Entropy

Jonathan L.F. King
University of Florida, Gainesville FL 32611-2082, USA
squash@math.ufl.edu
Webpage http://www.math.ufl.edu/~squash/
4 May, 2008 (at *0-900:20*)

Prolegomenon. The word *'entropy'* was coined by Rudolf Julius Emanuel Clausius in 1867, in [2], referring to the thermodynamic notion in physics.

Our focus here, however, will be the notion in measurable-dynamics and topological-dynamics. (Entropy in differentiable-dynamics,[♡1] would require an article by itself.) Shannon's 1948 paper [3] on Information Theory, then Kolmogorov's [4] and Sinai's [5] generalization to dynamical systems, will be our starting point. I will stay in the one-dimensional case, where the acting-group is $\mathbb{Z}$.

§Bird's-eye view

Glossary. Some of the following definitions refer to the "Notation" paragraph immediately below. Use ***mpt*** for *'measure-preserving transformation'*.

A ***measure space*** $(X, \mathscr{X}, \mu)$ is a set X, a ***field*** (that is, a σ-algebra) $\mathscr{X}$ of subsets of X, and a countably-additive measure $\mu: \mathscr{X} \to [0, \infty]$. (We often just write (X, μ), with the field implicit.) For a collection $\mathscr{C} \subset \mathscr{X}$, use Fld($\mathscr{C}$) for the smallest field $\supset \mathscr{C}$. The number $\mu(B)$ is the "μ-***mass*** of B".

A ***measure-preserving map*** $\psi: (X, \mathscr{X}, \mu) \to (Y, \mathscr{Y}, \nu)$ is a map $\psi: X \to Y$ such that the inverse image of each $B \in \mathscr{Y}$ is in $\mathscr{X}$, and $\mu(\psi^{-1}(B)) = \nu(B)$. A ***(measure-preserving) transformation*** is a measure-preserving map $T: (X, \mathscr{X}, \mu) \to (X, \mathscr{X}, \mu)$. Condense this notation to $(T : X, \mathscr{X}, \mu)$ or $(T : X, \mu)$.

A ***probability space*** is a measure space (X, μ) with $\mu(X) = 1$; this μ is a ***probability measure***. All our maps/transformations in this article are on probability spaces. A ***factor map***

$$\psi: (T : X, \mathscr{X}, \mu) \to (S : Y, \mathscr{Y}, \nu)$$

[♡1] For instance, see [24], [26], [18] and [15].

is a measure-preserving map $\psi: X \to Y$ which intertwines the transformations, $\psi \circ T = S \circ \psi$. And ψ is an ***isomorphism*** if –after deleting a nullset in each space– this ψ is a bijection and ψ^{-1} is also a factor map.

A measure-theoretic statement holds ***almost everywhere***, abbreviated ***a.e***, if it holds off of a nullset, mass-zero set.[♡2] For example, $B \overset{\text{a.e}}{\supset} A$ means that $\mu(B \setminus A)$ is zero. The *a.e* will usually be implicit.

A ***probability vector*** $\vec{v} = (v_1, v_2, \dots)$ is a list of non-negative reals whose sum is 1. We generally assume that probability vectors and partitions (see below) have *finitely* many components. We write "*countable* probability vector/partition", when finitely or denumerably many components are considered.

A ***partition*** $\mathsf{P} = (A_1, A_2, \dots)$ splits X into pairwise disjoint subsets $A_i \in \mathscr{X}$ so that the disjoint union $\bigsqcup_i A_i$ is all of X. Each A_i is an ***atom*** of P. Use $|\mathsf{P}|$ or $^{\#}\mathsf{P}$ for the number of atoms. When P partitions a probability space, then it yields a probability vector $\vec{v}$, where $v_j := \mu(A_j)$. Lastly, use $\mathsf{P}\langle x \rangle$ to denote the P-atom that owns x.

Fonts. We use the font $\mathcal{H}$, $\mathcal{E}$, $\mathcal{I}$ for *distribution-entropy*, *entropy* and the *information function*. In contrast, the script font $\mathscr{A}\mathscr{B}\mathscr{C}\dots$ will be used for collections of sets; usually subfields of $\mathscr{X}$. Use $\mathbb{E}(\cdot)$ for the (conditional) expectation operator.

Notation. $\mathbb{Z}$ = integers. $\mathbb{Z}_+$ = positive integers, and $\mathbb{N}$ = natural numbers[♡3] = $\{0, 1, 2, \dots\}$. Use $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ for the ***ceiling*** and ***floor*** functions; $\lfloor \cdot \rfloor$ is also called the "greatest-integer function". For an interval $J := [a, b) \subset [-\infty, +\infty]$, let $[a \,..\, b)$ denote the ***interval of integers*** $J \cap \mathbb{Z}$ (with a similar convention for closed and open intervals). E.g, $(\mathsf{e} \,..\, \pi] = (\mathsf{e} \,..\, \pi) = \{3\}$.

For subsets A and B of the same space, Ω, use $A \subset B$ for inclusion and $A \subsetneq B$ for *proper* inclusion. The difference set $B \setminus A$ is $\{\omega \in B \mid \omega \notin A\}$. Employ A^c for the complement $\Omega \setminus A$. Since we work in a probability space, if we let $x := \mu(A)$, then a convenient convention is to have

$$x^c \;:=\; 1 - x\,,$$

since then $\mu(A^c)$ equals x^c.

Use $A \triangle B$ for the ***symmetric difference*** $[A \setminus B] \cup [B \setminus A]$. For a collection $\mathscr{C} = \{E_j\}_j$ of sets in Ω, let the ***disjoint union*** $\bigsqcup_j E_j$ or $\bigsqcup(\mathscr{C})$ represent the union $\bigcup_j E_j$ <u>and</u> *also assert* that the sets are pairwise disjoint.

Use "$\forall_{\text{large}} n$" to mean: "$\exists n_0$ such that $\forall n > n_0$". To refer to lefthand side of an equation 17, use LhS(17); do analogously for RhS(17), the righthand side.

[♡2] Eugene Gutkin once remarked to me that the problem with Measure Theory is… that you have to say "almost everywhere", almost everywhere.

[♡3] Some well-meaning folk use $\mathbb{N}$ for $\mathbb{Z}_+$, saying *'Nothing could be more natural than the positive integers'*. And this is why $0 \in \mathbb{N}$.

Entropy example: *How many questions?* Imagine a dartboard, FIG. 1, split in five regions, $A, \ldots, E$, with known probabilities. Blindfolded, you throw a dart at the board. What is the expected number, V, of Yes/No questions needed to ascertain the region in which the dart landed?

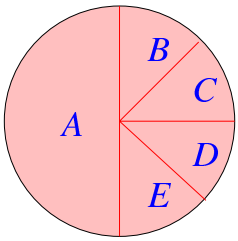

FIG. 1: *This dartboard is a probability space with a 5-set partition. The atoms have probabilities* $\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}$. *This probability distribution will be used later in Meshalkin's example on page 13.*

Solve this by always dividing the remaining probability in half. *'Is it A?'*; if Yes, then $V = 1$. Else: *'Is it B or C?'* —if Yes, then *'Is it B?'* —if No, then the dart landed in C, and V=3 was the number of questions. Evidently V=3 also for regions B, D, E. Using "log" to denote base-2 logarithm[♡4], the expected number of questions[♡5] is thus

$$\mathbf{2:} \qquad \mathbb{E}(V) \;=\; \tfrac{1}{2}\cdot 1 + \tfrac{1}{8}\cdot 3 + \tfrac{1}{8}\cdot 3 + \tfrac{1}{8}\cdot 3 + \tfrac{1}{8}\cdot 3 \;=\; \textstyle\sum_{j=0}^{4} p_j \cdot \log(\frac{1}{p_j}) \;\overset{\text{note}}{=\!=}\; 2\,.$$

Letting $\vec{v} := (\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ be the probability vector, we can write this expectation as

$$\mathbb{E}(V) \;=\; \sum\nolimits_{x\in\vec{v}} \boldsymbol{\eta}(x)\,.$$

Here, $\boldsymbol{\eta}\colon[0,1]\to[0,\infty)$ is the important function[♡6]

$$\mathbf{3:} \qquad \begin{aligned} \boldsymbol{\eta}(x) \;&:=\; x\cdot\log(1/x)\,; \quad \text{so extending by continuity gives} \\ \boldsymbol{\eta}(0) \;&=\; 0\,. \end{aligned}$$

An interpretation of "$\boldsymbol{\eta}(x)$" is the number of questions needed to winnow down to an event of probability x.

Distribution entropy

Given a probability vector $\vec{v}$, define its ***distribution entropy*** as

$$\mathbf{4:} \qquad \mathcal{H}(\vec{v}) \;:=\; \sum\nolimits_{x\in\vec{v}} \boldsymbol{\eta}(x)\,.$$

[♡4] In this paper, unmarked logs will be to base-2. In entropy theory, it does not matter much what base is used, but base-2 is convenient for computing entropy for messages described in bits.

When using the natural logarithm, some people refer to the unit of information as a ***nat***. In this paper, I have picked bits, rather than nats.

[♡5] This is holds when each probability p is a reciprocal power of two. For general probabilities, the "expected number of questions" interpretation holds in a weaker sense: Throw N darts independently at N copies of the dartboard. Efficiently ask Yes/No questions to determine where *all* N darts landed. Dividing by N, then sending $N\to\infty$, will be the $p\cdot\log(\frac{1}{p})$ sum of (2).

[♡6] There does not seem to be a standard name for this function. I use $\boldsymbol{\eta}$, since an uppercase η looks like an H, which is the letter that Shannon used to denote what I am calling distribution-entropy.

In this paper, I will use the term ***distropy*** for *'distribution entropy'* and will reserve ***entropy*** for the corresponding dynamical concept, when there is a notion of *time* involved. Getting ahead of ourselves, the *entropy* of a stationary process is the asymptotic average value that its distropy decays to, as we look at larger and larger finite portions of the process.

An equi-probable vector $\vec{v} := (\frac{1}{K}, \overset{K}{\ldots}, \frac{1}{K})$ evidently has $\mathcal{H}(\vec{v}) = \log(K)$. On a probability space, the "distropy of partition P", written $\mathcal{H}(\mathsf{P})$ or $\mathcal{H}(A_1, A_2, \ldots)$, shall mean the distropy of probability vector $j \mapsto \mu(A_j)$.

A (finite) partition necessarily has finite distropy. A *countable* partition can have finite distropy, e.g $\mathcal{H}(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots) = 2$. One could also have infinite distropy: Consider a piece $B \subset X$ of mass $1/2^N$. Splitting B into 2^k many equal-mass atoms gives an η-sum of $2^k \cdot \frac{k+N}{2^k 2^N}$. Setting $k = k_N := 2^N - N$ makes this η-sum equal 1; so splitting the pieces of $X = \bigsqcup_{N=1}^{\infty} B_N$, with $\mu(B_N) = \frac{1}{2^N}$, yields an ∞-distropy partition.

The η function. Our $\eta(x) = x \cdot \log(1/x)$ function[♡7] has vertical tangent at x=0, maximum at $1/\mathsf{e}$ and, when graphed in nats, slope -1 at x=1.

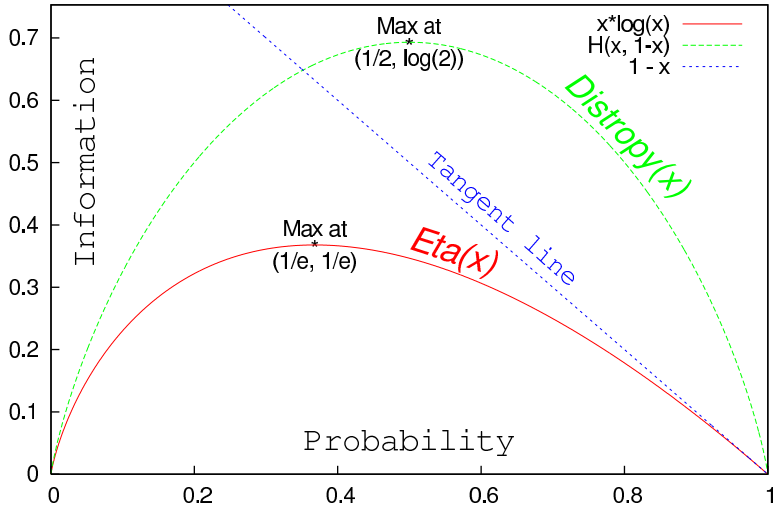

FIG. 5: *Using natural log, here are the graphs of:* $\eta(x)$ *in solid red.* $\mathcal{H}(x, x^c)$ *in dashed green.* $1 - x$ *in dotted blue.*

Both $\eta(x)$ *and* $\mathcal{H}(x, x^c)$ *are strictly convex-down. The* $1-x$ *line is tangent to* $\eta(x)$ *at* x=1.

Consider partitions P and Q on the same space (X, μ). Their ***join***, written $\mathsf{P} \vee \mathsf{Q}$, has atoms $A \cap B$, for each pair $A \in \mathsf{P}$ and $B \in \mathsf{Q}$. They are ***independent***, written $\mathsf{P} \perp \mathsf{Q}$, if $\mu(A \cap B) = \mu(A)\mu(B)$ for each A,B pair. We write $\mathsf{P} \succcurlyeq \mathsf{Q}$, and say that "P ***refines*** Q", if each P-atom is a subset of some Q-atom. Consequently, each Q-atom is a union of P-atoms.

Recall, for δ a real number, our convention that δ^c means $1 - \delta$, in analogy with $\mu(B^c)$ equaling $1 - \mu(B)$ on a probability space.

[♡7] Curiosity: Just in this footnote we compute distropy in ***nats***, that is, using natural logarithm. Given a small probability $p \in [0, 1]$ and setting $x := 1/p$, note that $\eta(p) = \frac{\log(x)}{x} \approx 1/\pi(x)$, where $\pi(x)$ denotes the number of prime numbers less-equal x. (This approximation is a weak form of the Prime Number Theorem.) Is there any actual connection between the *'approximate distropy'* function $\mathcal{H}_\pi(\vec{\mathbf{p}}) := \sum_{p \in \vec{\mathbf{p}}} 1/\pi(1/p)$ and Number Theory, other than a coincidence of growth rate?

6: Distropy fact. *For partitions* P, Q, R *on probability space* (X, μ)*:*

a: $\mathcal{H}(\mathsf{P}) \leqslant \log(\#\mathsf{P})$*, with equality IFF* P *is an equi-mass partition.*

b: $\mathcal{H}(\mathsf{Q} \vee \mathsf{R}) \leqslant \mathcal{H}(\mathsf{Q}) + \mathcal{H}(\mathsf{R})$*, with equality IFF* $\mathsf{Q} \perp \mathsf{R}$.

c: For $\delta \in [0, \frac{1}{2}]$*, the function* $\delta \mapsto \mathcal{H}(\delta, \delta^c)$ *is strictly increasing.*

d: $\mathsf{R} \preccurlyeq \mathsf{P}$ *implies* $\mathcal{H}(\mathsf{R}) \leqslant \mathcal{H}(\mathsf{P})$*, with equality IFF* $\mathsf{R} \overset{a.e}{=} \mathsf{P}$. ◊

Proof. Use the strict concavity of $\boldsymbol{\eta}()$, together with Jensen's Inequality. ♦

Remark. Although we will not discuss it in this paper, most distropy statements remain true with *'partition'* replaced by *'countable partition of finite distropy'*. □

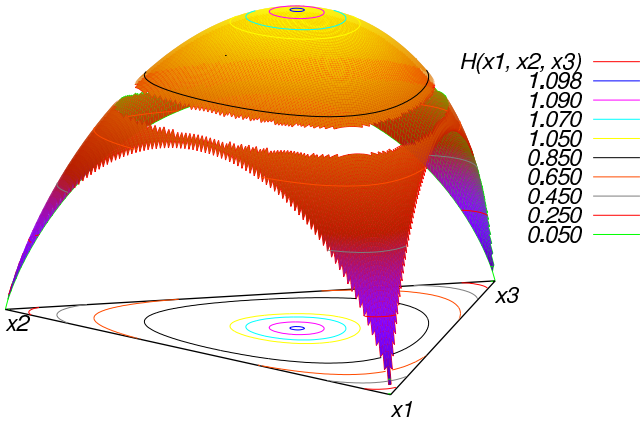

FIG. 7: Using natural log: *The graph of* $\mathcal{H}(x_1, x_2, x_3)$ *in barycentric coordinates; a slice has been removed, between* $z = 0.745$ *and* $z = 0.821$*. The three arches are copies of the distropy curve from* (5).

Binomial coefficients. The dartboard gave an example where distropy arises in a natural way. Here is a second example.

For a small δ>0, one might guess that the binomial coefficient $\binom{n}{\delta n}$ grows asymptotically (as $n \to \infty$) like $2^{\varepsilon n}$, for some small ε. But what is the correct relation between ε and δ? Well, Stirling's formula $n! \approx [n/\mathsf{e}]^n$ gives

$$\frac{n!}{[\delta n]! \cdot [\delta^c \cdot n]!} \approx \frac{n^n}{[\delta n]^{[\delta n]} \cdot [\delta^c \cdot n]^{[\delta^c \cdot n]}} = 1/[\delta^{\delta n} \cdot [\delta^c]^{\delta^c \cdot n}]\,. \quad (\text{Recall } \delta^c = 1 - \delta.)$$

Thus $\frac{1}{n} \cdot \log\binom{n}{\delta n} \approx \mathcal{H}(\delta, \delta^c)$. But by means of the above distropy inequalities, we get an inequality true for *all* n, not just asymptotically.

8: Binomial Lemma. *Fix a* $\delta \in [0, \frac{1}{2}]$ *and let* $\boldsymbol{H} := \mathcal{H}(\delta, \delta^c)$*. Then for each* $n \in \mathbb{Z}_+$*:*

9: $$\sum_{j \in [0\,..\,\delta n]} \binom{n}{j} \leqslant 2^{\boldsymbol{H} \cdot n}\,.$$ ◊

Proof. Let $X \subset \{0,1\}^n$ be the set of $\vec{x}$ with $^{\#}\{i \in [1\,..\,n] \mid x_i = 1\} \leqslant \delta \cdot n$. On X, let $\mathsf{P}_1, \mathsf{P}_2, \ldots$ be the coordinate partitions; e.g $\mathsf{P}_7 = (A_7, A_7{}^c)$, where $A_7 := \{\vec{x} \mid x_7 = 1\}$. Weighting each point by $\frac{1}{|X|}$, the uniform distribution on X, gives that $\mu(A_7) \leqslant \delta$. So $\mathcal{H}(\mathsf{P}_7) \leqslant \boldsymbol{H}$, by (6c). Finally, the join $\mathsf{P}_1 \vee \ldots \vee \mathsf{P}_n$ separates the points of X. So

$$\log(^{\#}X) \;=\; \mathcal{H}(\mathsf{P}_1 \vee \ldots \vee \mathsf{P}_n) \;\leqslant\; \mathcal{H}(\mathsf{P}_1) + \ldots + \mathcal{H}(\mathsf{P}_n) \;\leqslant\; \boldsymbol{H}n\,, \qquad \blacklozenge$$

making use of (6a,b). And $^{\#}X$ equals LhS(9).

A gander at Shannon's Noisy Channel theorem

We can restate the Binomial lemma using the ***Hamming metric*** on $\{0,1\}^n$,

$$\mathrm{Dist}(\vec{x},\vec{y}) \;:=\; ^{\#}\{i \in [1\,..\,n] \;\big|\; x_i \neq y_i\}\,.$$

Use $\mathrm{Bal}(\vec{x}, r)$ for the open radius-r ball centered at $\vec{x}$, and

$$\overline{\mathrm{Bal}}(\vec{x}, r) \;:=\; \{\vec{y} \mid \mathrm{Dist}(\vec{x},\vec{y}) \leqslant r\}$$

for the closed ball. The above lemma can be interpreted as saying that

9′: $$\left|\overline{\mathrm{Bal}}(\vec{x}, \delta n)\right| \;\leqslant\; 2^{\mathcal{H}(\delta,\delta^c)\cdot n}\,, \quad \text{for each } \vec{x} \in \{0,1\}^n.$$

10: Corollary. *Fix $n \in \mathbb{Z}_+$ and $\delta \in [0, \frac{1}{2}]$, and let $\boldsymbol{H} := \mathcal{H}(\delta, \delta^c)$. Then there is a set $C \subset \{0,1\}^n$, with $^{\#}C \geqslant 2^{[1-\boldsymbol{H}]n}$, that is **strongly** δn**-separated**. I.e, $\mathrm{Dist}(\vec{x},\vec{y}) > \delta n$ for each distinct pair $\vec{x},\vec{y} \in C$.* $\blacklozenge$

Noisy Channel. Shannon's theorem says that a noisy channel has a ***channel capacity***. Transmitting *above* this speed, there is a minimum error-rate (depending how much "*above*") that no error-correcting code can fix. Conversely, one can transmit *below* –but arbitrarily close to– the channel capacity, and encode the data so as to make the error-rate less than any given ε. We use (10) to show the existence of such codes, in the simplest case where the noise[♡8] is a binary independent-process (a "Bernoulli" process, in the language later in this article).

We have a channel which can pass one bit per second. Alas, there is a fixed noise-probability $\nu \in [0, \frac{1}{2})$ so that a bit in the channel is perturbed into the other value. Each perturbation is independent of all others. Let $\boldsymbol{H} := \mathcal{H}(\nu, \nu^c)$. The value $[1-\boldsymbol{H}]$ bits-per-second is the ***channel capacity*** of this noise-afflicted channel.

[♡8] The noise-process is assumed to be *independent* of the signal-process. In contrast, when the perturbation is highly dependent on the signal, then it is sometimes called ***distortion***.

Encoding/Decoding. Encode using an "***k,n-block-code***"; an injective map $F:\{0,1\}^k \to \{0,1\}^n$. The source text is split into consecutive k-bit blocks. A block $\vec{x} \in \{0,1\}^k$ is encoded to $F(\vec{x}) \in \{0,1\}^n$ and then sent through the channel, where it comes out perturbed to $\vec{\alpha} \in \{0,1\}^n$. The *transmission rate* is thus k/n bits-per-second.

For this example, we fix a radius $r>0$ to determine the decoding map,

$$D_r:\{0,1\}^n \to \{\mathit{Oops}\} \sqcup \{0,1\}^k .$$

We set $D_r(\vec{\alpha})$ to $\vec{z}$ if there is a *unique* $\vec{z}$ with $F(\vec{z}) \in \overline{\mathrm{Bal}}(\vec{\alpha}, r)$; else, set $D_r(\vec{\alpha}) := \mathit{Oops}$.

We can think of the noise as a $\{0,1\}$-independent-process, with $\mathrm{Prob}(1) = \nu$, which is added mod-2 to the signal-process. Suppose we can arrange that the set $\{F(\vec{x}) \mid \vec{x} \in \{0,1\}^k\}$ of codewords, is a strongly r-separated-set. Then

11: *The probability that a block is mis-decoded is the probability, flipping a ν-coin n times, that we get more than r many Heads.*

12: Theorem (Shannon). *Fix a noise-probability $\nu \in [0, \frac{1}{2})$ and let $\boldsymbol{H} := \mathcal{H}(\nu, \nu^c)$. Consider a rate $R < [1-\boldsymbol{H}]$ and an $\varepsilon>0$. Then $\forall_{\text{large}} n$ there exists a k and a code $F:\{0,1\}^k \to \{0,1\}^n$ so that: The F-code transmits bits at faster than R bits-per-second, and with error-rate $< \varepsilon$.* ◊

Proof. Let $\boldsymbol{H}' := \mathcal{H}(\delta, \delta^c)$, where $\delta > \nu$ was chosen so small that

13: $$\delta < \tfrac{1}{2} \quad \text{and} \quad 1 - \boldsymbol{H}' > R .$$

Pick a large n for which

14: $$\tfrac{k}{n} > R, \quad \text{where } k := \lfloor [1 - \boldsymbol{H}'] \cdot n \rfloor .$$

By (10), there is a strongly δn-separated-set $C \subset \{0,1\}^n$ with $\#C \geqslant 2^{[1-\boldsymbol{H}']n}$. So C is big enough to permit an injection $F : \{0,1\}^k \hookrightarrow C$. Courtesy (11), the probability of a decoding error is that of getting more than δn many Heads in flipping a ν-coin n times. Since $\delta > \nu$, the Weak Law of Large Numbers guarantees –once n is large enough– that this probability is less than the given ε. ♦

The information function

We use $\mathsf{P} = (A_1, \ldots)$, $\mathsf{Q} = (B_1, \ldots)$, $\mathsf{R} = (C_1, \ldots)$ for partitions, and $\mathscr{F}, \mathscr{G}$ for fields.

With $\mathfrak{C}$ a (finite or infinite) family of subfields of $\mathscr{X}$, their ***join*** $\bigvee_{\mathscr{G}\in\mathfrak{C}} \mathscr{G}$ is the smallest field $\mathscr{F}$ such that $\mathscr{G} \subset \mathscr{F}$, for each $\mathscr{G} \in \mathfrak{C}$. A partition Q can be interpreted also as a field; namely, the field of unions of its atoms. A join of denumerably many partitions will be interpreted as a field, but a join of *finitely* many, $\mathsf{P}_1 \vee \ldots \vee \mathsf{P}_N$, will be viewed as a partition *or* as a field, depending on context.

For an $A \subset X$, use $\mathbf{1}_A : X \to \{0, 1\}$ for its *indicator function*; $\mathbf{1}_A(x) = 1$ IFF $x \in A$. The ***information function*** of partition P, a map $\mathcal{I}_{\mathsf{P}} : X \to [0, \infty)$, is

15: $$\mathcal{I}_{\mathsf{P}}(\cdot) \;:=\; \sum_{A \in \mathsf{P}} \log\Big(\frac{1}{\mu(A)}\Big) \cdot \mathbf{1}_A(\cdot)\,.$$

It has been defined so that its expectation is the distropy of P.

$$\mathbb{E}(\mathcal{I}_{\mathsf{P}}) \;=\; \int_X \mathcal{I}_{\mathsf{P}}(\cdot)\, d\mu \;=\; \mathcal{H}(\mathsf{P})\,.$$

With respect to a subfield $\mathscr{F}$, let $\mu(A \mid \mathscr{F})$ be the ***conditional probability*** function; that is, the conditional expectation $\mathbb{E}(\mathbf{1}_A \mid \mathscr{F})$. This engenders the ***conditional information function***,

16: $$\mathcal{I}_{\mathsf{P}|\mathscr{F}}(x) \;:=\; \sum_{A \in \mathsf{P}} \log\Big(\frac{1}{\mu(A \mid \mathscr{F})(x)}\Big) \cdot \mathbf{1}_A(x)\,. \quad \text{Its integral}$$
$$\mathcal{H}(\mathsf{P} \mid \mathscr{F}) \;:=\; \int \mathcal{I}_{\mathsf{P}|\mathscr{F}}\, d\mu\,, \quad \text{is the } \textbf{\textit{conditional distropy}} \text{ of P on } \mathscr{F}.$$

Conditioning on a positive-mass set B, let $\mathsf{P}|B$ be the probability vector $A \mapsto \frac{\mu(A \cap B)}{\mu(B)}$. Conditional distropy, when conditioning on a partition, equals

17: $$\mathcal{H}(\mathsf{P} \mid \mathsf{Q}) \;=\; \sum_{B \in \mathsf{Q}} \mathcal{H}(\mathsf{P}|B) \cdot \mu(B) \;=\; \sum_{A \in \mathsf{P},\, B \in \mathsf{Q}} \log\Big(\tfrac{1}{\mu(A \cap B)/\mu(B)}\Big) \cdot \mu(A \cap B)\,.$$

Write $\mathscr{G}_j \nearrow \mathscr{F}$ to indicate that fields $\mathscr{G}_1 \subset \mathscr{G}_2 \subset \ldots$ are nested, and that $\mathrm{Fld}(\bigcup_1^\infty \mathscr{G}_j) = \mathscr{F}$, a.e. The Martingale Convergence Theorem, [20, P. 103], gives (c), below.

18: Conditional-distropy fact. *Consider partitions* P, Q, R *and fields* $\mathscr{F}$ *and* $\mathscr{G}_j$. *Then*

a: $0 \leqslant \mathcal{H}(\mathsf{P} \mid \mathscr{F}) \leqslant \mathcal{H}(\mathsf{P})$, *with equality IFF* $\mathsf{P} \overset{\text{a.e}}{\subset} \mathscr{F}$, *respectively,* $\mathsf{P} \perp \mathscr{F}$.

b: $\mathcal{H}(\mathsf{Q} \vee \mathsf{R} \mid \mathscr{F}) \leqslant \mathcal{H}(\mathsf{Q} \mid \mathscr{F}) + \mathcal{H}(\mathsf{R} \mid \mathscr{F})$.

c: *Suppose* $\mathscr{G}_j \nearrow \mathscr{F}$. *Then* $\mathcal{H}(\mathsf{P} \mid \mathscr{G}_j) \searrow \mathcal{H}(\mathsf{P} \mid \mathscr{F})$.

d: $\mathcal{H}(\mathsf{Q} \vee \mathsf{R}) \;=\; \mathcal{H}(\mathsf{Q} \mid \mathsf{R}) \;+\; \mathcal{H}(\mathsf{R})$.

d': $\mathcal{H}(\mathsf{Q} \vee \mathsf{R}_1 \mid \mathsf{R}_0) \;=\; \mathcal{H}(\mathsf{Q} \mid \mathsf{R}_1 \vee \mathsf{R}_0) + \mathcal{H}(\mathsf{R}_1 \mid \mathsf{R}_0)$. ◊

Imagining our dartboard (1) divided by superimposed partitions Q and R, equality (18d) can interpreted as saying: *'You can efficiently discover where the dart landed in both partitions, by first asking efficient questions about* R, *then –based on where you landed in* R– *asking intelligent questions about* Q.*'*

Entropy of a Process

Consider an transformation $(T : X, \mu)$ and partition $\mathsf{P} = (A_1, A_2, \ldots)$. Each "time" n determines a partition $\mathsf{P}_n := T^n\mathsf{P}$, whose j^{th}-atom is $T^{-n}(A_j)$. The ***process*** T,P refers to how T acts on the subfield $\bigvee_0^\infty \mathsf{P}_n \subset \mathscr{X}$. (An alternative view of a process is as a stationary sequence $V_0, V_1, \ldots$ of random variables $V_n : X \to \mathbb{Z}_+$, where $V_n(x) := j$ because x is in the j^{th}-atom of P_n.)

Write $\mathcal{E}(T,\mathsf{P})$ or $\mathcal{E}^T(\mathsf{P})$ for the **"*entropy* of the T,P process"**. It is the limit of the ***conditional-distropy-numbers***

$$c_n \quad := \quad \mathcal{H}(\mathsf{P}_0 \mid \mathsf{P}_1 \vee \mathsf{P}_2 \vee \ldots \vee \mathsf{P}_{n-1}) .$$

This limit exists since $\mathcal{H}(\mathsf{P}) = c_1 \geqslant c_2 \geqslant \ldots \geqslant 0$.

Define the ***average-distropy-number*** $\frac{1}{n}h_n$, where

$$h_n \quad := \quad \mathcal{H}(\mathsf{P}_0 \vee \mathsf{P}_1 \vee \ldots \vee \mathsf{P}_{n-1}) .$$

Certainly $h_n \;=\; c_n + \mathcal{H}(\mathsf{P}_1 \vee \ldots \vee \mathsf{P}_{n-1}) \;=\; c_n + h_{n-1}$, since T is measure preserving. Induction gives $h_n = \sum_{j=1}^n c_j$. So the Cesàro averages $\frac{1}{n}h_n$ converge to the entropy.

19: Theorem. *The entropy of process $(T,\mathsf{P} : X, \mathscr{X}, \mu)$ equals*

$$\lim_{n\to\infty} \tfrac{1}{n}\mathcal{H}(\mathsf{P}_0 \vee \ldots \vee \mathsf{P}_{n-1}) \;=\; \lim_{n\to\infty} \mathcal{H}(\mathsf{P}_0 \mid \bigvee\nolimits_1^n \mathsf{P}_j) \;=\; \mathcal{H}(\mathsf{P}_0 \mid \bigvee\nolimits_1^\infty \mathsf{P}_j) .$$

Both limits are non-increasing. The entropy $\mathcal{E}^T(\mathsf{P}) \geqslant 0$, with equality IFF $\mathsf{P} \overset{a.e}{\subset} \bigvee_1^\infty \mathsf{P}_j$. And $\mathcal{E}^T(\mathsf{P}) \leqslant \mathcal{H}(\mathsf{P})$, with equality IFF T,P is an independent process. ◊

Generators. We henceforth only discuss *invertible* mpts, that is, when T^{-1} is itself an mpt. Viewing the atoms of P as "letters", then, each $x \in X$ has a T,P-***name*** $\ldots \mathrm{x}_{-1}\,\mathrm{x}_0\,\mathrm{x}_1\,\mathrm{x}_2 \ldots$, where x_n is $\mathsf{P}\langle T^n(x)\rangle$, the P-letter owning $T^n(x)$.

A partition P ***generates*** (the whole field) under $(T : X, \mu)$, if $\bigvee_{-\infty}^\infty T^n\mathsf{P} =_\mu \mathscr{X}$. It turns out[♡9] that P generates IFF P ***separates points***. That is, after deleting a (T-invariant) nullset, distinct points of X have distinct T,P-names.

A finite set $[1\,..\,L]$ of integers, our ***alphabet***, yields the ***shift space*** $X := [1\,..\,L]^{\mathbb{Z}}$ of doubly-infinite sequences $x = (\ldots \mathrm{x}_{-1}\,\mathrm{x}_0\,\mathrm{x}_1 \ldots)$. The ***shift*** $T : X \circlearrowright$ acts on X by

$$T(x) \;:=\; [n \mapsto \mathrm{x}_{n+1}] .$$

Automatically, then, the ***time-zero partition*** P separates points, under the action of the shift. This L-atom partition has $\mathsf{P}\langle x\rangle = \mathsf{P}\langle y\rangle$ IFF $\mathrm{x}_0 = \mathrm{y}_0$. So no matter what shift-invariant measure is put on X, the time-zero partition will generate under the action of T.

[♡9] I am now at liberty to reveal that our X has always been a ***Lebesgue space***, that is, measure-isomorphic to an interval of $\mathbb{R}$ together with countably many point-atoms (points with positive mass). The equivalence of *generating* and *separating* is a technical theorem, due to Rokhlin.

Assuming μ to be Lebesgue is not much of a limitation. For instance, if μ is a finite measure on any Polish space, then μ extends to a Lebesgue measure on the μ-completion of the Borel sets. To not mince words: All spaces are Lebesgue spaces unless you are actively *looking* for trouble...

Time reversibility. A transformation need not be isomorphic to its inverse. Nonetheless, the average-distropy-numbers show that $\mathcal{E}(T^{-1}, \mathsf{P}) = \mathcal{E}(T, \mathsf{P})$; although this is not obvious from the conditioning-definition of entropy. Alternatively,

20:
$$\begin{aligned}\mathcal{H}(\mathsf{P}_0 \mid \textstyle\bigvee_1^n \mathsf{P}_j) &= \mathcal{H}(\mathsf{P}_0 \vee \ldots \vee \mathsf{P}_n) - \mathcal{H}(\mathsf{P}_1 \vee \ldots \vee \mathsf{P}_n) \\ &= \mathcal{H}(\mathsf{P}_{-n} \vee \ldots \vee \mathsf{P}_0) - \mathcal{H}(\mathsf{P}_{-n} \vee \ldots \vee \mathsf{P}_1) = \mathcal{H}(\mathsf{P}_0 \mid \textstyle\bigvee_{-1}^{-n} \mathsf{P}_j).\end{aligned}$$ □

Bernoulli processes. A probability vector $\vec{v} := (v_1, \ldots, v_L)$ can be viewed as a measure on alphabet $[1 \,..\, L]$. Let $\mu_{\vec{v}}$ be the resulting product measure on $X := [1 \,..\, L]^{\mathbb{Z}}$, with T the shift on X and P the time-zero partition. The independent process $(T, \mathsf{P} : X, \mu_{\vec{v}})$ is called, by ergodic theorists, a ***Bernoulli process***. Not necessarily consistently, we tend to refer to the underlying *transformation* as a ***Bernoulli shift***.

The $(\frac{1}{2}, \frac{1}{2})$-Bernoulli and the $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$-Bernoulli have different process-entropies, but perhaps their underlying transformations are isomorphic? Prior to the Kolmogorov-Sinai definition of entropy[♡10] of a *transformation*, this question remained unanswered.

Entropy of a Transformation

The Kolmogorov-Sinai definition of the entropy of an mpt is

$$\mathcal{E}(T) \quad := \quad \sup\{\, \mathcal{E}^T(\mathsf{Q}) \;\big|\; \mathsf{Q} \text{ a partition on } X \}\,.$$

Certainly entropy is an isomorphism invariant —but is it useful? After all, the supremum of *distropies* of partitions is always infinite (on non-atomic spaces) and one might fear that the same holds for entropies. The key observation (restated in (23c) and proved below) was this, from [4, Kol 1958] and [5, Sinai 1959].

21: Kolmogorov-Sinai theorem. *If* P *generates under* T*, then* $\mathcal{E}(T) = \mathcal{E}(T, \mathsf{P})$. ◊

Thereupon the $(\frac{1}{2}, \frac{1}{2})$ and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ Bernoulli-shifts are *not* isomorphic, since their respective entropies are $\log(2) \neq \log(3)$.

Wolfgang Krieger later proved a converse to the Kolmogorov-Sinai theorem.

22: Krieger Generator Thm (1970). *Suppose* T *ergodic. If* $\mathcal{E}(T) < \infty$*, then* T *has a generating partition. Indeed, letting* K *be the smallest integer* $K > \mathcal{E}(T)$*, there is a* K*-atom generator.*[♡11] ◊

Proof. See Rudolph [21], or [[de la Rue, §5.1]], where Krieger's theorem is stated in terms of joinings. ♦

[♡10] This is sometimes called ***measure(-theoretic) entropy*** or (perhaps unfortunately) ***metric entropy***, to distinguish it from topological entropy. Tools known prior to entropy, such as *spectral* properties, did *not* distinguish the two Bernoulli-shifts; see [[Lemanczyk]] for the definitions.

[♡11] It is an easier result, undoubtedly known much earlier, that every ergodic T has a *countable* generating partition —possibly of ∞-distropy.

Entropy is continuous. Given *ordered* partitions $\mathsf{Q} = (B_1, \dots)$ and $\mathsf{Q}' = (B'_1, \dots)$, extend the shorter by null-atoms until $|\mathsf{Q}| = |\mathsf{Q}'|$. Let $Fat := \bigsqcup_j [B_j \cap B'_j]$; this set should have mass close to 1 if Q and Q' are almost the same partition. Define a new partition

$$\mathsf{Q} \vartriangle \mathsf{Q}' \quad := \quad \{Fat\} \sqcup \{B_i \cap B'_j \mid \text{with } i \neq j\}.$$

(In other words, take $\mathsf{Q} \vee \mathsf{Q}'$ and coalesce, into a single atom, all the $B_k \cap B'_k$ sets.) Topologize the space of partitions by saying[♡12] that $\mathsf{Q}^{(L)} \to \mathsf{Q}$ when $\mathcal{H}(\mathsf{Q} \vartriangle \mathsf{Q}^{(L)}) \to 0$. Then (23b) says that process-entropy varies continuously with varying the partition.

23: Lemma. *Fix a mpt $(T : X, \mu)$. For partitions $\mathsf{P}, \mathsf{Q}, \mathsf{Q}'$, define $\mathsf{R} := \mathsf{Q} \vartriangle \mathsf{Q}'$ and let $\delta := \mathcal{H}(\mathsf{R})$. Then*

a: $\left|\mathcal{H}(\mathsf{Q}) - \mathcal{H}(\mathsf{Q}')\right| \leqslant \delta$. *(Distropy varies continuously with the partition.)*

b: $\left|\mathcal{E}^T(\mathsf{Q}) - \mathcal{E}^T(\mathsf{Q}')\right| \leqslant \delta$. *(Process-entropy varies continuously with the partition.)*

c: For all partitions $\mathsf{Q} \subset \mathrm{Fld}(T,\mathsf{P})$: $\mathcal{E}^T(\mathsf{Q}) \leqslant \mathcal{E}^T(\mathsf{P})$. ◊

Proof of (a). Evidently $\mathsf{Q}' \vee \mathsf{R} = \mathsf{Q}' \vee \mathsf{Q} = \mathsf{Q} \vee \mathsf{R}$. So $\mathcal{H}(\mathsf{Q}') \leqslant \mathcal{H}(\mathsf{Q} \vee \mathsf{R}) \leqslant \mathcal{H}(\mathsf{Q}) + \delta$. ♦

Proof of (b). As above, $\mathcal{H}(\bigvee_1^N \mathsf{Q}'_j) \leqslant \mathcal{H}(\bigvee_1^N \mathsf{Q}_j) + \mathcal{H}(\bigvee_1^N \mathsf{R}_j)$. Sending $N \to \infty$ gives $\mathcal{E}^T(\mathsf{Q}') \leqslant \mathcal{E}^T(\mathsf{Q}) + \mathcal{E}^T(\mathsf{R})$. Finally, $\mathcal{E}^T(\mathsf{R}) \leqslant \mathcal{H}(\mathsf{R})$ and so $\mathcal{E}^T(\mathsf{Q}') \leqslant \mathcal{E}^T(\mathsf{Q}) + \delta$. ♦

Proof of (c). Let $K := |\mathsf{Q}|$. Then there is a sequence of K-set partitions $\mathsf{Q}^{(L)} \to \mathsf{Q}$ with $\mathsf{Q}^{(L)} \preccurlyeq \bigvee_{-L}^{L} \mathsf{P}_\ell$. By above, $\mathcal{E}^T(\mathsf{Q}^{(L)}) \to \mathcal{E}^T(\mathsf{Q})$, so showing that $\mathcal{E}^T(\bigvee_{-L}^{L} \mathsf{P}_\ell) \overset{?}{\leqslant} \mathcal{E}^T(\mathsf{P})$ will suffice. Note that

$$h_N := \mathcal{H}\Big(\bigvee_{n=0}^{N-1} T^n\big(\bigvee_{-L}^{L} \mathsf{P}_\ell\big)\Big) = \mathcal{H}\Big(\bigvee_{j=-L}^{N-1+L} \mathsf{P}_j\Big).$$

So $\frac{1}{N} h_N \leqslant \frac{1}{N}\mathcal{H}\big(\bigvee_0^{N-1} \mathsf{P}_j\big) + \frac{1}{N} \cdot 2L \cdot \mathcal{H}(\mathsf{P})$. Now send $N \to \infty$. ♦

Entropy is not continuous. The most common topology placed on the space, Ω, of mpts is the *coarse topology*[♡13] that Halmos discusses in his "little red book", [14].

The Rokhlin lemma [21, P. 33] implies that the isomorphism-class of *each* ergodic mpt is *dense* in Ω, (e.g, see [14, P. 77]) disclosing that the $S \mapsto \mathcal{E}(S)$ map is exorbitantly discontinuous.

Indeed, the failure happens already for process-entropy with respect to a fixed partition. A Bernoulli process T,P has positive entropy. Take mpts $S_n \to T$, each isomorphic to an irrational rotation. Then each $\mathcal{E}(S_n, \mathsf{P})$ is zero, as shown in the later section on *Determinism and Zero-entropy*.

[♡12] On the set of ordered K-set partitions (with K fixed) this convergence is the same as: $\mathsf{Q}^{(L)} \to \mathsf{Q}$ when $\mu(Fat(\mathsf{Q}^{(L)}, \mathsf{Q})) \to 1$.

An alternative approach is the ***Rokhlin metric***, $\mathrm{Dist}(\mathsf{P}, \mathsf{Q}) := \mathcal{H}(\mathsf{P} \mid \mathsf{Q}) + \mathcal{H}(\mathsf{Q} \mid \mathsf{P})$, which has the advantage of working for *unordered* partitions.

[♡13] I.e, $S_n \to T$ IFF $\forall A \in \mathscr{X} : \mu(S_n^{-1}(A) \vartriangle T^{-1}(A)) \to 0$; this is a metric-topology, since our probability space is countably generated. This can be restated in terms of the unitary operator U_T on $\mathbb{L}^2(\mu)$, where $U_T(f) := f \circ T$. Namely, $S_n \to T$ in the coarse topology IFF $U_{S_n} \to U_T$ in the strong operator topology.

Further results. When $\mathscr{F}$ is a T-invariant subfield, agree to use $T\restriction_{\mathscr{F}}$ for "T restricted to $\mathscr{F}$", which is a factor (see Glossary) of T. Transformations T and S are ***weakly isomorphic*** if each is isomorphic to a factor of the other.

The foregoing entropy tools make short shrift of the following.

24: Entropy lemma. *Consider T-invariant subfields $\mathscr{G}_j$ and $\mathscr{F}$.*

a: *Suppose $\mathscr{G}_j \nearrow \mathscr{F}$. Then $\mathcal{E}(T\restriction_{\mathscr{G}_j}) \nearrow \mathcal{E}(T\restriction_{\mathscr{F}})$. In particular, $\mathscr{G} \subset \mathscr{F}$ implies that $\mathcal{E}(T\restriction_{\mathscr{G}}) \leqslant \mathcal{E}(T\restriction_{\mathscr{F}})$, so entropy is an invariant of weak-isomorphism.*

b: $\mathcal{E}(T\restriction_{\mathscr{G}_1\vee\mathscr{G}_2\vee\ldots}) \leqslant \sum_j \mathcal{E}(T\restriction_{\mathscr{G}_j})$. *And* $\mathcal{E}(T, \mathsf{Q}_1\vee\mathsf{Q}_2\vee\ldots) \leqslant \sum_j \mathcal{E}(T, \mathsf{Q}_j)$.

c: *For mpts $(S_j : Y_j, \nu_j)$:* $\mathcal{E}(S_1 \times S_2 \times \ldots) = \sum_j \mathcal{E}(S_j)$.

d: $\mathcal{E}(T^{-1}) = \mathcal{E}(T)$. *More generally,* $\mathcal{E}(T^n) = |n| \cdot \mathcal{E}(T)$. ◊

E.g: *Meshalkin's map.* In the wake of Kolmogorov's 1958 entropy paper, for two Bernoulli-shifts to be isomorphic one now knew that they had to have equal entropies. Meshalkin provided the first non-trivial example [44], in 1959.

Let $S : Y\circlearrowright$ be the Bernoulli-shift over the "letter" alphabet $\{\mathtt{E}, \mathtt{D}, \mathtt{P}, \mathtt{N}\}$, with probability distribution $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. The letters $\mathtt{E}, \mathtt{D}, \mathtt{P}, \mathtt{N}$ stand for *Even, oDd, Positive, Negative*, and will be used to describe the code (isomorphism) between the processes.

Use $T : X\circlearrowright$ for the Bernoulli-shift over "digit" alphabet $\{0, +1, -1, +2, -2\}$, with probability distribution $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$. Both distributions, $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$, have distropy log(4).

After deleting invariant nullsets from X and Y, we will construct a measure-preserving isomorphism $\psi : X \to Y$ so that $T \circ \psi = \psi \circ S$.

The Code. In X, consider this point x:

$$\ldots \quad \mathbf{0} \quad \mathbf{0} \quad \mathbf{0} \quad -1 \quad \mathbf{0} \quad \mathbf{0} \quad +1 \quad +2 \quad -1 \quad +1 \quad \mathbf{0} \quad \ldots$$

Regard each **0** as a left-parenthesis, and each non-zero as a right-parenthesis. Link them according to the legal way of matching parentheses, as shown in the top row, below:

$$\begin{array}{cccccccccccc}
\multicolumn{12}{c}{\rule{0.6\textwidth}{0.4pt}} \\
 & \multicolumn{10}{c}{\rule{0.5\textwidth}{0.4pt}} & \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & -1 & \mathbf{0} & \mathbf{0} & +1 & +2 & -1 & +1 & \mathbf{0} \\
\mathtt{P} & \mathtt{N} & \mathtt{N} & \mathtt{D} & \mathtt{P} & \mathtt{P} & \mathtt{D} & \mathtt{E} & \mathtt{D} & \mathtt{D} & ?
\end{array}$$

The leftmost **0** is linked to the rightmost +1, as indicated by the longest-overbar. The left/right-parentheses form a $(\frac{1}{2},\frac{1}{2})$-random-walk. Since this random walk is recurrent, we know that every position in x will be linked (except for a nullset of points x).

Below each **0**, write "P" or "N" as the **0** is linked to a *positive* or *negative* digit. And below the other digits, write "E" or "D" as the digit is *even* or *odd*. So the upper name in X is mapped to the lower name, a point $y \in Y$.

This map $\psi: X \to Y$ carries the upstairs $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ distribution to $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, downstairs. It takes some arguing to show that independence is preserved.

The inverse map, ψ^{-1}, views D and E as right-parentheses, and P and N as left. Above D, write the odd digit **+1** or **-1**, as this D is linked to Positive or Negative. □

E.g: *Markov Shifts.* A Bernoulli process T,P has independence, $\mathsf{P}_{(-\infty \,..\, 0]} \perp \mathsf{P}_1$, whereas a ***Markov process*** is a bit less aloof:

> *The infinite Past* $\mathsf{P}_{(-\infty \,..\, 0]}$ *doesn't provide any more information about Tomorrow than Today did.*

That is, the conditional distribution $\mathsf{P}_1|\mathsf{P}_{(-\infty \,..\, 0]}$ equals $\mathsf{P}_1|\mathsf{P}_0$. Equivalently,

25: $$\mathcal{H}(\mathsf{P}_1 \mid \mathsf{P}_0) \;=\; \mathcal{H}(\mathsf{P}_1 \mid \mathsf{P}_{(-\infty \,..\, 0]}) \overset{\text{note}}{=\!=} \mathcal{E}(T, \mathsf{P})\,.$$

The simplest non-trivial Markov process $(T,\mathsf{P} : X, \mu)$ is over a two-letter alphabet {a, b}, and has transition graph (26), for some choice of transition probabilities s and c. The

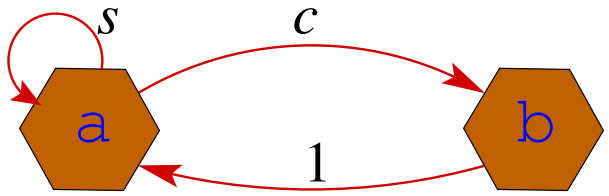

FIG. 26: *Call the transition probabilities* $s := Prob(a \to a)$ *for stay, and* $c := Prob(a \to b)$ *for change. These are non-negative reals, and* $s+c = 1$.

graph's Markov matrix is

$$\mathsf{M} \;=\; [m_{i,j}]_{i,j} \;=\; \begin{bmatrix} s & c \\ 1 & 0 \end{bmatrix},$$

where $c = 1 - s$, and $m_{i,j}$ denotes the probability of going from state i to state j.

If Today's distribution on the two states is the probability-vector $\vec{v} := \begin{bmatrix} p_a & p_b \end{bmatrix}$, then Tomorrow's is the product $\vec{v} \cdot \mathsf{M}$. So a *stationary* process needs $\vec{v}\mathsf{M} = \vec{v}$. This equation has the unique solution $p_a = \frac{1}{1+c}$ and $p_b = \frac{c}{1+c}$. An example of computing the probability of a word (or cylinder set; see [[Petersen, 5.1]]) in the process, is

$$\mu_s(\mathsf{baaaba}) \;=\; p_b \cdot m_{ba} \cdot m_{aa} \cdot m_{aa} \cdot m_{ab} \cdot m_{ba} \;=\; \tfrac{c}{1+c} \cdot 1 \cdot s \cdot s \cdot c \cdot 1\,.$$

The subscript on μ_s indicates the dependence on the transition probabilities; let's also mark the mpt and call it T_s. Using (25), the entropy of our Markov map is

27: $$\mathcal{E}(T_s) \;=\; p_a \cdot \mathcal{H}(s,c) + p_b \cdot \cancelto{0}{\mathcal{H}(1,0)} \;=\; \tfrac{-1}{1+c} \cdot [s \log(s) + c \log(c)]\,.$$

□

Determinism and Zero-entropy

Irrational rotations have zero-entropy; let's reveal this in two different ways.

Equip $X := [0,1)$ with "length" (Lebesgue) measure and wrap it into a circle. With "$\oplus$" denoting addition mod-1, have $T{:}X\circlearrowright$ be the rotation $T(x) := x \oplus \alpha$, where the rotation number α is irrational. Pick distinct points $y_0, z_0 \in X$, and let P be the partition whose two atoms are the intervals $[y_0, z_0)$ and $[z_0, y_0)$, wrapping around the circle.

The T-orbit of each point x is dense[♡14] in X. In particular, y_0 has dense orbit, so P separates points –hence generates– under T. Our goal, thus, is $\mathcal{E}(T,\mathsf{P}) \stackrel{?}{=} 0$.

E.g: *Rotations are deterministic.* The forward T-orbit of each point is dense. This is true for y_0, and so the *backward* T,P-name of each x actually tells us which point x is. I.e, $\mathsf{P} \subset \bigvee_{-\infty}^{-1} T^n\mathsf{P}$, which is our definition of "process T,P is ***deterministic***". Our P being finite, this determinism implies that $\mathcal{E}(T,\mathsf{P})$ is zero, by (19).

Counting names in a rotation. The $\mathsf{P}_0 \vee \ldots \vee \mathsf{P}_{n-1}$ partition places n translates of points y_0 and of z_0, cutting the circle into at most $2n$ intervals. Thus $\mathcal{H}(\mathsf{P}_0 \vee \ldots \vee \mathsf{P}_{n-1}) \leqslant \log(2n)$. And $\frac{1}{n}\log(2n) \to 0$.

Alternatively, the below SMB-theorem implies, for an ergodic process T,P, that the number of length-n names is approximately $2^{\mathcal{E}(T,\mathsf{P})\cdot n}$; this, after discarding small mass from the space. But the growth of $n \mapsto 2n$ is sub-exponential and so, for our rotation, $\mathcal{E}(T,\mathsf{P})$ must be zero. □

28: Shannon-McMillan-Breiman[♡15] Theorem (SMB-Thm). *Set $\boldsymbol{E} := \mathcal{E}(T, \mathsf{P})$, where tuple $(T,\mathsf{P} : X,\mu)$ is an ergodic process. Then the average information function*

28a: $$\frac{1}{n} \cdot \mathcal{I}_{\mathsf{P}_{[0\,..\,n)}}(x) \stackrel{n\to\infty}{\longrightarrow} \boldsymbol{E}, \quad \text{for a.e } x \in X.$$

The functions $f_n := \mathcal{I}_{\mathsf{P}_{[0\,..\,n)}}$ converge to the constant function $\boldsymbol{E}$ both in the $\mathbb{L}^1$-norm and in probability. ◊

Consequences. Recall that $\mathsf{P}_{[0\,..\,n)}$ means $\mathsf{P}_0 \vee \mathsf{P}_1 \vee \ldots \vee \mathsf{P}_{n-1}$, where $\mathsf{P}_j := T^j\mathsf{P}$. As usual, $\mathsf{P}_{[0\,..\,n)}\langle x\rangle$ denotes the $\mathsf{P}_{[0\,..\,n)}$-atom owning x.

Having deleted a nullset, we can restate (28a) to now say that $\forall\varepsilon, \forall x, \forall_{\text{large}} n$:

28b: $$1/2^{[\boldsymbol{E}+\varepsilon]n} \leqslant \mu(\mathsf{P}_{[0\,..\,n)}\langle x\rangle) \leqslant 1/2^{[\boldsymbol{E}-\varepsilon]n}.$$

This has the following consequence. Fixing a number $\delta > 0$, we consider any set with $\mu(B) \geqslant \delta$ and count the number of n-names of points in B. The SMB-Thm implies

28c: $$\forall\varepsilon,\ \forall_{\text{large}} n,\ \forall B \stackrel{\mu}{\geqslant} \delta : \quad \big|\{n\text{-names in } B\}\big| \geqslant 2^{[\boldsymbol{E}-\varepsilon]n}.$$

[♡14] Fix an $\varepsilon{>}0$ and an $N > 1/\varepsilon$. Points $x, T(x), \ldots, T^N(x)$ have some two at distance less than $\frac{1}{N}$; say, $\text{Dist}(T^i(x), T^j(x)) < \varepsilon$, for some $0 \leqslant i < j \leqslant N$. Since T is an isometry, $\varepsilon > \text{Dist}(x, T^k(x)) > 0$, where $k := j - i$. So the T^k-orbit of x is ε-dense.

[♡15] In engineering circles, this is called the Almost-everywhere equi-partition theorem.

E.g: *Rank-1 has zero-entropy.* There are several equivalent definitions for "rank-1 transformation", several of which are discussed in the introduction of [28]. (See [13, Chap. 6] and [47] and [27] for examples of stacking constructions.)

A ***rank-1 transformation*** $(T : X, \mu)$ admits a generating partition P and a sequence of Rokhlin stacks $S_n \subset X$, with heights going to ∞, and with $\mu(S_n) \to 1$. Moreover, each of these Rokhlin stacks is P-monochromatic, that is, each level of the stack lies entirely in some atom of P.

Taking a stack of some height $2n$, let $B{=}B_n$ be the union of the bottom n levels of the stack. There are at most n many length-n names starting in B_n, by monochromaticity. Finally, $\mu(B_n)$ is almost $\frac{1}{2}$, so is certainly larger than $\delta := \frac{1}{3}$. Thus (28c) shows that our rank-1 T has zero entropy. □

Cautions on determinism's relation to zero-entropy. A finite-valued process T,P has zero-entropy iff $\mathsf{P} \subset \bigvee_{-\infty}^{-1} \mathsf{P}_j$. Iterating gives

$$\bigvee_{0}^{\infty} \mathsf{P}_j \subset \bigvee_{-\infty}^{-1} \mathsf{P}_j, \quad \text{i.e, the future is measurable with respect to the past.}$$

This was the case with the rotation, where a point's past uniquely identified the point, thus telling us its future.

While determinism and zero-entropy mean the same thing for finite-valued processes, this fails catastrophically for real-valued (i.e, continuum-valued) processes, as shown by an example of the author's. A stationary real-valued process $\mathbf{V} = \ldots V_{-1}\ V_0\ V_1\ V_2\ \ldots$ is constructed in [39] which is simultaneously

strongly deterministic : The two valuess V_0, V_1 determine all of $\mathbf{V}$, future and past.

and ***non-consecutively independent***. This latter means that for each bi-infinite increasing integer sequence $\{n_j\}_{j=-\infty}^{\infty}$ with no consecutive pair (always $1 + n_j < n_{j+1}$), then the list of random variables $\ldots V_{n_{-1}}\ V_{n_0}\ V_{n_1}\ V_{n_2}\ \ldots$ is an independent process.

Restricting the random variables to be *countably*-valued, how much of the example survives? Joint work with Kalikow, [40], produced a countably-valued stationary $\mathbf{V}$ which is non-consecutively independent as well as deterministic. (Strong determinism is ruled out, due to cardinality considerations.) A side-effect of the construction is that $\mathbf{V}$'s time-reversal $n \mapsto V_{-n}$ is **not** deterministic.

Pinsker Field. Define a collection of sets (the script Z is for "zero")

29: $$\mathscr{Z} = \mathscr{Z}_T = \mathscr{Z}(T) := \left\{ D \in \mathscr{X} \;\middle|\; \mathcal{E}(T, (D, D^c)) = 0 \right\}.$$

Courtesy (24b), $\mathscr{Z}$ is a T-invariant *field*, and

29′: $$\forall \mathsf{Q} \subset \mathscr{Z} : \quad \mathcal{E}(T, \mathsf{Q}) = 0\,.$$

The ***Pinsker field*** of T[♡16] is this $\mathscr{Z}$. It is maximal with respect to (29′). Unsurprisingly, the ***Pinsker factor*** $T \restriction_{\mathscr{Z}}$ has zero entropy, that is, $\mathcal{E}(T \restriction_{\mathscr{Z}}) = 0$.

The asymptotic past of the T,P process is called its ***tail field***, where

$$\text{Tail}(T,\mathsf{P}) \;:=\; \bigcap_{L=1}^{\infty} \bigvee_{j=-\infty}^{-L} \mathsf{P}_j \,.$$

30**:** Theorem (Pinsker). *Suppose* P *is a generating partition for an ergodic* T*. Then Tail*(T,P) *equals* $\mathscr{Z}_T$*. In particular, all generating partitions for* T *have the same tail field. And the* ***future field*** *of* T,P*, which is* $\mathscr{Z}(T^{-1}) \overset{\text{note}}{=\!=} \mathscr{Z}(T)$*, equals its tail field.* ◊

Ornstein theory

In 1970, Don Ornstein solved the long-standing problem of showing that entropy was a complete isomorphism-invariant of Bernoulli transformations; that is, that two independent processes with same entropy necessarily have the same underlying transformation. (Earlier, Sinai had shown that two such Bernoulli maps were ***weakly isomorphic***, that is, each isomorphic to a factor of the other.)

Ornstein introduced the notion of a process being ***finitely determined***, see [46] for a definition, proved that a transformation T was Bernoulli IFF it had a finitely-determined generator IFF every partition was finitely-determined with respect to T, and showed that entropy completely classified the finitely-determined processes upto isomorphism.

The Pinsker-Field and K-automorphisms

Said differently, the zero-e trns are those whose Pinsker-field is everything. TBW • • •

[♡16] Traditionally, this called the *Pinsker algebra* where, in this context, "algebra" is understood to mean "σ-algebra".

Topological Entropy

Adler, Konheim and McAndrew, in 1965, published the first definition of *topological entropy* in the eponymous article [32]. Here, $T{:}X\circlearrowright$ is a continuous self-map of a compact topological space. The role of atoms is played by open sets. Instead of a finite partition, one uses a finite[♡17] ***open-cover*** $\mathscr{V} = \{U_j\}_{j=1}^{L}$, i.e each ***patch*** U_j is open, and their union $\bigcup(\mathscr{V}) = X$. (Henceforth, '*cover*' means "open cover".)

Let $\mathsf{Card}(\mathscr{V})$ be the minimum cardinality over all subcovers.

$$\mathsf{Card}(\mathscr{V}) \;:=\; \mathrm{Min}\,\{{}^{\#}\mathscr{V}' \;\big|\; \mathscr{V}' \subset \mathscr{V} \text{ and } \bigcup(\mathscr{V}') = X\}\,, \text{ and let}$$
$$\mathcal{H}(\mathscr{V}) \;=\; \mathcal{H}_{\mathrm{top}}(\mathscr{V}) \;:=\; \log(\mathsf{Card}(\mathscr{V}))\,.$$

Analogous to the definitions for partitions, we prescribe

$$\mathscr{V} \vee \mathscr{W} \;:=\; \{V \cap W \mid V \in \mathscr{V} \text{ and } W \in \mathscr{W}\}\,;$$
$$T\mathscr{V} \;:=\; \{T^{-1}(U) \mid U \in \mathscr{V}\} \quad \text{and} \quad \mathscr{V}_{[0\,..\,n)} \;:=\; \mathscr{V}_0 \vee \mathscr{V}_1 \vee \ldots \vee \mathscr{V}_{n-1}\,;$$
$$\mathscr{W} \succcurlyeq \mathscr{V}, \text{ if each } \mathscr{W}\text{-patch is a subset of some } \mathscr{V}\text{-patch.}$$

The $T,\mathscr{V}$***–entropy*** is

31: $$\mathcal{E}^{T}(\mathscr{V}) \;=\; \mathcal{E}(T,\mathscr{V}) \;=\; \mathcal{E}_{\mathrm{top}}(T,\mathscr{V}) \;:=\; \limsup_{n\to\infty} \tfrac{1}{n}\cdot\mathcal{H}_{\mathrm{top}}(\mathscr{V}_{[0\,..\,n)})\,.$$

And the ***topological entropy*** of T is

32: $$\mathcal{E}_{\mathrm{top}}(T) \;:=\; \sup_{\mathscr{V}} \mathcal{E}_{\mathrm{top}}(T,\mathscr{V})\,, \quad \text{taken over all open covers } \mathscr{V}.$$

Thus $\mathcal{E}_{\mathrm{top}}$ counts, in some sense, the growth rate in the number of T-orbits of length n.

Evidently, topological entropy is an isomorphism invariant. Two continuous maps $T{:}X\to X$ and $S{:}Y\to Y$ are ***topologically conjugate*** (as *isomorphism* is called in this category) if there exists a homeomorphism $\psi{:}X\to Y$ with $\psi T = S\psi$.

33: Subadditive Lemma. *Consider a sequence* $\mathbf{s} = (s_\ell)_1^\infty \subset [-\infty,\infty]$ *satisfying* $s_{k+\ell} \leqslant s_k + s_\ell$, *for all* $k,\ell \in \mathbb{Z}$. *Then the following limit exists in* $[-\infty,\infty]$, *and* $\lim_{n\to\infty} \frac{s_n}{n} = \inf_n \frac{s_n}{n}$. ◊

Topological entropy, or "top-ent" for short, satisfies many of the relations of measure-entropy.

34: Lemma.

a: $\mathscr{V} \preccurlyeq \mathscr{W}$ *implies* $\mathcal{H}(\mathscr{V}) \leqslant \mathcal{H}(\mathscr{W})$ *and* $\mathcal{E}(T,\mathscr{V}) \leqslant \mathcal{E}(T,\mathscr{W})$.

b: $\mathcal{H}(\mathscr{V} \vee \mathscr{W}) \leqslant \mathcal{H}(\mathscr{V}) + \mathcal{H}(\mathscr{W})$.

c: $\mathcal{H}(T(\mathscr{V})) \leqslant \mathcal{H}(\mathscr{V})$, *with equality if* T *is surjective. Also,* $\mathcal{E}(T,\mathscr{V}) \leqslant \mathcal{H}(\mathscr{V})$,

[♡17] Because we only work on a compact space, we can omit "finite". Some generalizations of topological entropy to non-compact spaces require that only *finite* open-covers be used; see [37].

d: *In* (31), *the* $\lim_{n\to\infty} \frac{1}{n}\cdot\mathcal{H}(\mathscr{V}_{[0\,..\,n)})$ *exists.*

e: *Suppose* T *is a homeomorphism. Then* $\mathcal{E}(T^{-1},\mathscr{V}) = \mathcal{E}(T,\mathscr{V})$, *for each cover* $\mathscr{V}$. *Consequently,* $\mathcal{E}_{\text{top}}(T^{-1}) = \mathcal{E}_{\text{top}}(T)$.

f: *Suppose* $\mathfrak{C}$ *is a collection of covers such that: For each cover* $\mathscr{W}$, *there exists a* $\mathscr{V}\in\mathfrak{C}$ *with* $\mathscr{V}\succcurlyeq\mathscr{W}$. *Then* $\mathcal{E}_{\text{top}}(T)$ *equals the supremum of* $\mathcal{E}_{\text{top}}(T,\mathscr{V})$, *just taken over those* $\mathscr{V}\in\mathfrak{C}$.

g: *For all* $\ell\in\mathbb{N}$: $\mathcal{E}_{\text{top}}(T^\ell) = \ell\cdot\mathcal{E}_{\text{top}}(T)$. ◊

Proof of (c). Let $\mathscr{C}\preccurlyeq\mathscr{V}$ be a min-cardinality subcover. Then $T\mathscr{C}$ is a subcover of $T\mathscr{V}$. So $\text{Card}(T\mathscr{V}) \leqslant |T\mathscr{C}| = |\mathscr{C}|$.

As for entropy, inequality (b) and the foregoing give $\mathcal{H}(\mathscr{V}_{[0\,..\,n)}) \leqslant \mathcal{H}(\mathscr{V})\cdot n$. ♦

Proof of (d). Set $s_n := \mathcal{H}(\mathscr{V}_{[0\,..\,n)})$. Then $s_{k+\ell} \leqslant s_k + \mathcal{H}(T^k(\mathscr{V}_{[0\,..\,\ell)})) \leqslant s_k + s_\ell$, by (b) and (c), and so the Subadditive Lemma (33), applies. ♦

Proof of (g). WLOG, $\ell = 3$. Given $\mathscr{V}$ a cover, triple it to $\widehat{\mathscr{V}} := \mathscr{V}\cap T\mathscr{V}\cap T^2\mathscr{V}$; so $\bigvee_{j\in[0\,..\,N)}[T^3]^j(\widehat{\mathscr{V}}) = \bigvee_{i\in[0\,..\,3N)} T^i(\mathscr{V})$. Thus $\mathcal{H}(T^3,\widehat{\mathscr{V}},N) = \mathcal{H}(T,\mathscr{V},3N)$, extending notation. Part (d) and sending $N\to\infty$, gives $\mathcal{E}(T^3,\widehat{\mathscr{V}}) = 3\cdot\mathcal{H}(T,\mathscr{V})$.

Lastly, take covers such that $\mathcal{E}(T^3,\mathscr{C}^{(k)})\to\mathcal{E}_{\text{top}}(T^3)$ and $\mathcal{E}(T,\mathscr{D}^{(k)})\to\mathcal{E}_{\text{top}}(T)$, as $k\to\infty$. Define $\mathscr{V}^{(k)} := \mathscr{C}^{(k)}\vee\mathscr{D}^{(k)}$. Apply the above to $\mathscr{V}^{(k)}$, then send $k\to\infty$. ♦

Using a metric

From now on, *our space is a compact metric space* (X,d). Dinaburg [33], and Bowen [34],[35], gave alternative, equivalent, definitions of top-ent, in the compact metric-space case, that are often easier to work with than covers. Bowen gave a definition also when X is not compact,[♡18] see [35] and [22, chap. 7].

Metric preliminaries. An ***ε-ball-cover*** comprises finitely many balls, all of radius ε. Since our space is compact, every cover $\mathscr{V}$ has a ***Lebesgue number*** $\varepsilon>0$. *I.e, for each* $z\in X$, *the* $\text{Bal}(z,\varepsilon)$ *lies entirely inside at least one* $\mathscr{V}$*-patch.* (In particular, there is an ε-ball-cover which *refines* $\mathscr{V}$.) Let $\text{Leb}(\mathscr{V})$ be the supremum of the Lebesgue numbers. Courtesy (34f) we can

35: *Fix a "universal" list* $\mathscr{V}^{(1)}\preccurlyeq\mathscr{V}^{(2)}\preccurlyeq\ldots$, *with* $\mathscr{V}^{(k)}$ *a* $\frac{1}{k}$*-ball-cover. For every* $T{:}X\to X$, *then, the* $\lim_k \mathcal{E}(T,\mathscr{V}^{(k)})$ *computes* $\mathcal{E}_{\text{top}}(T)$.

[♡18] When X is not compact, the definitions need not coincide; e.g [37]. And topologically-equivalent metrics, but which are not uniformly equivalent, may give the same T different entropies, [22, P. 171].

An ε-microscope. Three notions are useful in examining a metric space (X, m) at scale ε. Subset $A \subset X$ is an ***ε-separated-set***, if $\mathsf{m}(z, z') \geqslant \varepsilon$ for all distinct $z, z' \in A$. Subset $F \subset X$ is ***ε-spanning*** if $\forall x \in X, \exists z \in F$ with $\mathsf{m}(x, z) < \varepsilon$.

Lastly, a cover $\mathscr{V}$ is ***ε-small*** if $\mathrm{Diam}(U) < \varepsilon$, for each $U \in \mathscr{V}$.

You take the High Road and I'll take the Low Road. There are several routes to computing top-ent, some via maximization, others, minimization. Our foregoing discussion computed $\mathcal{E}_{\mathrm{top}}(T)$ by a family of ***sizes*** $f_k(n) = f_k^T(n)$, depending on a parameter k which specifies the fineness of scale. (In (35), this k is an integer; in the original definition, an open cover.) Define two numbers:

36: $$\overline{\mathsf{L}}^f(k) := \limsup_{n\to\infty} \quad \text{and} \quad \underline{\mathsf{L}}^f(k) := \liminf_{n\to\infty} \quad \text{of} \quad \frac{1}{n}\log(f_k(n)).$$

Finally, let $\mathcal{E}^f(T) := \sup_k \overline{\mathsf{L}}^f(k)$. If the limit exists in (36) then we write $\mathsf{L}^f(k)$ for the common value.

The A-K-M definition used the size $f_{\mathscr{V}}(n) := \mathsf{Card}(\mathscr{V}_{[0\,..\,n)})$, where

$$\mathsf{Card}(\mathscr{W}) := \text{Minimum cardinality of a subcover from } \mathscr{W}.$$

Here are three metric-space sizes $f_\varepsilon(n)$:

$$\begin{aligned}
\mathsf{Sep}(n,\varepsilon) &:= \textit{Maximum} \text{ cardinality of a } \mathsf{d}_n\text{-}\varepsilon\text{-separated set}.\\
\mathsf{Spn}(n,\varepsilon) &:= \text{Minimum cardinality of a } \mathsf{d}_n\text{-}\varepsilon\text{-spanning set}.\\
\mathsf{Cov}(n,\varepsilon) &:= \text{Minimum cardinality of a } \mathsf{d}_n\text{-}\varepsilon\text{-small cover}.
\end{aligned}$$

These use a list $(\mathsf{d}_n)_{n=1}^\infty$ of progressively finer metrics on X, where

$$\mathsf{d}_N(x,y) := \max_{j\in[0\,..\,N)} \mathsf{d}(T^j(x), T^j(y)).$$

37: All-Roads-lead-to-Rome Theorem. *Fix ε and let $\mathscr{W}$ be any d-ε-small cover. Then*

i: $\forall n$: $\mathsf{Cov}(n, 2\varepsilon) \leqslant \mathsf{Spn}(n,\varepsilon) \leqslant \mathsf{Sep}(n,\varepsilon) \leqslant \mathsf{Card}(\mathscr{W}_{[0\,..\,n)})$.

ii: *Take a cover $\mathscr{V}$ and a $\delta < \mathrm{Leb}(\mathscr{V})$. Then $\forall n$:* $\mathsf{Card}(\mathscr{V}_{[0\,..\,n)}) \leqslant \mathsf{Cov}(n,\delta)$.

iii: *The limit* $\mathsf{L}^{\mathsf{Cov}}(\varepsilon) = \lim_n \frac{1}{n}\log(\mathsf{Cov}(n,\varepsilon))$ *exists in* $[0,\infty)$.

iv: $\mathcal{E}^{\mathsf{Sep}}(T) = \mathcal{E}^{\mathsf{Spn}}(T) = \mathcal{E}^{\mathsf{Cov}}(T) = \mathcal{E}^{\mathsf{Card}}(T) \overset{\text{by defn}}{=\!=} \mathcal{E}_{\mathrm{top}}(T)$. ◊

Pf of (i). Take $F \subset X$, a min-cardinality d_n-ε-spanning set. So $\bigcup_{z\in F} D_z = X$, where

$$D_z := \mathsf{d}_n\text{-Bal}(z,\varepsilon) \overset{\text{note}}{=\!=} \bigcap_{j=0}^{n-1} T^{-j}(\mathrm{Bal}(T^j z\,,\,\varepsilon)).$$

This $\mathscr{D} := \{D_z\}_z$ is a cover, and it is d_n-2ε-small. Thus $\mathsf{Cov}(n, 2\varepsilon) \leqslant |\mathscr{D}| = |F|$.

For any metric, a *maximal* ε-separated-set is automatically ε-spanning; adjoin a putative unspanned point to get a larger separated set.

Let A be a max-cardinality d_n-ε-separated set. Take $\mathscr{C}$, a min-cardinality subcover of $\mathscr{W}_{[0\,..\,n)}$. For each $z \in A$, pick a $\mathscr{C}$-patch $C_z \ni z$. Could some pair $x, y \in A$ pick the same C? Well, write $C = \bigcap_{j=0}^{n-1} T^{-j}(W_j)$, with each $W_j \in \mathscr{W}$. For every $j \in [0\,..\,n)$, then, $\mathsf{d}(T^j(x), T^j(y)) \leqslant \mathrm{Diam}(W_j) < \varepsilon$. Hence $\mathsf{d}_n(x, y) < \varepsilon$; so $x = y$. Accordingly, the $z \mapsto C_z$ map is injective, whence $|A| \leqslant |\mathscr{C}|$. ♦

Pf of (ii). Choose a min-cardinality d_n-δ-small cover $\mathscr{C}$. For each $C \in \mathscr{C}$ and $j \in [0\,..\,n)$, the d-$\mathrm{Diam}(T^j(C)) < \delta$. So there is a $\mathscr{V}$-patch $V_{C,j} \supset T^j(C)$. Hence

$$\mathscr{V}_{[0\,..\,n)} \overset{\text{note}}{\ni} \bigcap\nolimits_{j=0}^{n-1} T^{-j}(V_{C,j}) \supset C\,.$$

Thus $\mathscr{V}_{[0\,..\,n)} \preccurlyeq \mathscr{C}$. So $\mathsf{Card}(\mathscr{V}_{[0\,..\,n)}) \leqslant \mathsf{Card}(\mathscr{C}) \leqslant |\mathscr{C}| = \mathsf{Cov}(n, \delta)$. ♦

Pf of (iii). To upper-bound $\mathsf{Cov}(k{+}\ell\,, \varepsilon)$ let $\mathscr{V}$ and $\mathscr{W}$ be min-cardinality ε-small covers, respectively, for metrics d_k and d_ℓ. Then $\mathscr{V} \cap T^\ell(\mathscr{W})$ is a ε-small for $\mathsf{d}_{k+\ell}$. Consequently $\mathsf{Cov}(k{+}\ell\,, \varepsilon) \leqslant \mathsf{Cov}(k, \varepsilon) \cdot \mathsf{Cov}(\ell, \varepsilon)$. Thus $n \mapsto \log(\mathsf{Cov}(n, \varepsilon))$ is subadditive. ♦

Pf of (iv). Pick a $\mathscr{V}$ from the list in (35), choose some $2\varepsilon < \textsc{Leb}(\mathscr{V})$ followed by an ε-small $\mathscr{W}$ from (35). Pushing $n \to \infty$ gives

38: $$\mathsf{L}^{\mathsf{Card}}(\mathscr{V}) \leqslant \mathsf{L}^{\mathsf{Cov}}(2\varepsilon) \leqslant \begin{matrix} \widehat{\mathsf{L}}^{\mathsf{Spn}}(\varepsilon) \leqslant \widehat{\mathsf{L}}^{\mathsf{Sep}}(\varepsilon) \\ \underline{\mathsf{L}}^{\mathsf{Spn}}(\varepsilon) \leqslant \underline{\mathsf{L}}^{\mathsf{Sep}}(\varepsilon) \end{matrix} \leqslant \mathsf{L}^{\mathsf{Card}}(\mathscr{W})\,.$$

Now send $\mathscr{V}$ and $\mathscr{W}$ along the (35) list. ♦

Pretension. Topological entropy takes its values in $[0, \infty]$. A useful corollary of (38) can be stated in terms of any Distance$(\cdot, \cdot)$ which topologizes $[0, \infty]$ as a compact interval.

39: *For each continuous $T{:}X{\to}X$ on a compact metric-space, the* Distance$(\widehat{\mathsf{L}}^{\mathsf{Sep}}(\varepsilon), \underline{\mathsf{L}}^{\mathsf{Sep}}(\varepsilon))$ *goes to zero as $\varepsilon \searrow 0$. Consequently, we can pretend that the*

$$\mathsf{L}^{\mathsf{Sep}}(\varepsilon) = \lim_{n\to\infty} \tfrac{1}{n} \log(\mathsf{Sep}(n, \varepsilon))$$

limit exists, in arguments that subsequently send $\varepsilon \searrow 0$. Ditto for $\mathsf{L}^{\mathsf{Spn}}(\varepsilon)$.

We'll use this during the proof of the Variational Principle. But first, here are two entropy computations which illustrate the efficacy in having several characterizations of top-ent.

E.g: $\mathcal{E}_{\text{top}}(\text{Isometry}) = 0$. Suppose $(T : X, \mathsf{d})$ is a distance-preserving map of a compact metric-space. Fixing ε, a set is d_n-ε-separated IFF it is d-ε-separated. Thus $\mathsf{Sep}(n, \varepsilon)$ does *not* grow with n. So each $\widehat{\mathsf{L}}^{\mathsf{Sep}}(\varepsilon)$ is zero. □

E.g: ***Topological Markov Shifts.*** Imagine ourselves back in the days when computer data is stored on large reels of fast-moving magnetic tape. One strategy to maximize the density of binary data stored is to *not* put timing-marks (which take up space) on the tape. This has the defect that if we write, say, 577 consecutive 1-bits, the tape-reader may erroneously count 578 copies of 1. We sidestep this flaw by first encoding our data so as to avoid the $11\overset{577}{\ldots}1$ word, then writing to tape.

Generalize this to a finite alphabet Q and a finite list $\mathscr{F}$ of disallowed Q-words. Extend each word to a common length K+1; now $\mathscr{F} \subset \mathsf{Q}^{K+1}$. The resulting "$K$-step TMS" (***topological Markov shift***) is the shift on the set of doubly-∞ Q-names having no substring in $\mathscr{F}$. In the above magnetic-tape example, $K = 576$. Making it more realistic, suppose some string of zeros, say $\mathtt{00}\overset{574}{\ldots}\mathtt{0}$, is also forbidden[♡19] Extending to length 577, we get 2^3=8 new disallowed words of form $\mathtt{00}\overset{574}{\ldots}\mathtt{0}b_1b_2b_3$.

We ***recode*** to a 1-step TMS (just called a TMS or a ***subshift of finite type***) over the alphabet $\mathsf{P} := \mathsf{Q}^K$. Each outlawed Q-word $w_0w_1\cdots w_K$ engenders a length-2 forbidden P-word $(w_0, \ldots, w_{K-1})(w_1, \ldots, w_K)$. The resulting TMS is topologically conjugate to the original K-step. The *allowed* length-2 words can be viewed as the edges in a directed-graph and the set of points $x \in X$ is the set of doubly-∞ paths through the graph. Once trivialities removed, this X is a Cantor set and the shift $T{:}X\circlearrowright$ is a homeomorphism.

The Golden Shift. As the simplest example, suppose our magnetic-tape is constrained by the Markov graph, FIG. 40, that we studied measure-theoretically in (26).

We want to store the text of *The Declaration of Independence* on our magnetic tape.[♡20] Imagining that English is a stationary process, we'd like to encode English into this Golden TMS as efficiently as possible. We seek a shift-invariant measure μ on X_{Gold} of *maximum entropy*, should such exist.

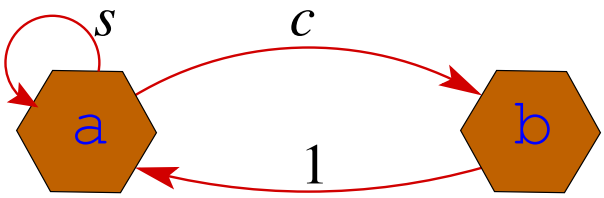

FIG. 40: *Ignoring the labels on the edges, for the moment, the* ***Golden shift****, T, acts on the space of doubly-infinite paths through this graph. The space can be represented as a subset $X_{\text{Gold}} \subset \{\mathsf{a}, \mathsf{b}\}^{\mathbb{Z}}$, namely, the set of sequences with no two consecutive* b *letters.*

View P={a, b} as the time-zero partition on X_{Gold}; that is, name x=…$\mathsf{x}_{-1}\mathsf{x}_0\mathsf{x}_1\mathsf{x}_2$…, is in

[♡19] Perhaps the 0-bad-length, 574, is shorter than the 1-bad-length because, say, 0s take less tape-space than 1s and so –being written more densely– cause ambiguity sooner.

[♡20] … which, by Rights, *should* be stored as a Bernoulli process…

atom b IFF letter x_0 is "b". Any measure μ gives conditional probabilities

$$\mu(\mathsf{a}\,|\,\mathsf{a}) =: s\,, \qquad \mu(\mathsf{b}\,|\,\mathsf{a}) =: c\,,$$
$$\mu(\mathsf{a}\,|\,\mathsf{b}) \overset{\text{note}}{=\!=} 1\,, \qquad \mu(\mathsf{b}\,|\,\mathsf{b}) \overset{\text{note}}{=\!=} 0\,.$$

But recall, $\mathcal{E}(T) = \mathcal{H}(\mathsf{P}_1\,|\,\mathsf{P}_{[-\infty\,..\,0)}) \leqslant \mathcal{H}(\mathsf{P}_1\,|\,\mathsf{P}_0)$. So among all measures that make the conditional distribution P|a equal (s, c), the *unique* one maximizing entropy is the (s, c)-Markov-process. Its entropy, derived in (27), is

41: $$f(s) := \tfrac{1}{2-s}\cdot\mathcal{H}(s,\,1-s) = \tfrac{-1}{2-s}\cdot[\,s\log(s) + [1-s]\log(1-s)]\,.$$

Certainly $f(0) = f(1) = 0$, so f's maximum occurs at the (it turns out) *unique* point $\widehat{s}$ where the derivative $f'(\widehat{s})$ equals zero. This $\widehat{s} = \frac{-1+\sqrt{5}}{2}$. Plugging in, the maximum entropy supportable by the Golden Shift is

42: $$\text{MaxEnt} = \tfrac{2}{5-\sqrt{5}}\cdot[\tfrac{-1+\sqrt{5}}{2}\log(\tfrac{2}{-1+\sqrt{5}}) + \tfrac{3-\sqrt{5}}{2}\log(\tfrac{2}{3-\sqrt{5}})]\,.$$

Exponentiating, the number of μ-typical n-names grows like G^n, where

42′: $$\mathrm{G} = \left[\frac{2}{-1+\sqrt{5}}\right]^{\frac{-1+\sqrt{5}}{5-\sqrt{5}}} \bullet \left[\frac{2}{3-\sqrt{5}}\right]^{\frac{3-\sqrt{5}}{5-\sqrt{5}}}.$$

This expression[♡21] looks unpleasant to simplify –it isn't even obviously an algebraic number– and yet topological entropy will reveal its familiar nature. This, because the Variational Principle (proved in the next section) says that the top-ent of a system is the supremum of measure-entropies supportable by the system.

Top-ent of the Golden Shift. For a moment, let's work more generally on an arbitrary subshift (a closed, shift-invariant subset) $X \subset \mathsf{Q}^{\mathbb{Z}}$, where Q is a finite alphabet. Here, the transformation is always the shift –but the *space* is varying– so agree to refer to the top-ent as $\mathcal{E}_{\text{top}}(X)$. Let $\text{Names}_X(n)$ be the number of distinct words in the set $\{x\restriction_{[0\,..\,n)} \mid x \in X\}$. Note that a metric inducing the product-topology on $\mathsf{Q}^{\mathbb{Z}}$ is

43: $\mathsf{d}(x, x') := \frac{1}{1+|m|}$, *for the smallest* $|m|$ *with* $x_m \neq x'_m$.

44: Lemma. *Consider a subshift X. Then the* $\lim_{n\to\infty}\frac{1}{n}\log(\text{Names}_X(n))$ *exists in* $[0, \infty]$, *and equals* $\mathcal{E}_{\text{top}}(X)$. ♢

Proof. With $\varepsilon \in (0, 1)$ fixed, two n-names are d_n-ε-separated IFF they are not the same name. Hence $\mathsf{Sep}(n, \varepsilon) = \text{Names}_X(n)$. ♦

[♡21] A popular computer-algebra-system was not, at least under my inexpert tutelage, able to simplify this. However, once top-ent gave the correct answer, it was able to detect the *equality*.

To compute $\mathcal{E}_{\text{top}}(X_{\text{Gold}})$, declare that a word is "golden" if it appears in some $x \in X_{\text{Gold}}$. Each $[n{+}1]$-golden word ending in a has form wa, where w is n-golden. An $[n{+}1]$-golden word ending in b, must end in ab and so has form $w$ab, where w is $[n{-}1]$-golden. Summing up,

$$\text{Names}_{X_{\text{Gold}}}(n{+}1) \;=\; \text{Names}_{X_{\text{Gold}}}(n) \;+\; \text{Names}_{X_{\text{Gold}}}(n{-}1)\,.$$

This is the Fibonacci recurrence, and indeed, these are the Fibonacci numbers, since $\text{Names}_{X_{\text{Gold}}}(0) = 1$ and $\text{Names}_{X_{\text{Gold}}}(1) = 2$. Consequently, we have that

$$\text{Names}_{X_{\text{Gold}}}(n) \;\sim\; \text{Const}\cdot\lambda^n\,, \qquad \text{where } \lambda = \tfrac{1+\sqrt{5}}{2} \text{ is the Golden Ratio.}$$

So the sesquipedalian number G from (42′) is simply λ, and $\mathcal{E}_{\text{top}}(X_{\text{Gold}}) = \log(\lambda)$.

Since $\log(\lambda) \approx 0.694$, each thousand bits written on tape (subject to the "no bb substrings" constraint) can carry at most 694 bits of information.

Top-ent of a general TMS. A (finite) digraph G engenders a TMS $T{:}X_G{\to}X_G$, as well as a $\{0,1\}$-valued adjacency matrix $\mathsf{A}{=}\mathsf{A}_G$, where $a_{i,j}$ is the number of directed-edges from state i to j. (Here, each $a_{i,j}$ is 0 or 1.) The (i,j)-entry in power A^n is automatically the number of length-n paths from i to j. Employing the matrix-norm $\|\mathsf{M}\| := \sum_{i,j}|m_{i,j}|$, then,

$$\|\mathsf{A}^n\| \;=\; \text{Names}_X(n)\,.$$

Happily Gelfand's formula[♡22] applies: *For an arbitrary (square) complex matrix,*

45: $$\lim_{n\to\infty} \|\mathsf{A}^n\|^{1/n} \;=\; \text{SpecRad}(\mathsf{A})\,.$$

This righthand side, the ***spectral radius*** of A, means the maximum of the absolute values of A's eigenvalues. So the top-ent of a TMS is thus the

46: $$\mathcal{E}_{\text{top}}(X_G) \;=\; \text{SpecRad}(\mathsf{A}_G) \;:=\; \text{Max}\{|e| \;\big|\; e \text{ is an eigenvalue of } \mathsf{A}_G\}\,.$$

The (a,b)-adjacency matrix of Fig. 40 is $\left[\begin{smallmatrix}1&1\\1&0\end{smallmatrix}\right]$, whose eigenvalues are λ and $-1/\lambda$.

Labeling edges. Interpret $(s,c,1)$ simply as edge-labels in (40). The set of doubly-∞ paths can also be viewed as a subset $Y_{\text{Gold}} \subset \{s,c,1\}^{\mathbb{Z}}$, and it too is a TMS. The shift on Y_{Gold} is conjugate (topologically isomorphic) to the shift on X_{Gold}, so they *a fortiori* have the same top-ent, $\log(\lambda)$. The $(s,c,1)$-adjacency matrix is $\left[\begin{smallmatrix}1&1&0\\0&0&1\\1&1&0\end{smallmatrix}\right]$. Its $|\cdot|$-largest eigenvalue is still λ, as it must.

Now we make a new graph. We modify (40) by manufacturing a total of two s-edges, seven c-edges, and three edges $1_1, 1_2, 1_3$. Give these 2+7+3 edges twelve distinct labels. We *could* compute the resulting TMS-entropy from the corresponding 12×12 adjacency matrix. Alternatively, look at the (a,b)-adjacency matrix $\mathsf{A} := \left[\begin{smallmatrix}2&7\\3&0\end{smallmatrix}\right]$. The roots of its characteristic polynomial are $1 \pm \sqrt{22}$. Hence $\mathcal{E}_{\text{top}}$ of this 12-symbol TMS is $\log(1+\sqrt{22})$. □

[♡22] See [54, *10.13*] or [53, Spectral_radius].

The variational principle

Let $\mathfrak{M} := \mathfrak{M}(X, \mathsf{d})$ be the set of Borel probability measures, and $\mathfrak{M}(T) := \mathfrak{M}(T : X, \mathsf{d})$ the set of T-invariant $\mu \in \mathfrak{M}$. Assign

$$\text{EntSup}(T) \;:=\; \sup\{\mathcal{E}_\mu(T) \mid \mu \in \mathfrak{M}(T)\}\,.$$

47: Variational Principle (Goodson). $\text{EntSup}(T) \;=\; \mathcal{E}_{\text{top}}(T)$. ◊

This says that top-ent is the top entropy —*if* there is a measure μ which realizes the supremum. There doesn't have to be. Choose a sequence of metric-systems $(S_k\colon Y_k, \mathsf{m}_k)$ whose entropies *strictly* increase $\mathcal{E}_{\text{top}}(S_k) \nearrow L$ to some limit in $(0, \infty]$. Let $(S_\infty\colon Y_\infty, \mathsf{m}_\infty)$ be the identity-map on a 1-point space. Define a new system $(T\colon X, \mathsf{d})$, where $X := \bigsqcup_{k\in[1\,..\,\infty]} Y_k$. Have $T(x) := S_k(x)$, for the unique k with $Y_k \ni x$. As for the metric, on Y_k let d be a scaled version of m_k, so that the d-Diam(Y_k) is less than $1/2^k$. Finally, for points in *distinct* components, $x \in Y_k$ and $z \in Y_\ell$, decree that $\mathsf{d}(x, z) := |2^{-k} - 2^{-\ell}|$. Our T is continuous, and is a homeomorphism if each of the S_k is. Certainly $\mathcal{E}_{\text{top}}(T) \;=\; L \;>\; \mathcal{E}_{\text{top}}(S_k)$, for every $k \in [1\,..\,\infty]$.

If L is *finite* then there is *no* measure μ of maximal entropy; for μ must give mass to some Y_k; this pulls the entropy below L, since there are no compensatory components with entropy exceeding L.

In contrast, when $L = \infty$ then there *is* a maximal-entropy measure (put mass $1/2^j$ on some component Y_{k_j}, where $k_j \nearrow \infty$ swiftly); indeed, there are continuum-many maximal-entropy measures. But there is no[♡23] *ergodic* measure of maximal entropy.

For a concrete $L{=}\infty$ example, let S_k be the shift on $[1\,..\,k]^{\mathbb{Z}}$.

Topology on $\mathfrak{M}$. Let's arrange our tools for establishing the Variational Principle. I follow Misiurewicz's proof, adapted from the presentations in [22] and [11].

Equip $\mathfrak{M}$ with the ***weak-**** topology.[♡24] An $A \subset X$ is ***μ-nice*** if its topological boundary $\partial(A)$ is μ-null. And a *partition* is ***μ-nice*** if each atom is.

48: Prop'n. *If $\alpha_L \to \mu$ and $A \subset X$ is μ-nice, then $\alpha_L(A) \to \mu(A)$.* ◊

Proof. Define operator $\mathscr{U}(D) := \text{limsup}_L\, \alpha_L(D)$. It suffices to show that $\boxed{\mathscr{U}(A) \leqslant \mu(A)}$. For since A^c is μ-nice too, then $\mathscr{U}(A^c) \leqslant \mu(A^c)$. Thus $\lim_L \alpha_L(A)$ exists, and equals $\mu(A)$.

Because $C := \overline{A}$ is closed, the continuous functions $f_N \searrow \mathbf{1}_C$ pointwise, where

$$f_N(x) \;:=\; 1 \;-\; \text{Min}(N\cdot \mathsf{d}(x, C)\,,\, 1)\,.$$

[♡23] The ergodic measures are the extreme points of $\mathfrak{M}(T)$; call them $\mathfrak{M}_{\text{Erg}}(T)$. This $\mathfrak{M}(T)$ is the set of barycenters obtained from Borel probability measures on $\mathfrak{M}_{\text{Erg}}(T)$; see [53, Krein-Milman_theorem, Choquet_theory]. In this instance, what explains the failure to have an *ergodic* maximal-entropy measure? Let μ_k be an invariant ergodic measure on Y_k. These measures *do* converge to the one-point (ergodic) probability measure, μ_∞, on Y_∞. But the map $\mu \mapsto \mathcal{E}_\mu(T)$ is not continuous at μ_∞.

[♡24] Measures $\alpha_L \to \mu$ IFF $\int f\, d\alpha_L \to \int f\, d\mu$, for each continuous $f\colon X \to \mathbb{R}$. This metrizable topology makes $\mathfrak{M}$ compact. Always, $\mathfrak{M}(T)$ is a non-void compact subset; see [[Petersen, §6]].

By the Monotone Convergence theorem, then, $\int f_N \, d\mu \xrightarrow{N} \mu(C)$. And $\mu(C) = \mu(A)$, since A is nice. Fixing N, then, it suffices to establish $\boxed{\mathscr{U}(A) \leqslant \int f_N \, d\mu}$. But f_N is continuous, so

$$\int f_N \, d\mu \;=\; \limsup_{L\to\infty} \int f_N \, d\alpha_L \;\geqslant\; \limsup_{L\to\infty} \int \mathbf{1}_A \, d\alpha_L \;=\; \mathscr{U}(A)\,. \qquad \blacklozenge$$

49: Corollary. *Suppose $\alpha_L \to \mu$ and partition P is μ-nice. Then $\mathcal{H}_{\alpha_L}(\mathsf{P}) \to \mathcal{H}_\mu(\mathsf{P})$.* $\diamond$

The ***diameter*** of partition P is $\mathrm{Max}_{A\in\mathsf{P}} \mathrm{Diam}(A)$.

50: Prop'n. *Take $\mu \in \mathfrak{M}$ and $\varepsilon > 0$. Then there exists a μ-nice partition with* $\mathrm{Diam}(\mathsf{P}) < \varepsilon$.$\diamond$

Proof. Centered at an x, the uncountably many balls $\{\mathrm{Bal}(x,r) \mid r \in (0,\varepsilon)\}$ have disjoint boundaries. So all but countably many are μ-nice; pick one and call it B_x. Compactness gives a finite nice cover, say, $\{B_1, \ldots, B_7\}$, at different centers. Then the partition $\mathsf{P} := (A_1, \ldots, A_7)$ is nice,[♡25] where $A_k := B_k \smallsetminus \bigcup_{j=1}^{k-1} B_j$. $\blacklozenge$

Here is a consequence of Jensen's Inequality.

51: Distropy-averaging Lemma. *For $\mu, \nu \in \mathfrak{M}$, a partition R, and a number $t \in [0,1]$:*

$$t \cdot \mathcal{H}_\mu(\mathsf{R}) + t^c \cdot \mathcal{H}_\nu(\mathsf{R}) \quad\leqslant\quad \mathcal{H}_{t\cdot\mu + t^c\cdot\nu}(\mathsf{R})\,. \qquad \diamond$$

Strategy for $\mathrm{EntSup}(T) \geqslant \mathcal{E}_{\mathrm{top}}(T)$. Choose an ε>0. For $L = 1, 2, 3, \ldots$, take a maximal (L,ε)-separated-set $F_L \subset X$, then define

$$\mathbf{F} \;=\; \mathbf{F}_\varepsilon \;:=\; \limsup_{L\to\infty} \tfrac{1}{L} \cdot \log(|F_L|)\,.$$

Let $\varphi_L()$ be the equi-probable measure on F_L; each point has weight $1/|F_L|$. We will construct our desired invariant measure μ from the Cesàro averages

$$\alpha_L \;:=\; \tfrac{1}{L} \cdot \sum\nolimits_{\ell\in[0\,..\,L)} T^\ell \varphi_L\,,$$

which get more and more invariant.

52: Lemma. *Let μ be any weak-$*$ accumulation point of the above $\{\alpha_L\}_1^\infty$. (Automatically, μ is T-invariant.) Then $\mathcal{E}_\mu(T) \geqslant \mathbf{F}$. Indeed, if Q is any μ-nice partition with* $\mathrm{Diam}(\mathsf{Q}) < \varepsilon$, *then $\mathcal{E}_\mu(T, \mathsf{Q}) \geqslant \mathbf{F}$.* $\diamond$

[♡25] For any two sets $B, B' \subset X$, the union $\partial B \cup \partial B'$ is a superset of the three boundaries $\partial(B \cup B')$, $\partial(B \cap B')$, $\partial(B \smallsetminus B')$.

Tactics. As usual, $\mathsf{Q}_{[0\,..\,N)}$ means $\mathsf{Q}_0 \vee \mathsf{Q}_1 \vee \ldots \vee \mathsf{Q}_{N-1}$. Our goal is

$*$: $$\forall N: \quad \mathbf{F} \overset{?}{\leqslant} \tfrac{1}{N} \cdot \mathcal{H}_\mu(\mathsf{Q}_{[0\,..\,N)})\,.$$

Fix N and $\mathsf{P} := \mathsf{Q}_{[0\,..\,N)}$, and a $\delta{>}0$. It suffices to verify: $\forall_{\text{large}} L \gg N$,

52′: $$\tfrac{1}{L}\log(|F_L|) \overset{?}{\leqslant} \delta + \tfrac{1}{N} \cdot \mathcal{H}_{\alpha_L}(\mathsf{P})\,,$$

since this and (49) will prove $(*)$: Pushing $L \to \infty$ along the sequence that produced μ essentially sends LhS(52′) to $\mathbf{F}$, courtesy (39). And RhS(52′) goes to $\delta + \frac{1}{N} \cdot \mathcal{H}_\mu(\mathsf{P})$, by (49), since P is μ-nice. Descending $\delta \searrow 0$, hands us the needed $(*)$. □

Proof[♡26] *of (52′).* Since L is frozen, agree to use φ for the φ_L probability measure.

Our d_L-ε-separated set F_L has at most *one* point in any given atom of $\mathsf{Q}_{[0\,..\,L)}$, thereupon

$$\log(|F_L|) \;=\; \mathcal{H}_\varphi(\mathsf{Q}_{[0\,..\,L)})\,.$$

Regardless of the "offset" $K \in [0\,..\,N)$, we can always fit $C := \lfloor \frac{L-N}{N} \rfloor$ many N-blocks into $[0\,..\,L)$. Denote by $\mathscr{G}(K) := [K\,..\,K{+}CN)$, this union of N-blocks, the ***good*** set of indices. Unsurprisingly, $\mathscr{B}(K) := [0\,..\,L) \setminus \mathscr{G}(K)$ is the ***bad*** index-set. Therefore,

53: $$\mathcal{H}_\varphi(\mathsf{Q}_{[0\,..\,L)}) \;\leqslant\; \overbrace{\mathcal{H}_\varphi(\textstyle\bigvee_{j\in\mathscr{B}(K)} \mathsf{Q}_j)}^{\text{Bad}(K)} + \overbrace{\mathcal{H}_\varphi(\textstyle\bigvee_{j\in\mathscr{G}(K)} \mathsf{Q}_j)}^{\text{Good}(K)}\,.$$

Certainly $\text{Bad}(K) \leqslant 3N\log(|\mathsf{Q}|)$. So $\frac{1}{NL}\sum_{K\in[0\,..\,N)} \text{Bad}(K) \leqslant \frac{3N}{L}\log(|\mathsf{Q}|)$. This is less than δ, since L is large. Applying $\frac{1}{NL}\sum_{K\in[0\,..\,N)}$ to (53) now produces

54: $$\tfrac{1}{L} \cdot \log(|F_L|) \;\leqslant\; \delta + \tfrac{1}{NL}\sum\nolimits_K \text{Good}(K)\,.$$

Note $\bigvee_{j\in\mathscr{G}(K)} T^j(\mathsf{Q}) \;=\; \bigvee_{c\in[0\,..\,C)} T^{K+cN}(\mathsf{P})$. So $\text{Good}(K) \leqslant \sum_c \mathcal{H}_\varphi(T^{K+cN}\mathsf{P})$. This latter, by definition, equals $\sum_c \mathcal{H}_{T^{K+cN}(\varphi)}(\mathsf{P})$. We conclude that

$$\begin{aligned}
\tfrac{1}{NL}\sum_K \text{Good}(K) \;&\leqslant\; \tfrac{1}{NL}\sum_K\sum_c \mathcal{H}_{T^{K+cN}\varphi}(\mathsf{P}) \\
&\leqslant\; \tfrac{1}{NL}\sum_{\ell\in[0\,..\,L)} \mathcal{H}_{T^\ell\varphi}(\mathsf{P})\,, \qquad \text{by adjoining a few translates of } \mathsf{P}, \\
&\leqslant\; \tfrac{1}{N} \cdot \mathcal{H}_{\alpha_L}(\mathsf{P})\,, \qquad \text{by the Distropy-averaging Lemma, (51),}
\end{aligned}$$

since α_L is the average $\frac{1}{L}\sum_\ell T^\ell\varphi$. Thus (54) implies (52′), our goal. ♦

[♡26] The idea is to mostly fill $[0\,..\,L)$ with N-blocks, starting with a offset $K \in [0\,..\,N)$. Averaging over the offset will create a Cesàro average over each N-block. Averaging over the N-blocks will allow us to compute distropy with respect to the averaged measure, α_L.

Proof of $\mathrm{EntSup}(T) \leqslant \mathcal{E}_{\mathrm{top}}(T)$. Fix a T-invariant μ. For partition $\mathsf{Q} = (B_1, \ldots, B_K)$, choose a compact set $A_k \subset B_k$ with $\mu(B_k \smallsetminus A_k)$ small.[♡27] Letting $D := [\bigsqcup_i A_i]^c$ and $\mathsf{P} := (D, A_1, \ldots, A_K)$, we can have made $\mathcal{H}(\mathsf{P} \,|\, \mathsf{Q})$ as small as desired. Courtesy (23b), then, we only need consider partitions of the form that P has.

Open-cover $\mathscr{V} := (U_1, \ldots, U_K)$ has patches $U_k := D \cup A_k$. What atoms of, say, $\mathsf{P}_{[0..3)}$, can the intersection $U_9 \cap T^{-1}(U_2) \cap T^{-2}(U_5)$ touch? Only the eight atoms

$$(D \text{ or } A_9) \;\cap\; T^{-1}(D \text{ or } A_2) \;\cap\; T^{-2}(D \text{ or } A_5)\,.$$

Thus $^{\#}\mathsf{P}_{[0..n)} \;\leqslant\; 2^n \cdot {}^{\#}\mathscr{V}_{[0..n)}$. (Here, $^{\#}()$ counts the number of non-void atoms/patches.) So

$$\tfrac{1}{n}\mathcal{H}_\mu(\mathsf{P}_{[0..n)}) \quad\leqslant\quad 1 \,+\, \tfrac{1}{n}\log({}^{\#}\mathscr{V}_{[0..n)}) \quad\leqslant\quad 1 \,+\, 1 + \mathcal{E}_{\mathrm{top}}(T)\,;$$

this last inequality, when n is large. The upshot: $\mathcal{E}_\mu(T) \leqslant 2 + \mathcal{E}_{\mathrm{top}}(T)$.

Applied to a power T^ℓ, this asserts that $\mathcal{E}_\mu(T^\ell) \leqslant 2 + \mathcal{E}_{\mathrm{top}}(T^\ell)$. Thus

$$\mathcal{E}_\mu(T) \quad\leqslant\quad \tfrac{2}{\ell} + \mathcal{E}_{\mathrm{top}}(T)\,,$$

using (24d) and using (34g). Now coax $\ell \to \infty$. ♦

Three recent results

Having given an survey of older results in measure-theoretic entropy and in topological entropy, let us end this survey with a brief discussion of a few recent results, chosen from many.

Ornstein-Weiss: Finitely-observable invariant. In a landmark paper [8, 2007], Ornstein and Weiss show that all "finitely observable" properties of ergodic processes are secretly entropy; indeed, they are continouous functions of entropy. This was generalized by Gutman and Hochman [9]; some of the notation below is from their paper.

Here is the setting. Consider an ergodic process, on a non-atomic space, taking on only finitely many values in $\mathbb{N}$; let $\mathscr{C}$ be some family of such processes. An ***observation scheme*** is a metric space (Ω, d) and a sequence of functions $\mathbf{S} = (S_n)_1^\infty$, where S_n maps $\mathbb{N}\times\overset{n}{\ldots}\times\mathbb{N}$ into Ω. On a point $\vec{x} \in \mathbb{N}^\infty$, the scheme ***converges*** if

$$55\text{:} \qquad n \;\mapsto\; S_n(x_1, x_2, \ldots, x_n)$$

converges in Ω. And on a particular process $\mathcal{X}$, say that $\mathbf{S}$ ***converges***, if $\mathbf{S}$ converges on a.e $\vec{x}$ in $\mathcal{X}$.

A function $J{:}\mathscr{C}\to\Omega$ is isomorphism invariant if, whenver the underlying transformations of two processes $\mathcal{X}, \mathcal{X}' \in \mathscr{C}$ are isomorphic, then $J(\mathcal{X}) = J(\mathcal{X}')$. Lastly, say that $\mathbf{S}$ "converges to J", if for each $\mathcal{X} \in \mathscr{C}$, scheme $\mathbf{S}$ converges to the value $J(\mathcal{X})$.

[♡27] This can be done, since μ is automatically a regular measure.

The work of David Bailey [38, 1976], a student of Ornstein, produced an observation scheme for entropy. The Lempel-Ziv algorithm [43] was another entropy observer, with practical application.

Ornstein and Weiss provided entropy schemes in [41] and [42]. Their recent paper "*Entropy is the only finitely-observable invariant*" [8, 2007], give a converse, a uniqueness result.

56: Theorem (Ornstein, Weiss). *Suppose J is a finitely observable function, defined on all ergodic finite-valued processes. If J is an isomorphism invariant, then J is a continouous function of the entropy.* ◊

Yonatan Gutman and Michael Hochman, in [9], significantly extend the Ornstein-Weiss result, by proving that it holds even when the isomorphism invariant, J, is well-defined only on certain subclasses of the set of all ergodic processes. In particular they obtain the following result on three classes of zero-entropy transformations.

57: Theorem (Gutman, Hochman). *Suppose $J()$ is a finitely observable invariant on one of the following classes:*

i: *The Kronecker systems; the class of systems with pure point spectrum.*

ii: *The zero-entropy mild mixing processes.*

iii: *The zero-entropy strong mixing processes.*

Then $J()$ is constant. ◊

Entropy of actions of free groups. Consider $(G, \mathscr{G})$, a topological group and its Borel field (sigma-algebra). Let $\mathscr{G} \times \mathscr{X}$ be the field on $G \times X$ generated by the two coordinate-subfields. A map

$*$: $\psi: G \times X \to X$ is ***measurable*** if $\psi^{-1}(\mathscr{X}) \subset \mathscr{G} \times \mathscr{X}$. Use $\psi^g(x)$ for $\psi(g, x)$.

This map $(*)$ is a (measure-preserving) ***group action*** if $\forall g, h \in G$: $\psi^g \circ \psi^h = \psi^{gh}$, and each $\psi^g: X \to X$ is measure preserving.

This encyclopedia article has only discussed entropy for $\mathbb{Z}$-actions, i.e, when $G = \mathbb{Z}$. The ergodic theorem, our definition of entropy, and large parts of ergodic theory, involve taking averages (of some quantity of interest) over larger and larger "pieces of Time". In $\mathbb{Z}$, we typically use the intervals $I_n := [0 \,..\, n)$. When G is $\mathbb{Z} \times \mathbb{Z}$, we might average over squares $I_n \times I_n$.

The *amenable groups* are those which possess, in a certain sense, larger and larger averaging sets. Parts of ergodic theory have been carried over to actions of amenable groups, e.g [49] and [51]. Indeed, much of the Bernoulli theory was extended to certain amenable groups by Ornstein and Weiss, [50].

The stereotypical example of a *non*-amenable group, is a free group (on more than one generator). But recently, Lewis Bowen [10] succeeded in extending the definition of entropy to actions of finite-rank free groups.

58: Theorem (Lewis Bowen). *Let G by a finite-rank free group. Then two Bernoulli G-actions are isomorphic IFF they have the same entropy.* ◊

The paper introduces a new isomorphism invariant, the "f invariant", and shows that, for Bernoulli actions, the f invariant agrees with entropy, that is, with the distropy of the independent generating partition.

Conclusion

Ever since the pioneering work of Shannon, and of Kolmogorov and Sinai, entropy has been front and center as a major tool in Ergodic Theory. Simply *mentioning* all the substantial results in entropy theory would dwarf the length of this encyclopedia article many times over. And, as the above three results (cherry-picked out of many) show, Entropy shows no sign of fading away...

§BIBLIOGRAPHY

[1] Citations to other articles in this Encyclopedia are made with the author's name in double-brackets, e.g, [[Lemanczyk]], [[Petersen]].

Historical

[2] Rudolf Clausius, *Abhandlungen ueber die mechanische Waermetheorie*, vol. **1**, (F.Vieweg, Braunschweig, 1864); vol. **2**, (1867).

[3] Claude Elwood Shannon, *A Mathematical Theory of Communication*, Bell Syst. Tech. J., 27, pp. 379–423, pp. 623–656, (1948).

[4] A.N. Kolmogorov, *A New Metric Invariant of Transitive Automorphisms of Lebesgue Spaces*, Dokl. Akad. Nauk SSSR119, no. 5, pp. 861–864, (1958).

[5] Ya. Sinai, *On the Concept of Entropy of a Dynamical System*, Dokl. Akad. Nauk SSSR124, pp. 768–771, (1959).

[6] B. McMillan, *The Basic Theorems of Information Theory*, Ann. Math. Stat.24, pp. 196–219, (1953).

[7] Roy Adler and Benjamin Weiss, *Entropy, a complete metric invariant for automorphisms of the torus*, Proc. Nat. Acad. Sci. USA, vol. **57**, pp. 1573–1576, (1967).

Recent Results

[8] D.S. Ornstein and B. Weiss, *Entropy is the only finitely-observable invariant* J. of Modern Dynamics, vol. **1**, (2007), pp. 93-105. http://www.math.psu.edu/jmd

[9] Yonatan Gutman and Michael Hochman, *On processes which cannot be distinguished by finitary observation* to appear in Israel Journal of Mathematics. Preprint at http://arxiv.org/pdf/math/0608310

[10] Lewis Bowen, *A new measure-conjugacy invariant for actions of free groups* Preprint at http://www.math.hawaii.edu/%7Elpbowen/notes11.pdf

Ergodic Theory books

[11] Michael Brin, Garrett Stuck, *Introduction to dynamical systems*, Cambridge University Press, (2002).

[12] I. Cornfeld, S. Fomin, Ya. Sinai, *Ergodic theory*, Grundlehren der Mathematischen Wissenschaften, 245. Springer-Verlag, New York, (1982).

[13] Nathaniel A. Friedman, *Introduction to Ergodic Theory*, Van Nostrand Reinhold, (1970).

[14] Paul R. Halmos, *Lectures on Ergodic Theory*. The Mathematical Society of Japan, (1956).

[15] Anatole Katok, Boris Hasselblatt, *Introduction to the modern theory of dynamical systems. (With a supplementary chapter by Katok and Leonardo Mendoza)*, Encyclopedia of Mathematics and its Applications, 54. Cambridge University Press, (1995).

[16] G. Keller, A. Greven and G. Warnecke (eds), *Entropy*, Princeton Series in Applied Mathematics, Princeton University Press (2003).

[17] Doug Lind, Brian Marcus, *An introduction to symbolic dynamics and coding*, Cambridge University Press, Cambridge, (1995).

[18] R. Mané. *Ergodic theory and differentiable dynamics*. Ergebnisse der Mathematik und ihrer Grenzgebiete; ser.3, Bd. 8. Springer-Verlag, Berlin, (1987).

[19] William Parry, *Entropy and Generators in Ergodic Theory*, W.A. Benjamin, (1969).

[20] Karl Petersen, *Ergodic theory*. Cambridge Univ. Press, Cambridge, (1983).

[21] Daniel J. Rudolph, *Fundamentals of Measurable Dynamics*, Clarendon Press, (1990).

[22] Peter Walters, *An introduction to ergodic theory*, Graduate Texts in Mathematics, no. 79, Springer, (1982).

[23] Ya.G. Sinai. *Topics in ergodic theory*, volume 44 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, (1994).

Differentiable entropy

[24] Ya.B. Pesin. *Characteristic Lyapunov exponents and smooth ergodic theory*. Russ. Math. Surveys, 32:55–114, (1977).

[25] L.-S. Young, *Dimension, entropy and Lyapunov exponents*, Ergodic Theory and Dynamical Systems, 2, (1982), no. 1, pp. 109–124.

[26] F. Ledrappier and L.-S. Young, *The metric entropy of diffeomorphisms*, Ann. of Math., (2) 122, (1985), pp. 509–539 (I) and pp. 540–574 (II).

Finite rank

[27] S. Ferenczi, *Systems of finite rank*, Colloq. Math., 73, (1997), no. 1, pp. 35–65.

[28] J.L.F. King, *Joining-rank and the structure of finite rank mixing transformations*, *J. Analyse Math.*, vol. **51**, (1988), pp. 182–227.

Maximal-entropy measures

[29] J. Buzzi and S. Ruette, *Large entropy implies existence of a maximal entropy measure for interval maps*, Discrete Contin. Dyn. Syst., 14, (2006), no. 4, pp. 673–688.

[30] M. Denker, *Measures with maximal entropy*, Théorie ergodique Actes Journées Ergodiques, Rennes, 1973/1974, pp. 70–112. Lecture Notes in Math., vol. 532, Springer, Berlin, (1976).

[31] M. Misiurewicz, *Diffeomorphism without any measure with maximal entropy*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys. 21, (1973), pp. 903–910.

Topological entropy

[32] R.L. Adler, A.G. Konheim, M.H. McAndrew, *Topological Entropy*, Transactions of the American Mathematical Society, vol. **114**, no. 2, (Feb., 1965), pp. 309–319.

[33] E.I. Dinaburg, *The relation between topological entropy and metric entropy*, Soviet Math. Dokl. vol. **11**, (1970), pp. 13–16.

[34] Rufus Bowen, *Entropy for group endomorphisms and homogeneous spaces*, Transactions, American Mathematical Society, vol. **153**, (1971), pp. 401–414, erratum, **181**(1973) pp. 509–510.

[35] Rufus Bowen, *Topological entropy for noncompact sets*, Transactions, American Mathematical Society, vol. **184**, (1973), pp. 125–136.

[36] Roy Adler, Brian Marcus, *Topological entropy and equivalence of dynamical systems*, Mem. Amer. Math. Soc., vol. **20**, (1979), no. 219.

[37] Boris Hasselblatt, Zbigniew Nitecki, James Propp, *Topological entropy for non-uniformly continuous maps*, http://www.citebase.org/abstract?id=oai:arXiv.org:math/0511495, (2005).

Determinism and Zero-entropy, and entropy observation

[38] David Bailey. *Sequential schemes for clasifying and predicting ergodic processes.* Stanford University, (1976). Ph.D. Dissertation.

[39] J.L.F. King, *Dilemma of the sleeping stockbroker*, The American Mathematical Monthly, **99**, no. 4, (1992), pp. 335–338.

[40] S. Kalikow and J.L.F. King, *A countably-valued sleeping stockbroker process*, Journal of Theoretical Probability, vol. **7**, no. 4, (1994), pp. 703–708.

[41] Donald S. Ornstein and Benjamin Weiss, *How sampling reveals a process.* Ann. Probab., 18(3):905–930, (1990).

[42] Donald S. Ornstein and Benjamin Weiss, *Entropy and data compression schemes.* IEEE Trans. Inform. Theory, 39(1):78–83, (1993).

[43] Jacob Ziv and Abraham Lempel. *A universal algorithm for sequential data compression.* IEEE Trans. Information Theory, IT-23(3):337–343, (1977).

Bernoulli Transformations, K-automorphisms, Amenable groups

[44] L.D. Meshalkin, *A Case of Isomorphism of Bernoulli Schemes*, Dokl. Akad. Nauk SSSR128, pp. 41–44, (1959).

[45] Ya.G. Sinai, *A Weak Isomorphism of Transformations Having an Invariant Measure*, Dokl.Akad.Nauk. SSSR 147, pp. 797–800, (1962), *MR 161960*, Zbl 0205.13501.

[46] Donald S. Ornstein, *Bernoulli Shifts with the Same Entropy Are Isomorphic*, Adv. Math., 5, pp. 337–352, (1970).

[47] Paul Shields. *The Theory of Bernoulli Shifts*, University of Chicago Press, Chicago and London, (1973).

[48] Donald S. Ornstein, *Ergodic Theory Randomness and Dynamical Systems*, Yale Math. Monographs, vol. **5**, Yale University Press, (1974), MR 447525, Zbl 0296.28016.

[49] D.S. Ornstein and B. Weiss, *The Shannon-Mc Millan-Briman Theorem For A Class Of Amenable Groups*, Isr. J. Math., vol. **44**, no. 3, (1983), pp. 53–60.

[50] D.S. Ornstein and B. Weiss, *Entropy and isomorphism theorems for actions of amenable groups*, *J. Analyse Math.*, vol. **48**, (1987) pp. 1–141.

Abramov Formula

[51] T. Ward and Q. Zhang, *The Abramov-Rohlin entropy addition formula for amenable group actions*, *Monatsh. Math*, **114**, (1992), pp. 317–329.

Miscellaneous

[52] S.E. Newhouse. *Continuity properties of entropy*, *Annals of Mathematics*, 129:215–235, (1989). Also Ann. of Math. **131**, (1990), pp. 409–410.

[53] *Wikipedia*, http://en.wikipedia.org/wiki/ pages: Spectral_radius, Information_entropy.

[54] Walter Rudin, *Functional Analysis*, McGraw-Hill, (1973).