ALL ENTROPIES AGREE FOR AN SFT

CHRIS HILLMAN

1. Introduction

In this paper I discuss a number of "entropies" which have definitions which are respectively probabilistic, topological, algebraic, and algorithmic. I shall explain how these entropies are all defined in the setting of shift dynamical systems. The main result of the paper, which should be regarded as part of the folklore, is the fact that for topologically transitive shifts of finite type (the definitions of these terms will be found below) all these entropies agree *numerically*.[1] The precise statement of this result is fairly involved, since none of the entropies involved are defined at equivalent levels of mathematical structure, and thus are not directly comparable. For reasons of space, I shall confine myself here to explaining what the result says, rather than attempting to prove it. However, interested readers should have no difficulty in constructing a proof by filling out my sketch of the easy parts of this result and by consulting the cited references for the hard parts.

This is an expository paper, as self contained as I could make it, given the inevitable limitations of time and space. References are given for all quoted results. I do not assume any prior exposure to ergodic theory or dynamical systems, but I do assume some familiarity with the most basic concepts of measure and topology (see, for instance, the appropriate chapters of [5]).

2. Dynamical Systems

A **topological dynamical system** (X, S) is a compact metric space X with a continuous mapping S taking X into itself. A Borel measure μ on X is said to be S-**invariant** if for every Borel set E, $\mu\{S^{-1}(E)\} = \mu\{E\}$; in this case, (X, S, μ) defines a **measure-theoretic dynamical system**.

Let (X, S) be a TDS. Then each S-invariant Borel probability measure μ makes (X, S) into an MTDS in a different way. The set of regular S-invariant Borel probability measures forms a *compact convex subset*, analogous to a simplex, in the set of regular Borel probability measures. The "vertices" of this "simplex" are special S-invariant measures called **ergodic measures**. Ergodic measures are precisely the ones having the useful property that for every integrable Borel measureable function f on X, and for almost every $x \in X$, the time averaged values of f on the **trajectory** of x; that is, the sequence of points $x, S(x), S^2(x) \ldots$, will converge to

Date: February 18, 1998.

1991 *Mathematics Subject Classification.* Primary 94A17; Secondary 06A15.

Key words and phrases. measures of information, entropy.

I wish to thank Peter White (Motorola) for pointing out a typo in a previous draft.

[1]In choosing the title I have taken a certain alliterative license, since there are many, many entropies we shall not even mention in this paper.

the spatial average $\int_X f\, d\mu$. In symbols:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(S^j(x)) = \int_X f\, d\mu$$

This extremely important property is known as the **ergodic theorem**. Alternatively, if μ is an S-invariant ergodic Borel probability measure, then S eventually moves every subset of X "all over X". Every S-invariant measure can be decomposed into an "average" of (actually an integral over) the ergodic ones; this corresponds geometrically to the fact that every vector inside a tetrahedron can be written as a weighted sum of the four vertices. Thus, there is close relationship between the geometric and the analytic features of the set of ways of turning a given TDS into an MTDS (see [24] for more detail).

3. Dynamical Invariants

The basic problem of the theory of dynamical systems is to classify them. To formulate this problem precisely, we need a notion of when two TDS's or two MTDS's are equivalent. Let (X, S) and (Y, T) be two TDS's, and suppose F is a continuous mapping from X to Y such that $F \circ S = T \circ F$. If F is a homeomorphism, then (X, S) and (Y, T) are said to be **conjugate**. If F is a one-one mapping, we say (X, S) is a **sub-system** of (Y, T), because F maps X into a homeomorphic image of X inside Y, such that the effect of S on X is exactly the same as the effect of T on the image of X, so that we have found a copy of (X, S) sitting inside (Y, T). If F is an onto mapping, we say that (Y, T) is a **factor** of (X, S); this means that (Y, T) provides a "simplified model" of the global behaviour (X, S). Similar definitions hold for two MTDS's (X, S, μ) and (Y, T, ν), except that now F must be a **measure preserving mapping** from X to Y; that is, for every Borel set E in Y we must have $\mu\{F^{-1}(E)\} = \nu\{E\}$.

One fundamental approach to the problem of classifying TDS's (or MTDS's) is to attach to each TDS (or MTDS) a number (or more complicated object such as a group), defined in terms of the dynamics of the TDS (or MTDS), in such a way that whenever (X, S) is conjugate to (Y, T), the numbers agree. Such a number (or other object) is called a **dynamical invariant**. Provided we can actually compute these numbers, if we obtain different numbers for two TDS's we know they cannot be conjugate. The entropies we will define for TDS's and MTDS's will be numerical invariants.

Here are some examples of invariant properties which are not numbers. The TDS (X, S) is **topologically transitive** if every open set is eventually moved "all over X"; this is the topological analogue of ergodicity. (X, S) is **chaotic** if it is topologically transitive and the periodic points of S are dense in X. This definition is equivalent [1] to the one given in the influential textbook [3]; the two stated conditions imply a third condition, **sensitivity to initial conditions**, which some other authors take as the definition of chaos. The TDS (X, S) is **topologically mixing** if, given two subsets A, B of X, S eventually "mixes them up". Topologically mixing implies topologically transitive.

4. Shift Dynamical Systems

Sequence spaces form dynamical systems with a particularly simple structure. It is convenient to write $\mathbb{B} = \{0, 1\}$, and to think of a sequence or string as a function from an index set to $\mathbb{B}$. Let $\mathbb{B}^n$ denote strings of length n, $\mathbb{B}^*$ denote all finite length strings, $\mathbb{B}^{\mathbb{N}}$ denote all strings which start at the zero-th symbol and extend infinitely far to the right, and let $\mathbb{B}^{\mathbb{Z}}$ denote all strings which extend infinitely far in both directions. If s in $\mathbb{B}^{\mathbb{Z}}$, i.e. $s : \mathbb{Z} \to \mathbb{B}$, then $s(n)$ is the symbol appearing in the n-th place of s, and similarly for $\mathbb{B}^{\mathbb{N}}, \mathbb{B}^*$, etc.

If x, y are two strings in $\mathbb{B}^{\mathbb{Z}}$, let $n = \inf\{|k| : x(k) \neq y(k)\}$, and put $d(x, y) = 2^{-n}$. This defines a metric on $\mathbb{B}^{\mathbb{Z}}$ making it into a compact metric space; in fact, into a homeomorphic copy of the Cantor set. Note that the ball of radius n about x is the set of all y which agree with x on the set $\{-n, -n+1, \ldots n\}$ (and possibly on a larger set).

We can define an operator on $\mathbb{B}^{\mathbb{Z}}$, the **shift operator**, by taking each s in $\mathbb{B}^{\mathbb{Z}}$ to $\sigma(s)$ where $\sigma(s)(n) = s(n+1)$. In other words, σ simply shifts each sequence one place to the left. This defines a homeomorphism on $\mathbb{B}^{\mathbb{Z}}$, so $(\mathbb{B}^{\mathbb{Z}}, \sigma)$ forms an (invertible) TDS, called the **full two-shift**. By a classic result, the Curtis-Lyndon-Hedlund Theorem [18], the closed shift invariant subsets of $\mathbb{B}^{\mathbb{Z}}$ are precisely those subsets which can be described by listing "forbidden blocks" of symbols. Such subsets are called **shift spaces**, and the topological dynamical systems of the form (A, σ) where A is a closed shift-invariant subset of $\mathbb{B}^{\mathbb{Z}}$, are called **shift dynamical systems**, or shifts for short. Conjugacies and factors between shifts correspond to **sliding block codes** [18].

A **shift of finite type** is a shift which can be defined by a *finite* list of "forbidden blocks". If A is an SFT, with none of the forbidden blocks longer than n, list all blocks of length $n-1$ which do occur in A. Say there are m of these. Define an m by m matrix $\tilde{A}$ by putting a one in the (j, k) entry if and only if the k-th block can follow the j-th block (overlapping by all but one symbol, of course). For example, the set of all strings containing no occurrences of '11' is an SFT called the **golden mean shift** (the reason for the name will become apparent later). In this case, the matrix $\tilde{A}$ is

$$\tilde{A} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

The matrix $\tilde{A}$ is more than just a convenient way to encode the definition of an SFT, as we'll see.

We can make the shift (A, σ) into an MTDS by putting a shift invariant Borel probability measure on A (or, without loss of generality, on $\mathbb{B}^{\mathbb{Z}}$). Such measures can often be defined most conveniently in terms of the **cylinder sets**. For example: $Z(5, 0)$ is the set of $x \in \mathbb{B}^{\mathbb{N}}$ such that $x(5) = 0$, $Z(2, 101)$ is the set of all x such that $x(2) = 1, x(3) = 0, x(4) = 1$, and so forth. The cylinders generate the sigma-algebra of Borel sets. The **(p, q)-Bernoulli shift**, where $q = 1 - p$, is defined by giving $\mathbb{B}^{\mathbb{Z}}$ the shift invariant Borel probability measure defined by $\mu\{Z(n, w)\} = p^m q^{n-m}$, where w is a word containing m zeros and $n - m$ ones. The (p, q)-Bernoulli shift models a process which produces zeros and ones independently of whatever symbol was produced in the previous step, according to the probabilities p, q. Somewhat more generally, a **Markov shift** is defined by a matrix P of transition probabilities, where P_{01} is the probability that a zero will be immediately followed by a one, and so forth, and an eigenvector π giving a probability measure on $\mathbb{B}$ which is invariant under P. For example, we might have

$$P = \begin{pmatrix} 1/3 & 2/3 \\ 1 & 0 \end{pmatrix}, \qquad \pi = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$

The measure μ is now defined by

$$\mu\{Z(n, 01001)\} = \pi_0 \cdot P_{01} \cdot P_{10} \cdot P_{00} \cdot P_{01} = 1/9$$

and so forth. More generally still, an **n-step Markov shift** is defined by giving $\mathbb{B}^{\mathbb{N}}$ a shift invariant Borel probability measure such that the probability of producing a zero (one) in each step depends only on the symbols produced in the n immediately previous steps.

We've been discussing two-sided shifts; we will also need the concept of the one sided shift map defined on $\mathbb{B}^{\mathbb{N}}$ the same way as above, except that now σ is two-to-one and thus not invertible. One sided shift spaces, Markov shifts, and so forth, are now defined exactly as above.

The importance of shift dynamical systems lies in the fact that they provide relatively simple models of a suprisingly wide variety of behavior. For instance, it turns out that SFTs are hiding inside horshoes and Julia sets, among other places (see [3]). Whenever a shift can be found inside some more complicated dynamical system, because dynamical invariants for these shifts can be easily computed (as we will see) they often provide useful insight into the full dynamical system.

5. Stationary Information Sources

Let (X, S, μ) be an invertible MTDS. A **finite partition** of X, written $\mathcal{A}$, is a collection of measurable subsets $A_1, A_2, \ldots A_r$ (known as the **atoms** of $\mathcal{A}$) which are disjoint and whose union is X. Now the sequence space $\mathcal{A}^{\mathbb{N}}$ consists of (one-sided) sequences of symbols $A_1, \ldots A_r$ which can be identified with base r strings of digits. We define a metric on $(\mathbb{Z}_r)^{\mathbb{N}}$ by setting

$$d(x, y) = \sum_{n=0}^{\infty} \frac{|x(n) - y(n)|}{r^n}$$

This makes $(\mathbb{Z}_r)^{\mathbb{N}}$ into a compact metric space (the case $r = 2$ agrees with the topology already defined for $\mathbb{B}^{\mathbb{N}}$). The upshot is that we can consider $\mathcal{A}^{\mathbb{N}}$ a compact metric space.

Given two finite partitions on X, we can make a more complicated one as follows. Suppose $\mathcal{A}$ denotes the partition $X = \biguplus_j A_j$ while $\mathcal{B}$ denotes the partition $X = \biguplus_k B_k$. Then their **join**, written $\mathcal{A} \vee \mathcal{B}$, is the partition $X = \biguplus_j \biguplus_k A_j \cap B_k$. This partition is "more complex" than either $\mathcal{A}$ or $\mathcal{B}$ in the sense that it **refines** each of them. $\mathcal{A}, \mathcal{B}$ can be considered *quotient objects* of $\mathcal{A} \vee \mathcal{B}$, since by forgetting the distinction between the atoms $A_j \cap B_k$, for each A_j, we can recover $\mathcal{A}$ from $\mathcal{A} \vee \mathcal{B}$, and similarly for $\mathcal{B}$.

The **pullback** of $\mathcal{A}$, written $S^{\#}\mathcal{A}$, is the partition $X = \biguplus_j S^{-1}(A_j)$. To understand the meaning of the pullback, which we can think of as an operation on partitions, consider the following process. Choose an arbitrary x in X, and consider the trajectory $x, S(x), S^2(x) \ldots$. Each of these points is in a unique A_j, so each trajectory defines a symbolic sequence in $\mathcal{A}^{\mathbb{N}}$. To define a stationary source, we need to say how to compute the probability of a given finite block of consecutive symbols, and to show that this probability does not depend upon the starting place ("time of occurrence") of the block. (The term stationary refers to a process whose behaviour does not change over time.) For example, suppose we want to compute the probability of the block of symbols

$$A_2, A_1, A_3$$

where A_2 occurs at time k. This means we must compute the probability that an arbritary x satisfies $S^k(x)$ in A_2, S^{k+1} in A_1, S^{k+2} in A_3, or equivalently the probability of the set

$$S^{-k}(A_2) \cap S^{-k-1}(A_1) \cap S^{-k-2}(A_3)$$

which is one atom of the partition $S^{\#k}\mathcal{A} \vee S^{\#k+1}\mathcal{A} \vee S^{\#k+2}(\mathcal{A})$. This shows how using pullback and join we can express the idea of a sequence of events occuring over time.

Since our measure μ is S-invariant, the set

$$S^{-k}(A_2) \cap S^{-k-1}(A_1) \cap S^{-k-2}(A_3) = S^{-k}\left(A_2 \cap S^{-1}(A_1) \cap S^{-2}(A_3)\right)$$

has the same measure as

$$A_2 \cap S^{-1}(A_1) \cap S^{-2}(A_3)$$

which is one atom of the partition $\mathcal{A} \vee S^{\#}\mathcal{A} \vee S^{\#2}\mathcal{A}$. This means that the probability that A_3 follows A_1 follows A_2 at time k is the same as the probability that this happens at time zero. Thus, we are justified in referring to the quadruple $(X, S, \mu, \mathcal{A})$ as a **stationary information source**. See [22] to see how Shannon thought of this concept.

6. Source Entropy

Given a MTDS (X, S, μ) and a finite partition $\mathcal{A}$, define the **entropy** of $\mathcal{A}$ to be

$$H_\mu(\mathcal{A}) = -\sum \mu(A_j) \log \mu(A_j)$$

where $\mathcal{A}$ denotes the finite partition $X = \uplus_j A_j$. Note that the numbers $\mu(A_j)$ are simply the probability of being in the j-th atom of $\mathcal{A}$. $H_\mu(\mathcal{A})$ vanishes if only zeros (or only ones) are produced, and is maximized when they are produced in equal proportions. If we take all logarithms base two, the units of $H_\mu(\mathcal{A})$ will be bits. In the mathematical literature, where all logarithms are natural logarithms, the units will be "nats"; one bit is equal to $1/(\log 2)$ nats.

$H_\mu(\mathcal{A})$ can be considered to also define the **first order entropy** of the source $(X, \mathcal{M}, S, \mathcal{A})$, because it fails to take account of statistical correlations between successive digits, but does take account of the relative frequency of zeros and ones.

It is not hard to see that $H_\mu(\mathcal{B}) \leq H_\mu(\mathcal{A} \vee \mathcal{B})$; the difference $H_\mu(\mathcal{A} \vee \mathcal{B}) - H_\mu(\mathcal{B})$ is called the **conditional entropy** of $\mathcal{A}$ given $\mathcal{B}$ and written $H_\mu(\mathcal{A}/\mathcal{B})$. Likewise for the entropy of $\mathcal{B}$ given $\mathcal{A}$. One important formal property of conditional entropy is that $H_\mu(\mathcal{A}/\mathcal{B}) = 0$ iff $\mathcal{B}$ refines $\mathcal{A}$ up to sets of measure zero. This means that we can interpret $H_\mu(\mathcal{A})$ as a measure of the information needed to located an unknown point $x \in X$ with respect to $\mathcal{A}$ (that is, to name the atom in which x lies), while $H_\mu(\mathcal{A}/\mathcal{B})$ is a measure of the information needed to locate x with respect to $\mathcal{A}$ if its location with respect to $\mathcal{B}$ is known.

Entropy is subadditive in the sense that $H_\mu(\mathcal{A} \vee \mathcal{B}) \leq H_\mu(\mathcal{A}) + H_\mu(\mathcal{B})$, with equality iff $\mathcal{A}, \mathcal{B}$ are **statistically independent** in the sense that for each pair of atoms we have $\mu(A_j \cap B_k) = \mu(A_j)\mu(B_k)$. Furthermore, we have $H_\mu(S^{\#}\mathcal{A}) = H_\mu(\mathcal{A})$, so $(1/2)\, H_\mu(\mathcal{A} \vee S^{\#}\mathcal{A}) \leq H_\mu(\mathcal{A})$, with equality iff $\mathcal{A}, S^{\#}\mathcal{A}$ are statistically independent; that is, with strict inequality if there are any statistical correlations between adjacent symbols $A_j, S^{-1}(A_k)$ produced by the source. This suggests defining the **n-th order entropy** of $(X, S, \mu, \mathcal{A})$ to be

$$\frac{1}{n} H_\mu(\mathcal{A} \vee S^{\#}\mathcal{A} \ldots \vee S^{\# n-1}\mathcal{A})$$

This quantity can be interpreted as a measure of the information per time unit needed to define the partial trajectory $x, S(x), \ldots S^{n-1}(x)$ of an unknown point $x \in X$, relative to $\mathcal{A}$. It takes account of all statistical correlations between digits separated by no more than n places and is maximized when all 2^n blocks of n digits are equally likely, and vanishes if only one such block is produced (that, if the source is periodic.)

Continuing this line of thought leads to the conclusion that the entropy of the join of $\mathcal{A}$ and its first $n-1$ pullbacks, *divided by n*, will decrease to a measure of the "average entropy per symbol", or better yet, the entropy per symbol of a typical sequence produced by the source, which takes account of statistical correlations of *all* orders. We define the **source entropy** of $(X, S, \mu, \mathcal{A})$ to be

$$h_\mu(S, \mathcal{A}) = \lim_{n \to \infty} \frac{H_\mu(\mathcal{A} \vee S^{\#}\mathcal{A} \cdots \vee S^{\# n-1}\mathcal{A})}{n}$$

Note that if we take the units of $H_\mu(\mathcal{A})$ to be bits (that is, take all logarithms base two), then the units of h will be bits per symbol.

The **zero-time partition** of the sequence space $\mathbb{B}^{\mathbb{Z}}$, written $\mathcal{E}$, is $\mathbb{B}^{\mathbb{Z}} = Z(0,0) \uplus Z(0,1)$. Note that $\sigma^{\#}\mathcal{E}$ is the partition

$$\mathbb{B}^{\mathbb{Z}} = Z(0,00) \uplus Z(0,01) \uplus Z(0,10) \uplus Z(0,11)$$

and similarly for $\sigma^{\# n}\mathcal{E}$. For the (p,q)-Bernoulli shift $(\mathbb{B}^{\mathbb{Z}}, \sigma, \mu)$, $\mathcal{E}$, $S^{\#}\mathcal{E}$, and so forth are all statistically independent, so we obtain the simple formula

$$h_\mu(\sigma, \mathcal{E}) = H_\mu(\mathcal{E}) = -p \log p - q \log q$$

Similarly, for the Markov shift $(\mathbb{B}^{\mathbb{Z}}, \sigma, \nu)$ defined by (P, π), because $S^{\#2}\mathcal{A}$ is statistically independent of $\mathcal{A}$, we have

$$h_\mu(\sigma, \mathcal{E}) = H_\nu(\mathcal{E}/S^{\#}\mathcal{E}) = -\sum_{j=0}^{1}\sum_{k=0}^{1} \pi_j P_{jk} \log P_{jk}$$

Since the n-th order entropy of the source depends only on correlations of length up to n, for purposes of computing it we can replace the actual source by an n-step Markov shift with the same n-step transition probabilities. This suggests trying to approximate the source itself by an n-step Markov shift. If the n-th order approximation to the source entropy of the actual source is a good one, the statistical characteristics of the Markov shift will be virtually indistinguishable from those of the actual source. See [22] for Shannon's striking illustration of this idea, using Markov approximations to natural English text.

7. Metric Entropy and Topological Entropy

Taking the supremum of the entropies over the sources $(X, S, \mu, \mathcal{A})$ for fixed (X, S, μ) gives the **metric entropy** of S:

$$h_\mu(S) = \sup_{\mathcal{A}} h_\mu(S, \mathcal{A})$$

This turns out to be a dynamical invariant of the MTDS (X, S, μ). (See [24] or [20].)

A classic theorem of Kolmogorov and Sinai says that if $\mathcal{E}$ is a **generating partition** [24][20], then

$$h_\mu(S) = h(S, \mathcal{E})$$

The precise definition of a generating partition need not concern us here; suffice it to say that the zero-time partition is a generating partition of any information source, so we immediately obtain that for the (p, q)-Bernoulli shift

$$h_\mu(\sigma) = h_\mu(\sigma, \mathcal{E}) = H_\mu(\mathcal{E}) = -p \log p - q \log q$$

while for the Markov shift defined by (P, π) we have

$$h_\nu(\sigma) = h_\nu(\sigma, \mathcal{E}) = H_\nu(\mathcal{E}) = -\sum_{j=0}^{1} \sum_{k=0}^{1} \pi_j P_{jk} \log P_{jk}$$

The Kolmogorov-Sinai theorem constituted a landmark in ergodic theory precisely because it provided an effective method of computing a nontrivial dynamical invariant (namely the metric entropy). To see the power of the theorem, consider the question of whether the Markov shifts defined

$$\tilde{A}_1 = \begin{pmatrix} 1/3 & 2/3 \\ 1 & 0 \end{pmatrix}, \ \pi_1 = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$
$$\tilde{A}_2 = \begin{pmatrix} 1/4 & 3/4 \\ 3/4 & 1/4 \end{pmatrix}, \ \pi_2 = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$

are isomorphic as MTDS's. Computing the entropies

$$h_{\mu_1}(\sigma) = (1/2) \log 3 - (1/3) \log 2 \approx 0.318257$$
$$h_{\mu_2}(\sigma) = \log 4 - (3/4) \log 3 \approx 0.562335$$

shows that they are not; the behaviour of the second shift is considerably more unpredictable, in a way which is detected by the metric entropy. Neither of these Markov shifts is isomorphic to the full two shift, for that has an entropy $h_{\mu_3}(\sigma) = \log 2 \approx 0.693147$.

The **topological entropy** of S, written $h(S)$, is defined in analogy with the above three-step procedure, except that one replaces "finite partitions" with "finite open covers" (a cover is a collection of open sets whose union is all of X), and defines $H(\mathcal{U}) = \log N(\mathcal{U})$, where $N(\mathcal{U})$ is the size of a minimal subcover of $\mathcal{U}$. The topological entropy is a dynamical invariant of the TDS (X, S). It was introduced by Adler; a second formulation of this concept (based upon ideas of Kolmogorov!) was introduced later by Bowen (see [24]).

Note well that the topological entropy of S is defined independently of the measures, while the metric entropy of S is only defined relative to an S-invariant Borel measure. The Variational Principle says that the supremum over all S-invariant Borel measures on X of the metric entropies is precisely the topological entropy of S. I think it is fair to say that this is a truly amazing theorem: there is absolutely no reason to expect such that any relationship at all between these two quantities would exist, much less such a simple one! This theorem was conjectured by Adler and proven in two halves by Goodman and Goodwyn about 1968.

See [24] for a proof of the Variational Principle and its generalization to the notion of topological pressure and equilibrium measures. See [14] for a survey paper which discusses a number of relations between metric entropy and quantities such as Lyapunov coefficients and correlation dimensions. See [15] for connections between topological and other entropies, and between topological entropy and the growth rate of various geometric quantities.

It can be shown [24] that the topological entropy of a shift is given by the following expression:

$$h(A) = \lim_{n \to \infty} \frac{\log B_n(A)}{n}$$

Here $B_n(A)$ is the number of blocks of length n which *do* occur in A (note this is never greater than 2^n). The numbers $B_n(A)$ in the case of the golden mean shift happen to form a Fibonnacci sequence, incidently.

One nice property of SFTs is that they possess a special shift invariant Borel measure, π, which actually achieves the supremum in the Variational Principle. The MTDS (A, σ, π) is a Markov shift and the metric entropy $h_\pi(A)$ can be computed using a simple generating partition $\mathcal{E}$, so whenever A is an SFT we have

$$h(A) = h_\pi(A) = h_\pi(A, \mathcal{E})$$

i.e. the topological entropy agrees with the entropy of the source $(A, \sigma, \pi, \mathcal{E})$. If (A, σ) is topologically transitive, there is a unique measure of maximal entropy, π, called **Parry measure**, and this measure is ergodic.

If A is a SFT, then the corresponding matrix, $\tilde{A}$, always has a unique positive real eigenvalue which is larger in modulus than any other eigenvalue (some of these might be complex). This special eigenvalue is called the **Perron eigenvalue** of $\tilde{A}$, and we have the simple formula

$$h(A) = \log \lambda$$

where λ is the Perron eigenvalue of $\tilde{A}$. You can check that in the case of the golden mean shift, $\lambda = (1 + \sqrt{5})/2$, whence $h(A) \approx 0.481212$.

If A is an SFT, then the periodic points of (A, σ); that is, the periodic strings in A, are dense in A. This implies that the following are equivalent:

(1) the characteristic polynomial of $\tilde{A}$ is irreducible (doesn't factor),
(2) For every (j, k) there is some n such that the (j, k)-th entry of $\tilde{A}^n$ is positive (where $\tilde{A}^n$ refers to matrix multiplication of $\tilde{A}$ with itself),
(3) (A, σ) is topologically transitive,
(4) (A, σ) is chaotic.

Moreover, if A is an SFT, then the following are equivalent:

(1) For some $n > 0$, $\tilde{A}^n$ has all positive entries,
(2) (A, σ) is a topologically mixing TDS.

These results should show what I meant earlier by saying that $\tilde{A}$ is more than just a convenient representation of a shift of finite type.

See [18] for the equivalent criteria just given, as well as details of entropy computations for SFT's and their cousins, the sofic shifts. For Parry measure and some basic information on shifts, see [4].

8. Algorithmic Complexity

In this section, we consider a notion of entropy which is completely different from the ones so far discussed.

Let U be a fixed Universal Turing Machine (no harm done if you think of this as your personal computer or your local mainframe, idealized to have an infinite memory, together with its operating system). Then given any *finite* binary string, x, we can define $K_U(x)$, the **algorithmic complexity** of x, to be the minimum size of any program which produces x when run on U. This is certainly well defined since, because x is finite, we can always write a program which says in essence "print the following string" and then continues with the string x. That is, there is some constant C such that $K_U(x) \leq length(x) + C$ for all x.

One important point to notice about this definition is that it depends upon which Universal Turing Machine you are using. Fortunately, if V is another such machine, it turns out that there is some constant C' such that for *all* finite strings x, we have

$$||K_U(x) - K_V(x)|| \leq C'$$

This implies that while it is true that if you change to another Universal Turing Machine, all your previous entropy computations will be off, they will be off by only a bounded function of x.

It is obviously very difficult to compute $K_U(x)$ directly from the definition. Indeed, the algorithmic complexity of strings is uncomputable in the technical sense that there exists no algorithm which, when run with any finite string s as input, is guaranteed to terminate in a finite time with output $K_U(s)$. However, this does not imply that one can *never* find exact values or even sharp estimates of complexity. For one thing, given a particular string, one may very well be able to prove that its algorithmic complexity has a certain value; certainly any program which actually produces that string gives an upper bound on the complexity.

There is, I think, no reason to expect any particular relationship between the algorithmic entropies of finite strings produced by an information source and the source entropy, which makes the following fact come as something of a suprise.

Consider a Bernoulli process which produces, every second, either a one, with probability $p(1)$, or zero, with probability $p(0)$, independently of the previous symbol. If x is a length n sequence produced by the process, the probability of observing x is

$$p^n(x) = p(x_1)p(x_2)...p(x_n)$$

where x_j is the j-th bit of x. There are 2^n possible binary strings of length n, and each of them has some complexity $K_U(x)$. Let E_n be the average of $K_U(x)$ over these strings, namely

$$E_n = \frac{1}{n} \sum_x p^n(x) K_U(x)$$

where the sum is taken over all sequences of length n. Then there is some constant C'' such that for all n, the difference between E_n and the usual entropy of a Bernoulli process

$$-p(0) \log p(0) - p(1) \log p(1)$$

is at most $(C'' + 2 \log n)/n$, which goes to zero as n increases. (See [2] for an elementary proof.)

This means that the complexity of "typical" long binary strings produced by a Bernoulli process will agree "pretty nearly" with the entropy of the source, although certain exceptional strings might have a large discrepancy. Since the source entropy carries information only about the characteristics of "typical" long sequences produced by the source, this is clearly the best we could hope for.

Given s in $\mathbb{B}^{\mathbb{N}}$, let s_n denote the first n digits of s. Then define the **upper** and **lower complexity** of s by

$$UK(s) = \limsup_{n\to\infty} \frac{K_U(s_n)}{n}$$
$$LK(s) = \liminf_{n\to\infty} \frac{K_U(s_n)}{n}$$

respectively. If these agree we write simply $K(s)$ for the complexity of s.

Note that UK, LK do not depend on U anymore, because as n grows, the difference $K_U(s_n)/n - K_V(s_n)/n$ must vanish. If they agree, it makes sense to call the common value, $K(s)$, the complexity of s. Next, given s in A, let $s+$ be the string in $\mathbb{B}^{\mathbb{N}}$ obtained by truncating everything before the zero-th place of s. Then $UK(s+)$ and $LK(s+)$ are well defined, and if they agree, we can call $K(s+)$ the complexity of s.

Since I said earlier that $K_U(x)$ cannot be too much larger than n if x is a string of length n, it might seem that the sequence $K_U(s_n)/n$ is likely to converge to zero. Actually, one of the basic properties of algorithmic complexity is that most strings have complexities comparable to their length, and thus near the upper limit of possible complexity. This means that $UK(s)$ and $LK(s)$ are in fact generally nonzero.

White [25] has obtained a number of results connecting algorithmic entropy and various of the entropies previously discussed. One of his results implies that if A is a topologically transitive SFT with Parry measure π, then for π-almost every string s in A, we have the identities

$$UK(s) = LK(s) = h(A)$$

See [23] or [21] for more information about algorithmic complexity.

9. Galois Entropy

We have so far considered examples of probabilistic, topological, and algorithmic entropies. In this section, we consider an entropy which is defined by purely algebraic means.

Consider the string 01011011 of $\mathbb{B}^8$. The **symmetric group** S_8 (the group of all permutations on the eight "places") **acts** on $\mathbb{B}^8$ in a natural way: for instance, the permutation $(x_1, x_3, x_4)(x_2, x_5)$ takes 01011011 to 1000111 (the zero in the first place moves into third place, the zero in third place moves into fourth place, the one in fourth place moves into first place, while the one in second place is exchanged with the zero in fifth place.) Now the classic way to measure the symmetries of a set is find a group which acts on that set, and to declare the **stabilizer** (the subgroup which does not move a particular element) to be the **symmetry group** of that element. In the case of S_8 acting on $\mathbb{B}^8$, the stabilizer of 01011011 is isomorphic to $S_3 \times S_5$ since any permutation which permutes the first, third and sixth places among themselves, and also the second, fourth, fifth, seventh and eigth places among themselves, will not alter 01011011. By a theorem of Lagrange, the size of the **orbit** of an element under a group acting on a set is the size of the group divided by the size of the stabilizer. In the case of S_n acting on $\mathbb{B}^n$, this will always yield a multinomial coefficient. For instance, in our example we have $8!/(3!5!) = 56$. That means there are 56 strings (including 01011011 itself) which you can obtain from 01011011 by permuting the digits. A little thought shows that this is the same as saying that there are 56 strings with precisely three zeros and five ones.

It turns out that the log of the orbit size, called the **galois entropy**, obeys all the same formal properties as does Shannon's probabilistic entropy. (Some readers might recognize that the orbit size for a finite string has the form a **thermodynamic probability** and that therefore galois entropies have the same form as Planck's notion of the entropy of a macrostate. Indeed, the notion of symmetry groups generalizes his distinction of macrostates from microstates; an individual element corresponds to a microstate and its orbit corresponds to the macrostate.)

Let A be a shift. Given x in A, let x^n be the string in $\mathbb{B}^n$ obtained by restricting x to $\{-n, -n+1, \ldots n\}$. We can the define **upper** and **lower galois entropy** of x by

$$UG(x) = \limsup_{n\to\infty} \frac{\log |S_n \cdot x_n|}{2n+1}$$
$$LG(x) = \liminf_{n\to\infty} \frac{\log |S_n \cdot x_n|}{2n+1}$$

respectively, where $S_n \cdot x_n$ denotes the orbit of S_n acting on x_n.

$UG(x)$ and $LG(x)$ may be interpreted as measures of the asymmetries of x. These are first order entropies, since in effect they simply count the relative frequency of zeros and ones. To construct second order entropies, we consider pairs $\{x_n, S(x)_n\}$. If we visualize the pair for (say) the string 010110111 like this:

10110111
01011011

it is easy to see how S_n acts on the set of all such pairs; we simply permute the columns of the array. Note these columns give the digrams of adjacent symbols occuring in the string 010110111. The stabilizer group of this particular pair is isomorphic to $S_3 \times S_2 \times S_3$, since the blocks 01, 10, 11 occur three, two, and three times respectively, which gives an orbit of size $8!/(3!2!3!) = 560$. We can define upper and lower second order galois entropies by

$$UG^2(x) = \limsup_{n\to\infty} \frac{\log |S_n \cdot \{x_n, S(x)_n\}|}{2n+1}$$
$$LG^2(x) = \liminf_{n\to\infty} \frac{\log |S_n \cdot \{x_n, S(x)_n\}|}{2n+1}$$

If these agree we write simply $G^2(x)$.

Note that $1/2 \log 560 \approx 3.14$ is less than $\log 56 \approx 4.02$; the first order approximation gives a smaller result, in general, because it takes account of more structure than does the zero-th order approximation. This behavior is familiar from the definition of the source entropy.

In the same way, we can define galois block entropies (per symbol) for all $k > 0$; the k-th order galois entropies measure the statistical asymmetry of sequences, taking into account correlations between as many as k adjacent digits.

If (A, σ) is topologically transitive, then, because the Parry measure π is then ergodic, the frequencies of k-grams will approach the correct probabilities as n grows, for all k and for π-almost every sequence. An application of Stirling's formula then shows that $UG^k(x)$, $LG^k(x)$, and Shannon's k-th order block entropies will all agree, for all $k > 0$. If we define the **galois entropy** by

$$g(x) = \lim_{k \to \infty} \frac{G^k(x)}{k}$$

then $g(x)$ will agree with the source entropy $h_\pi(A)$, for π-almost every x in A.

See [12][11] for more information about galois entropies.

10. Conclusion

Suppose A is a topologically transitive SFT, with Parry measure π, and let $\mathcal{E}$ be a generating partition for π. We can sum up the preceeding discussion by saying that the following entropies, defined in terms of different aspects of the structure of A, must all agree:

(1) the conditional entropy $H_\pi(\mathcal{E}/\sigma^\# \mathcal{E}) = H_\pi(\sigma^\# \mathcal{E}/\mathcal{E})$,
(2) the source entropy $h_\pi(\sigma, \mathcal{E})$,
(3) the metric entropy $h_\pi(\sigma)$,
(4) the topological entropy $h(\sigma)$,
(5) the algorithmic complexity $K(x)$, for π-a.e. x in A,
(6) the galois entropy $G(x)$, for π-a.e. x in A.

Thus, we have here six distinct notions of entropy defined in the very different contexts of probability, topology, the theory of algorithms, and algebra, which nevertheless all happen to agree!

I stress that when measure μ distinct from π is used, we will in general have $h_\mu(A) < h(A)$. When π is used with a non-generating partition $\mathcal{A}$, we will have $h_\pi(\sigma, \mathcal{A}) < h_\pi(\sigma, \mathcal{E})$. An "atypical" string x in A might have $K(x) < h(A)$, which means the trajectory is unusually easy to compute, or $G(x) < h(A)$, which means the trajectory is unusually symmetrical.

References

1. J. Banks, et al., *On Devaney's definition of chaos*, Amer. Math. Monthly, April 1992, 332–334.
2. Thomas M. Cover and Joy A. Thomas, *Elements of information theory.* Wiley, New York, 1991.
3. Robert L. Devaney, *Introduction to chaotic dynamical systems.* New York: Addison-Wesley, 1989.
4. Manfred Denker, Christian Grillenberger, and Karl Sigmund, *Ergodic theory on compact spaces.* New York: Springer, 1976.
5. Gerald B. Folland, *Real analysis: modern techniques and their applications.* New York: Wiley, 1984.
6. Robert Geroch, *Mathematical physics.* Chicago: University of Chicago Press, 1985.
7. Silviu Guiasu and Abe Shenitzer, *The principle of maximal entropy.* Math. Intelligencer 7:1 (1985), pp. 42–48.
8. Chris Hillman, *what is Hausdorff dimension?*, preprint, 1995.
9. —, *What is information?*, preprint, 1995.
10. —, *An entropy primer*, preprint, 1995.
11. —, *Symmetry and information*, preprint, 1995.
12. —, *A formal theory of information: statics*, preprint, 1995.
13. —, *An outline of the theory of G-sets*, preprint, 1995. To obtain copies of my preprints, see my home page at the URL:

    `http://www.math.washington.edu/~hillman/personal.html`

14. Jans Ledet Jensen, *Chaotic dynamical systems with a view towards statistics: a review.* In *Networks and chaos : statistical and probabilistic aspects.* Edited by O.E. Barndorff-Nielsen, J.L. Jensen, and W.S Kendall. New York : Chapman & Hall, 1993.

15. Anatole Katok and Boris Hasselblatt, *Introduction to the modern theory of dynamical systems.* Cambridge: Cambridge University Press, 1995.
16. A.I. Khinchin, *The mathematical foundations of information theory.* New York: Dover, 1957.
17. Andrej Lasota and Michael C. Mackey, *Chaos, fractals, and noise: stochastic aspects of dynamics.* New York: Springer, 1994.
18. Douglas Lind and Brian Marcus, *Introduction to Symbolic Dynamics and Coding.* Cambridge: Cambridge University Press, 1995.
19. Peter M. Neumann, Gabrielle A. Stoy, and the late Edward C. Thompson, *Groups and geometry.* New York : Oxford University Press, 1994.
20. Karl Petersen, *Ergodic theory.* Cambridge: Cambridge University Press, 1983.
21. Gregorz Rozenberg and Arto Salomaa, *Cornerstones of undecidability.* New York: Prentice Hall, 1994.
22. C. E. Shannon, *A mathematical theory of communication*, in C. E. Shannon and Warren Weaver, *The mathematical theory of communication.* University of Illinois Press, Urbana, 1949.
23. V. V. Vyugin, *Algorithmic entropy (complexity) of finite objects and its application to defining randomness and amount of information*, Selecta Mathematica (1994) 13 (4), 357–389.
24. Peter Walters, *Introduction to ergodic theory.* New York: Springer, 1981.
25. Homer S. White, *Algorithmic complexity of points in dynamical systems*, Ergodic Th. and Dyn. Systems (1993) 13, 807–830.

University of Washington, Department of Mathematics, Box 354350, Seattle, WA 98195-4350