AN ENTROPY PRIMER

CHRIS HILLMAN

Contents

1. Introduction 1
2. Probability Measure Spaces 2
3. The Entropy of a Partition 4
4. Conditional Entropy and Mutual Information 5
5. Formal Properties of Entropy 8
6. Probability Operators 10
7. A Very Brief Tour of Ergodic Theory 18
8. The Entropy of a Message Source 20
9. The Asymptotic Equipartition Theorem 21
10. Data Compression 24
11. Conditional Entropy and Information Rates 25
References 27

1. Introduction

In this expository paper, I discuss the theory of discrete probabilistic entropy, which was introduced in Parts I and II of the now classic 1948 paper by Shannon [20].

There are two major approaches to this subject. One approach is via probability theory and the other is via ergodic theory. I have followed the second approach here, for a number of reasons. First, I happen to find it easier to think about measureable functions than about random variables. Second, the ergodic theory approach is more closely connected with the recent explosion of work on the mathematical theory of chaotic dynamical systems, but is considerably less well known than the probabilistic viewpoint. Third, presenting the theory in terms of measureable functions (rather than random variables) makes it much easier to understand how the classical notion of entropy fits into the standard paradigm of twentieth century mathematics; this paradigm might be briefly expressed by the slogan "in order to understand some phenomenom of interest, one should define a category and then use invariants to classify the objects in this category". (See [11] for an extended discussion of how information theory fits into this paradigm.) Fourth, whereas the probability theory approach has been described in an excellent undergraduate level text [1]— not to mention the original paper by Shannon [20], which is still well

Date: May 21, 1996.

I am grateful to Geon Ho Choe and his ergodic theory students for correcting an error and spotting numerous typos in an earlier draft of this paper.

worth reading— the only comprehensive introduction known to me of the ergodic-theoretic viewpoint is the invaluable (but out of print) graduate level text [21]. Thus, I hope that this paper will help to fill a gap in the literature.

I will not discuss the Noisy Coding Theorem (see [1]) or metric entropy (see [21]) in this paper, but I *will* state and prove considerably more general versions of the Asymptotic Equipartition Theorem and Joint Asymptotic Equipartition Theorem than those given in [1]. Throughout the paper I will discuss at least one, and often more than one, intuitive interpretation of each quantitity we will define, illustrating them by a running example which I call "the Weatherman's Gambling Game". Proofs will be kept to a minumum.

The paper is largely self contained, although some familarity with the basic facts of measure-theoretical life will be very helpful.

2. Probability Measure Spaces

Definition 2.1. *Suppose X is a set equipped with a collection $\mathcal{M}$ of subsets $E \subset X$ such that $\mathcal{M}$ is closed under countable unions and complements; that is, whenever $E \in \mathcal{M}$, so is $X \setminus E$, and whenever we have a sequence E_j of sets in $\mathcal{M}$, then $\cup_{j=1}^{\infty} E_j$ is in $\mathcal{M}$. Then $\mathcal{M}$ is said to be a* **sigma-algebra** *on X.*

(The word "sigma" in "sigma-algebra" refers to sum, meaning union, while the word "algebra" indicates that $\mathcal{M}$ is defined in terms of certain formal operations, in this case unions and complements of sets.)

If $\mathcal{M}$ is a sigma-algebra on X, then taking any $E \in \mathcal{M}$ we see that $E \cup (X \setminus E) = X$ and $X \setminus X = \emptyset$ must be in $\mathcal{M}$; that is, sigma-algebras always contain both the largest and smallest subsets of X, namely X itself and the empty set $\emptyset$.

Definition 2.2. *Suppose $\mathcal{M}$ is a sigma-algebra on X, and suppose that $\mu : \mathcal{M} \longrightarrow [0, 1]$ is a function such that*

1 $\mu(\emptyset) = 0$,
2 $\mu(X) = 1$,
3 *given any sequence of* disjoint *sets E_j in $\mathcal{M}$, we have*

$$\mu\{\cup_{j=1}^{\infty} E_j\} = \sum_{j=1}^{\infty} \mu(E_j)$$

Then μ is said to be a **probability measure** *on $\mathcal{M}$, and $(X, \mathcal{M}, \mu)$ is called a* **probability measure space***.*

In this situation, sets $E \in \mathcal{M}$ are called **measureable sets**. If μ is a measure on $\mathcal{M}$, then whenever $E \subset F$ are measureable, we we have $\mu(E) \leq \mu(F)$. It turns out that $\mathcal{M}$ can only rarely be chosen to be the entire **power set** 2^X (the power set is the set of all subsets of X); that is, non-measureable subsets usually exist. However, they are so pathological that we will not need to worry about them.

If E is measureable, $\mu(E)$ can be considered a measure of the "size" of E. For instance, the usual notion of length on the unit interval $[0, 1]$ defines a probabilty measure, roughly as follows. Let $\mathcal{M}$ be the smallest sigma algebra containing all the closed sub-intervals of $[0, 1]$, and define the measure of each subinterval $I = [a, b]$ by $\lambda(I) = b - a$. It turns out that this can be extended to give a well defined probability measure on every set in $\mathcal{M}$; the resulting measure λ is called **Lebesgue measure** (on $[0, 1]$). Similarly, the usual notion of area on the unit square $[0, 1] \times [0, 1]$ defines a probability measure, also called Lebesgue measure (on the square).

If f is a non-negative function on $\mathbb{R}$ such that $\int f(x)dx = 1$, then we say that f is a **probability density**; classically, the probability of observing a value in the interval $[a,b]$ is $\int_a^b f(x)dx$. In measure theory this idea is extended to define the probability of any "event" $E \in \mathcal{M}$ by putting $\mu(E) = \int_E f(x)dx$, and it turns out that this defines a probability measure on $\mathbb{R}$. (We'll see some examples later in the paper.)

Unfortunately, it is not at all easy to visualize the details of how all this happens. One would be tempted to ignore the sophisticated concepts of measure theory and stick to the classical ideas of Laplace and Euler, which involve nothing more complicated than calculus, were it not for the following example, which is so important that it alone justifies using a measure theoretical approach, albeit at the expense of some handwaving in this section.

Let $X = \mathbb{B}^{\mathbb{N}} = \{x : \mathbb{N} \to \mathbb{B}\}$ be the set of all binary sequences indexed from zero to infinity. A **cylinder set** is a set such as

$$Z(01001) = \{x \in X : x(0) = 0, x(1) = 1, x(2) = 0, x(3) = 0, x(4) = 1\}$$

That is, choose some block of n digits and collect all sequences whose zeroth through $(n-1)$-st digits are given by corresponding digit of the given block; the result is a cylinder of **length** n. Note that we can decompose X into a disjoint union of 2^n cylinders of length n as follows

$$X = \cup_{k=0}^{k=2^n-1} Z([k]_2)$$

where $[k]_2$ denotes the base two representation of k. Let $\mathcal{M}$ be the sigma-algebra generated by combining (using countably many unions, intersections, and complements) the cylinder sets.

Example 2.3. *Suppose* $0 \le p, q \le 1$ *with* $p + q = 1$. *Then the* (p,q) Bernoulli measure *of any cylinder* $Z(w)$ *defined by a word* w *containing* m *zeros and* $n - m$ *ones is* $\mu\{Z(w)\} = p^j q^k$. *(We can extend this definition to define the* μ *measure of any set in* $\mathcal{M}$*, but the details won't concern us).*

It often happens that a probability measure space $(X, \mathcal{M}, \mu)$ is also a topological space; that is, has some notion of continuity, open and closed sets, and the like. In this situation, there is a natural sigma-algebra, generated by the closed (or open) subsets of X, which is called the **Borel sigma-algebra** $\mathcal{B}$. Any measure on $\mathcal{B}$ is called a **Borel** measure.

In particular, we can make the sequence space $\mathbb{B}^{\mathbb{N}}$ into a metric space (a topological space in which a notion of distance is defined) by declaring that the sequences x, y have a mutual distance of 2^{-n} if n is the smallest place where they first differ. It follows that the open ball around x of radius 2^{-n} consists of all sequences which agree with x at least up to the $(n-1)$-th place– in short, the balls in this topology are precisely the cylinder sets already defined. Thus, the sigma-algebra generated by the cylinders is precisely the Borel sigma-algebra for the sequence space. (Warning! This space is rather strange— it is homeomorphic to the Cantor set.)

The importance of sequence spaces will only become evident later in the paper, when we see how certain maps defined on sequence spaces give interesting and suprisingly simple models of extremely complicated dynamical situations.

For a readable and efficient introduction to the rigorous theory of measures, see the second chapter of the textbook [4]. The article [6] contains a detailed explanation of the relationship between probability measures and the classical notion of probability.

3. The Entropy of a Partition

Definition 3.1. *Suppose $(X, \mathcal{M}, p)$ is a probability measure space. Let Ω_X be the collection of all partitions of X into finitely many measureable subsets. If $\mathcal{A} \in \Omega_X$ denotes the partition $X = \cup_{j=1}^r A_j$, then each set A_j is called an* **atom** *of $\mathcal{A}$. Now let $\mathcal{B} \in \Omega_X$ denote the partition $X = \cup_{k=1}^s B_k$. Then the* **join** *of $\mathcal{A}$ and $\mathcal{B}$, written $\mathcal{A} \vee \mathcal{B}$, is the partition $X = \cup_{j=1}^r \cup_{k=1}^s A_j \cap B_k$. The* **trivial partition** *$X = X$ is denoted $\mathcal{Z}$. We say that $\mathcal{B}$* **refines** *$\mathcal{A}$, written $\mathcal{A} \leq \mathcal{B}$, whenever every $B \in \mathcal{B}$ is included in some $A \in \mathcal{A}$.*

In the partial ordering of Ω_X by $\leq$, $\mathcal{Z}$ is the smallest partition, and $\mathcal{A} \vee \mathcal{B}$ is the least upper bound of $\mathcal{A}, \mathcal{B}$; that is, the smallest partition refining both $\mathcal{A}$ and $\mathcal{B}$.

We can understand the intuitive significance of these definitions by considering the following example. Let us imagine that there is some huge but definite number n of atoms (idealized to be mutually indistinguishable) in the atmosphere over Omaha, NE. Each atom has a definite velocity and momentum (six real parameters in all) and the n atoms together define a point in a $6n$ dimensional **phase space** which represents the **microscopic state**, or microstate, of the atmosphere over Omaha. Naturally, human observers cannot hope to determine the velocities or momenta of individual atoms even approximately, but we can imagine that at a given time the detailed state of the atmosphere is nonetheless represented by such a point.

Now imagine a function α which takes each $x \in X$ to either 0 or 1 depending on whether or not the microstate x is associated with a state of rain in Omaha. Next, imagine a function β which takes each $x \in X$ to a temperature range. For instance, the range of β might consist of the three (Fahrenheit) temperature ranges

$$T < 30, \qquad 30 \leq T < 40, \qquad T \geq 40$$

Then, saying that $\beta(x_1)$ equals the temperature range $30 \leq T < 40$ is equivalent to saying that the microstate x_1 is associated with a ground temperature in that range. We can also imagine a function γ which takes each x to the windspeed associated with x. (It is understood that all these macroscopic parameters are measured at a particular "official" weather station in Omaha.)

Definition 3.2. *If $\sigma : X \to S$, where S is some* finite *set, is a measureable function, then σ is called a* **simple** *function. The* **kernel** *of σ is the partition of X into preimages under σ.*

In probability theory, a simple function is known as a **random variable**. Note that the preimages are *measureable sets* (that is the meaning of saying that σ is a measureable function.)

To resume our example: suppose that α is the precipitation function, β is the temperature function, and γ is the windspeed function. If kernels of α (precipitation), β (temperature), and γ (windspeed), respectively, are denoted $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, then $\mathcal{A} \vee \mathcal{B} \vee \mathcal{C}$ denotes a set of observable **macroscopic states** (or macrostates) of the weather over Omaha. For instance, the atom $A_3 \cap B_1 \cap C_2$ might consist of all those microstates which yield states of precipitation with temperature in the range $40 \leq T < 50^\circ$ F and windspeed in the range $0 \leq v < 5$ mph. More generally, we can consider statements like $x \in E_3$, where $E_3 \in \mathcal{E} \in \Omega_X$, to be denote **partial information** concerning the weather in Omaha.

Definition 3.3. *The* **entropy** *of $\mathcal{A} \in \Omega_X$, where $\mathcal{A}$ is the partition $X = \cup_{j=1}^r A_j$, is*

$$H(\mathcal{A}) = -\sum_{j=1}^r p(A_j) \log p(A_j)$$

$H(\mathcal{A})$ increases, all other things being equal, with the "fine-ness" of the partition, in the sense that $H(\mathcal{Z}) = 0$, and $H(\mathcal{A}) \leq H(\mathcal{B})$ whenever $\mathcal{A} \leq \mathcal{B}$. Moreover, $H(\mathcal{A})$ tends to increase as, holding the number of sets A_j fixed, we adjust the atoms A_j so as to make their probabilities more nearly equal. Let us abbreviate $p(A_1)$ by p_1 and so forth. Suppose that we increase p_1 by ε and decrease p_2 by the same amount, where we assume without loss of generality that $p_2 - p_1 > 2\varepsilon$. This has the effect of making the p_j very slightly more homogeneous, and the effect on the entropy is to increase it by about

$$\begin{aligned} dH &= \frac{\partial H}{\partial p_1} \cdot \epsilon + \frac{\partial H}{\partial p_2}(-\epsilon) \\ &= (-1 - \log p_1) \cdot \epsilon + (1 + \log p_2) \cdot \epsilon \\ &= \epsilon \log \frac{p_2}{p_1} \end{aligned}$$

It is not hard to see that for a fixed number r of sets A_j, H is maximized when all the p_j are equal; the maximum value is $\log r$. Thus, $H(\mathcal{A})$ measures some combination of "fine-ness" and "homogeneity".

A word about the units of measurement. Changing the base of the logarithm in the definition of entropy has the effect of multiplying H by a positive constant; therefore this can be considered "a change in units of entropy". In the engineering literature, it is customary to use logarithms base two, with the corresponding "units on information" being called **bits**. In the mathematical literature, it is customary to use natural logarithms, and that is the convention I follow here. The corresponding units of information are sometimes called **nats**.

We can obtain an "operational" interpretation of $H(\mathcal{A})$ as follows. Suppose that the Devil challenges us to guess whether or not it currently it is raining in Omaha, and promises us a penny if we guess correctly. In this situation, the two atoms of the partition $\mathcal{A}$ represent our two alternative guesses ("yes" or "no"), and $H(\mathcal{A})$ can be interpreted as a measure of the difficulty of guessing whether or not it is currently raining in Omaha. The point is that if the Devil had challenged us instead to guess whether it is currently raining in Death Valley, we would no doubt be much more confident of winning the penny.

We can sum up this discussion by saying that $H(\mathcal{A})$ may be considered a measure of the difficulty of guessing which atom an unknown point x lies in. Alternatively, we can say that $H(\mathcal{A})$ *measures, in a probabilistic sense, the variety of alternatives contained in $\mathcal{A}$.*

4. Conditional Entropy and Mutual Information

Definition 4.1. *If $\mathcal{B}$ is the partition $X = \cup_{k=1}^s B_k$, then the* **conditional entropy** *of $\mathcal{A}$ given $\mathcal{B}$ is $H(\mathcal{A}/\mathcal{B}) = H(\mathcal{A} \vee \mathcal{B}) - H(\mathcal{B})$.*

It is not hard to see that

$$\begin{aligned} H(\mathcal{A}/\mathcal{B}) &= -\sum_{j=1}^{r}\sum_{k=1}^{s} p(A_j \cap B_k) \log \frac{p(A_j \cap B_k)}{p(B_k)} \\ &= \sum_{k=1}^{s} p(B_k) H(B_k \cap \mathcal{A}) \end{aligned}$$

Here $B_k \cap \mathcal{A}$ is the partition $B_k = \cap_{j=1}^{r} A_j \cap B_k$, and B_k is given the **conditional probability** measure $p_k(E) = p(E \cap A_j)/p(B_k)$, so that

$$H(B_k \cap \mathcal{A}) = -\sum_{j=1}^{r} \frac{p(A_j \cap B_k)}{p(B_k)} \log \frac{p(A_j \cap B_k)}{p(B_k)}$$

Thus $H(\mathcal{A}/\mathcal{B})$ is the average, over the atoms of $\mathcal{B}$, of the entropies $H(B_k \cap \mathcal{A})$. $H(\mathcal{A}/\mathcal{B})$ can be interpreted as a measure of the difficulty of guessing which atom of $\mathcal{A}$ and unknown point x lies in, given that it lies in a specific atom of $\mathcal{B}$. Alternatively, we can say that *the conditional entropy $H(\mathcal{A}/\mathcal{B})$ measures the variety of alternatives left in $\mathcal{A}$ if a unique alternative in $\mathcal{B}$ is specified.*

Note that should "most" of every atom $B_k \in \mathcal{B}$ lie inside an atom $A_j \in \mathcal{A}$ then $H(\mathcal{A}/\mathcal{B})$ will be close to zero. On the other hand, should the regions A_j and B_k bear no particular relation to one another, then $H(\mathcal{A}/\mathcal{B})$ will be not much smaller than $H(\mathcal{A})$. Any change in $\mathcal{B}$ which tends to "de-homogenize" the conditional probabilites $p_k(A_j)$ will decrease the $H(B_k \cap \mathcal{B})$; this will decrease $H(\mathcal{A}/\mathcal{B})$ and increase our average earnings. This happens because such modifications of $\mathcal{B}$ permit more reliable estimation of which atom of $\mathcal{A}$ a point x lies in, given the knowledge that x is in a particular atom of $\mathcal{B}$.

We can obtain an "operational" interpretation of conditional entropy by imagining a modified guessing game. Let $\mathcal{A}$ denote the kernel of the precipitation function on the microstates of the atmosphere over Omaha, and let $\mathcal{E} = \mathcal{B} \vee \mathcal{C}$ denote current temperature and windspeed data for Omaha, as outlined above. Suppose that the Devil challenges us to guess whether or not it is presently raining in Omaha, but is now willing to inform us (truthfully) of the current temperature and windspeed. Our difficulty of guessing whether or not it is raining in Omaha, given the information available to us, is measured by $H(\mathcal{A}/\mathcal{E})$.

Additional motivation for the definition of conditional entropy is suggested by an idea due to Resnikoff [19]. Consider the situation of an experimenter making numerical measurements of some physical quantity x. Suppose the first measurement is a "coarse" one which determines the value of x to lie within the range $x_1 < x < y_1$. Next, perhaps by adjusting the settings of his intstruments on the basis of this knowledge, the experimenter is able to determine that the value of x lies within the smaller range $x_2 < x < y_2$, where $x_1 \leq x_2$ and $y_2 \leq y_1$. In this situation, it is reasonable to seek a measure of the information gained about the value x from this procedure. Resnikoff proves that if we demand that this measure be invariant under affine transformations of the real line (that is, it should not be altered by a change of scale or by translating the nested intervals $(x_2, y_2) \subset (x_1, y_1)$ to $(x_2+t, y_2+t) \subset (x_1+t, y_1+t)$), then we must take as our information measure some positive constant multiple of the quantity

$$I = \log \frac{|y_1 - x_1|}{|y_2 - x_2|}$$

Note that if we take Lebesgue measure m on the real line, then this quantity has the form $m(A_j \cap B_k)/m(A_j)$, where the A_j and B_k are finite length intervals; that is, it has the same form as the terms in $H(\mathcal{A}/\mathcal{B})$, except that m is not a probability measure.

Now consider what happens when we have two partitions $\mathcal{A}$ and $\mathcal{B}$ of the real line. Suppose $\mathcal{A}$ has the atoms

$$A_1 = (-\infty, a_1], A_2 = (a_1, a_2], \ldots A_r = (a_{r-1}, \infty)$$

while $\mathcal{B}$ has the atoms

$$B_1 = (-\infty, b_1], B_2 = (b_1, b_2], \ldots B_r = (b_{s-1}, \infty)$$

Suppose that $b_1 = a_1$ and $b_{s-1} = a_{r-1}$. Define $\mathcal{A} \vee \mathcal{B}$ as before; note that two atoms of $\mathcal{A} \vee \mathcal{B}$ will be infinite half rays (namely $A_1 = B_1$ and $A_r = B_r$) and the remainder will be finite intervals. If we ignore the infinite intervals and average the information gain determined by each inclusion $A_j \cap B_k \subset A_j$, we obtain

$$-\sum_{i=2}^{r-1}\sum_{j=2}^{s-1} \frac{m(A_j \cap B_k)}{L} \log \frac{m(A_j \cap B_k)}{m(A_j)}$$

where

$$L = a_{r-1} - a_1 = \sum_{j=2}^{r-1} m(A_j) = \sum_{k=2}^{s-1} m(B_k)$$

The former expression represents the expected information gain about the value of $a_1 < x < a_{r-1}$ when we refine the precision of our measurements. If we set $p(E) = m(E)/L$, then we have precisely the same form as obtained above for the conditional entropy $H(\mathcal{B}/\mathcal{A})$, where $\mathcal{A} \ll \mathcal{B}$.

Definition 4.2. *With $\mathcal{A}, \mathcal{B} \in \Omega_X$ as before, the* **mutual information** *of $\mathcal{A}, \mathcal{B}$ is*

$$I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A} \vee \mathcal{B})$$

Note this quantity is symmetric in $\mathcal{A}, \mathcal{B}$, unlike $H(\mathcal{A}/\mathcal{B})$. It is not hard to see that

$$\begin{aligned} I(\mathcal{A}, \mathcal{B}) &= H(\mathcal{A}) - H(\mathcal{A}/\mathcal{B}) = H(\mathcal{B}) - H(\mathcal{B}/\mathcal{A}) \\ &= -\sum_{j=1}^{r}\sum_{k=1}^{s} p(A_j \cap B_k) \log \frac{p(A_j)p(B_k)}{p(A_j \cap B_k)} \end{aligned}$$

The first two equalities show that $I(\mathcal{A}, \mathcal{B})$ can be interpreted in terms of our weather game as the information that we expect to gain about whether or not is raining in Omaha when the Devil tells us the current temperature range. The striking fact is that neither player need object if the roles of $\mathcal{A}$ and $\mathcal{B}$ are interchanged, since neither one stands to benefit from the change! This is because $\mathcal{A}$ gives the same amount of information about $\mathcal{B}$ as knowledge of $\mathcal{B}$ gives about $\mathcal{A}$.

It is easy to see that $I(\mathcal{A}, \mathcal{B}) = 0$ iff for all $A \in \mathcal{A}$ and all $B \in \mathcal{B}$, the sets A, B are **statistically independent** in the sense that $p(A \cap B) = p(A)p(B)$. At the opposite extreme, $I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A})$ iff knowledge of $\mathcal{B}$ determines $\mathcal{A}$ (up to measure zero). Thus, $I(\mathcal{A}, \mathcal{B})$ can be considered a measure of the mutual dependence, or better yet, the *statistical correlation* of $\mathcal{A}$ and $\mathcal{B}$. It would be quite wrong to infer from this that $\mathcal{A}$ is "causing" $\mathcal{B}$ or vice versa.

There is an interesting interpretation (due to J. Kelly) of $I(\mathcal{A},\mathcal{B})$ in terms of the expected financial gain in gambling games where "side information" is available; see [1] for details.

5. Formal Properties of Entropy

By "formal properties" I mean properties which do not depend upon the mathematical nature of the symbols $\mathcal{A},\mathcal{B}$, etc. The idea is probably best understood by considering some examples. The following are among the more important formal properties satisfied by discrete entropy and its brethern, conditional entropy and mutual information. Bear in mind that $H(\mathcal{A}/\mathcal{Z}) = H(\mathcal{A})$, where $\mathcal{Z}$ is the trivial partition.

1 *Quotient Rule.* $H(\mathcal{A}\vee\mathcal{B}/\mathcal{C}) = H(\mathcal{A}/\mathcal{C}) + H(\mathcal{B}/\mathcal{A}\vee\mathcal{C})$.

2 *Chain Rule.*[1] The entropy of the multiple join $H(\vee_{j=1}^{n}\mathcal{A}_j)$ can be expanded as follows:

$$H(\mathcal{A}_1) + H(\mathcal{A}_2/\mathcal{A}_1) + H(\mathcal{A}_3/\mathcal{A}_2\vee\mathcal{A}_1) + \ldots + H(\mathcal{A}_n/\vee_{j=1}^{n-1}\mathcal{A}_j)$$

3 *Entropy Balance.*

$$\begin{aligned} I(\mathcal{A},\mathcal{B}) &= H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}\vee\mathcal{B}) \\ &= H(\mathcal{A}) - H(\mathcal{A}/\mathcal{B}) \\ &= H(\mathcal{B}) - H(\mathcal{B}/\mathcal{A}) \end{aligned}$$

4 *Order Properties.* If $\mathcal{A}\le\mathcal{B}$, then

$$H(\mathcal{A}/\mathcal{C}) \le H(\mathcal{B}/\mathcal{C})$$

$$H(\mathcal{C}/\mathcal{A}) \ge H(\mathcal{C}/\mathcal{B})$$

5 *Redundancy.* $H(\mathcal{A}\vee\mathcal{B}/\mathcal{A}\vee\mathcal{C}) = H(\mathcal{B}/\mathcal{A}\vee\mathcal{C})$.

6 *Subadditivity.* $H(\mathcal{A}\vee\mathcal{B}/\mathcal{C}) \le H(\mathcal{A}/\mathcal{C}) + H(\mathcal{B}/\mathcal{C})$.

7 *Dependency Criteria.* We say that $\mathcal{A}$ depends on $\mathcal{B}$, written $\mathcal{A}\ll\mathcal{B}$, iff every atom of $\mathcal{A}$ is included "up to null set" in some atom of $\mathcal{B}$. (Note that $\mathcal{A}\le\mathcal{B}$ implies $\mathcal{A}\ll\mathcal{B}$ but not conversely). The following are equivalent:
- $\mathcal{A}\ll\mathcal{B}$.
- $H(\mathcal{A}/\mathcal{B}) = 0$.
- $H(\mathcal{A}\vee\mathcal{B}) = H(\mathcal{B})$.

8 *Codependency Criteria.* We say that $\mathcal{A},\mathcal{B}$ are **codependent**, written $\mathcal{A}\approx\mathcal{B}$, if their atoms can be paired off so that they agree up to null sets. This is the natural notion of when two partitions of X are "indistinguishable". The following are equivalent:
- $\mathcal{A}\approx\mathcal{B}$.
- $H(\mathcal{A}/\mathcal{B}) = H(\mathcal{B}/\mathcal{A}) = 0$.
- $H(\mathcal{A}\vee\mathcal{B}) = H(\mathcal{A}) = H(\mathcal{B})$.

9 *Independence Criteria.* We say that $\mathcal{A},\mathcal{B}$ are independent iff for every $A_j\in\mathcal{A}$ and $B_k\in\mathcal{B}$ we have $p(A_j\cap B_k) = p(A_j)p(B_k)$. The following are equivalent:
- $I(\mathcal{A},\mathcal{B}) = 0$.
- $H(\mathcal{A}/\mathcal{B}) = H(\mathcal{A})$.
- $H(\mathcal{B}/\mathcal{A}) = H(\mathcal{B})$.
- $H(\mathcal{A}\vee\mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B})$.

[1] This vivid and appropriate name comes from [1].

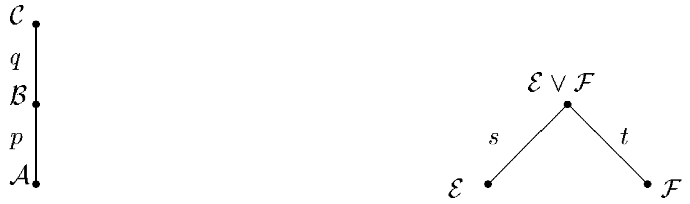

FIGURE 1. Left: $D(\mathcal{A},\mathcal{C}) = D(\mathcal{A},\mathcal{B}) + D(\mathcal{B},\mathcal{C}) = p + q$. Right: $D(\mathcal{E},\mathcal{F}) = D(\mathcal{E},\mathcal{E} \vee \mathcal{F}) + D(\mathcal{E} \vee \mathcal{F},\mathcal{F}) = s + t$.

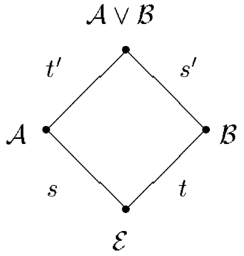

FIGURE 2. The Diamond Lemma says that $s' \leq s$, $t' \leq t$, and $s + t' = t + s'$.

10 *Class Functions.* $H(\cdot)$ is a **class function** in the sense that it is constant on each codependency class. Similarly, $H(\cdot/\cdot)$ and $I(\cdot,\cdot)$ are constant on codependency classes, and $\ll$ forms a partial order on the codependency classes.

11 *Metric Space.* $D(\mathcal{A},\mathcal{B}) = H(\mathcal{A}/\mathcal{B}) + H(\mathcal{B}/\mathcal{A})$ defines a **metric** on the set of codependency classes. In [8] I develop the **entropy geometry** defined in this way on the poset formed by the classes of paritions $\mathcal{A} \in \Omega_X$ together with the partial order $\ll$. The following three items give a taste of this geometric picture— note that for convenience of notation we do not distinguish between $\mathcal{A}$ and the codependency class of $\mathcal{A}$.

12 *Chain Addivity.* (See Figure 1a.) If $\mathcal{A} \ll \mathcal{B} \ll \mathcal{C}$, then $D(\mathcal{A},\mathcal{C}) = D(\mathcal{A},\mathcal{B}) + D(\mathcal{B},\mathcal{C})$.

13 *Lambda Property.* (See Figure 1b.) We have the identity $D(\mathcal{E},\mathcal{F}) = D(\mathcal{E},\mathcal{E} \vee \mathcal{F}) + D(\mathcal{E} \vee \mathcal{F},\mathcal{F})$.

14 *Diamond Lemma.* (See Figure 2.) Suppose $\mathcal{E} \ll \mathcal{A},\mathcal{B}$. Then $D(\mathcal{E},\mathcal{A}) \leq D(\mathcal{B},\mathcal{A} \vee \mathcal{B})$ and $D(\mathcal{E},\mathcal{B}) \leq D(\mathcal{A},\mathcal{A} \vee \mathcal{B})$. Moreover,

$$D(\mathcal{E},\mathcal{A}) + D(\mathcal{A},\mathcal{A} \vee \mathcal{B}) = D(\mathcal{E},\mathcal{B}) + D(\mathcal{B},\mathcal{A} \vee \mathcal{B})$$

The Diamond Lemma is analogous to well-known property of group indices. This is no accident; see [8][7].

15 *Lipschitz Continuity.* Entropy is Lipschitz continuous with respect to the entropy distance. Specifically, $|H(\mathcal{A}) - H(\mathcal{B})| \leq D(\mathcal{A},\mathcal{B})$. Similarly for conditional entropy and mutual information.

See [8] for proofs and an extensive discussion (in a more abstract context) of the intuitive significance of the Quotient Rule and other formal properties of entropy.

6. Probability Operators

In this section we begin to consider dynamical processes.

Definition 6.1. *Let $(X, \mathcal{M}, p)$ and $(Y, \mathcal{N}, q)$ be probability measure spaces. The map $\varphi : X \to Y$ is a* **measure-preserving transformation** *if for all $E \in \mathcal{M}$, we have $p\{\varphi^{-1}(E)\} = p(E)$. A measure-preserving map S from $(X, \mathcal{M}, p)$ to itself, written $(X, \mathcal{M}, p, S)$, is a* **probability operator**.

This terminology is intended to suggest an analogy with linear operators. A linear operator T is a mapping $V \to V$, where V is an object possessing a certain mathematical structure (linear structure) which is preserved by T. Similarly, a probability operator S is a mapping $X \to X$, where X is an object possessing a certain structure (a probability measure) which is preserved by S.

Given a microstate $x \in X$, the sequence $x, S(x), S^2(x) \ldots$ is called the **trajectory of x under** S, or the **microhistory** of x. Of course, we should regard microhistories as unknowable in practice. In the next section we will consider "macrohistories" which *are* observable.

Example 6.2. *Let $D(\theta) = 2\theta$ (taken modulo 2π) denote the angle doubling map on the unit circle S^1. We use the natural probability measure on the Borel sets, $\mathcal{B}$, of the unit circle, namely*

$$q(E) = \frac{1}{2\pi} \int_E d\theta$$

If $0 < A < B < 2\pi$, so that (A, B) is a typical interval on the circle, then

$$D^{-1}(A, B) = (A/2, B/2) \cup (\pi + A/2, \pi + B/2)$$

is the union of two intervals of half the length of the original. Thus, $(S^1, \mathcal{B}, q, D)$ is a probability operator. It is called the **doubling map**.

Note that if I is a small interval on the circle, D takes I to an interval twice as long, so that $q(D(I) = 2q(I)$, whereas the *preimage* of I is *two* intervals each half as long as I itself, so that $q\{D^{-1}(I)\} = q(I)$. This may help to clarify the definition of a measure preserving map.

Another example of a probability operator which is easy to study is the following.

Example 6.3. *Let $T : [0, 1] \to [0, 1]$ be defined by*

$$T(x) = \begin{cases} 2x & 0 \leq x < 1/2 \\ 2 - 2x & 1/2 \leq x \leq 1 \end{cases}$$

We use the natural probability measure given by ordinary integration (or Lebesgue measure); $p(E) = \int_E dx$. It is easy to see that the preimage of a small interval is the union of two disjoint intervals half as long, so $([0, 1], \mathcal{B}, p, T)$ is a probability operator. It is called the **tent map** *(see Figure 3).*

Here is an example which is quite different from the preceeding two examples.

Example 6.4. *Let $X = \{x : \mathbb{N} \to \mathbb{B}\}$ be the set of all binary sequences indexed from zero to infinity, with the (p, q) Bernoulli measure. Define the* **left shift map** *$S : X \to X$ by $(Sx)(n) = x(n + 1)$. Note that S shifts every sequence one place to the left, "erasing" the first symbol as it does so. Thus, it is a two-to-one map, and the shift of the cylinder of length $n > 0$ $Z(0, n, x)$ is the disjoint union of two cylinders of length $n - 1$. Therefore the p measure of the shift preimage of any*

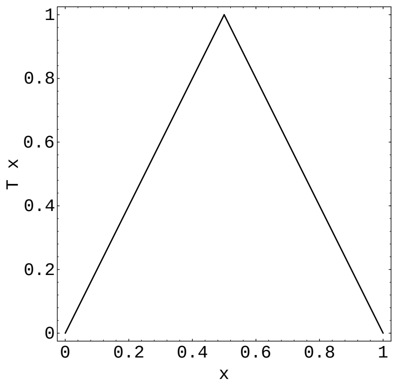

FIGURE 3. The tent map $T(x)$.

cylinder Z agrees with the measure of Z; this is enough to guarantee that S is a measure preserving map. Therefore $(X, \mathcal{M}, p, S)$ is a probability operator. It is called the **one-sided (p, q) Bernoulli shift**.

The shift map is continuous in the metric topology introduced earlier. It "inflates" the ball of radius 2^{-n} around x to the larger ball of radius 2^{1-n} around x, and iterating S inflates the ball to the entire space in a finite time. The preimage $S^{-1}(B)$ of a ball B of radius 2^{-n} consists of *two* smaller balls of radius 2^{-n-1}.

The $(1/2, 1/2)$ Bernoulli shift can be used to model the process of tossing a fair coin. Each $x \in X$ determines a trajectory which corresponds to one possible result of an infinite sequence of coin tosses, where 0 is interpreted as "tails" and 1 as "heads".

We would like to have some way of comparing the "complexity" of the behavior of various probability operators, and in particular identifying those which are as complex as coin tossing. The following definition provides the neccessary concept.

Definition 6.5. *Suppose $(X, \mathcal{M}, p, S)$ and $(Y, \mathcal{N}, q, T)$ are probability operators and suppose $\varphi : X \to Y$ is a measure preserving map such that $\varphi \circ S = T \circ \varphi$; that is, such that the diagram*

$$\begin{tikzcd}
X \arrow[r, "S"] \arrow[d, "\varphi"'] & X \arrow[d, "\varphi"] \\
Y \arrow[r, "T"'] & Y
\end{tikzcd}$$

commutes. Then φ is a **probability operator homomorphism** *or* **factor map** *from $(X, \mathcal{M}, p, S)$ to $(Y, \mathcal{N}, q, T)$.*

If φ has an inverse map which is also an operator homomorphism, it is said to be a **metric conjugacy**. In this case, we can think of the two probability operators as being "equivalent" from the point of view of probabilistic phenomena. These definitions give a category [18] in which objects are probability operators, morphisms are operator homomorphisms, and composition is the ordinary composition of mappings.

Because an operator homomorphism ϕ from $(X, \mathcal{M}, p, S)$ to $(Y, \mathcal{N}, q, T)$ is onto a.e., by definition, we can think of $(Y, \mathcal{N}, q, T)$ as a "simplified model" of $(X, \mathcal{M}, p, S)$;

hence the term "factor map". (In this situation, we can also say that $(Y, \mathcal{N}, q, T)$ is a **factor** of $(X, \mathcal{M}, p, S)$.) Incidently, the above definitions are not completely satisfactory because of the somewhat inconsistent way in which they "ignore sets of measure zero"; therefore many authors prefer to set up the theory in terms of slightly different categories; see [6][3] for two alternative categories. However, these defects are too subtle to worry about here, so we have adopted the above definitions because of their simplicity and clear analogy with linear operators.

It often happens that two very different appearing probability operators are in fact conjugate; for instance, the $(1,1)$ Bernoulli shift is conjugate to the doubling map! To see this, consider the map $\xi : X \to S^1$ defined by

$$\xi(x) = \pi \sum_{k=0}^{\infty} \frac{x(k)}{2^k}$$

Since $x(k)$ is either zero or one for every k, we see that $0 \leq \xi(k) \leq 2\pi$. Moreover, if x, y disagree in the n-th place, then $|\xi(x) - \xi(y)| \geq 2^{-n}$, so ξ is one-one (except perhaps at the endpoints, but this doesn't matter). Now suppose $[A, B)$ is an interval on the circle such that $A/(2\pi) = j2^{-n}$ and $B/(2\pi) = k2^{-n}$ are "dyadic rationals"; then $\xi^{-1}[A, B)$ is a cylinder of length n. This shows that $p\{\xi^{-1}[A, B)\} = q[A, B)$, so ξ is a measure preserving map. Finally, by considering separately the cases $x(0) = 0$ and $x(0) = 1$, it is not hard to see that $\xi(Sx) = D(\xi x)$, so ξ gives the desired conjugacy. (Here, ξ is also a homeomorphism, so in fact S and D are conjugate as **topological operators** as well; see [9].) This shows that the doubling map is precisely as "chaotic" as coin tossing!

The following construction is very useful for constructing new probability operators which are factors of a given operator.

Definition 6.6. *Let $(X, \mathcal{M}, p)$ be a probability measure space and suppose $\varphi : X \to Y$ is some map. Define $\mathcal{N}$ to be the sigma-algebra generated by all sets $E \subset Y$ such that $\varphi^{-1}(E) \in \mathcal{M}$, and define q by $q(E) = p\{\varphi^{-1}(E)\}$. Then $(Y, \mathcal{N}, q)$ is a new probability measure space, called the* **pushout** *of $(X, \mathcal{M}, p)$ via φ.*

The point is that if we have a given operator $(X, \mathcal{M}, p, S)$ and can find a map $\varphi : X \to Y$, where Y is just some set, such that $\varphi \circ S = T \circ \varphi$ for some map $T : Y \to Y$, then when we give Y the pushout structure via φ, T becomes a new probability operator which is automatically a factor of $(X, \mathcal{M}, p, S)$! This is easy to see, since

$$q\{T^{-1}(E)\} = p\{\varphi^{-1}T^{-1}(E)\} = p\{S^{-1}\varphi^{-1}(E)\} = p\varphi^{-1}(E) = q(E)$$

where I used (left to right) the definition of q, the fact that $\varphi \circ S = T \circ \varphi$, the fact that $p\{S^{-1}(F)\} = p(F)$, and the definition of q. If X, Y are topological spaces and $\mathcal{M}$ is the Borel sets on X, and if φ happens to continuous μ-a.e., then the pushout is a Borel measure on Y. However, although φ is automatically "almost onto", even if it is also "almost one-one", φ need not be a conjugacy! (This is one of the subtle deficiencies earlier alluded to).

I shall illustrate this method of finding factors of a given probability operator by constructing two interesting new operators which are factors of the doubling map, and thus of the $(1/2, 1/2)$ Bernoulli shift.

Example 6.7. *Let $L(x) = 4x(1-x)$; since L maps the unit interval to itself we may consider L to be a map $L : [0,1] \to [0,1]$. It is called the* **logistic map** *(see*

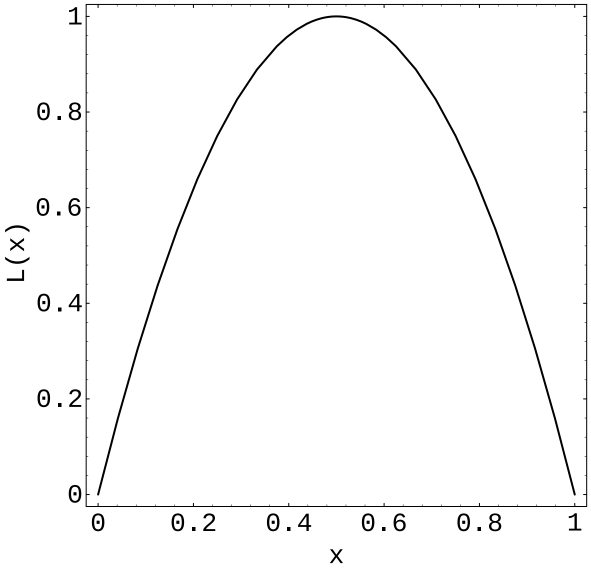

FIGURE 4. The Logistic Map $L(x) = 4x(1-x)$.

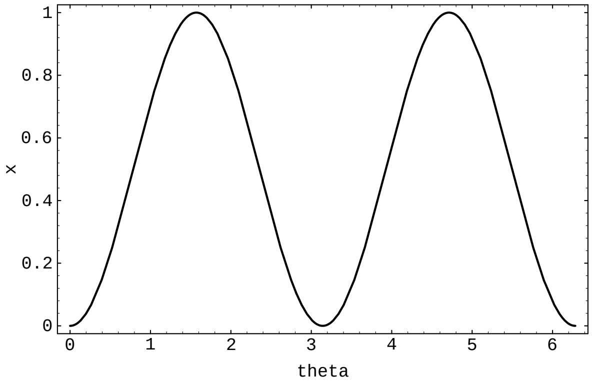

FIGURE 5. The transformation $\psi(\theta) = \sin^2(\theta)$.

Figure 4). If we define $\psi : S^1 \to [0,1]$ *by*

$$\psi(\theta) = \sin^2(\theta)$$

(see Figure 5) then ψ *is continuous and*

$$\begin{aligned}(L \circ \psi)(\theta) &= 4\,\sin^2(\theta)\,\cos^2(\theta)\\ &= \sin^2(2\theta)\\ &= (\psi \circ D)(\theta)\end{aligned}$$

(This shows that L *is a factor of* D *as a topological operator; see* [9]*.) Therefore we can define a new probability operator on* $[0,1]$ *which is a factor of the doubling map, as follows. Let* $I = [a,b]$ *be some small subinterval of* $[0,1]$*. Then the pushout*

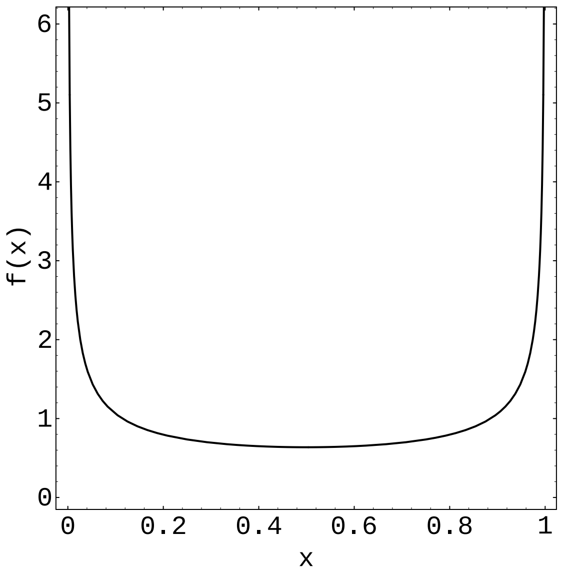

FIGURE 6. The L-invariant density on the unit interval.

measure of I is

$$\begin{aligned} \nu(I) &= \lambda\{\psi^{-1}(I)\} \\ &= \frac{1}{2\pi}\int_{\arcsin\sqrt{a}}^{\arcsin\sqrt{b}} d\theta \\ &= \int_a^b f(x)\,dx \end{aligned}$$

where

$$f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$$

is the desired L-invariant density (see Figure 6).

This implies [17] that if we iterate L on a computer, starting with various initial points, the trajectories of each point may be very complicated (and essentially unpredictable in the long term), but if we divide $[0,1]$ into small subintervals and make a histogram showing how many of the first one million iterates fell in a given range, we will see something that looks very much like the graph of $f(x)$; in particular, more iterates will fall near zero or one than fall near the center of the interval. This regular and predictable statistical behavior constrasts strongly with the chaotic behavior of the trajectories, each of which can be said to be a slightly simplified model of a random coin toss.

Note that ψ is a four-to-one map except at $\theta = n\pi/2, n = 0,1,2,3$ (where it is two-to-one). Thus, it cannot give a conjugacy. Indeed, there is no conjugacy between L and D.

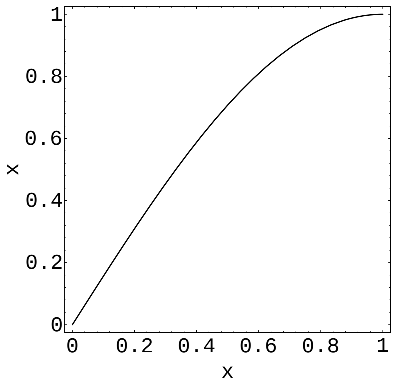

FIGURE 7. The transformation $\xi(x) = \sin^2(\pi x/2)$.

To see that L is conjugate to T, let $\xi(x) = \sin^2(\pi x/2)$. (See Figure 7.) Then

$$\begin{aligned}(\xi \circ T)(x) &= \begin{cases} \sin^2(\pi x), & 0 \le x < 1/2 \\ \sin^2(\pi - \pi x), & 1/2 \le x \le 1 \end{cases} \\ &= \sin^2(\pi x) \\ &= 4 \sin^2(\pi x/2) \cos^2(\pi x/2) \\ &= (L \circ \xi)(x)\end{aligned}$$

It is not hard to check that ξ is measure preserving, so it is a factor map. (Indeed, ξ is a homeomorphism, so that L and T are also conjugate as topological operators; see [9].)

Example 6.8 (Doug Lind)**.** *Define*

$$N(x) = \begin{cases} 0 & x = 0 \\ \frac{x - 1/x}{2} & x \neq 0 \end{cases}$$

on the real line. Then

$$\varphi(\theta) = \begin{cases} 0 & \theta = 0 \\ -\cot(\theta/2) & 0 < \theta < 2\pi \end{cases}$$

(see Figure 9) defines a map from the circle to the real line, which is continuous almost everywhere, and it is not hard to see (consider the case $\theta = \pi$ seperately) that $N \circ \varphi = \varphi \circ D$, so we can define a new probability operator on the real line which is conjugate to the doubling map, as follows. Let $I = [a, b]$ be some subinterval of $[0, 1]$ and define the pushout measure of I by

$$\begin{aligned}\mu(I) &= \lambda\{\varphi^{-1}(I)\} \\ &= \frac{1}{2\pi} \int_{-2/\arctan(a)}^{-2/\arctan(b)} d\theta \\ &= \int_a^b g(x) dx\end{aligned}$$

where

$$g(x) = \frac{1}{\pi(1 + x^2)}$$

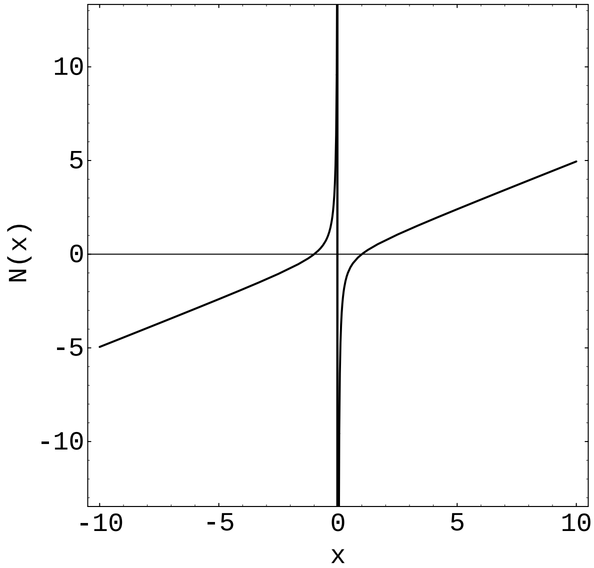

FIGURE 8. The map $N(x)$.

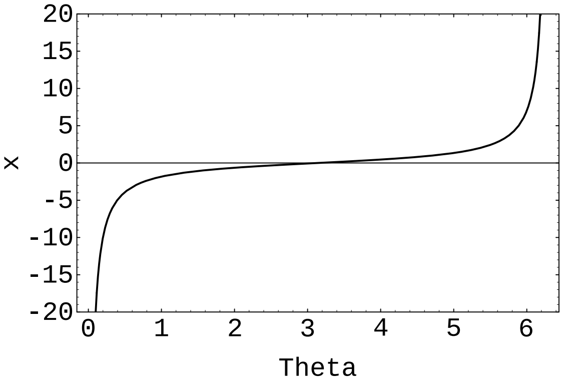

FIGURE 9. The transformation $\varphi(\theta)$.

is the desired N-invariant measure on $\mathbb{R}$ (see 10).

N is the map which arises in applying Newton's Method to the problem of evaluating $\sqrt{-1}$, that is finding a root of $h(x) = x^2 + 1$, with an initial guess which is real. To see this, recall that Newton's method for finding a root of $h(x) = 0$, where h is differentiable, replaces the guess x_n by x_{n+1}, where $h(x)/(x_{n+1} - x_n) = h'(x)$. This amounts to iterating the function $h^*(x) = x - h(x)/h'(x)$. In the case $h(x) = x^2 + 1$ we have $h^*(x) = \frac{x - 1/x}{2} = N(x)$.

Thus, we see that in this example, Newton's method not only fails to converge either of the actual roots (which are of course imaginary), but is essentially as "random" in the details of the trajectory as is a coin toss! Nevertheless, if you iterate some initial guess on the computer and make a histogram as above, we will find something resembling the graph of the N-invariant density g; in particular,

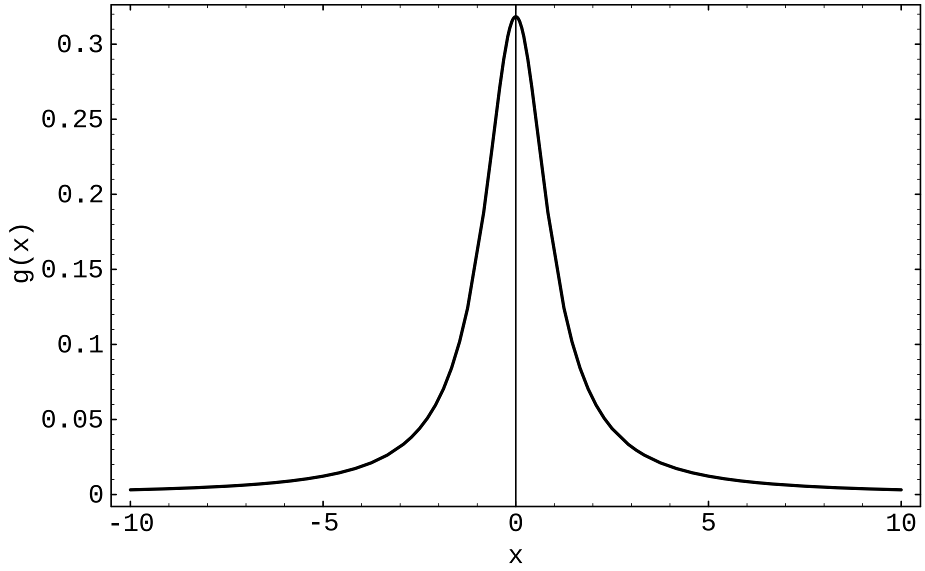

FIGURE 10. The N-invariant density g.

iterates tend to remain near the origin, and rarely stray very far away from the origin. The density g is sometimes called "the Poor Man's Bell Curve", because its Taylor expansion begins the same way as that of the standard normal distribution. This phenomenom well illustrates a fundamental feature [17] of ergodic probability operators: *while their individual trajectories may be very badly behaved, they nonetheless exhibit a rigorous* statistical *regularity.*

To sum up: the one-sided $(1/2, 1/2)$ Bernoulli shift and the angle doubling map on the circle are conjugate probability operators. The tent and logistic maps on the unit interval and the Newton map $N(x) = x - 1/x$ on $\mathbb{R}$ are mutually conjugate probability operators; these are all factors of the angle doubling map and thus, of the one sided $(1/2, 1/2)$ Bernoulli shift.

Suppose P is a matrix whose entries are all real numbers between zero and one, such that the row sums are always exactly one. We can consider P to define **transition probabilities** between states of a stochastic process. Suppose that we happen to know a row vector π, whose entries are non-negative numbers adding up to one, such that $P\,\pi = \pi$. 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 **Markov measure** on $\mathbb{B}^{\mathbb{N}}$ defined by the pair (P, π) gives the cylinder $Z(01001)$ the measure

$$\mu\{Z\} = \pi(0)\,P(0,1)\,P(1,0)\,P(0,0)\,P(0,1)$$

and similarly for other cylinders. Note that Markov measures are defined in a way which ensures that they are shift invariant.

Definition 6.9. *Let the sequence space $\mathbb{B}^{\mathbb{N}}$ be equipped with the Markov measure μ defined by some (P, π), and let S denote the shift map. Then the probability operator $(\mathbb{B}^{\mathbb{N}}, \mu, S)$ is called a* **Markov shift**.

Unlike Bernoulli shifts, sequences in Markov shifts possess statistical correlations between pairs of adjacent digits, but no higher order correlations.

7. A Very Brief Tour of Ergodic Theory

We begin our tour with the very first theorem in this subject, proven by Poincare in 1890.

Theorem 7.1 (Recurrence Theorem). *Let $(X, \mathcal{M}, p, S)$ be any probability operator. Then for all $A \in \mathcal{M}$, almost every $x \in A$ is* **recurrent** *in A, in the sense that for some $k > 0$, $S^k(x) \in A$.*

This theorem has been interpreted as implying that the Second Law of Thermodynamics must hold only "on average". Supposedly, if you simply wait long enough, you will eventually observe small and temporary accretions of "order" (with resulting decreases in physical entropy), because the Recurrence Theorem implies that every dynamical system must eventually return to states arbitrarily "close" to its initial state[2].

Proof: Let $F \subset A$ be the nonrecurrent points; that is,

$$F = \{x \in A : \forall j > 0, \; S^j \notin A\}.$$

CLAIM: the sets $S^{-j}(F)$ are mutually disjoint.

Reason: if $x \in S^{-k}(F)$, then $S^k(x) \in A$, so $x \notin F$. Thus $S^{-k}(F) \cap F = \emptyset$; pulling back both sides by S^j shows that $S^{-k-j}(F) \cap S^{-j}(F) = \emptyset$.

CONSEQUENCE: since S is measure preserving, we have a countably infinite collection of disjoint measureable sets sharing the same measure, so they must all be null sets. In particular, F is null, which is what we wanted to show. ∎

Definition 7.2. *Suppose $(X, \mathcal{M}, p, S)$ is a probability operator. Then $E \in \mathcal{M}$ is said to be an S-***invariant set** *if $S^{-1}(E) \Delta E$ is a null set. Similarly, $f \in L^1(X, \mathcal{M}, p)$ is an S-***invariant function** *if $f(S(x)) = f(x)$ for almost every $x \in X$.*

Here $A \Delta B = (A \cup B) \setminus (A \cap B)$ is the **symmetric difference** of A, B (the set of all x in one or the other but not both of A, B). A **null set** is any set of measure zero, and "for almost every x" means for all x not in some null set.

We are now in a position to introduce the central notion in ergodic theory.

Definition 7.3. $(X, \mathcal{M}, p, S)$ *is an* **ergodic operator** *if every S-invariant set has measure either zero or one.*

We will need two more technical definitions. We say that the sequence of functions f_n **converges a.e.** to f means that for all $x \in E$, where E is some set such that $p(E) = 1$, the numbers $f_n(x)$ converge to $f(x)$ in the usual sense of calculus. More generally, saying that something "holds true a.e." means "except on a set of measure zero". Similarly, f_n **converges in mean** to f if the integrals

$$\int |f_n - f| dp \to 0$$

[2] The average time interval between recurrences is inversely proportional to $p(A)$, according to another famous theorem, due to Kac; see [12]. In practice this means that any real isolated physical system will usually be destroyed by other physical processes long before any truly striking entropy decreases have time to occur. This is why chemical engineers need not concern themselves with such "thermodynamic anomalies"!

(Neither of these definitions of convergence implies the other; see [4].)

Theorem 7.4 (Basic Ergodicity Criteria). *Let $(X, \mathcal{M}, p, S)$ be a probability operator. The following are equivalent:*

1 *LAW OF AVERAGES: For any integrable function α, the sequence of time averages*

$$\frac{1}{n}\sum_{j=0}^{n-1} \alpha(S^j(x))$$

converges a.e. and in mean to a constant. (If so, the constant in question can only be the phase space average $\int_X \alpha dp$.)

2 *LAW OF FREQUENCIES: For each $E \in \mathcal{M}$, for almost every x,*

$$\frac{1}{n}\sum_{j=0}^{n-1} \chi_E(S^j(x)),$$

the frequency with which the orbit of x enters E, approaches $p(E)$.

3 *ZERO-ONE LAW: Every S-invariant set has measure zero or one; i.e., the operator is ergodic.*

4 *LAW OF CONSTANT INVARIANTS: Every S-invariant integrable function is constant a.e.*

5 *TRANSITIVITY: For each pair of non-null $A, B \in \mathcal{M}$, for some $n > 0$, we have $S^{-n}(A) \cap B$ non-null.*

The Law of Averages is the famous pointwise ergodic theorem of Birkhoff (1936). The Law of Frequencies says that we can reliably estimate the probability of an event by looking at the empirical frequencies for almost any orbit; the expression $\sum_{j=0}^{n-1} \chi_E(S^j(x))$ simply counts the number of times the trajectory of x enters E for $0 \leq j \leq n-1$. (Here χ_E is the **characteristic function** of E, which is defined by setting $\chi_E(x) = 1$ if $x \in E$ and $\chi_E(x) = 0$ otherwise.) The Zero-One Law is, of course, simply our original definition of what it means for $(X, \mathcal{M}, p, S)$ to be an ergodic operator. The Law of Constant Invariants says that the only invariant functions are the ones (constant functions) that are neccessarily invariant under *any* measureable mapping $X \to X$. Transitivity means that if one fixes some region of phase space, i.e. some B non-null, then the probability for the iterates of a point $x \in A$ non-null to visit that region is nonzero. Note how each of these equivalent criteria adds new insight into the intuitive meaning of ergodicity.

Despite their apparently very restrictive properties, ergodic probability operators often arise in applications, although the ergodicity may be difficult to prove rigorously (as, for instance, in the case of billiards.) It is often easier to establish certain "mixing" properties which imply ergodicity; see [12] or [21]. The fact that the ZERO-ONE LAW implies the LAW OF AVERAGES is a special case of the Individual Ergodic Theorem, a classical "hard" theorem due to G. D. Birkhoff. For a proof of this theorem and of Theorem 7.4, see [12].

We will need one more "hard" theorem concerning ergodic operators.

Theorem 7.5 (Shannon-McMillan-Breiman Theorem). *Let $(X, \mathcal{M}, p, S)$ be an ergodic operator and let $\mathcal{A} \in \Omega_X$. Let*

$$\alpha_n(x) = \frac{-1}{n} \sum_{E \in \mathcal{A}_0^{n-1}} \log p(E)\, \chi_E(x)$$

Then the sequence of functions α_n converges to the constant function $h_S(\mathcal{A})$ almost everywhere and in mean.

Note that $\int_X \alpha_n dp = (1/n)H(\mathcal{A}_0^{n-1})$. Note also that because Theorem 7.5 says that a certain sequence of functions converges both almost everywhere and in mean to a certain constant function, it has the same form as the Individual Ergodic Theorem. It is *not* a special case of that theorem because the α_n are not of the form $f \circ S^n$. Two different proofs of Theorem 7.5 are given in [12].

8. The Entropy of a Message Source

Definition 8.1. *Let $(X, \mathcal{M}, p)$ and $(Y, \mathcal{N}, q)$ be probability measure spaces and suppose $\varphi : X \to Y$ is a measure-preserving transformation. Given $\mathcal{B} \in \Omega_Y$, with atoms B_k, the* **pullback** *of $\mathcal{B}$ is the element $\varphi^\# \mathcal{B} \in \Omega_X$ whose atoms are the sets $\varphi^{-1}(B_k)$.*

Notice that if φ is many-to-one, several atoms of $\mathcal{B}$ might pull back to the same atom of $\varphi^\# \mathcal{B}$. That is, we might have (for example) $\varphi^{-1}(B_2) = \varphi^{-1}(B_5) = A_3$.

Think again of X as the set of microstates for the atmosphere over Omaha. We can imagine that some probability operator $(X, \mathcal{M}, p, S)$ can be interpreted as the "evolution operator", such that if the weather has microstate x at time $t = n$, it has microstate $S(x)$ at time $t = n + 1$. As we have already remarked, the trajectory $x, S(x), S^2(x) \dots$ is not observable. However, we *can* hope to determine the trajectory *relative to* $\mathcal{A}$; that is, the sequence of atoms in which the trajectory lies. We will call this sequence the **macrohistory** of x relative to $\mathcal{A}$.

Observe that $x \in A_2$, $S(x) \in A_1$, and $S^2(x) \in A_4$ exactly in case $x \in A_2 \cap S^{-1}(A_1) \cap S^{-2}(A_4)$. In short, the atoms of $\mathcal{A} \vee S^\# \mathcal{A} \vee S^{\#2} \mathcal{A}$ correspond to alternative sequences of three successive macrostates. It is convenient to abbreviate the join

$$\mathcal{A} \vee S^\# \mathcal{A} \vee S^{2\#} \mathcal{A} \cdots \vee S^{(n-1)\#} \mathcal{A}$$

as $\mathcal{A}_0^{n-1}$. Atoms of $\mathcal{A}_0^{n-1}$ represent **partial macrohistories**.

This idea is adaptable to any complex process, for instance the process by which human beings produce "messages" in natural language, which suggests the following definition.

Definition 8.2. *Let $(X, \mathcal{M}, p, S)$ be a probability operator and let $\mathcal{A} \in \Omega_X$. Then $(X, \mathcal{M}, p, S, \mathcal{A})$ is a (stationary)* **message source***.*

Note that every $x \in X$ defines a sequence in the finite set $\mathcal{A}$ by associating the sequence $A_3, A_2, A_4, A_1, A_2, A_2 \dots$ to every x such that $x \in A_3$, $S(x) \in A_2$, $S^2(x) \in A_4$, $S^3(x) \in A_1$, $S^4(x) \in A_2$, $S^5(x) \in A_2$, and so forth. The set of such sequences is the set of messaages which can be produced by the source $(X, \mathcal{M}, p, S, \mathcal{A})$. Note that the probability that the sequence defined by an arbitrary $x \in X$ will begin with a given finite sequence of length n is precisely $p(E)$, where $E \in \mathcal{A}_0^{n-1}$ is the atom corresponding to the finite sequence in question.

In the case of the weather in Omaha, if $\mathcal{A}$ is the kernel of the precipitation function, then $(X, \mathcal{M}, p, S, \mathcal{A})$ is a source which produces "messages" according to whether or not it is raining at a given hour in Omaha.

Definition 8.3. *The* **source entropy** *of $(X, \mathcal{M}, p, S, \mathcal{A})$ is*

$$h_S(\mathcal{A}) = \lim_{n \to \infty} (1/n) \cdot H(\mathcal{A}_0^{n-1})$$

It is not very difficult to show [21][1][8] that this limit always exists and that we always have $h_S(\mathcal{A}) \leq H(\mathcal{A})$. The source entropy $h_S(\mathcal{A})$ can be interpreted as a measure of the information about the microhistory of a typical microhistory which is gained from knowledge of the corresponding macrohistory. Alternatively, $h_S(\mathcal{A})$ *measures the average variety of alternatives at each place in the sequences produced by the source.*

We can obtain an operational interpretation of $h_S(\mathcal{A})$ by imagining a third version of our guessing game, in which every hour on the hour we must guess whether or not it is currently raining in Omaha. Clearly $(1/n) \cdot H(\mathcal{A}_0^{n-1})$ measures the variety of alternative outcomes, and thus the difficulty per round of guessing, averaged over the first n rounds. As n grows, we are taking account of longer and longer range statistical correlations between macrostates of precipitation in Omaha. In the limit, we are taking account of even the most subtle and longest range correlations between the weather at present and in the distant past (and future).

Definition 8.4. *An information source $(X, \mathcal{M}, p, \mathcal{A})$ such that $S^{\#}\mathcal{A}$ is statistically independent of $\mathcal{A}$ is called a* **Bernoulli source***. If $S^{\#2}\mathcal{A}$ is statistically independent of $\mathcal{A}$, $(X, \mathcal{M}, p, \mathcal{A})$ is called a* **Markov source***. More generally, if $S^{\#n}\mathcal{A}$ is statistically independent of $\mathcal{A}$, then $(X, \mathcal{M}, p, \mathcal{A})$ is called an n*-**step Markov source***.*

It is easy to obtain simple formulas for the entropies of Bernoulli and Markov sources. For a Bernoulli source we have

$$\begin{aligned} H(\mathcal{A}_0^{n-1}) &= H(\mathcal{A}) + H(S^{\#}\mathcal{A}/\mathcal{A}) + H(S^{\#2}\mathcal{A}/\mathcal{A}_0^1) + \ldots H(S^{\#n-1}\mathcal{A}/\mathcal{A}_0^{n-2}) \\ &= n \cdot H(\mathcal{A}) \end{aligned}$$

whence

$$h_S(\mathcal{A}) = H(\mathcal{A}) = -\sum_{j=1}^{r} p(A_j) \log p(A_j)$$

For the (one-step) Markov source defined by (P, π) in the obvious way, we have

$$\begin{aligned} H(\mathcal{A}_0^{n-1}) &= H(\mathcal{A}) + H(S^{\#}\mathcal{A}/\mathcal{A}) + H(S^{\#2}\mathcal{A}/\mathcal{A}_0^1) + \ldots H(S^{\#n-1}\mathcal{A}/\mathcal{A}_0^{n-2}) \\ &= H(\mathcal{A}) + (n-1) \cdot H(S^{\#}\mathcal{A}/\mathcal{A}) \end{aligned}$$

whence

$$\begin{aligned} h_S(\mathcal{A}) &= H(S^{\#}\mathcal{A}/\mathcal{A}) \\ &= -\sum_{j=1}^{r}\sum_{k=1}^{r} p(A_j) \frac{p(A_j \cap S^{-1}A_k)}{p(A_j)} \log \frac{p(A_j \cap S^{-1}A_k)}{p(A_j)} \\ &= -\sum_{j=1}^{r}\sum_{k=1}^{r} \pi(j) P(j,k) \log P(j,k) \end{aligned}$$

Similarly, for a two-step Markov source we have $h_S(\mathcal{A}) = H(S^{\#2}\mathcal{A}/\mathcal{A}_0^1)$, and so forth.

9. The Asymptotic Equipartition Theorem

A sequence of functions f_n is said to **converge in measure** to f if for every $\varepsilon > 0$, there is some $N > 0$ such that for all $n \geq N$, the measure of the set $\{x \in X : |f_n(x) - f(x)| \geq \varepsilon\}$ is at most ε.

Lemma 9.1. *If f_n converges in mean to f, then f_n also converges in measure to f.*

Proof: Fix $\varepsilon > 0$. Put $E_n = \{x \in X : |f_n(x) - f(x)| \geq \varepsilon\}$. Then $\int_X |f_n - f| dp \geq \int_{E_n} |f_n - f| dp \geq \varepsilon \, p(E_n)$. Now we can see that if f_n converges in mean to f, so that the left hand side approaches zero as n grows without bound, then the right hand side also must approach zero; that is, f_n converges in measure to f. ∎

(See [4] for an excellent discussion of the meaning and inter-relationships of various notions of convergence useful in analysis.)

Theorem 9.2 (Asymptotic Equipartition Theorem)**.** *Suppose $(X, \mathcal{M}, p, S)$ is an ergodic probality operator. Given $\mathcal{A} \in \Omega_X$ and $\varepsilon > 0$, define for all $n > 0$ the collections*

$$T(n) = \{E \in \mathcal{A}_0^{n-1} : |(-1/n) \log p(E) - h_S(\mathcal{A})| \leq \varepsilon\}$$

Then for all $\epsilon > 0$, there is some N such that for all $n \geq N$,

1 *we have, for all $E \in T(n)$,*

$$e^{-n\,h_S(\mathcal{A}) - n\varepsilon} \leq p(E) \leq e^{-n\,h_S(\mathcal{A}) + n\varepsilon}$$

2 *we have*

$$p\{x \in X : x \textit{ is in some } E \in T(n)\} > 1 - \varepsilon$$

3 *we have*

$$(1 - \varepsilon)\, e^{n\,h_S(\mathcal{A}) - n\varepsilon} \leq |T(n)| \leq e^{n\,h_S(\mathcal{A}) + n\varepsilon}$$

The $E \in T(n)$ are called the **typical atoms** of $\mathcal{A}_0^{n-1}$.

The AEP seems counterintuitive at first sight, but it actually has a very graphic interpretation. For large n, we can visualize $\mathcal{A}_0^{n-1}$ as a "beach" of total volume one cubic mile (say). (See Figure 11.) Most of the volume is made up of about $e^{n\,h_S(\mathcal{A})}$ "pebbles" each having a volume of about $e^{-n\,h_S(\mathcal{A})}$ cubic miles. (Since n is very large, this will indeed be a pebble-sized volume, despite the unusual units.) The beach also contains on the order of $e^{n\,\log|\mathcal{A}|}$ of tiny "sand grains", which are so small that collectively they account for at most ϵ cubic miles of the total volume of the beach; nevertheless, their teeming numbers completely overwhelm the number of pebbles (provided that $\log|\mathcal{A}| > h_S(\mathcal{A})$). The beach might contain the occasional "rock" having a volume considerably larger than $e^{nh_S(\mathcal{A})}$ cubic miles, but there are certainly no "boulders" with volume on the order of ϵ cubic miles.

The AEP is a suprisingly easy consequence of the Shannon-McMillan-Breiman Theorem.

Proof: (1) holds for all $n > 0$ by definition of $T(n)$. Because the α_n converge in mean to $h_S(\mathcal{A})$, they also converge in measure to $h_S(\mathcal{A})$, which means that for all $\varepsilon > 0$, there is some $N > 0$ such that for all $n \geq N$, the measure of the set

$$\{x \in X : x \text{ is in some } E \in \mathcal{A}_0^{n-1} \text{ such that } |(-1/n) \log p(E) - h_S(\mathcal{A})| > \varepsilon\}$$

is at most ε, which gives (2).

Now (3) follows by combining (1) and (2). Fix $n \geq N$ and let E_- minimize $p(E)$ over $T(n)$. Then by (1),

$$e^{-n\,h_S(\mathcal{A}) - n\varepsilon} \leq p(E_-)$$

Let E_+ maximize $p(E)$ over $T(n)$. Then by (1),

$$p(E_+) \leq e^{-n\,h_S(\mathcal{A}) + n\epsilon}$$

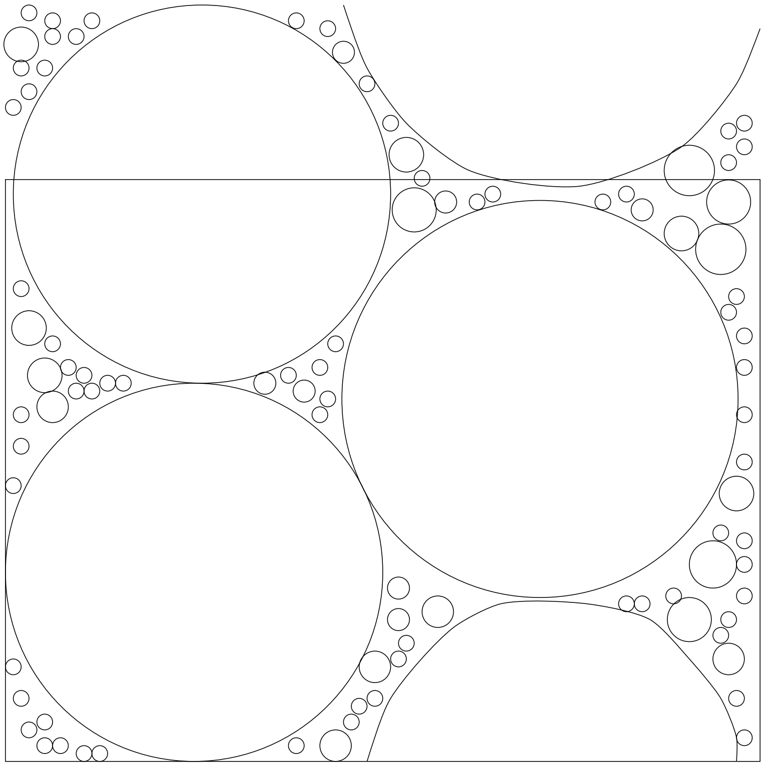

FIGURE 11. The pebbly beach.

Let $M_n = |T(n)|$ be the number of typical atoms, and let

$$X_n = \{x \in X : x \text{ is in some } E \in T(n)\}$$

be the union over those atoms. Clearly

$$M_n\, p(E_-) \leq p(X_n) \leq M_n\, p(E_+)$$

But by (2), $1 - \epsilon \leq p(X_n)$, so on the one hand we have

$$1 - \epsilon \leq p(X_n) \leq M_n\, p(E_+) \leq M_n \cdot e^{-nh_S(\mathcal{A}) + n\epsilon}$$

whence

$$(1 - \epsilon)\, e^{nh_S(\mathcal{A}) - n\epsilon} \leq M_n$$

and on the other hand, we have

$$M_n\, e^{-nh_S(\mathcal{A}) - n\epsilon} \leq M_n\, p(E_-) \leq p(X_n) \leq 1$$

whence

$$M_n \leq e^{nh_S(\mathcal{A}) + n\epsilon}$$

∎

10. Data Compression

The AEP gives one interpretation of the "meaning" of the source entropy. To explain an alternative interpretation, we require the concept of *block coding.*

Definition 10.1. *Let $(X, \mathcal{M}, p, S, \mathcal{A})$ be a message source. Recall that $\mathcal{A}_0^{n-1}$ may be identified with the set of alternative sequences of length n which may be produced by the source. Suppose $\phi : \mathcal{A}_0^{n-1} \to W$, where W is some finite set of codewords, perhaps $W = \{01302, 2201, \ldots\}$. The corresponding* **block code** *assigns a sequence in W to every $x \in X$, as follows. Suppose $x \in E_1 \in \mathcal{A}_0^{n-1}, S^{n+1}(x) \in E_2 \in \mathcal{A}_0^{n-1}$, and so forth. Then assign to x the sequence $\phi(E_1), \phi(E_2), \ldots$. This code is said to have* **block length** *n.*

Generally speaking, larger n allow a block code to take advantage of "longer range" statistical correlations between earlier and later letters in sequences produced by the source, and thus the efficiency of optimal codes of length n generally increases (perhaps very slowly) as n increases. The key idea, due to Shannon [20], is that as long as we have enough codewords to code up the *typical* atoms in $\mathcal{A}_0^{n-1}$, with at least one left over to serve as a "flag", we can obtain a workable n-block code.

Theorem 10.2 (Data Compression Theorem). *Suppose $(X, \mathcal{M}, p, S)$ is an ergodic operator and $\mathcal{A} \in \Omega_X$. Let $d = h_S(\mathcal{A})/\log r$ where $r = |\mathcal{A}|$ is the number of atoms in the partition $\mathcal{A}$. Then for all $\epsilon > 0$, there exists N such that for all $n \geq N$, we can find a block coding with block length n of sequences produced by $(X, \mathcal{M}, p, S, \mathcal{A})$ such that the mean codeword length,*

$$\overline{\lambda} = \sum_{E \in \mathcal{A}_0^{n-1}} p(E) \cdot \lambda(E)$$

(where $\lambda(E)$ is the length of the code word assigned to $E \in \mathcal{A}_0^{n-1}$) satisfies

$$\overline{\lambda} \leq n\,(d + 2\epsilon)$$

In short, the fraction by which typical sequences produced by $(X, \mathcal{M}, p, S, \mathcal{A})$ can be "compressed" by using block coding is $d = h_S(\mathcal{A})/\log r$. The following ingenious proof, which involves the idea of a "randomly chosen coding", is due to Shannon [20].

Proof: By the AEP we can find N such that for all $n \geq N$, $p(X_n) > 1 - \epsilon$ and

$$M_n \leq e^{n\,h_S(\mathcal{A}) + n\epsilon} = r^{n(d+\epsilon)/\log r}$$

where X_n is the union of the typical atoms in $\mathcal{A}_0^{n-1}$ and M_n is the number of typical atoms in $\mathcal{A}_0^{n-1}$. Now we can encode the typical atoms in $\mathcal{A}_0^{n-1}$ by assigning each an arbitrary word of length $n(d+\epsilon)$; if n is sufficiently large, this will leave one word of length $n(d+\epsilon)$ left over to serve as a "flag". We can encode the atypical atoms by prefixing the flag followed by a literal quote of the base r sequence corresponding to the atom. This means that each typical atom will be encoded by a codeword of length $nd + n\epsilon/\log r \leq n(d+\epsilon)$, whereas atypical atoms will be encoded by a codeword of length at most $n(1 + d + (\epsilon/\log r)) \leq n(1 + d + \epsilon)$. Therefore,

$$\begin{aligned} \overline{\lambda} &\leq (1-\epsilon) \cdot n(d+\epsilon) + \epsilon \cdot n(1 + d + \epsilon) \\ &= n(d + 2\epsilon) \end{aligned}$$

∎

11. Conditional Entropy and Information Rates

Just as we took limits of entropies in order to define the source entropy, we can take limits of conditional entropies and mutual informations.

Definition 11.1. *Let $(X, \mathcal{M}, p, S)$ be a probability operator and let $\mathcal{A}, \mathcal{B} \in \Omega_X$. Then the* **conditional entropy rate** *of $\mathcal{A}$ given $\mathcal{B}$ is*

$$h_S(\mathcal{A}/\mathcal{B}) = \lim_{n\to\infty} (1/n) \cdot H(\mathcal{A}_0^{n-1}/\mathcal{B}_0^{n-1})$$

Similarly, the **information rate** *between $\mathcal{A}, \mathcal{B}$ is*

$$i_S(\mathcal{A}, \mathcal{B}) = \lim_{n\to\infty} (1/n) \cdot I(\mathcal{A}_0^{n-1}, \mathcal{B}_0^{n-1})$$

It is not hard to see that $h_S(\mathcal{A}/\mathcal{B}) = h_S(\mathcal{A} \vee \mathcal{B}) - h_S(\mathcal{B})$ and that

$$\begin{aligned} i_S(\mathcal{A}, \mathcal{B}) &= h_S(\mathcal{A}) + h_S(\mathcal{B}) - h_S(\mathcal{A} \vee \mathcal{B}) \\ &= h_S(\mathcal{A}) - h_S(\mathcal{A}/\mathcal{B}) \\ &= h_S(\mathcal{B}) - h_S(\mathcal{B}/\mathcal{A}) \end{aligned}$$

Both conditional entropy and information rates can be understood by considering a fourth version of our guessing game. In this version, every hour on the hour, the Devil informs us (truthfully) about the current temperature range in Omaha, and we must guess whether or not it is presently raining there. Now $h_S(\mathcal{A}/\mathcal{B})$ measures the variety of outcomes, or the difficulty of guessing correctly, averaged over many rounds, and $i_S(\mathcal{A}, \mathcal{B})$ measures the average information gained per hour from what the Devil tells us concerning the current temperature.

The quantities $h_S(\mathcal{A})$, $h_S(\mathcal{A}/\mathcal{B})$, $i_S(\mathcal{A}, \mathcal{B})$, and so forth, satisfy numerous formal properties analogous to those satisfied by $H(\mathcal{A})$, $H(\mathcal{A}/\mathcal{B})$, $I(\mathcal{A}, \mathcal{B})$ and so forth. For more information about these formal properties see [8].

The following result is a fundamental extension of the Equipartition Theorem. It was first found by Shannon [20] but the first correct proof did not appear for some years after that.

Theorem 11.2 (Joint Asymptotic Equipartition Theorem). *Suppose $(X, \mathcal{M}, p, S)$ is an ergodic probability operator. Given $\mathcal{A}, \mathcal{B} \in \Omega_X$ and $\varepsilon > 0$, define for $n > 0$ the following collections of atoms:*

1 *$J(n)$ is the set of all $E \cap F \in (\mathcal{A} \vee \mathcal{B})_0^{n-1}$ such that*

$$\begin{aligned} |(-1/n)\log p(E \cap F) - h_S(\mathcal{A} \vee \mathcal{B})| &\leq \epsilon, \\ |(-1/n)\log p(E) - h_S(\mathcal{A})| &\leq \epsilon, \\ |(-1/n)\log p(F) - h_S(\mathcal{B})| &\leq \epsilon, \end{aligned}$$

2 *$K(n) = \{E \in \mathcal{A}_0^{n-1} : E$ contains some $E \cap F \in J(n)\}$,*

3 *$L(n) = \{F \in \mathcal{B}_0^{n-1} : F$ contains some $E \cap F \in J(n)\}$,*

Then there exists $N > 0$ such that for all $n \geq N$,

1 *we have that for all $E \cap F \in J(n)$,*

(a) $e^{-n\, h_S(\mathcal{A}\vee\mathcal{B}) - n\varepsilon} \leq p(E \cap F) \leq e^{-n\, h_S(\mathcal{A}\vee\mathcal{B}) + n\varepsilon}$,

(b) $e^{-n\, h_S(\mathcal{A}) - n\varepsilon} \leq p(E) \leq e^{-n\, h_S(\mathcal{A}) + n\varepsilon}$,

(c) $e^{-n\, h_S(\mathcal{B}) - n\varepsilon} \leq p(F) \leq e^{-n\, h_S(\mathcal{B}) + n\varepsilon}$,

(d) $e^{-n\, h_S(\mathcal{A}/\mathcal{B}) - 2n\varepsilon} \leq \frac{p(E\cap F)}{p(F)} \leq e^{-n\, h_S(\mathcal{A}/\mathcal{B}) + 2n\varepsilon}$,

(e) $e^{-n\, h_S(\mathcal{B}/\mathcal{A}) - 2n\varepsilon} \leq \frac{p(E\cap F)}{p(E)} \leq e^{-n\, h_S(\mathcal{B}/\mathcal{A}) + 2n\varepsilon}$,

(f) $e^{-n\, i_S(\mathcal{A},\mathcal{B})-3n\varepsilon} \le \frac{p(E)\,p(F)}{p(E\cap F)} \le e^{-n\, i_S(\mathcal{A},\mathcal{B})+3n\varepsilon}$,
where $E \in \mathcal{A}_0^{n-1}$ *and* $F \in \mathcal{B}_0^{n-1}$ *are the unique atoms in* $\mathcal{A}_0^{n-1}$ *and* $\mathcal{B}_0^{n-1}$, *respectively, which contain the given atom* $E \cap F \in J(n)$,

2 *we have*
(a) $p\{x \in X : x \in E \cap F \in J(n)\} > 1-\varepsilon$,
(b) $p\{x \in X : x \in E \in K(n)\} > 1-\varepsilon$,
(c) $p\{x \in X : x \in F \in L(n)\} > 1-\varepsilon$,

3 *we have*
(a) $(1-\varepsilon)\, e^{n\, h_S(\mathcal{A}\vee\mathcal{B})-n\varepsilon} \le |J(n)| \le e^{n\, h_S(\mathcal{A}\vee\mathcal{B})+n\varepsilon}$,
(b) $(1-\varepsilon)\, e^{n\, h_S(\mathcal{A})-n\varepsilon} \le |K(n)| \le e^{n\, h_S(\mathcal{A})+n\varepsilon}$,
(c) $(1-\varepsilon)\, e^{n\, h_S(\mathcal{B})-n\varepsilon} \le |L(n)| \le e^{n\, h_S(\mathcal{B})+n\varepsilon}$.

The $E \cap F \in J(n)$ are called **jointly typical atoms**. Pairs (E,F), where $E \in \mathcal{A}_0^{n-1}$ and $F \in \mathcal{B}_0^{n-1}$ and $E \cap F \in J(n)$, are called **jointly typical pairs**.
Proof: Note first that (1abc) follow directly from the definition and (1def) follow easily from (1abc); for instance

$$\begin{aligned} \frac{P(E\cap F)}{p(F)} &\le e^{-nh_S(\mathcal{A}\vee\mathcal{B})+n\varepsilon}\, e^{nh_S(\mathcal{A})+n\varepsilon} \\ &= e^{-nh_S(\mathcal{A}/\mathcal{B})+2n\varepsilon} \end{aligned}$$

By the Shannon-McMillan-Breiman Theorem, the functions

$$\alpha_n = \frac{-1}{n} \sum_{E \in (\mathcal{A}\vee\mathcal{B})_0^{n-1}} \log p(E) \chi_E$$

converge to $h_S(\mathcal{A}\vee\mathcal{B})$ in mean, and therefore also in measure. Thus there is some N_1 such that for all $n \ge N_1$, the measure of the set

$$\{x \in X : x \in E \cap F \in (\mathcal{A}\vee\mathcal{B})_0^{n-1} \text{such that} |(-1/n)\log p(E\cap F) - h_S(\mathcal{A}\vee\mathcal{B})| \ge \varepsilon\}$$

is at most $\varepsilon/3$. Similarly, there is some N_2 such that for all $n \ge N_2$, the measure of the set

$$\{x \in X : x \in E \in \mathcal{A}_0^{n-1} \text{such that} |(-1/n)\log p(E) - h_S(\mathcal{A})| \ge \varepsilon\}$$

is less than $\varepsilon/3$, and there is some N_3 such that for all $n \ge N_3$, the measure of the set

$$\{x \in X : x \in F \in \mathcal{B}_0^{n-1} \text{such that} |(-1/n)\log p(F) - h_S(\mathcal{B})| \ge \varepsilon\}$$

is less than $\varepsilon/3$. Now take N to be the maximum of N_1, N_2, and N_3. Now the set

$$\{x \in X : \text{ for all } E\cap F \in J(n), x \notin E \cap F\}$$

evidently has measure at most ε, for all $n \ge N$, so the set

$$\{x \in X : \text{ for some } E\cap F \in J(n), x \in E \cap F\}$$

has measure at least $1-\varepsilon$ for all $n \ge N$, as claimed in (2a). This set is included in the sets mentioned in (2bc), so we have now proven (2abc). Finally, (3abc) follow by the same argument used in the Equipartition Theorem. ∎

This theorem shows that for jointly typical pairs (E, F), we have

$$\begin{aligned} p(E) &\approx e^{-n\, h_S(\mathcal{A})} \\ p(F) &\approx e^{-n\, h_S(\mathcal{A})} \\ p(E \cap F) &\approx e^{-n\, h_S(\mathcal{A} \vee \mathcal{B})} \\ \frac{p(E \cap F)}{p(F)} &\approx e^{-n\, h_S(\mathcal{A}/\mathcal{B})} \\ \frac{p(E \cap F)}{p(E)} &\approx e^{-n\, h_S(\mathcal{B}/\mathcal{A})} \\ \frac{p(E)p(F)}{p(E \cap F)} &\approx e^{-n\, i_S(\mathcal{A},\mathcal{B})} \end{aligned}$$

This gives a very intuitive significance to these dynamical entropies. For instance, the information source $(X, \mathcal{M}, S, \mathcal{A})$ might be used as the input into a noisy communication channel, whereas $(X, \mathcal{M}, S, \mathcal{B})$ could denote the output from the channel (assuming that the noise sometimes alters zeros to ones and vice versa, but does not introduce new symbols). Then $i_S(\mathcal{A}, \mathcal{B})$ denotes the asymptotic rate at which we can extract information about the source $(X, \mathcal{M}, S, \mathcal{A})$ using the source $(X, \mathcal{M}, S, \mathcal{B})$.

References

1. Thomas M. Cover and Joy A. Thomas, *Elements of information theory.* Wiley, New York, 1991.
2. Robert L. Devaney, *An introduction to chaotic dynamical systems.* Second edition. New York: Addison-Wesley, 1989.
3. Manfred Denker, Christian Grillenberger, and Karl Sigmund, *Ergodic theory on compact spaces.* Springer, New York, 1976.
4. Gerald B. Folland, *Real analysis: modern techniques and their applications.* New York: Wiley, 1984.
5. Silviu Guiasu and Abe Shenitzer, 'The principle of maximal entropy'. *Math. Intelligencer* 7:1 (1985), pp. 42–48.
6. P. R. Halmos, 'The foundations of probability', in The Chauvenet papers. Volume I. J. C. Abott, Ed. Mathematical Association of American, 1978.
7. Chris Hillman, 'Symmetry and information', forthcoming.
8. —, 'A formal theory of information'. Preprint, 1995.
9. —, 'All entropies agree for an SFT', Preprint, 1995.
10. —, 'What is Hausdorff dimension?' Preprint, 1995.
11. —, 'What is information?' Preprint, 1995.
12. —, 'Measure-theoretic entropy'. Preprint, 1993. All of my preprints are available at the URL `<http://www.math.washington.edu/~hillman/personal.html>`
13. Nathan Jacobson, *Basic algebra.* Two volumes. Second Edition. W.H. Freeman, N.Y. 1989.
14. Anatole Katok and Boris Hasselblatt. *Introduction to the modern theory of dynamical systems.* Cambridge: Cambridge University Press, 1995.
15. A.I. Khinchin, *The mathematical foundations of information theory.* Dover, N.Y., 1957.
16. Douglas Lind and Brian Marcus, *Introduction to Symbolic Dynamics and Coding*, to appear.
17. Andrzej Lasota and Michael C. Mackey, *Chaos, fractals, and noise: stochastic aspects of dynamics.* New York: Springer-Verlag, 1994.
18. Saunders Mac Lane, *Categories for the working mathematician.* Second Edition. Springer-Verlag, N.Y., 1971.
19. Howard L. Resnikoff, *The illusion of reality.* New York: Springer, 1989.
20. 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.

21. Peter Walters, *Introduction to ergodic theory.* Springer, New York, 1981.

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

E-mail address: hillman@math.washington.edu