Information theory, entropy (and statistical mechanics)

Vojkan Jaksic
McGill University

INTRODUCTION

This course is a sketch of certain aspects of the research program ”Beyond Gibbsianity” currently developed by A. Shirikyan (Paris), C-A. Pillet (Toulon/Luminy), and V.J. (Montreal/Milan).

The aspects we will discuss are at the interface between information theory and statistical mechanics.

Part I: ENTROPY

Part II: RELATIVE ENTROPY

Unfortunately, there is no time for discussion of Renyi entropies, which play a central role in the research program, or non-equilibrium statistical mechanics in the Hamiltonian setting.
The quantum aspects of the program also will not be discussed.

Lecture notes are available:

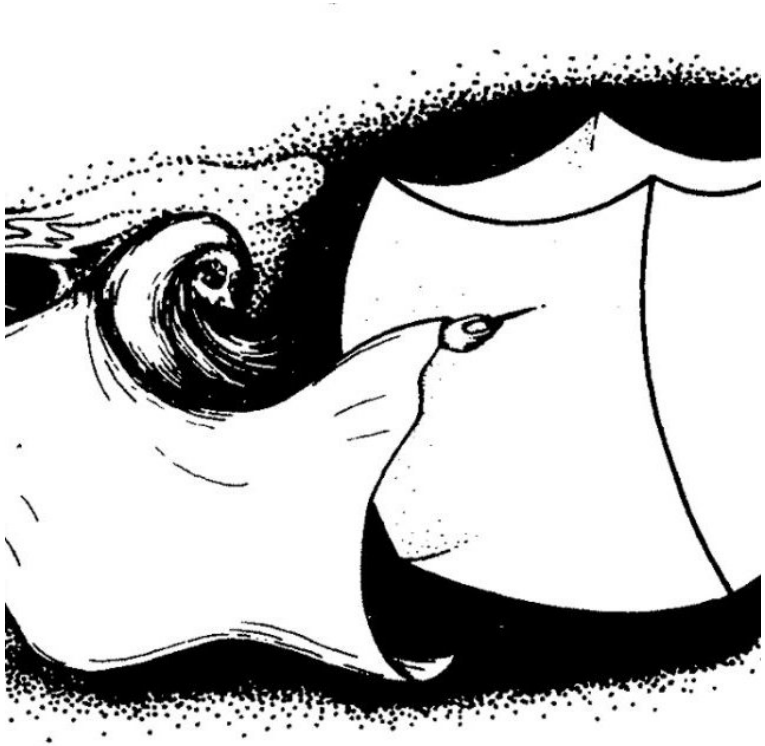
I. Lectures on Entropy. https://arxiv.org/pdf/1806.07249 Based on undergraduate course taught at McGill. Absolutely minimal first year mathematical background is assumed.

II. Entropic Information Theory. $60 + 12$ hours graduate course taught at McGill the Fall 2020. Link to the dropbox directory is available (videos and pdf files of lectures, tutorials, and additional material).

III. Entropy in Toulouse. Unfinished lecture notes of the course taught in the Winter 2024 school in Toulouse. Hopefully useful.

IV. Additional references available upon request (see also my McGill web site).

PART I: ENTROPY



God picking out the special (low-entropy) initial conditions of our universe. Penrose (1999).

NOTATION

$\mathcal{A} = \{a_1, \cdots, a_l\}$ finite alphabet (often $\mathcal{A} = \{0,1\}^N$).

$\mathcal{P}(\mathcal{A})$–the set of all probability measures on $\mathcal{A}$
$P(a) \geq 0$, $\sum_{a \in \mathcal{A}} P(a) = 1$.

The support of P is $\{a : P(a) > 0\}$ and we will often implicitly assume that it is equal to $\mathcal{A}$.
P is pure if $P(a) = 1$ for some a. P is chaotic (uniform) if $P(a) = 1/|\mathcal{A}|$ for all a (denoted P_{ch}).

If $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ ($\{0,1\}^{N+M} = \{0,1\}^N \times \{0,1\}^M$,
P_1, P_2 denote the respective marginals of $P \in \mathcal{P}(\mathcal{A})$,

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

ENTROPY

In **thermodynamics** it goes back to 1850 and work of Clausius (Entropie=transformation). Clausius was motivated by Sadi Carnot's work on efficency of thermal engines.

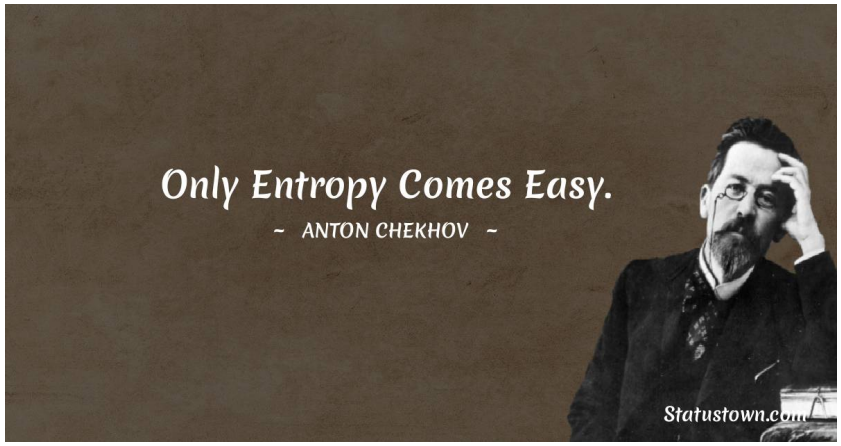
Entropy is the measure of a system's thermal energy per unit temperature that is unavailable for doing useful work (EB).

Its statistical interpretation is due to Boltzmann, Gibbs, and Maxwell.

In **information theory** it goes back to Shannon and his fundamental 1948 paper ”A Mathematical Theory of Communication”

Arguably one of the most used (and abused) scientific notions...

In **popular culture...** ”Entropy”– a song about Discord and instrumental vocals by AwkwardMarina... In literature



We will take Shannon's road

Reprinted with corrections from *The Bell System Technical Journal,*
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base 2 logarithm of this number. Doubling the time roughly squares the number of possible messages, or doubles the logarithm, etc.

2. It is nearer to our intuitive feeling as to the proper measure. This is closely related to (1) since we intuitively measures entities by linear comparison with common standards. One feels, for example, that two punched cards should have twice the capacity of one for information storage, and two identical channels twice the capacity of one for transmitting information.

3. It is mathematically more suitable. Many of the limiting operations are simple in terms of the logarithm but would require clumsy restatement in terms of the number of possibilities.

The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits*, a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information. N such devices can store N bits, since the total number of possible states is 2^N and $\log_2 2^N = N$. If the base 10 is used the units may be called decimal digits. Since

$$\begin{aligned}\log_2 M &= \log_{10} M / \log_{10} 2 \\ &= 3.32 \log_{10} M,\end{aligned}$$

[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal,* April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A.I.E.E. Trans.*, v. 47, April 1928, p. 617.

[2] Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal,* July 1928, p. 535.

The entropy of $P \in \mathcal{P}(\mathcal{A})$ is the non-negative number

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

Base of the logarithm=choice of units (shannons= 2, nats = e, hartleys = 10).

Entropy = the measure of "randomness" of the stochastic experiment described by P.

Entropy = the measure of "informational content" of the stochastic experiment described by P.

Probabilistic experiment: tossing a coin. Outcome: H(ead) or T(ail)

Fair coin, $P(H) = P(T) = 1/2$.

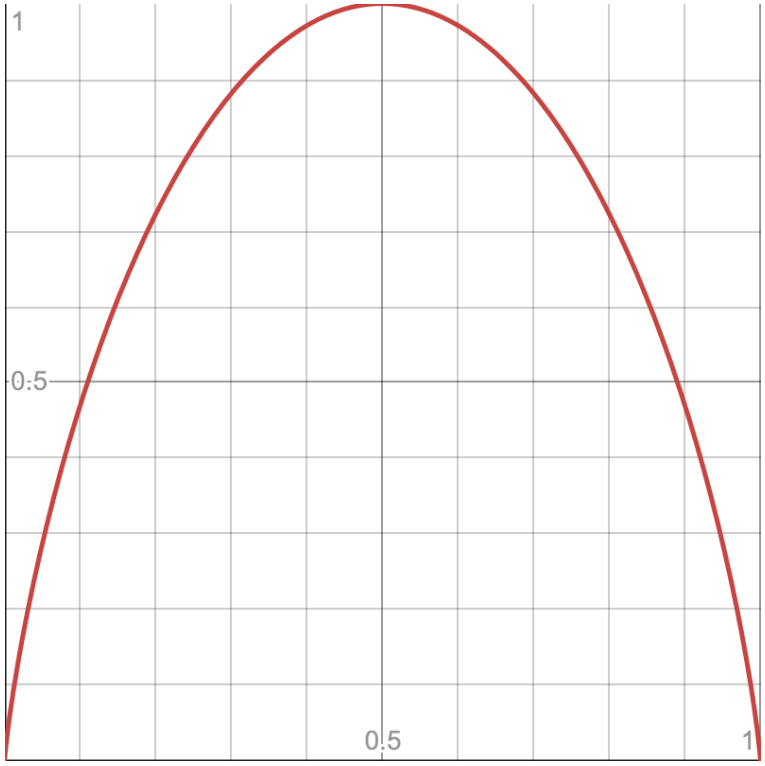
$$S(P) = \log_2 2 = 1.$$

Unfair coin with two heads: $P(H) = 1$, $P(T) = 0$,

$$S(P) = 0.$$

Biased coin, $P(H) = 1/3$, $P(T) = 2/3$,

$$S(P) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.917...$$



$P(H) = p$, $P(T) = 1 - p$, plot of $S(p) = -p \log_2 p - (1-p) \log_2 (1-p)$.

BASIC PROPERTIES

1. $0 \leq S(P) \leq \log |\mathcal{A}|$. $S(P) = 0$ iff P is pure and $S(P) = \log |\mathcal{A}|$ iff P is chaotic.

2. The entropy map $P \mapsto S(P)$ is continuous.

3. The entropy map is concave:

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

for any n, $p_k \geq 0$, $\sum_{k=1}^{n} p_k = 1$.

4. The entropy map is "almost convex":

$$S(p_1P_1+\cdots+p_nP_n) \leq p_1S(P_1)+\cdots+p_nS(P_n)+S(p_1,\cdots,p_n).$$

The equality holds if the supports of P_j's are disjoint:

entropy of mixture= mixtures of entropies

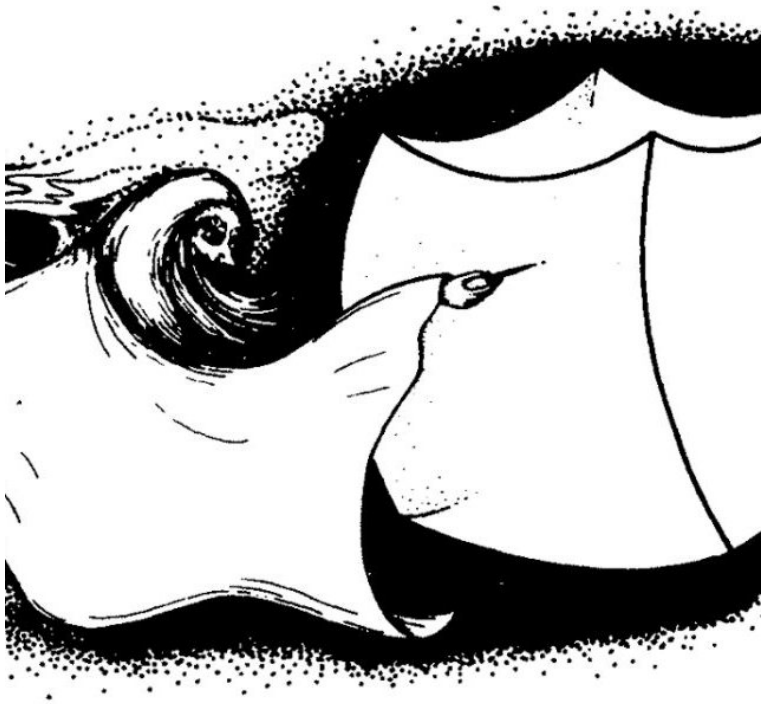
5. The entropy map is subadditive: if $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$, then

$$S(P) \leq S(P_1 \times P_2) = S(P_1) + S(P_2)$$

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

These properties are fundamental and easy to prove; see Shannon (1948)

BUT WHY THE FORMULA $S(P) = -\sum P(a) \log P(a)$**?**



UNIVERSALITY I: AXIOMATIZATION

The axiomatization idea is that the formula $S(P) = -\sum P(a) \log P(a)$ is forced by a few basic intuitive properties that any "entropy" function should have. It is due to Shannon (1948).

Let $\mathcal{P} = \cup_{\mathcal{A}} \mathcal{P}(\mathcal{A})$ and let $\mathfrak{S} : \mathcal{P} \to [0, \infty)$ be a putative entropy function. We assume the obvious:

Requirement 1. $\mathfrak{S}(P)$ does not depend on the enumeration of the alphabet $\mathcal{A}$.

Requirement 2. If $\mathcal{A}' = \mathcal{A} \cup \{b\}$ and $P'(a) = P(a)$, $P'(b) = 0$, then $\mathfrak{S}(P') = \mathfrak{S}(P)$.

Requirement 3. The map $P \mapsto \mathfrak{S}(P)$ is continuous.

FIRST AXIOMATIZATION

Theorem Suppose that the function $\mathfrak{S}$ satisfies

$$\mathfrak{S}(p_1 P_1 + \cdots p_n P_N) = p_1 \mathfrak{S}(P_1) + \cdots + p_n \mathfrak{S}(P_n) + \mathfrak{S}(p_1, \cdots, p_n)$$

whenever the support of P_j's are disjoint. Then, up to the choice of the base of the logarithm, for all $P \in \mathcal{P}$,

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

This axiomatization is (essentially) due to Shannon (1948), with further contributions from Fadeev, Khinchine...The proof is not hard.

entropy of mixture= mixture of entropies

fixes uniquely the entropy function.

SECOND AXIOMATIZATION

Theorem Suppose that the function $\mathfrak{S}$ is subadditive:
If $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$, then for any $P \in \mathcal{P}(\mathcal{A})$,

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

with equality iff $P = P_1 \times P_2$. Then for all $P \in \mathcal{P}$,

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

This result is due to Aczél, Forte, and Ng (1973). The proof is deep.

Book: J. Aczél and Z. Daróczy.: On Measures of Information and Their Characterizations. Academic Press, 1975.

SHANNON'S COMMENT ON AXIOMATIZATION

This theorem [axiomatization], and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications.

Revolutionary implication: the universal optimal bound on the compression of the information.

The main point at issue is the effect of statistical knowledge about the source in reducing the required capacity of the channel, by the use of proper encoding of the information.

FURTHER NOTATION

$\mathcal{A}^n = \mathcal{A} \times \cdots \times \mathcal{A}$.

$x_1^n := (x_1, x_2, \cdots, x_n) \equiv x_1 x_2 \cdots x_n$.

$\Omega = \mathcal{A}^{\mathbb{N}} = \{x = (x_k)_{k \geq 1} \mid x_k \in \mathcal{A}\}$.

$x = x_1 x_2 \cdots x_n x_{n+1} \cdots$.

$\varphi : \Omega \to \Omega$ the shift map,

$$x_1 x_2 x_3 \cdots \mapsto x_2 x_3 \cdots$$

Dynamical system (Ω, φ).

Ergodic discrete time statistical sources with values in $\mathcal{A}$ are described by ergodic probability measures P on (Ω, φ).

Such P is uniquely determined by the sequence $(P_n)_{n \geq 1}$ of its $\mathcal{A}^n$-marginals.

NOTE ON ERGODICITY

P is shift-invariant (or stationary) if $P(\varphi^{-1}(E)) = P(E)$.

A shift-invariant P is ergodic if $\varphi^{-1}(E) = E \Rightarrow P(E) \in \{0, 1\}$.

A shift-invariant P is ergodic iff for any continuous $F : \Omega \to \mathbb{R}$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=0}^{N-1} F(\varphi^j(x)) = \int_\Omega F \mathrm{d}P$$

for P-a.e. x.

SPECIFIC ENTROPY

Let $P \sim (P_n)_{n\geq 1}$ be an ergodic source. The stationarity gives

$$S(P_{n+m}) \leq S(P_n) + S(P_m),$$

and so

$$s(P) := \lim_{n\to\infty} \frac{1}{n} S(P_n) = \inf_{n\geq 1} \frac{1}{n} S(P_n).$$

$s(P)$-specific entropy of the source P.

For any P, $0 \leq s(P) \leq \log |\mathcal{A}|$.

The specific entropy is the basic notion in information theory, dynamical systems theory, probability, statistical mechanics...

FIRST SHANNON'S THEOREM

The ergodic source $P \sim (P_n)_{n\geq 1}$ is given.The compression alphabet is $\{0,1\}$. Let $0 < \epsilon < 1$ be fixed "allowed coding error". Let $N, M \geq 1$.

Coding pair (C_N, D_N). Coder $C_N : \mathcal{A}^N \to \{0,1\}^M$. Decoder $D_N : \{0,1\}^M \to \mathcal{A}^N$.
Compression coefficient = M/N.

The error probability of the coding pair (C_N, D_N) is

$$P_N \left\{ x_1^N \in \mathcal{A}^N \mid D_N \circ C_N(x_1^N) \neq x_1^N \right\}.$$

If this probability is $< \epsilon$, the pair (N, M) is called ϵ-good.

For given N, let $M(N)$ be smallest number such that the pair $(N, M(N))$ is ϵ-good.

$M(N)/N$ is the **best possible compression** subject to the allowed ϵ-error probability.

Shannon Source Coding Theorem
For any $0 < \epsilon < 1$,

$$\lim_{N\to\infty} \frac{M(N)}{N} = s(P).$$

The specific entropy = universal optimal asymptotic bound on the compression of information.

The deep link with Hypothesis Testing will be discussed in the Lecture II.

SHANNON-MCMILLAN-BREIMAN THEOREM

Information functions $I_n : \Omega \to [0, \infty]$,

$$I_n(x) = I_n(x_1^n) = -\log P_n(x_1^n).$$

$$S(P_n) = \int_\Omega I_n(x) \mathrm{d}P.$$

$$\lim_{n\to\infty} \frac{1}{n} S(P_n) = \lim_{n\to\infty} \frac{1}{n} \int_\Omega -\log P_n(x_1^n) \mathrm{d}P$$

$$= s(P).$$

Theorem.

$$\lim_{n\to\infty} -\frac{1}{n} \log P_n(x_1^n) = s(P)$$

P-a.s and in $L^1(\Omega, dP)$.

Central to Shannon's theory.

Three classical proofs:
Martingale proof.
Derreniec subadditivity proof (Kingman subadditive ergodic theorem)
Ornstein-Weiss covering intervals argument.

All three proofs are available in the "Entropic Information Theory" course dropbox directory together with the background material.

SECOND SHANNON'S THEOREM

Lossless variable length coding of the ergodic source $P \sim (P_n)_{n\geq 1}$.

$\{0,1\}^* = \cup_{M\geq 1}\{0,1\}^M$ = the set of all finite length words with letters from the compression alphabet $\{0,1\}$:

$$\{0,1\}^* = \{0,1,00,01,10,11,000,001,\cdots\}$$

A code sequence $(C_N)_{N\geq 1}$ is the collection of one-one maps

$$C_N : \mathcal{A}^N \to \{0,1\}^*.$$

Length function $L_N(x_1^N)$–the length of the code word $C_N(x_1^N)$. The expected length is

$$\langle L_N \rangle = \sum_{x_1^N \in \mathcal{A}^N} L_N(x_1^N) P_N(x_1^N).$$

Compression is measured by the ratios $\langle L_N \rangle / N$ or L_N / N.

Theorem.

$$\liminf_{N \to \infty} \langle L_N \rangle / N \geq s(P)$$

and there are **optimal codes** for which

$$\lim_{N \to \infty} \langle L_N \rangle / N = s(P)$$

This result is essentially proven in Lecture II.

Theorem.

$$\liminf_{N \to \infty} L_N(x_1^N)/N \geq s(P),$$

P-a.s. and there are **optimal codes** for which

$$\lim_{N \to \infty} L_N(x_1^N)/N = s(P)$$

P-a.s.

The interest turns to the **optimal** codes...

One naturally expects that the construction of optimal codes (examples are Shannon's, Huffman...) depends on P.

> *The main point at issue is the effect of statistical knowledge about the source in reducing the required capacity of the channel, by the use of proper encoding of the information.*

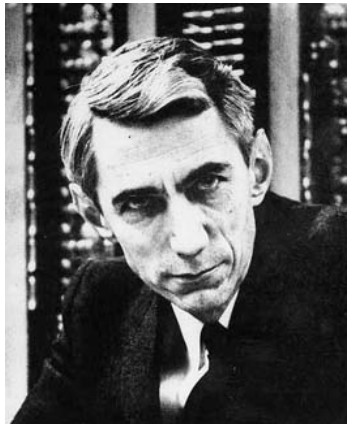
But what if have have **no** statistical knowledge of the source?

UNIVERSALITY II: COMPRESSION OPTIMALITY

Do there exist **deterministic universally optimal** code sequences $(C_N)_{N\geq 1}$ such that for **any** ergodic source P,

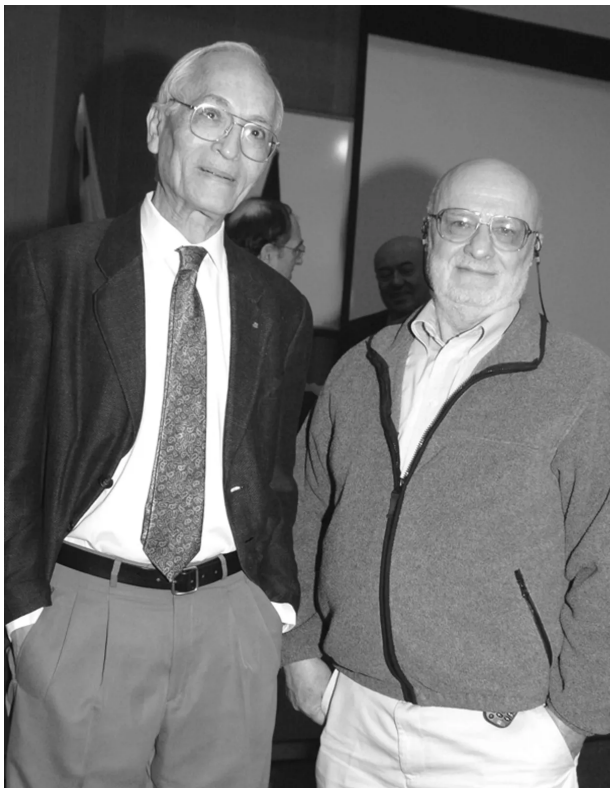
$$\lim_{N\to\infty} \langle L_N \rangle / N = s(P)$$

$$\lim_{N\to\infty} L_N(x_1^N)/N = s(P) \qquad P-\text{a.s.}$$



Astonishing!?

LEMPEL-ZIV UNIVERSALITY



For historical account see the recent article N. Merhav ”On Jacob Ziv’s Individual-Sequence Approach to Information Theory” Arxiv https://arxiv.org/pdf/2406.02904.

Lempel-Ziv universal codes (1977, 1978)—based on LZ parsing.

$x_1^N = x_1 x_2 \cdots x_N$ is parsed so that every new word is the shortest word not seen before.

Example. $N = 11$,

$$10110001001$$

$$1|0|11|00|01|001$$

$\mathcal{P}(x_1^N)$ is the number of parsed words in x_1^N. In the above example,

$$\mathcal{P}(10110001001) = 6.$$

The LZ parsing algorithm is completely **deterministic.**

LZ-code builds on the LZ-parsing algorithm.

Theorem. For **any** ergodic source P,

$$\lim_{N\to\infty} \frac{\mathcal{P}(x_1^N)\log N}{N} = s(P)$$

P-a.s. and in $L^1(\Omega, \mathrm{d}P)$.

Apart from its important practical applications, this is mathematically a deep foundational result!

It gives the universal optimality of the resulting LZ code sequences.

The studies of the LZ parsing have lead to further deep and unexpected universal characterizations of the specific entropy. We will discuss two of them.

RETURN TIMES CHARACTERIZATION

Due to Wyner and Ziv (1989), Ornstein and Weiss (1993).

$x = x_1 x_2 \cdots \in \Omega$, $x_k^m = x_k x_{k+1} \cdots x_m$.

The recurrence time function is

$$R_n(x) := \inf\{k \geq 1 : x_{n+k}^{2n+k-1} = x_1^n\}$$

The first time the string x_1^n reappears in x.

Example:

$$x = 0100110101001101001110100100101001010\ldots$$

The first time the string $x_1^4 = 0100$ reappears in x:

$$x = \underbrace{0100110101001}_{5}10100111010010010100101\ldots$$

$R_4(x) = 5$.

Theorem. For **any** ergodic source P,

$$\lim_{N\to\infty} \frac{1}{N} \log R_N(x) = s(P)$$

P-a.s. and in $L^1(\Omega, dP)$.

Again, a mathematically deep foundational result.

Simpler proof with completely novel strategy: Kontoyiannis' 1998 PhD thesis.

WAITING TIMES CHARACTERIZATION

Wyner and Ziv (1989), Morton and Shields (1995), Kontoyiannis (1998).

The waiting times functions are defined in terms of a pair (x, y) of elements of Ω. For $n \in \mathbb{N}$ and $(x, y) \in \Omega \times \Omega$

$$W_n(x, y) = \inf\{k \geq 1 : y_k^{k+n-1} = x_1^n\}.$$

$W_n(x, y)$ is the first time the string x_1^n of x appears in y.

$$x = 0100110101001101001110100100101001010\ldots,$$
$$y = 1101010001001010011010100100101010100\ldots$$

$W_4(x, y)$ = the first time the string $x_1^4 = 0100$ appears in y

$$x = \underline{0100}110101001101001110100100101001010\ldots,$$
$$y = \underbrace{1101\underline{0100}}_{5}0100101001101010010010101010\ldots$$

$$W_4(x, y) = 5$$

Theorem. Under suitable (mixing) regularity assumptions on P,

$$\lim_{N\to\infty} \frac{1}{N} \log W_N(x,y) = s(P)$$

for $P \times P$-almost all (x,y).

The technically best result (β-mixing) is due to Morton and Shields. The proof is hard. An elegant simple argument in the case of ψ-mixing is due to Kontoyiannis.

Important: This is the first result that is not completely universal and some regularity is required (counterexample due to Shields (1993)). There are some important open questions in this respect.

RESEARCH PROGRAM

The fundamental entropic laws of large numbers (P-a.s. sure convergence) we discussed are:

$$\lim_{N\to\infty} -\frac{1}{N}\log P_N(x_1^N) = s(P)$$

$$\lim_{N\to\infty} \frac{\mathcal{P}(x_1^N)\log N}{N} = s(P)$$

$$\lim_{N\to\infty} \frac{1}{N}\log R_N(x) = s(P)$$

$$\lim_{N\to\infty} \frac{1}{N}\log W_N(x,y) = s(P)$$

Beyond the laws of large numbers:

1. Study of fluctuations–Large Deviation Principle.

2. Study of the fractal dimensions of entropic level sets.

3. Other entropies: Renyi, relative entropies.

4. Applications to non-equilibrium statistical mechanics, linguistics, and biology.

5. Long term research program with involvement of students and postdocs. A small but important part of the general ”Beyond Gibbsianity” program.

First results:

1. Cuneo N., VJ., Pillet C-A, Shirikyan A.: Large deviations and fluctuation theorem for selectively decoupled measures on shift spaces, Rev. Math. Phys. 31 (2019).

2. Cristadoro G., Degli Esposti M., JV, Raquépas R.: Recurrence times, waiting times and universal entropy production estimators, Lett. Math. Phys., 113 (2023)

3. Cristadoro G., Degli Esposti M., JV, Raquépas R.: On a waiting-time result of Kontoyiannis: mixing or decoupling?, Stoch. Process. their Appl., 166 (2023)

4. Cuneo N., Raquépas R.: Large deviations of return times and related entropy estimators on shift spaces. Commun. Math. Phys. 405: Article 135 (2024)

The works of students and postdocs on Merhav-Ziv universal cross-entropy estimators.

1.Barnfield N., Grondin R., Pozzoli G., Raquépas R.: On the Ziv–Merhav theorem beyond Markovianity I. To appear in Canadian Journal of Mathematics.

2. Barnfield N., Grondin R., Pozzoli G., Raquépas R.: On the Ziv–Merhav theorem beyond Markovianity II. Submitted, on Arxiv.