The Boltzmann Entropy and Randomness Tests (Extended Abstract)

Péter Gács*
Computer Science Department
Boston University
Boston, MA 02192
email: gacs@cs.bu.edu

Abstract

In the context of the dynamical systems of classical mechanics, we introduce two new notions called "algorithmic fine-grain and coarse-grain entropy". The fine-grain algorithmic entropy is, on the one hand, a simple variant of the randomness tests of Martin-Löf (and others) and is, on the other hand, a connecting link between description (Kolmogorov) complexity, Gibbs entropy and Boltzmann entropy.

The coarse-grain entropy is a slight correction to Boltzmann's coarse-grain entropy. Its main advantage is its less partition-dependence, due to the fact that algorithmic entropies for different coarse-grainings are approximations of one and the same fine-grain entropy. It has the desirable properties of Boltzmann entropy in a somewhat wider range of systems, including those of interest in the "thermodynamics of computation".

1 Introduction

This paper assumes some familiarity with the notions of computabilitly and description (Kolmogorov) complexity. A good survey of the latter is [2].

Coarse-graining The physical model considered in the present paper is that of classical mechanics: it is defined by a phase space Ω and a dynamics U^t giving the point $U^t\omega$ on an orbit at time t, where the transformation U^t preserves the volume measure L (like in Liouville's Theorem). In case of a container of "ideal" gas consisting of n simple molecules, the state space is the $6n$-dimensional Euclidean space given by the positions and moments of each molecule.

A thermodynamical system is characterized by a relatively small number of parameters (functions of the state) called **macroscopic** parameters $u_1, \ldots, u_m$. The canonical example is a certain quantity of gas in a container, with the macroscopic parameters of volume, temperature, energy and pressure. A microscopic state is the completely specified state. A macroscopic state is determined by the macroscopic parameters and it determines only the (by far) most probable behavior and properties of the system and only when the system is in equilibrium.

We will assume that each macroscopical parameter u_i takes only a finite number of values: a macroscopical parameter that is originally a real number will only be taken to a certain precision agreed in advance. This gives a partition of the state space into cells

$$\Omega = \bigcup_u \Gamma_u$$

corresponding to **macrostates** where $u = (u_1, \ldots, u_m)$. Finer and finer partitions $\mathcal{P}_i$ of the phase space can be introduced by adding more macroscopic parameters and more precision. The partition interpretation of a macrostate is called **coarse-graining**.

Coarse-graining solves the **paradox of irreversibility**: in a mechanical system, any evolution seems just as possible as the corresponding reverse evolution, and at the same time, the world seems to be full of irreversible phenomena (examples omitted). To reconcile the two pictures when we say that a certain transformation from state a to state b is reversible this statement refers to *macrostates*; what is meant is that the reverse transformation exists for *most* microstates within the macrostate Γ_b, as measured by volume. Now asymmetries are quite possible.

Another possible interpretation of a macrostate is as a certain distribution ν over microscopic states. It is possible (but not always done) to require ν to be a probability distribution, given by a density function $p(\omega)$ with $\int p(\omega)L(d\omega) = 1$. Gibbs called such a distribution an **ensemble**. The ensemble $p_\Gamma(\omega)$ corresponding to a macrostate Γ is defined as $1/L(\Gamma)$ for $\omega \in \Gamma$ and 0 elsewhere.

Coarse-grained entropy The Boltzmann entropy of a cell Γ is $\log L(\Gamma)$. This quantity depends on the choice of the partition: indeed, another digit of precision will decrease it by about $\log 10$. In the typical classical examples, this difference is negligible compared to the volumes in question. In nontypical systems, partition dependence can lead to paradoxes: we

*Part of this research was done during the author's stay at IBM Watson Research Center. Partially supported by NSF grant CCR-9002614, and by NWO through NFI Project ALADDIN under Contract number NF 62-376 and Scientific Visitor Award B 62-394.

0-8186-6715-X/94 $4.00 © 1994 IEEE

will show that the Maxwell demon paradox is one of these.

We propose a new quantity

$$H(\Gamma) = K(\Gamma) + \log L(\Gamma) \qquad (1.1)$$

called *coarse-grained algorithmic entropy* to replace Boltzmann entropy. Here, $K(\Gamma)$ is the description (Kolmogorov) complexity of the cell Γ (to be defined later). This quantity is closely related to the one introduced by Zurek, and we will return to their relation. The term $K(\Gamma)$ is typically negligible compared to the Boltzmann entropy $\log L(\Gamma)$; however, the new entropy is less partition-dependent since it is an approximation to a certain partition-independent quantity $H(\omega) = H_L(\omega)$ which we call **(fine-grained) algorithmic entropy**.

The paper explores the basic properties of fine-grained and coarse-grained algorithmic entropy. Fine-grain algorithmic entropy is a simple variant of the randomness tests of Martin-Löf (and others). Its integral over a Gibbs ensemble is close to the so-called Gibbs entropy. It possesses a simple conservation property that, together with coarse-grained algorithmic entropy, is helpful in explaining the Maxwell demon paradox as well as some other physical situations and models not handled well by Boltzmann entropy. Explanation of the entropy increase property for the new coarse-grained entropy does not rely on the huge volume differences in cells the way it does in the case of Boltzmann's entropy.

We hope that the new quantity extends the possibilities of the kind of reasoning involving entropy to a wider range of systems (in particular, large computers in which it is not clear whether the whole memory content should be considered macroscopic or microscopic information).

Before introducing the new notions we recall the basic facts of the theory of complexity and randomness.

2 Complexity

Given some computer F, let $K_F(y)$ be the length of the shortest program (measured in bits) that causes F to output string y. We will require the program to be self-delimiting: no endmarker is allowed. The machine-dependence of this concept is limited since there is a machine G on which the function $K_G(y)$ is optimal within an additive constant: i.e. for every other machine F there is a constant c_F such that for all x we have $K_G(x) \leq K_F(x) + c_F$. The notation

$$f(x) \stackrel{+}{<} g(x).$$

will mean that for some constant c and for all x we have $f(x) \leq g(x) + c$. The notation $<$ will mean the same with a multiplicative constant. The notation $\stackrel{\pm}{=}$ means that both $\stackrel{+}{<}$ and $\stackrel{+}{>}$ hold. With this notation, the invariance theorem's formula can be written as $K_G(x) \stackrel{+}{<} K_F(x)$. The function $K(x) = K_G(x)$ is called the **complexity** of the natural number x conditional on the information y. (This is the modified version of the Kolmogorov-Solomonoff complexity invented by Levin and Chaitin.)

The conditional complexity $K(x \mid y)$ is defined by leaving y everywhere as a parameter. The elementary properties of the complexity function are discussed in several expositions, and we will not dwell on them. Let us just mention that $K(n) \stackrel{+}{<} \log n + 2\log\log n$, and $K(f(x) \mid y) \stackrel{+}{<} K(x \mid g(y)) \stackrel{+}{<} K(x)$ for any computable functions f, g.

The function $K(x)$ is not computable, but it has a certain weaker property. Let $\mathbf{Q}$ be the set of rational numbers. We call a function $f(x)$ from natural numbers to real numbers **upper semicomputable** if there is a computable sequence $f_n(x)$ with rational values such that $f_n(x) \searrow f(x)$. It is easy to see that $K(x \mid y)$ is upper semicomputable. We will need the following important theorem of Levin.

(2.1) **Coding Theorem** *Let us consider the class of lower semicomputable functions $f(x, y)$ with the property that $\sum_x f(x, y) \leq 1$. The function $2^{-K(x|y)}$ is an element of this class and is maximal in it, to within a multiplicative constant. In other words, for each element f of this class, we have $2^{-K(x|y)} \stackrel{+}{>} f(x, y)$.*

3 Randomness

Notions of computability can be introduced over the space Ω if some sequence of neighborhoods, called **open cells** (e.g. Cartesian products of rational intervals) is fixed. We will assume that the transformation group U^t as well as the invariant volume measure L are computable.

For a probability space Ω and a fixed increasing sequence Ω_n of subsets of Ω whose union is the whole space, define the set $\mathcal{M}(\Omega)$ as the set of measures μ over Ω such that $\mu(\Omega_n) < \infty$. A nonnegative lower semicomputable function $f_\mu(\omega, y)$ over the space $\mathcal{M}(\Omega) \times \Omega \times Y$ is a (parametrized) **test of randomness** or, shortly, test with respect to a parameter y, if for all μ, y we have

$$\int f_\mu(\omega, y)\mu(d\omega) \leq 1.$$

Here is some motivation for the case of probability measures. For a moment, forget the parameter y. If a certain casino claims that it draws elements from Ω according to the distribution μ then it must accept the following deal:

1. I prove that $f_\mu(\omega)$ is a test of randomness;
2. I offer two dollars for a game, and ask for ω;
3. my payoff is $f_\mu(\omega)$.

If the casino owner indeed draws according to μ then the test property implies that my expected payoff is at most a dollar, so she even makes more than a dollar of profit on average. My strategy is to try to find some nonrandomness in ω, (without seeing ω first) by

making an acceptable test function $f_\mu(\omega)$ as large as possible.

It can be shown that among all randomness tests, there is a certain one, denoted by $t_\mu(\omega \mid y)$ and called a **universal test**, that takes only values of the form 2^n for (possibly negative) integers n and is maximal to within a multiplicative constant, i.e. that has the property that for all other tests $f_\mu(\omega, y)$, we have

$$f_\mu(\omega, y) < t_\mu(\omega \mid y).$$

This test is a close relative of the universal tests of Martin-Löf and Levin and can be used as a criterion of randomness. (The property that t takes only values 2^n is only for convenience.)

In the gambling interpretation, the universal test is, in some sense, optimal. Its existence is surprising. E.g., if ω is supposed to be a sequence of coin tosses then I could make my test function large for those ω's in which the frequency of 1's is at least 60% (since their probability is small): this way, I would profit from a certain kind of cheating the casino might attempt. Alternatively, I can make f larger on many other improbable sets of sequences. The universal test anticipates and combines all these strategies.

The **algorithmic (fine-grained) entropy** of ω with respect to μ is defined as

$$H_\mu(\omega \mid y) = -\log t_\mu(\omega \mid y).$$

We will delete μ from the subscript when it is obvious from the context.

H can take arbitrarily large negative values, even $-\infty$. In other words, an object can be infinitely non-random, though the measure of such objects has probability 0. For a finite measure μ, the function $H_\mu(\omega)$ is bounded from above. For infinite measures, it can also take arbitrarily large positive values; but it will never be ∞.

Let $H_\mu(\omega) = H_\mu(\omega \mid 0)$ where Y is chosen as the one-element set $\{0\}$. If both Ω and Y have measures μ, ν then we define

$$H_{\mu,\nu}(\omega, y) = H_{\mu\times\nu}((\omega, y))$$

where $\mu \times \nu$ is the product measure.

Here are some additional easy properties of $H(\omega)$. If f is a computable function then the following holds:

$$H_\mu(\omega \mid y) \stackrel{+}{<} H_\mu(\omega \mid f(y)).$$

We have

$$\mu\{\omega : H_\mu(\omega) < m\} < 2^m \ (-\infty < m < \infty), \quad (3.1)$$

$$H_\nu(y \mid \mu) \stackrel{+}{<} -\log \int 2^{-H_{\mu,\nu}(\omega,y)} \mu(d\omega).$$

The first inequality states that $H(\omega)$ is large only with small probability. The second one is needed for the addition theorem, stated later. For the volume measure L in the phase space of a dynamical system, we define $H(\omega) = H_L(\omega)$ as the fine-grained algorithmic entropy of a state ω. Let

$$E_m = \{\omega : H(\omega) < m\}.$$

In (3.1), we have shown $L(E_m) < 2^m$ for all m. For finite space volume $L(\Omega)$, this implies that

$$L(E_{\log L(\Omega)-m})/L(\Omega) < 2^{-m}, \quad (3.2)$$

i.e. the proportial volume of the set of those points where algorithmic (fine-grained) entropy is not close to its maximum is very small.

4 Additivity

Complexity has the following additivity property, due in various forms to Kolmogorov, Levin, Gacs, Chaitin:

$$K(x, y) \stackrel{+}{=} K(y) + K(x \mid y, K(y)).$$

This property, in an appropriate form, generalizes for algorithmic entropy, as shown by Vovk and Vyugin. We spare the reader the general form in the present abstract. A simple corollary is

$$H_{\mu,\nu}(x, y) \stackrel{+}{<} H_\nu(y) + H_\mu(x \mid y).$$

The symmetric quantity

$$I(x, y) = K(x) + K(y) - K(x, y)$$

is called the **mutual information** between the objects x, y. It generalizes to

$$I_{\mu,\nu}(x, y) = H_\mu(x \mid \nu) + H_\nu(y \mid \mu) - H_{\mu,\nu}(x, y).$$

(with respect to the measures μ, ν). A related quantity is

$$I_\mu(y : x) = H_\mu(x) - H_\mu(x \mid y)$$

which is the **information** that y carries about x with respect to μ. When μ is obvious from the context we omit it from the subscript. The complexity addition theorem is equivalent to the following relation between the two kinds of information:

$$I(x, y) \stackrel{+}{=} I(x, K(x) : y)$$

which also generalizes to algorithmic entropy. The right-hand side can be interpreted as the information that the pair $(x, K(x))$ carries about y. For elements y of a countable set Y, the inequality

$$I_\mu(y : x) \stackrel{+}{<} K(y) \quad (4.1)$$

can be proved. It says that an object cannot carry more information about a string than its own complexity.

5 Randomness, complexity, coarse-grained entropy

Let Ω^0 be the set of of all elements of Ω that do not belong to the boundary of any open cell. Let $\mathcal{M}^0(\Omega)$ be the set of measures μ such that $\mu(\Omega^0) = 0$. If Ω is the interval $[0,1)$ and cells are rational subintervals then elements ω of Ω^0 can be expressed as infinite strings of bits: just use the binary representation. There is a family of cells corresponding to this representation: for binary string s of length n, the cell Γ_s is the set of states ω first n bits ω_i are those in s.

This representation can be generalized to arbitrary phase spaces Ω, by choosing the appropriate cells Γ_s: divide Ω in half, then each half again in half, etc., in such a way that each point of Ω_0 is eventually the intersection of all the subdivisions containing it. If s has length n then Γ_s will be called a **canonical n-cell**, or simply canonical cell, or n-cell. For each n, the canonical n-cells form a partition $\mathcal{P}_n$ of Ω_0 with 2^n elements.

We will generally assume that the partition chosen is also "natural". The bits $\omega_1, \omega_2, \ldots$ could contain information about the point ω in decreasing order of importance from a macroscopic point of view. For example, for a container of gas, the first few bits may describe, to a reasonable degree of precision, the amount of gas in each left half of the container, the next few bits may describe the amounts in each quarter, the next few bits may describe the temperature in each half, the next few bits may describe again the amount of gas in each half, but now to more precision, etc. From now on, whenever Γ denotes a subset of Ω, it means a canonical cell. From now on, for elements of Ω^0, we can talk about the n-th bit ω_n of the description of ω: it is uniquely determined. Let

$$\omega^n = \omega_1 \cdots \omega_n.$$

The Coding Theorem 2.1 implies that if x runs on a discrete space then $H_\#(x) \stackrel{+}{=} K(x)$. More generally,

$$H_\mu(x) \stackrel{+}{=} K(x \mid \mu) + \log \mu(x).$$

There is a similar characterization of tests over arbitrary spaces. Let us denote

$$H_\mu(\Gamma) = K(\Gamma \mid \mu) + \log \mu(\Gamma)$$

for canonical cells Γ.

(5.1) **Test Characterization Theorem** *For a computable measure μ in $\mathcal{M}^0(\Omega)$, we have*

$$H_\mu(\omega) \stackrel{+}{=} \inf_{\omega \in \Gamma} H_\mu(\Gamma). \tag{5.2}$$

The constant in $\stackrel{+}{=}$ here depends on μ. This is the promised connection between the coarse-grained algorithmic entropy $K(\Gamma) + \log L(\Gamma)$ and the (partition-independent) fine-grained algorithmic entropy $H_L(\omega)$. In the special case when Ω is the set of natural numbers and μ is the counting measure this theorem is equivalent to the Coding Theorem 2.1.

(5.3) **Remark** An interesting application where a related formula is used is the "minimum description length" principle (MDL) theory of statistics. There, instead of the description complexity $K(x)$, often the codeword length $C(x)$ of some other coding (universal over some class of measures) is considered, and the quantity $C(\Gamma) + \log \mu(\Gamma)$ is called the **redundancy**. In these statistical applications, the presence or absence of μ in the condition makes a difference. $\diamondsuit$

The Test Characterization Theorem, in the form

$$H_L(\omega) \stackrel{+}{=} \inf_n K(\omega^n) + \log L(\Gamma_{\omega^n}) \tag{5.4}$$

says that the fine-grained algorithmic entropy $H(\omega) = H_L(\omega)$ with respect to the invariant volume measure L can be essentially expressed as $K(\omega^n) + \log L(\Gamma_{\omega^n})$ for a certain n: so, it is the sum of the Boltzmann entropy for the partition $\mathcal{P}_n$ plus the description complexity of the macroscopic description ω^n. For the systems and partitions of interest in physics, the additive term $K(\omega^n) \stackrel{+}{<} 2n$ is typically negligible compared to the other one since the total number of macroscopic cells is typically small compared to the volume of the large cells.

In "practice", to find the "right" n we should keep increasing it, i.e. include (the program for) more and more bits of ω into the macroscopic description as long as the complexity increase buys greater decrease in the Boltzmann entropy $\log L(\Gamma_{\omega^n})$ (our *a priori* uncertainty about ω).

The following theorem says that, for most elements ω of a cell Γ, the value of $H_\mu(\omega)$ cannot be much higher than $H_\mu(\Gamma)$.

(5.5) **Stability Theorem**

$$\mu\{\, \omega \in \Gamma : H_\mu(\omega) < H_\mu(\Gamma) - K(l(\Gamma)) - m \,\} < 2^{-m}\mu(\Gamma).$$

We can also interpret this theorem as saying that if some elements of the cell are (sufficiently) random then most of them are (sufficiently) random. Note that the difference $K(l(\Gamma))$ is less than $2 \log n$ for Γ_{ω^n}.

6 Entropy increase properties

Fine-grain nondecrease As stated in the Introduction, we are considering an isolated physical system with state space Ω whose development is described by a transformation group U^t. We also assume that $U^t\omega$ is computable as a function of the pair ω and t. We assume the existence of a computable invariant measure L (the "Liouville measure"): it has the property that $L(U^t A) = L(A)$ for all t and all measurable sets A. Under suitable conditions, the existence, computability and even uniqueness of L can be proven.

These properties imply that the function $2^{-H(U^t\omega)}$ is a parametrized randomness test. From this, it can be derived that

$$H(U^t\omega) \stackrel{+}{>} H(\omega \mid t) = H(\omega) - I(t : \omega).$$

This is, in essence, our entropy nondecrease formula since the term $I(t:\omega)$ can be shown to be generally very small. Alas, it is also an entropy nonincrease formula, giving

$$-I(t:\omega) \stackrel{+}{<} H(U^t\omega) - H(\omega) \stackrel{+}{<} I(t:U^t\omega). \quad (6.1)$$

According to this, the only amount of decrease we will ever see in $H(U^t\omega)$ is due to the information that the value of the time t may have on ω, which is very small for all simple moments of time. But the amount of increase is also only due to the information that t may have on $U^t\omega$. Technically, the nondecrease property can be better seen from the following theorem, which says that entropy is smaller by $m + K(T)$ than its present value only in a fraction 2^{-m} of all times between 0 and T.

(6.2) **Entropy Nondecrease Theorem** *Let λ be the length (Lebesgue) measure, and let T be a rational value of time. We have*

$$\lambda\{t \in [0,T] : H(U^t\omega) < H(\omega) - K(T) - m\} < 2^{-m}T.$$

Gibbs ensembles Another accepted model of a macrostate is a certain distribution ν over microscopic states given by a density function (ensemble) $p(\omega)$ with respect to the volume measure L. Let us require $\int p(\omega)L(d\omega) = 1$. We can ask what is the probability density to find it in state ω at time $t + t_0$? Let us call this new ensemble p^t.

The classical definition of Gibbs entropy of a probability distribution with density function $p(\omega)$ over L is

$$G(p) = -\int p(\omega)\log p(\omega)L(d\omega).$$

In the special case when p is the macrostate-ensemble we have $G(p_\Gamma) = \log L(\Gamma)$, i.e. the Gibbs entropy is the same as the Boltzmann entropy. Liouville's Theorem implies $G(p^t) = G(p)$, i.e. that the Gibbs entropy of an ensemble does not change at all in an isolated system during evolution. This shows that in case of the evolution of isolated nonequilibrium systems, the evolution of a Gibbs ensemble does not express adequately what we consider thermodynamic behavior. The problem is that even if at the starting time t_0 the Gibbs ensemble was something simple, it can develop in time t into a very complicated density function that does not correspond to any reasonable macroscopic description. Ensembles that are invariant in time retain their usefulness, however, for equilibrium systems.

The following relation shows that the Gibbs entropy is the average of algorithmic entropy:

$$\begin{aligned}\int H_\mu(\omega)\nu(d\omega) &\stackrel{+}{<} -\int -p(\omega)\log p(\omega)\mu(d\omega) \\ &\stackrel{+}{<} \int H_\mu(\omega)\nu(d\omega) + K(\lfloor \log \nu(\Omega) \rfloor).\end{aligned}$$

The increase of coarse-grained entropy Let us now consider the much more speculative problem of approach to equilibrium, i.e. the argument that the algorithmic Boltzmann entropy $H^n(U^t\omega)$ must indeed increase fast if it is far from its upper bound $\log L(\Omega)$.

There is a classical argument to show that in a nonequilibrium system, Boltzmann entropy can be expected to increase fast until it almost reaches its upper bound $\log L(\Omega)$. The argument relies on two properties. The property that we will call Noncompensation Condition says that in the partitions of interest, the "large cells" dominate the space so much that even the union of all "small" cells taken together is small. It holds for typical systems simply because the total number of cells is small compared to the volumes of interest. With algorithmic coarse-grain entropy, in this condition, "large" must be replaced by "small-entropy", and then it is always satisfied. Indeed, the set $\{\omega : H^n(\omega) < m\}$ is contained in $E_m = \{\omega : H(\omega) < m\}$, which, according to (3.2), has volume at most $2^{-m}L(\Omega)$.

The second property needed for entropy increase, which we will call the Weak Mixing Condition, says that if the system starts from a state in a not very small cell then after a time t, it is unlikely to end up in any small union of cells. However plausible, this condition can be rigorously proved only in some special cases. For coarse-grained algorithmic entropy, the condition remains just as difficult to prove but our framework may allow to formulate the mixing property in a sharper way.

In the following two (arguably exotic) examples the coarse-grained algorithmic entropy is preferable to Boltzmann entropy.

(6.3) **Example** Take a large container filled with ideal gas and a few large balloons. At start, the balloons are fixed. Then we release them. They gain energy from collisions with the gas molecules until they achieve the average energy appropriate to their number of degrees of freedom.

It is reasonable to count the positions of the balloons to the macrostate of the system. The volume of the cell will be essentially determined by the energy of the system consisting of the gas alone. This energy, and hence the Boltzmann entropy, becomes smaller by the amount transferred to the balls. $\Diamond$

This example becomes less ridiculous if we replace balloons by the memory of a computer. For a while, it will still be reasonable to count the content of the memory as part of the macroscopic description since bits can be individually observed and manipulated. However, as the size of a site storing an individual bit decreases, there will come a point where it is not reasonable to consider the memory state as part of the macroscopical description. The communication of two computers, one with "macroscopic" memory and the one with "microscopic" memory, leads to the Maxwell demon paradox. This shows the necessity of the smooth transition between macroscopic and microscopic transitions exhibited by algorithmic entropy.

In terms of our scheme, we are talking about increasing n (refining the partition). The additive term $K(\omega^n)$ which is so insignificant for small values of n, gains in significance in this process and makes the transition continuous. Ignoring it, by defining entropy just as $\log L(\omega^n)$, is bound to lead to paradoxes.

The following system also defies Boltzmann entropy but submits to algorithmic coarse-grain entropy.

(6.4) **Example: the baker's map** Let Ω be the set of doubly infinite binary sequences $\omega = \ldots\omega_{-1}\omega_0\omega_1\omega_2\ldots$ with the shift transformation $(U^t\omega)_i = \omega_{i+t}$ over discrete time. Let us write $\omega^n = \omega_{-\lfloor n/2\rfloor}\cdots\omega_{\lceil n/2\rceil-1}$. The n-cells have the form Γ_{ω^n}. Let the volume L be such that all n-cells have the same volume 2^{-n}.

Since all n-cells have the same measure no matter what fixed precision we choose, the Boltzmann entropy of $U^t z$ does not increase with t. The quantity $H^n(U^t z)$, however, can be shown to increase fast for all typical sequences, between times time 0 and n, linearly from $-n$ to 0.

We can also use independent biased coin tossings or any stationary process for the measure, and obtain similar results. For such processes, there is an "asymptotic equidistribution property" guaranteeing that most volume will be taken up by n-cells of about the same size 2^{-hn} (where h is the so-called "entropy rate"). $\Diamond$

In typical physical systems, the partitions given by the canonical cells have no simple connection with the computable transformation group U^t of our dynamical system. In particular, they are not "generated" from the first partition into Γ_0 and Γ_1 by U^t the way they are in the baker's map.

Algorithmic entropy can also be used to relate the growth the the Kolmogorov-Sinai entropy of a dynamic system to the growth rate of complexity.

The example of the shift transformation suggests that in chaotic systems, the parameter n can often be made a function of t as long as it grows slower than linearly with t. Thus, if $\lim_{t\to\infty} n(t)/t = 0$ then in the baker's map with the uniform distribution,

$$H^{n(t)}(U^t\omega)$$

will approach $\log\mu(\Omega)$ almost as fast as if we held n constant. The growth of $n(t)$ seems a good measure of the mixing of U^t.

The paradox of typicality The notion of a typical object is an informal one, and the present remark calls attention to the fact that our intuition concerning the properties of typical objects may be misleading. Consider the space of infinite 0-1 sequences obtained by tossing a biased coin, with probabilities 1/3, 2/3. We would consider typical those sequences in which the relative frequency of 0's tends to 1/3. On more reflection, we would consider those sequences ω typical that also satisfy all other criteria or randomness, and are random according to Martin-Löf's definition, or, equivalently, which have $H_p(\omega) > -\infty$ where p is the appropriate coin-tossing measure.

Consider now a dynamical system with the volume measure L. The (fine-grained) entropy nondecrease property, which can also be considered a randomness-conservation property, guarantees that the above defined typicality is "conserved": the evolution of a system takes typical states into typical ones.

There is, however, another, similarly attractive idea of typicality, which we will call "local typicality", which is not conserved. Consider a given partition $\mathcal{P}_n$ and a given cell Γ in this partition. Let us call those states ω of the cell Γ "locally typical" whose fine-grained entropy $H(\omega) = H_L(\omega)$ is close to the coarse-grained entropy $H^n(\omega) = \log L(\Gamma) + K(\Gamma)$. We know from the Stability Theorem 5.5, that most points of each cell (in terms of the measure L) are typical in this sense. However, local typicality is not conserved. Indeed, assume that coarse-grained entropy increases for most points, and that the coarse-grained entropy of Γ is low. Then for most points ω in Γ (and hence for most locally typical points), their coarse-grained entropy increases, so $U^t\omega$ belongs to a cell Γ' with much higher coarse-grained entropy. At the same time, the fine-grained entropy $H(U^t\omega)$ does not change too much with respect to $H(\omega)$. So, the locally typical state ω turns into a locally nontypical point $U^t\omega$. The reason is that ω was a locally typical point of a "nontypical" cell, and it is still carrying this history. (However, as long as only the macroscopic information embodied in the cell Γ' is available for inspection and manipulation, this history is inaccessible to later observers.)

The non-conservation of local typicality eliminates a potentially attractive "principle": namely that the state we have at present is locally typical. Consider the present state of a container of gas after a wall was removed that had confined the gas to one half. In a usual macroscopic description (partition), the coarse-grained entropy of the present state will be much larger than what it was before the wall removal. Since the fine-grained entropy is approximately the same (since it did not change much), the present state is actually highly nontypical.

The refuted principle is attractive since, together with the Boltzmann Entropy Increase Property (or its counterpart using algorithmic coarse-grained entropy) it could be used to prove that entropy increase is not only likely to occur but will occur. One possible substitute of the principle is the introduction of probabilistic perturbations, see e.g. [?].

We prefer to say that *the entropy increase property relates strictly only to the present macroscopic state of our system, and does not assert directly anything about the present microscopic state.*

7 Maxwell's demon

Entropy balance Let $\mathcal{X}$ and $\mathcal{Y}$, be two systems where $\mathcal{Y}$ is considered the environment from which $\mathcal{X}$ is temporarily isolated. In order to to "do something" to $\mathcal{X}$, we couple it with $\mathcal{Y}$, giving rise to a joint Hamiltonian and a joint transformation $U^t(\xi,\eta)$. Let us assume that, being in classical mechanics, the impulses

and momenta of the joint system are simply the impulses and momenta of the two subsystems, therefore the Liouville measure on $X \times Y$ is, even in the coupled system, the product of the original Liouville measures L_X, L_Y in the subsystems. We will denote the transformations in the joint system again by $U^t(\xi, \eta)$. Let $(\xi_t, \eta_t) = U^t(\xi, \eta)$, and

$$\Delta H(\xi) = H(\xi_t) - H(\xi).$$

(7.1) **Entropy Balance Theorem**

$$\Delta H(\xi) + \Delta H(\eta) \stackrel{+}{>} I(\xi_t, \eta_t) - I(\xi, \eta) - I(t : \xi, \eta).$$

Since the last term is generally negligible this theorem says that if the two systems were originally independent (i.e. $I(\xi, \eta) \approx 0$) then a decrease in the entropy of ξ must be accompanied by an increase in the entropy of η.

The entropy balance theorem is not new, of course, for Boltzmann entropy. But its present form makes it useful for the treatment of Maxwell's demon.

Maxwell's demon and Landauer's thesis Maxwell's demon is a being sitting at a tiny door between two compartments of gas and letting the molecules through selectively in a way as to achieve entropy decrease in the container.

The typical explanations assume either that the door will heat up and begin to work randomly after a while, or that in order to make its observations, the demon must descend into this world more than she cares to and interact energetically with the molecules; this heats her up, making it harder and harder for her to concentrate. These explanations introduce additional physical assumptions which are alien to the general mathematical nature of the second law (increase of disorder). Several such explanations are refuted by more refined models (see [1]).

A convincing modern solution emerged in a principle announced by Landauer (see [1]). Let us model the demon as some computer-controlled device interacting with the gas. She seems to be able to decrease the Boltzmann entropy of the gas only at the expense of the increase of her own information content. Landauer introduced a principle saying that in order to erase a bit of information, a certain minimal amount ($kT \log 2$) of heat dissipation into the environment (and, of course, investment of the corresponding amount of work into the system) is necessary: hence entropy decrease occurs only as long as the demon keeps filling up with more and more information.

(7.2) **Remark** In order to ~~argue~~ prove that the erasure results in heat dissipation, Landauer argues that the erasure must be a general operation that decreases the phase space of the computer memory.

I find it difficult to interpret the increase of $H(\eta)$ universally as heat dissipation. Consider a memory ξ consisting of a row of pendulums swinging transversally. The bit 1 means that the pendulum swings while the bit 0 means it does not. Let the "environment" η be an identical row of pendulums, each of which hangs motionless originally. Now we can use η to erase the memory by moving it next to ξ in the right moment. The change in η is obviously reversible, so there is no heat dissipation. $\Diamond$

Using our framework, the demon paradox occurs since the state of demon's memory was implicitly considered part of the macroscopic description of the joint system gas-demon: the quantity of information in it failed to contribute to the classical Boltzmann entropy. Since this device is able to decrease the Boltzmann entropy of the gas at the expense of the increase of the information content of its memory (without increasing its own Boltzmann entropy), it is able to decrease the Boltzmann entropy of the total system. Our solution eliminates the paradox by including the information content (complexity) of the macroscopic description into the expression for entropy. It should be considered as the rigorous formulation of the more special and informal principles of Landauer and Zurek.

Zurek [8] saw that the Maxwell's demon paradox and Landauer's thesis are two sides of the same interaction between an information-processing machine (the demon) and a classical thermodynamic system. The demon turns entropy into information, the information-erasure operation turns information into entropy. Zurek constructed an entropy-like quantity specifically for this situation and argued that it is non-increasing. He created a special macroscopic variable d (without actually distinguishing macroscopic and microscopic), whose value is equal to the demon's memory state. He defined then a quantity called "physical entropy" associated with such a system that is essentially $Z(a, d) = B(a) + K(d)$ where $B(a)$ is the Boltzmann entropy of the classical part. This can be seen as essentially the same as $B(a, d) + K(a, d)$. Indeed, $B(a, d) = B(a)$ since the demon's macroscopic and microscopic states are the same. Also, we can delete a from $K(a, d)$ since we are interested in situations in which d contains much more information than a.

Zurek argues that if, at constant temperature T, the system is brought from state (a_1, d_1) to state (a_1, d_2) then the amount of work obtained is at most $Z(a_2, d_2) - Z(a_1, d_1)$. For this, he tacitly assumes that the work gained from the operation of the system can be separated into the work obtained from bringing the classical machine from a_1 to a_2 and into the work bringing the memory from d_1 to d_2. With this assumption, the second law indeed implies the upper bound $T(B(a_2) - B(a_1))$ on the first kind of work and Landauer's principle implies the bound $T(K(d_2) - K(d_1))$ for the second kind of work.

Formally, our coarse-grained algorithmic entropy looks similar to Zurek's but is defined more generally, and has many connections to various different definitions of entropy (for ensembles as well as cells) and also to the theory of randomness. In particular, the above bound can be proven without the tacit assumptions.

Let ξ be the gas whose entropy the demon is trying to decrease. We also count the whole state η of the

demon into her macroscopic description. As it is usual with classical machines, we can assume that there is an n such that $H^n(\xi)$ is close to $H(\xi)$ but n (and therefore $K(\xi^n)$) is still negligibly small with respect to $H(\xi)$, and therefore

$$H(\xi) \approx \log L(\Gamma_{\xi^n}).$$

The entropy balance theorem guarantees that going from (ξ, η) to (ξ^t, η^t), the decrease in the sum $H(\xi) + H(\eta)$ will be small. Since $H^n(\xi^t) \stackrel{+}{>} H(\xi^t)$ this implies that any decrease in $H^n(\xi)$, i.e. the Boltzmann entropy of the machine, must be compensated by an increase in $H(\eta) \approx K(\eta)$, i.e. the information content of the demon's memory (this is the Maxwell's demon direction) and vice versa (this is the Landauer thesis direction).

References

[1] C.H. Bennett. Demons, engines and the second law. *Scientific American*, 255(11):108–116, November 1987.

[2] Carlton M. Caves. Information, entropy, chaos. In J. Pérez-Mercader and W. H. Zurek, editors, *Physical Origins of Time Asymmetry*. Cambridge University Press, Cambridge, England, 1993.

[3] Peter Gács. Exact expressions for some randomness tests. *Z. Math. Log. Grdl. M.*, 26:385–394, 1980. Short version: Springer Lecture Notes in Computer Science 67 (1979) 124-131.

[4] Joel L. Lebowitz. Boltzmann's entropy and time's arrow. *Physics Today*, pages 32–38, Sept 1993.

[5] L. A. Levin. Randomness conservation inequalities: Information and independence in mathematical theories. *Information and Control*, 61(1):15–37, 1984.

[6] M. Li and P. M. B. Vitányi. *Introduction to Kolmogorov Complexity and its Applications*. Springer Verlag, New York, 1993.

[7] V. G. Vovk and V. V. Vyugin. On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society B*, 55(1):253–266, 1993.

[8] W. H. Zurek. Algorithmic randomness and physical entropy. *Physical Review A*, 40(8):4731–4751, October 1989.

[9] A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Math. Surveys*, 25(6):83–124, 1970.