lim_{N\to\infty} P_N \left\{\omega \in \mathcal{A}^N \,|\, d_{\rm var}(P_\omega, P) \geq \epsilon\right\} = 0. \tag{2.8}$$

The Boltzmann-Sanov theorem is a deep refinement of the limit (2.8).

Theorem 2.5 *For any* $\Gamma \subset \mathcal{P}(\Omega)$*,*

$$-\inf_{Q\in{\rm int}(\Gamma)} S(Q|P) \liminf_{N\to\infty} \frac{1}{N} \log P_N \left\{\omega \in \mathcal{A}^N \,|\, P_\omega \in \Gamma\right\}$$

$$\limsup_{N\to\infty} \frac{1}{N} \log P_N \left\{\omega \in \mathcal{A}^N \,|\, P_\omega \in \Gamma\right\} \leq -\inf_{Q\in{\rm cl}(\Gamma)} S(Q|P),$$

where ${\rm int}/{\rm cl}$ *stands for the interior/closure.*

Remark 2.4 *If* Γ *is an open subset of* $\mathcal{P}(\mathcal{A})$ *or a convex subset with non-empty interior, then*

$$\inf_{Q\in{\rm int}(\Gamma)} S(Q|P) = \inf_{Q\in{\rm cl}(\Gamma)} S(Q|P).$$

In this case

$$\lim_{N\to\infty} \frac{1}{N} \log P_N \left\{\omega \in \mathcal{A}^N \,|\, P_\omega \in \Gamma\right\} = -\inf_{Q\in\Gamma} S(Q|P).$$

Remark 2.5 *Taking* $P = P_{\rm ch}$ *and* $\Gamma = \{Q : d_{\rm var}(P_{\rm ch}, Q) \leq \epsilon\}$*, Theorem 2.5 reduces to Proposition 1.4. Moreover, by the previous remark,* $\limsup$ *and* $\liminf$ *in Proposition 1.4 can be replaced with* $\lim$.

Boltzmann-Sanov theorem has many important consequences, one of which is Cramer's theorem[5] We proceed to describe Cramer's theorem and contraction principle that allows to deduce it from the Boltzmann-Sanov theorem.

Let $X : \mathcal{A} \to \mathbb{R}$[6] and let $m = \min_a X(a)$, $M = \max_a X(s)$. We assume $m < M$. The cumulant generating function of X is

$$C(\alpha) = \log \int_{\mathcal{A}} \mathrm{e}^{\alpha X} \mathrm{d}P, \qquad \alpha \in \mathbb{R}.$$

[5]Of course, Cramer's theorem can be also proven by independent means.

[6]To avoid trivialities we assume that X is not a constant function to avoid trivialities.

The so called rate function of the random variable X is the Legendre transform of C,

$$I(\theta) = \sup(\alpha\theta - C(\alpha)).$$

The function I is real-analytic on (m, M)

$$I(m) = \log|\{a : X(a) = m\}|, \qquad I(M) = \log|\{a : X(a) = M\}|,$$

and $I(\theta) = \infty$ for $\theta \notin [m, M]$. The function I is strictly convex on $[m, M]$ and $I(\theta) = 0$ iff $\theta = \int_{\mathcal{A}} X \mathrm{d}P$.

For $\omega = \omega_1 \cdots \omega_N \in \mathcal{A}^N$ we set

$$\mathcal{S}_N(\omega) = \sum_{k=1}^{N} X(\omega_k).$$

Note that

$$\frac{\mathcal{S}_N(\omega)}{N} = \int_{\mathcal{A}} X \mathrm{d}P_\omega.$$

For any $S \subset \mathbb{R}$,

$$\frac{\mathcal{S}_N(\omega)}{N} \in S \;\Leftrightarrow\; P_\omega \in \Gamma_S,$$

where

$$\Gamma_S = \left\{ Q \in \mathcal{P}(\mathcal{A}) \,\middle|\, \int_{\mathcal{A}} X \mathrm{d}Q \in S \right\}.$$

One has

$$\mathrm{int}(\Gamma_S) = \Gamma_{\mathrm{int}(S)}, \qquad \mathrm{cl}\,(\Gamma_S) = \Gamma_{\mathrm{cl}(S)}.$$

We now have:

Theorem 2.6 *Let $S \subset \mathbb{R}$,*

(1)

$$-\inf_{Q \in \Gamma_{\mathrm{int}(S)}} S(Q|P) \le \liminf_{N\to\infty} \frac{1}{N} \log P_N \left\{ \omega \in \mathcal{A}^N \,\middle|\, \frac{\mathcal{S}_N(\omega)}{N} \in S \right\}$$
$$\le \limsup_{N\to\infty} \frac{1}{N} \log P_N \left\{ \omega \in \mathcal{A}^N \,\middle|\, \frac{\mathcal{S}_N(\omega)}{N} \in S \right\} \le -\inf_{Q \in \Gamma_{\mathrm{cl}(S)}} S(Q|P).$$

(2)

$$\inf_{\theta \in S} I(\theta) = \inf_{Q \in \Gamma_S} S(Q|P)$$

(3)

$$-\inf_{\theta \in \mathrm{int}(S)} I(\theta) \le \liminf_{N\to\infty} \frac{1}{N} \log P_N \left\{ \omega \in \mathcal{A}^N \,\middle|\, \frac{\mathcal{S}_N(\omega)}{N} \in S \right\}$$
$$\le \limsup_{N\to\infty} \frac{1}{N} \log P_N \left\{ \omega \in \mathcal{A}^N \,\middle|\, \frac{\mathcal{S}_N(\omega)}{N} \in S \right\} \le -\inf_{\theta \in \mathrm{cl}(S)} I(\theta).$$

Remark 2.6 *Part* (1) *follows from the Boltzmann-Sanov theorem. Part* (2) *is the contraction principle. Part* (3) *follows from* (1) $\wedge$ (2).

We will not discuss in these lecture notes the axiomatizations of relative entropy; see [LONE19, Chapter 5].

2.5 Back to Boltzmann entropy

Going back to the discussion of Boltzmann's entropy in Section 1.1 **(c)** and taking $P = \mathrm{P}_{\mathrm{ch}}$, one derives from Part (3) that (1.10) holds with

$$S_B(e) = \log_{\mathrm{e}} |\mathcal{A}| - I(e).$$

Part (2) with $S = \{e\}$ gives (1.12).

2.6 Back to covering exponents

2.7 Prefix-free coding

To set the stage, we denote by $\{0,1\}^*$ the set of all finite length words from the alphabet $\{0,1\}$. If $w = w_1 \cdots w_n$ and $u = u_1 \cdots u_m$ are two words in $\{0,1\}^*$. their concatenation is the word

$$wu = w_1 \cdots w_n u_1 \cdots u_m.$$

The length of a word is defined in the obvious sense, if $w = w_1 \cdots w_n$, then $\ell(w) = n$. Let F be a finite set. Latter in this section we will take $F = \mathcal{A}^N$, but at the moment it is natural to keep F general. A (binary) code on F is a map

$$C : F \to \{0,1\}^*.$$

$C(a) \in \{0,1\}^*$ is call the codeword of $a \in F$ and the image set $C(F)$ is called the codebook. If $P \in \mathcal{P}(F)$ is the source statistics, the expected length of the code C wrt P is

$$\langle C \rangle_P = \sum_{a \in F} \ell(C(a)) P(a).$$

Given P and reasonable regularity assumption on C, the goal is to minimize the expected code length $\langle C \rangle_P$. To each code C we associate the Kraft-MacMillan pressure

$$\mathsf{P}_{\mathrm{KM}} = \log_2 \left(\sum_{a \in F} 2^{-\ell(C(a))} \right),$$

and the Kraft-McMillan probability distribution

$$P_{\mathrm{KM}}(a) = \frac{2^{-\ell(C(a)}}{\sum_{a \in F} 2^{-\ell(C(a))}}.$$

We will be only interested in the codes that are one-one; such codes are called faithful. In fact we will consider stringer requirement then faithfulness, namely we will focus on a special class of faithful codes that are called prefix-free codes.We shall see latter that regarding relevant asymptotic properties of codes, there is no difference between faithful and prefix-free codes. However, in finite setting prefix-free codes have some special properties that simplify their analysis. They are also of considerable practical importance and we will comment further on this point in Section 2.14.

A word $u \in \{0,1\}^*$ is a prefix of a word $w \in \{0,1\}^*$ if $w = uv$ for some $u \in \{0,1\}^*$. A set $W \in \{0,1\}^*$ is called prefix-free if no element of W is a prefix of some other element of W. From perspective of coding, the fundamental property of prefix-free sets is:

Proposition 2.7 *Let $W = \{w_1, \cdots, w_n\}$ be a prefix-free subset of $\{0,1\}^*$. Then*

$$\sum_{j=1}^{n} 2^{-\ell(w_j)} \le 1. \tag{2.9}$$

We will refer to (2.9) as the Kraft-McMillan inequality.

Proof. The result can be proven in a several different ways, each proof shedding a different light on the inequality (2.9). We present one proof, two others are outlined in exercises.

Let W_N be set of concatenations $w_{i_1} \cdots w_{i_N}$ of the N words chosen from the set W. The prefix-free assumption gives that

$$w_{i_1} \cdots w_{i_N} = w_{j_1} \cdots w_{j_N} \;\Rightarrow\; w_{i_k} = w_{j_k} \text{ for all } k. \tag{2.10}$$

It follows that

$$\left(\sum_{j=1}^{n} 2^{-\ell(w_j)} \right)^N = \sum_{u \in W_N} 2^{-\ell(u)}.$$

Now, if $l_{\max} = \max_j l_j$, and, for $1 \le m \le N l_{\max}$,

$$a(m) = |\{u \in W_N \,|\, \ell(u) = m\}|,$$

we have

$$\sum_{u \in W_N} \mathrm{e}^{-\ell(u)} = \sum_{m=1}^{N l_{max}} a(m) 2^{-m}.$$

Obviously, $a(m) \le 2^m$ and so $\sum_{u \in W_N} \mathrm{e}^{-\ell(u)} \le N l_{\max}$. This gives that for all $N \ge 1$,

$$\sum_{j=1}^{n} 2^{-\ell(w_j)} \le (N l_{\max})^{1/N}, \tag{2.11}$$

The Kraft-McMillan inequality follows by taking $N \to \infty$ in the inequality (2.11). □

Proposition (2.7) has a converse.

Proposition 2.8 *If $1 \le l_1 \le \cdots \le l_n$ is a sequence of integers satisfying*

$$\sum_{j=1}^{n} 2^{-l_j} \le 1,$$

then there exists a prefix-free set $W = \{w_1, \cdots, w_n\}$ in $\{0,1\}^*$ such that $\ell(w_j) = l_j$.

Proof. Again, one can argue in a several different ways. A perhaps shortest proof goes as follows. We start with w_1 which is the the word of l_1 zeros,

$$w_1 = \underbrace{0 \cdots 0}_{\text{length } l_1}.$$

After that, take for w_j the first l_j digits of $\sum_{k=1}^{j-1} 2^{-l_k}$ written in the binary form.[7] A moment's thought leads to the conclusion that the set $\{w_1, \cdots, w_n\}$ is prefix-free.□

A faithful code C is called prefix-free if its codebook $C(F)$ is a prefix free subset of $\{0,1\}^*$. Proposition 2.7 and 2.8 then yield the following. First, the Kraft-McMillan pressure of C satisfies $\mathsf{P}_{\mathrm{KM}} \le 0$. Seecond, to each faithful $P \in \mathcal{P}(F)$ we can associate a prefix-free code in the following way. Set

$$l(a) = -\lceil \log_2 P(a) \rceil.$$

Then $l(a) \ge 1$ and

$$\sum_{a \in F} 2^{-l(a)} \le \sum_{a \in F} P(a) = 1,$$

and so that there exists prefix-free code C such that

$$\ell(C(a)) = -\lceil \log_2 P(a)) \rceil.$$

A code with these properties is not unique and any such code is called a *Shannon's code* for P. Note that for any Shannon code C,

$$\langle C \rangle_P \le S_{\mathrm{Sh}}(P) + 1.$$

A basic result of the prefix-free coding is:

Theorem 2.9 *Let $P \in \mathcal{P}(F)$.*

(1) *The expected length of any prefix-free code $C : F \mapsto \{0,1\}^*$ satisfies*

$$\langle C \rangle_P \ge S_{\mathrm{Sh}}(P).$$

(2) *There exists a prefix-free code $C : F \mapsto \{0,1\}^*$ such that*

$$\langle C \rangle_P \le S_{\mathrm{Sh}}(P) + 2.$$

If P is faithful, 2 in the above inequality can be replaced with 1.

[7] For example, $\frac{1}{2} + \frac{1}{8} = 101000 \cdots$.

Proof. (1) Let $C : F \mapsto \{0,1\}^*$ be a prefix-free code. Let $P_{\rm Kraft} \in \mathcal{P}(F)$ defined by

$$P_{\rm Kraft}(a) = \frac{2^{-\ell(C(a))}}{\sum_{b\in F} 2^{-\ell(C(b))}}.$$

Taking logarithm in the base 2,

$$S(P|P_{\rm Kraft}) = \sum_{a\in F} \ell(C(a))P(a) - S_{\rm sh}(P) - \log_2\left(\sum_{a\in F} 2^{-\ell(C(a))}\right) \geq 0,$$

which gives

$$\sum_{a\in F} \ell(C(a))P(a) \geq S_{\rm Sh}(P) + \log_2\left(\sum_{a\in F} 2^{-\ell(C(a))}\right) \geq S_{\rm Sh}(P),$$

where we have used the Kraft inequality.

(2) If P is faithful, one can take for C any Shannon code for P. If P is not faithful, for $a \in {\rm supp}P$ set

$$\widehat{C}(a) = -\lceil P(a) \rceil$$

and extend $\widehat{C}$ to F by setting it to be any prefix-free free code on $F \setminus {\rm supp}P$. The code $\widehat{C}$ may not be prefix-free or faithful. Let now $C : F \to \{0,1\}^*$ be defined by

$$C(a) = \begin{cases} 0\widehat{C}(a) & \text{if } a \in {\rm supp}P \\ 1\widehat{C}(a) & \text{if } a \notin {\rm supp}P. \end{cases}$$

Then C is a prefix-free code and

$$\langle C \rangle_P = \sum_{a\in {\rm supp}P} (-\lceil \log_2 P(a) \rceil + 1)P(a) \leq S_{\rm Sh}(P) + 2.$$

□

To re-iterate, the proof of Theorem 2.9 stems from the identity

$$\langle C \rangle_P - S_{\rm Sh}(P) = S(P|P_{\rm KM}) - \mathsf{P}_{KM}. \tag{2.12}$$

The inequality $\langle C \rangle_P - S_{\rm Sh}(P) \geq 0$ then follows from the sign of the relative entropy and the Kraft-McMillan inequality. The identity (2.12) gives much more and indicates the mechanism that leads to saturation of the Shannon bound $\langle C \rangle_P \geq S_{\rm Sh}(P)$ in the asymptotic setting, to which we turn now,

For each $N \geq 1$ let $C_N : \mathcal{A}^N \to \{0,1\}^*$ be a prefix-free code. Its Kraft-McMillan pressure and probability distribution are

$$\mathsf{P}_{N,{\rm KM}} = \log_2\left(\sum_{\omega\in\mathcal{A}^N} 2^{-\ell(C_N(\omega))}\right),$$

$$P_{N,{\rm KM}}(\omega) = \frac{2^{-\ell(C(\omega)}}{\sum_{\omega\in\mathcal{A}} 2^{-\ell(C_N(\omega)))}}.$$

The equality (2.12) turns to

$$\langle C_N \rangle_{P_N} - N S_{\mathrm{Sh}}(P) = S(P_N | P_{N,\mathrm{KM}}) - \mathsf{P}_{N,\mathrm{KM}}. \tag{2.13}$$

This gives the asymptotic result

$$\liminf_{N\to\infty} \frac{1}{N}\langle C_N \rangle_{P_N} \geq S_{\mathrm{Sh}}(P)$$

and that

$$\lim_{N\to\infty} \frac{1}{N}\langle C_N \rangle_{P_N} = S_{\mathrm{Sh}}(P) \tag{2.14}$$

iff

$$\lim_{N\to\infty} \frac{1}{N} S(P_N | P_{N,\mathrm{KM}}) = 0 \qquad \text{and} \qquad \lim_{N\to\infty} \frac{1}{N} \mathsf{P}_{N,\mathrm{KM}} = 0. \tag{2.15}$$

A code sequence $(C_N)_{N\geq 1}$ is called Shannon-optimal if (2.14) holds. An example is a Shannon sequence where each C_N is a Shannon code. The relations (2.14) give characterization of Shannon-optimality to which we will return latter in the notes.

2.8 Lempel-Ziv parsing/coding and entropy

A parsing of $\omega = \omega_1 \cdots \omega_N \in \mathcal{A}^N$ is an ordered set of words

$$\{w_1(\omega), \cdots, w_k(\omega)\}$$

such that

$$\omega = w_1(\omega) \cdots w_k(\omega). \tag{2.16}$$

We denote by $\mathcal{F}(\omega)$ the number of parsing words. If $w_i(\omega) \neq w_j(\omega)$ for $i \neq j$, we say that (2.16) is parsing into distinct words (abbreviated PDW). When the meaning is clear within the context, we write w_j for $w_j(\omega)$.

Proposition 2.10 *There exists a sequence $(\epsilon_N)_{N\geq 1}$ in $(0,1)$ with $\lim_{N\to\infty} \epsilon_N = 0$, such that for any $N \geq 1$, $\omega \in \mathcal{A}^N$, and any PDW of ω,*

$$\mathcal{F}(\omega) \leq \frac{1}{1-\epsilon_N} \frac{N}{\log_{\mathrm{e}} N} \log_{\mathrm{e}} |\mathcal{A}|.$$

There are several different version of Lempel-Ziv (LZ) parsing. We will deal only with the perhaps best known one in which the next word is the shortest new word. More precisely, for $\omega = \omega_1 \cdots \omega_N$, $w_1 = \omega_1$, and if $w_1, \cdots, w_k$ are chosen, w_{k+1} is the shortest word such that w_{k+1} is different from the previous words and

$$w_1 \cdots w_k w_{k+1}$$

is either prefix of ω or is equal to ω. If such w_{k+1} does not exist, then the last word of the parsing is u such that

$$\omega = w_1 \cdots w_k u.$$

Note that in the second case $u = w_j$ for some $j \leq k$. We also remark if j is such that $\ell(w_j) > 1$, then $w_j = w_i a$ for some $j < i$ and $a \in \mathcal{A}$. Finally, if the ending word u is non-empty, then $\ell(u) \leq \sqrt{2n}$. To see that, let j be such that $u = w_j$.

Exercise 1. Prove that $\ell(w_j) \leq \sqrt{2N}$ for all j.
Hint. Obviously, $\ell(w_j) \leq j$. If $\ell(w_j) = j$, then $\ell(w_i) = i$ for $i < j$, and the statement follows from $\ell(w_1) + \cdots + \ell(w_j) \leq N$.

The Lempel-Ziv code sequence $(C_N)_{N\geq 1}$ is based on the Lempel-Ziv parsing. First choose one-one functions

$$F : \{1, \cdots, N\} \to \{0,1\}^{\lceil \log_2 N \rceil}, \qquad G : \mathcal{A} \to \{0,1\}^{\lceil \log_2 |\mathcal{A}| \rceil}.$$

Let

$$\omega = w_1 \cdots w_k u$$

be the LZ-parsing of $\omega \in \mathcal{A}^N$. Then

$$C_N(\omega) = \overline{w}_1 \cdots \overline{w}_k \overline{u}$$

where $\overline{w}_j, \overline{u} \in \{0,1\}^*$ are defined as follows:

(1) If $\ell(w_j) = 1$, $\overline{w}_j = 0G(w_j)$.

(2) If $\ell(w_j) > 1$ and i is the smallest integer such that $w_j = w_i a$ for some $a \in \mathcal{A}$, then

$$\overline{w}_j = 1F(i)0G(a).$$

(3) $\overline{u}$ is empty word if u is empty word. Otherwise, if i is the smallest integer such that $u = w_i$, $\overline{u} = 1F(i)$.

Exercise 2. Verify that the code C_N is prefix-free. Show that

$$\ell(C_N(\omega)) \leq (\mathcal{F}(\omega) + 1)\log_2 N + K_1 \mathcal{F}(\omega) + K_2,$$

where K_0, K_1 are constants that depend only on $|\mathcal{A}|$.

Theorem 2.11 *Let $P \in \mathcal{P}(\mathcal{A})$.*

(1)

$$\lim_{N\to\infty} \int_{\mathcal{A}^N} \frac{1}{N} \mathcal{F}(\omega) \log_2 \mathcal{F}(\omega) \mathrm{d}P_N(\omega) = S_{\mathrm{Sh}}(P).$$

(2)

$$\lim_{N\to\infty} \frac{1}{N} \langle C_N \rangle_{P_N} = S_{\mathrm{Sh}}(P).$$

The stunning aspect of this result is that the Lempel-Ziv code sequence is *universal* in a sense that that it is Shannon-optimal for any P. A far reaching generalization will be presented later in the notes.

Proof. Since the code sequence (C_N)

2.9 Merhav-Ziv parsing and relative entropy

2.10 Properties of Renyi entropy

2.11 Renyi entropy and prefix-free coding

G In [Cam65] Campbell introduced a family of exponential cost functions

$$\mathcal{L}_\mu^{(\alpha)}(C) = \frac{1}{\alpha}\log_2\left(\sum_b 2^{\alpha\ell(C(b))}\mu(b)\right), \qquad \alpha > 0.$$

The function $\alpha \mapsto \mathcal{L}_\mu^{(\alpha)}$ is increasing and is strictly increasing unless all the code words $C(b)$ have the same length. Moreover,

$$\lim_{\alpha\downarrow 0}\mathcal{L}_\mu^{(\alpha)}(C) = \mathcal{L}_\mu(C), \qquad \lim_{\alpha\to\infty}\mathcal{L}_\mu^{(\alpha)}(C) = \max_b \ell(C(b)).$$

Concavity of the logarithm gives $\mathcal{L}_\mu^{(\alpha)}(C) \geq \alpha\mathcal{L}_\mu(C)$. Campbell proves

Proposition 2.12 *For $\alpha > 0$,*

$$\mathcal{L}_\mu^{(\alpha)}(C) \geq \frac{\alpha+1}{\alpha}S_{\frac{1}{\alpha+1}}(\mu). \tag{2.17}$$

Proof. Set $x_b = [\mu(b)]^{1/\alpha}$, $y_b = [\mu(b)]^{-1/\alpha}2^{-\ell(C(b))}$, $p = \frac{\alpha}{\alpha+1}$, $q = -\alpha$. By the Kraft inequality and the reverse Hölder inequality[8]

$$1 \geq \sum_b 2^{-\ell(C(b))} = \sum_b x_b y_b \geq \left(\sum_b 2^{\alpha\ell(C(b))}\right)^{-1/\alpha}\left(\sum_b [\mu(b)]^{\frac{1}{\alpha+1}}\right)^{\frac{\alpha+1}{\alpha}},$$

and so

$$\left(\sum_b 2^{\alpha\ell(C(b))}\right)^{1/\alpha} \geq \left(\sum_b [\mu(b)]^{\frac{1}{\alpha+1}}\right)^{\frac{\alpha+1}{\alpha}}.$$

Taking $\log_2$ of both sides yields the statement. □

The inequality (2.17) yields the Shannon bound (**??**) since

$$\mathcal{L}_\mu(C) = \lim_{\alpha\downarrow 0}\frac{1}{\alpha}\mathcal{L}_\mu^{(\alpha)}(C) \geq \lim_{\alpha\downarrow 0}\frac{\alpha+1}{\alpha}S_{\frac{1}{\alpha+1}}(\mu) = S(\mu).$$

[8] This inequality states the following. Let $x_1, \cdots, x_n, y_1, \cdots, y_n$ be strictly positive real numbers. Let p, q be real numbers such that $p^{-1} + q^{-1} = 1$ and suppose that $p < 1$. Then

$$\sum_{i=1}^n x_i y_i \geq \left(\sum_{i=1}^n x_i^p\right)^{1/p}\left(\sum_{i=1}^n y_i^q\right)^{1/q}.$$

Regarding the saturation of (2.17), set

$$l_b := \left\lceil -\frac{1}{\alpha+1}\log_2 \mu(b) + S_{\frac{1}{\alpha+1}}(\mu)\right\rceil. \tag{2.18}$$

Since

$$\sum_b 2^{-l_b} \leq 2^{\frac{1}{\alpha+1}\log_2 \mu(b) - S_{\frac{1}{\alpha+1}}(\mu)} = 1, \tag{2.19}$$

there exist a prefix-free code C such that $\ell(C(b)) = l_b$. For this code we have

$$\begin{aligned}\sum_b 2^{\alpha\ell(C(b))}\mu(b) &\leq 2^\alpha \sum_b 2^{\frac{1}{\alpha+1}\log_2 \mu(b) + \alpha S_{\frac{1}{\alpha+1}}(\mu)} \\ &= 2^\alpha 2^{(\alpha+1)S_{\frac{1}{\alpha+1}}(\mu)},\end{aligned}$$

which gives

$$\mathcal{L}_\mu^{(\alpha)}(C) \leq 1 + \frac{\alpha+1}{\alpha} S_{\frac{1}{\alpha+1}}(\mu). \tag{2.20}$$

In the limit $\alpha \downarrow 0$ this inequality reduces to (**??**).

For later purposes we introduce a family of functionals

$$Q_\mu(\alpha) = \log_2\left(\sum_b 2^{\alpha\ell(C(b))}\mu(b)\right), \qquad \alpha \in \mathbb{R},$$

where, unlike in the Campbell cost function $\mathcal{L}_\mu^{(\alpha)}(C)$, our emphasis will be on the α dependence. The function $\alpha \mapsto Q_\mu(\alpha)$ is real-analytic, increasing and convex.[9] Obviously, $Q_\mu(0) = 0$ and

$$\lim_{\alpha\to\infty} \frac{Q_\mu(\alpha)}{\alpha} = \max_b \ell(C(b)), \qquad \lim_{\alpha\to-\infty} \frac{Q_\mu(\alpha)}{\alpha} = \min_b \ell(C(b)).$$

Proposition 2.13 (1) *For* $\alpha < -1$,

$$Q_\mu(\alpha) \geq (\alpha+1)S_{\frac{1}{\alpha+1}}(\mu) - \alpha\log_2\left(\sum_b 2^{-\ell(C(b))}\right).$$

(2) *For* $-1 < \alpha < 0$,

$$Q_\mu(\alpha) \leq (\alpha+1)S_{\frac{1}{\alpha+1}}(\mu).$$

(3) *For* $\alpha \geq 0$,

$$Q_\mu(\alpha) \geq (\alpha+1)S_{\frac{1}{\alpha+1}}(\mu).$$

[9] This function is strictly convex unless all the code words have the same length.

Proof. Part (3) follows from Proposition 2.12, and the same proof yields Part (2). To prove Part (1), let x_b, y_b, p, q, be as in the proof of Proposition 2.12. The Hölder inequality

$$\sum_b^n x_b y_b \le \left(\sum_b^n x_b^p\right)^{1/p} \left(\sum_b y_b^q\right)^{1/q}$$

gives that

$$\sum_b 2^{-\ell(C(b))} \le \left(\sum_b [\mu(b)]^{\frac{1}{\alpha+1}}\right)^{\frac{\alpha+1}{\alpha}} \left(\sum_b \mathrm{e}^{\alpha\ell(C(b))}\mu(b)\right)^{-\frac{1}{\alpha}}.$$

Rearranging, we get

$$\left(\sum_b \mathrm{e}^{\alpha\ell(C(b))}\mu(b)\right)^{\frac{1}{\alpha}} \le \left(\sum_b 2^{-\ell(C(b))}\right)^{-1} \left(\sum_b [\mu(b)]^{\frac{1}{\alpha+1}}\right)^{\frac{\alpha+1}{\alpha}}$$

Taking $\log_2$ gives

$$Q_\mu(\alpha) \ge (\alpha+1)S_{\frac{1}{\alpha+1}}(\mu) - \alpha\log_2\left(\sum_b \mathrm{e}^{-\ell(C(b))}\right),$$

and Part (1) follows. □

Turning to the optimality of Proposition 2.13, (2.18) is defined for $\alpha \neq -1$ and the inequality (2.19) remains valid. Let C is a prefix-free code satisfying $\ell(C(b)) = l_b$. The inequality (2.20) gives that for $\alpha \ge 0$,

$$Q_\mu(\alpha) \le \alpha + (\alpha+1)S_{\frac{1}{\alpha+1}}(\mu).$$

The computation that gives (2.20) also yields that $\alpha < 0$, $\alpha \neq -1$,

$$Q_\mu(\alpha) \ge \alpha + (\alpha+1)S_{\frac{1}{\alpha+1}}(\mu).$$

The last bound (compare with Part (1) of Proposition 2.13) is not effective in the regime $\alpha < -1$. For $\alpha < -1$ we estimate

$$\sum_b 2^{\alpha\ell(C(b))}\mu(b) \le \sum_b 2^{\frac{1}{\alpha+1}\log_2\mu(b)+\alpha S_{\frac{1}{\alpha+1}}(\mu)} = 2^{(\alpha+1)S_{\frac{1}{\alpha+1}}(\mu)},$$

which leads to

$$Q_\mu(\alpha) \le (\alpha+1)S_{\frac{1}{\alpha+1}}(\mu).$$

The above discussion singles out the function

$$F(\alpha) = (\alpha+1)S_{\frac{1}{\alpha+1}}(\mu),$$

defined for $\alpha \neq -1$. If μ is the uniform measure, $\mu(b) = 1/|\mathcal{B}|$ for all $b \in \mathcal{B}$, and $F(\alpha) = \alpha\log_2|\mathcal{B}|$. Since

$$\lim_{\alpha\downarrow -1} F(\alpha) = \max_b \log_2\mu(b), \qquad \lim_{\alpha\uparrow -1} F(\alpha) = \min_b \log_2\mu(b),$$

F is discontinuous at -1 unless μ is the uniform measure. F is an increasing function, concave on $(-\infty,-1)$ and convex on $(-1,\infty)$. Moreover, if μ is not the uniform measure, F is strictly concave on $(-\infty,-1)$ and is strictly convex on $(-1,\infty)$. To see this, one computes

$$F''(\alpha)=\frac{1}{(\alpha+1)^3\ln 2}\left(\sum_{b\in\mathcal{B}}\frac{\mu(b)^{\frac{1}{\alpha+1}}}{\sum_{b\in\mathcal{B}}\mu(b)^{\frac{1}{\alpha+1}}}[\ln\mu(b)]^2-\left(\sum_{b\in\mathcal{B}}\frac{\mu(b)^{\frac{1}{\alpha+1}}}{\sum_{b\in\mathcal{B}}\mu(b)^{\frac{1}{\alpha+1}}}\ln\mu(b)\right)^2\right)$$

and observes that by Jensen's inequality for $f(x)=x^2$,

$$\sum_{b\in\mathcal{B}}\frac{\mu(b)^{\frac{1}{\alpha+1}}}{\sum_{b\in\mathcal{B}}\mu(b)^{\frac{1}{\alpha+1}}}[\ln\mu(b)]^2-\left(\sum_{b\in\mathcal{B}}\frac{\mu(b)^{\frac{1}{\alpha+1}}}{\sum_{b\in\mathcal{B}}\mu(b)^{\frac{1}{\alpha+1}}}\ln\mu(b)\right)^2\geq 0,$$

with equality iff μ is the uniform measure.

2.12 Properties of relative Renyi entropy

2.13 First rumination

2.14 Notes

3 One-sided shift

References

[Cam65] Campbell, L.L. A coding theorem and Renyi's entropy. Information and Control **8**, 423-420 (1965).

[CDEJR22-1] Cristadoro, G., Degli Esposti, M., Jaksic, V., and Raquepas, R.: Recurrence times, waiting times and universal entropy production estimators. Submitted.

[CDEJR22-2] Cristadoro, G., Degli Esposti, M., Jaksic, V., and Raquepas, R.: On a waiting-time result of Kontoyiannis: mixing or decoupling?. Submitted.

[CJPS19] Cuneo, N., Jaksic, V., Pillet, C-A, and Shirikyan, A.: Large deviations and fluctuation theorem for selectively decoupled measures on shift spaces. Rev. Math. Phys. **3** (2019), 1950036.

[CR23] Cuneo, N., and Raquépas, R.: Large deviations of return times and related entropy estimators on shift spaces. Preprint.

[CJPS] Jaksic, V., Pillet, C.-A., and Shirikyan, A.: Beyond Gibbsianity. In preparation.

[Ja20] Jaksic, V.: Lectures on entropic information theory. McGill Fall 2020 course.

[LONE19] Jakšić, V.: Lectures on entropy.I: Information-theoretic notions. Bahns et al (Eds): Dynamical Methods in Open Quantum Systems, Tutorials, Schools and Workshops in the Mathematical Sciences, (2019), 141-268, Springer.

[JB15] Jain, S., and Bansal, R.K.: On match lengths, zero entropy and large deviations–with application to sliding window Lempel-Ziv algorithm. IEEE Transactions on Information Theory 61 (2015), 120–132.

[JB13] Jain, S., and Bansal, R.K.: On large deviation property of recurrence times. 2013 IEEE International Symposium on Information Theory, Istanbul (2013), 2880–2884.

[JZ95] Jacquet, P., and Szpankowski, W.: Asymptotic behaviour of the Lempel–Ziv parsing scheme and digital search trees. Theoretical Computer Science 144 (1995), 161–197.

[JZ11-1] Jacquet, P., and Szpankowski, W.: Limiting distribution of Lempel Ziv'78 redundancy. 2011 IEEE International Symposium on Information Theory Proceedings, St. Petersburg (2011), 1509–1513.

[JZ11-2] Jacquet, P., Szpankowski, W., and Tang, J.: Average profile of the Lempel-Ziv parsing scheme for a Markovian source. Algorithmica 31 (2011), 318–360.

[LZ77] Ziv, L., and Lempel, A.: A universal algorithm for sequential data compression. IEEE Trans. Inf. Theory **23**, 327-343 (1977).

[LZ78] Ziv, L., and Lempel, A.: Compression of individual sequences via variable-rate coding. IEEE Trans. Inf. Theory **24**, 530-536 (1978)

[Sha48] Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27** (1948), 379-423.

[Shi96] Shields, P.C.: The Ergodic Theory of Discrete Sample Paths. Graduate Studies in Mathematics, AMS 1996.

[MZ93] Ziv, J., and Merhav, N.: A measure of relative entropy between individual sequences with application to universal classification. IEEE Transactions on Information Theory **39**, 1270-1279 (1993).