The Boltzmann entropy and randomness tests

P. Gács

Computer Science/Department of Algorithmics and Architecture

Report CS-R9342 June 1993

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications. SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

P.O. Box 4079, 1009 AB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

The Boltzmann Entropy and Randomness Tests

Péter Gács*
Boston University
Computer Science Department
Boston, MA 02215, USA

Abstract

Descriptonal (Kolmogorov) complexity has a certain well-understood relation to the information-theoretical entropy introduced by Shannon. The latter is, by its form, related to the Gibbs entropy of statistical mechanics. Entropy in physics has several different definitions, and a history of philosophical controversy concerning the related notion of irreversibility.

After a fast-paced introduction to the necessary concepts of thermodynamics, complexity and randomness tests, we develop some of the connections to physics by introducing a new notion called "algorithmic entropy". This, on the one hand, is in a simple relation to randomness tests (of Martin-Löf and others), and on the other hand, is a connecting link between description complexity, Gibbs entropy and Boltzmann entropy. Its coarse-grained approximation will be shown to have the desirable properties of Boltzmann entropy in a somewhat wider range of systems, including those of interest in the "thermodynamics of computation". It is hoped that eventually these new concepts will prove helpful not only for computer science but also for physics and dynamical system theory, by extending thermodynamical reasoning to some new classes of systems.

1991 Mathematics Subject Classification: 60A99, 68Q30, 80A05, 82B03, 82C03
Keywords and Phrases: entropy, description complexity, Kolmogorov complexity, algorithmic entropy, randomness test.

1. Introduction

In its present form, the intended audience of this paper is mainly computer-scientists, who want to understand the relation of description complexity to thermodynamics. This determines the form of exposition somewhat: Section 2 is a (very brief) introduction to some thermodynamical concepts and problems. But we hope that the actual technical content of the paper will ultimately be also considered a contribution to physics and dynamical system theory: mainly in providing some tools to extend thermodynamical reasoning to some new classes of systems.

In the present form, this paper cannot be used as an introduction to Kolmogorov complexity even if formally all definitions are given. There are several survey papers serving that purpose, *e.g.* [2] or the forthcoming book by the same authors. The concept of computability is assumed to be known, even though computability in continuous spaces is developed in the Appendix.

The rest of the introduction presents the main ideas but some readers may need to read the first few sections of the paper to understand them.

The central concept introduced in this paper is what we call the **algorithmic entropy**

*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. Part of the work was done while the author was visiting Rutgers University and DIMACS.

Report CS-R9342
ISSN 0169-118X
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

$$H(\omega) = H_L(\omega)$$

of an element ω of a state space Ω with respect to the volume measure L. This quantity will be shown to be nearly equal to the negative of Martin-Löf's randomness tests. The entropy of a Gibbs ensemble (a probability distribution) $p(\omega)$ will be shown to be nearly equal to the average of $H_L(\omega)$ with respect the distribution $p(\omega)$.

In order to talk about coarse-grained entropy, a series of finer and finer partitions $\mathcal{P}_n$ of the state space is introduced. Each microscopic state ω is the intersection of all the partition cells Γ with $\omega \in \Gamma$. These cells can be identified with the macroscopic descriptions of various levels of detail. Let

$$H(\Gamma) = K(\Gamma) + \log L(\Gamma). \tag{1.1}$$

Here, $K(\Gamma)$ is the description complexity of the cell Γ (to be defined later). For a state $\omega \in \Gamma$, the interpretation of this formula is a separation of algorithmic entropy into a part relating to the information contained in the macroscopic description of ω and the remaining uncertainty $\log L(\Gamma)$ which is the same as the Boltzmann entropy of ω (corresponding to the partition $\mathcal{P}_n$ with $\Gamma \in \mathcal{P}_n$).

Description complexity itself will be shown to be a special case of algorithmic entropy.

First, some properties of the new quantity (in particular, the so-called Noncompensation Condition and the Landauer principle, see later) are now derivable. Second, it 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).

Section 2 is a very cursory introduction to some thermodynamical ideas and the Maxwell demon problematic.

Section 3 defines description complexity $K(x)$ (the self-delimiting version) of a finite objext x and shows its main properties.

Section 4 introduces randomness tests. The algorithmic (fine-grained) entropy $H_\mu(\omega)$ of a state ω of the system with respect to an underlying measure μ (typically, the volume measure over a finite-dimensional space or the counting measure over a discrete space) is defined as a common generalization of the description complexity $K(x)$ and the randomness test of Martin-Löf (as modified by Levin). It is the negative logarithm of the maximal (to within an additive constant) lower semicomputable function $f(x) \geq 0$ with the property that

$$\int f(\omega)\mu(d\omega) \leq 1.$$

It is known that the description complexity $K(x)$ is within constant distance from $H_\#(x)$ where $\#$ is the counting measure.

Section 5 shows how to express $H_\mu(\omega)$ in terms of complexity, with the help of a formula

$$K(\Gamma|\mu) + \log \mu(\Gamma)$$

similar to (1.1). The appearence of μ in the condition is not important if μ is computable but makes some difference otherwise. In particular, an interesting and related 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.

The same section also proves an addition property for algorithmic entropy, which is the generalization of the information-theoretic addition property of complexity.

Section 6 develops the connection in more detail to Boltzmann entropy and Gibbs entropy.

The classical definition of Gibbs entropy of a probability distribution with density function $p(\omega)$ over L is $-\int p(\omega)\log p(\omega)L(d\omega)$. We will show that this quantity is very close to the average algorithmic (fine-grained) entropy. Therefore the algorithmic entropy of an individual state is really an "algorithmic Gibbs entropy", and the algorithmic coarse-grained entropy (1.1) is some approximation of it.

For "typical" systems of classical statistical mechanics, and for typical cells, the complexity term here is negligible in size compared to other one, therefore coarse-grained algorithmic entropy is close to Boltzmann entropy. But we hope we can justify the new concept by the theoretical clarity, mathematical manageability and its position in shedding light on the connection of several different concepts. We will give some examples that are better handled by our definition, partly due to the fact that our entropy automatically satisfies the so-called Noncompensation Condition (2.6). These examples do not belong to mainstream statistical mechanics, but are borderline cases that may become more important with further miniaturization of computers. It would also be interesting to see other chaotic systems in which the extension of the notion of entropy increase is useful.

Our definition resembles Zurek's entropy defined in [3] but there are important differences. It is defined more generally (not just for systems consisting of a classical and a "demon" part). Its properties are proved with fewer assumptions. It connects randomness tests, fine-grain and coarse-grain entropy (these notions are not distinguished by Zurek) and complexity.

In Section 7, we will prove a strong nondecrease and slow increase property for algorithmic entropy, and show that on the other hand, classical Boltzmann entropy can actually slightly decrease; in some other systems where intuition tells it should increase very fast, it stays constant.

Strong heuristic arguments exist to show that in a nonequilibrium system, Boltzmann entropy can be expected to increase strongly. These arguments consist of an easy part, which we will call Noncompensation Condition (2.6), and a difficult part, the Weak Mixing Condition (2.8), which, however plausible, can be rigorously proved only in some special cases. With our algorithmic Boltzmann entropy, the Noncompensation Condition will come free. The nontrivial, mixing part of the argument remains just as difficult and conditional as for Boltzmann entropy but our framework allows at least to formulate the mixing property in a sharp way.

In Section 8, we will show an additivity property and will discuss its relation to the so-called Landauer thesis which says that erasing a bit of information from, say, computer memory (even if it is done indirectly) requires the dissipation of a certain minimal amount of heat. Maxwell's demon will also be treated in this context.

The Appendix gives some of the longer proofs and is also used to define the various notions of computability in a continuous space (like the phase space of a system of particles). It is quite technical, and I recommend to refer to it only when some definition is really needed, and to rely on an intuitive understanding of computability otherwise (and, of course, to trust the author that such a notion can be naturally defined for a sufficiently nice topological space).

In what follows log is the logarithm to base 2. The quantities

$$\lfloor x\rfloor, \lceil x\rceil$$

are the largest integer $\leq x$ and the smallest integer $\geq x$ respectively.

Acknowledgement I am indebted to Shelly Goldstein for absolutely indispensable instruction on Boltzmann entropy without implying, however, his agreement with the ideas expressed here.

2. Ten minutes on thermodynamics

Thermodynamical systems Entropy was first introduced in classical, "phenomenological" thermodynamics. This theory, which is also the form of thermodynamics most widely used in engineering, is concerned with a physical system when the latter is in a state called "equilibrium".

An **equilibrium state** of the system is characterized by the fact that for all practical purposes, its properties relevant for interaction with the rest of the world are determined by a relatively small number of parameters (functions of the state) called **macroscopic parameters** $u_1, \ldots, u_m$. The simplest system to consider is a certain quantity of gas in a container, with just a few macroscopic parameters: volume, temperature energy and pressure. Two of these can actually be deleted, since they are a function of the other two, but it is not necessary for our purposes to minimize the number of parameters.

Let us agree that energy will always be included among the parameters. The **first law of thermodynamics** is a consequence of a more general law of physics: it says that the energy of an isolated system does not change. The interaction of the system with the outside world involves some exchange of energy.

Macroscopic state The careful distinction between macroscopic and microscopic states originates, as far as we know, with Boltzmann. A macroscopic state is determined by just a few 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. A microscopic state is the complete state, which determines the future (and past) behavior of the system completely whether it is in equilibrium or not.

Thus, with the physical systems we consider there will always be associated a certain space Ω, the set of all possible states of the system. In case of a container of "ideal" gas consisting of n simple molecules, the state space is the $6n$-dimensional Euclidean space since the state of the system is determined if the positions and the velocities of each molecule are known. (It turns out that it is more convenient to use the impulses—mass times velocity—in place of the velocities.)

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. We can agree in advance on some reasonable precision (4 digits are probably more than sufficient). In this way, we obtain a partition of the state space into cells

$$\Omega = \bigcup_x \Gamma_x$$

where $x = x_1, \ldots, x_m$ runs over all possible values of the parameters $u_1, \ldots, u_m$. We will identify a **macroscopic state**, or **macrostate** with a cell Γ_x, *i.e.* with a certain set of microscopic states, or **microstates**. The partition interpretation of a macrostate is also called **coarse-graining**.

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

$$p_\Gamma(\omega) = \begin{cases} 1/L(\Gamma) & \omega \in \Gamma \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Dynamics and volume According to the laws of mechanics, an isolated system undergoes an evolution described by a transformation group U^t. If at time t_0, the system was in state ω then at time $t + t_0$ it will be in state $U^t\omega$. The group U^t is generally given by a system of differential equations which, at least in the example of ideal gas given with coordinates and impulses, are called the Hamiltonian

equations. For this case, Liouville's Theorem holds saying that the volume of a domain remains invariant under transformation under U^t. In most other cases also, a natural measure is found on Ω that remains invariant under U^t, and we will call this measure the **volume**, and denote the volume of a set A by $L(A)$.

The law of energy conservation means that during an evolution of an isolated system, it is confined to a surface of the state space determined by the requirement that the energy is equal to a certain value. Therefore the volume measure to use will be actually obtained by restricting the original volume measure to a thin layer determined by the requirement that the value of energy is in a certain small interval, and normalizing. This measure has, in the limit, the intimidating name **microcanonical ensemble.**

The paradox of irreversibility If a mechanical system has a certain trajectory $t \to U^t\omega$ of its evolution from state $\omega = a$ to state $U^{t_1}\omega = b$ then there is also a trajectory of evolution from state b to state a. It is sufficient to reverse all the final velocities of all particles and the system will trace its trajectory backward. Therefore any evolution seems just as possible as the corresponding reverse evolution.

At the same time, the world seems to be full of irreversible phenomena. Imagine a container A of gas separated from an empty container B by a wall. After the removal of the wall much of the gas will occupy container B and the reverse physical process, when the gas collects itself voluntarily in part A, will never be seen. The reversibility of the equations seems to be in contradiction to irreversibilities of this kind.

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. It is easy to imagine macrostates a, b such that in a time unit, 99.99% of the elements of Γ_a end up in Γ_b but only 0.01% of the elements of Γ_b end up in Γ_a. This is exactly what happens with the gas, except that the percentages are much more extreme.

Equilibrium For an isolated system, we may want to call **equilibrium states** those macrostates that are relatively stable: the values of the macroscopic variables will undergo only small fluctuations in time. On reflection, however, this requirement must be weakened to hold only for *most* microstates within the macrostate (as measured by volume).

How can an equilibrium state be transformed into a different equilibrium state at all? Suppose that our system is a container of gas. The system can be combined with some other systems like a heat reservoir or a piston connedted to a lever. Then some constraint is removed (*e.g.* an insulation is removed or a piston is allowed to move) changing the nature (the equations of motion) of the joint system by making its parts indeed interdependent. When a new equilibrium is reached the constraints can be restored.

Boltzmann entropy Boltzmann defined the **entropy** of a macroscopic state a as the logarithm of the volume of Γ_a:

$$B(a) = \log L(\Gamma_a). \tag{2.2}$$

The most obvious problem with this definition seems to be its dependence on the particular choice and number of macroscopic variables and the precision with which we want to determine them. Indeed, another digit of precision will decrease the Boltzmann entropy of most states by about $\log 10$. The volumes in question are, however, typically so large, that such a small difference is negligible (especially, if we take into consideration that the actual definition also multiplies $B(a)$ by the very small Boltzmann constant k). We will also use the notation

$$B(\Gamma_a) = B(a)$$

since it is harmless here to identify a cell with its description.

(2.3) Example Let us n identical molecules of gas in a container with rigid walls, and the following cell of phase space. The positions of the molecules are restricted to the container, of volume V. The total kinetic energy of the molecules is confined between the values K and $K + \Delta K$. The kinetic energy is $\frac{1}{2m}\sum_{i=1}^{3n} p_i^2$ where p_i are the impulses mv_i and m is the mass of a molecule. The entropy for this "cell" consists of two terms. The first term is coming from the amount of freedom in choosing the positions of the molecules and is $n \log V$. The second term is the log volume of the set of points $p = (p_1, \ldots, p_{3n})$ such that $K - \Delta K \leq \frac{1}{2m}\sum_i p_i^2 < K$. For "reasonable" values of ΔK this is the same as the log volume of the set of points p with $\sum_i p_i^2 < 2mK$, which is $C_{3n}(2mK)^{3n/2}$ where C_n is the volume of an n-dimensional unit ball. This gives for entropy the value

$$n(\log V + \frac{3}{2}\log(2mK)) + \log C_{3n}.$$

It turns out that the correct volume measure to use is $n!$ smaller due to the fact that molecules are not distinguishable. Therefore $\log n!$ must be subtracted from here to get the correct value. This can be best seen when we consider the act of mixing of two quantities of gas, with the same pressure and temperature. Depending on whether the two gases are of different kind or not the entropy of the mixture will or will not be greater than the sum of the two original entropies. (This is the so-called Gibbs paradox.) □

(2.4) Example Consider a container of ideal gas consisting of n molecules of the same kind which is partitioned, in our mind, into m compartments $C_1, \ldots, C_m$ where m is much smaller than n. Forget about velocities, for simplicity. Let the macro variable n_i give the number of molecules in compartment C_i. Let $\mathbf{n} = (n_1, \ldots, n_m)$. Let us also consider a second set of variables, the numbers $i_1, \ldots, i_n$ telling that molecule j is in compartment i_j.

The description $\mathbf{i} = (i_1, \ldots, i_n)$ is, of course, more detailed than the description $\mathbf{n}$, which is a function $\mathbf{n}(\mathbf{i})$ of $\mathbf{i}$. Therefore $\Gamma_{\mathbf{n}} = \bigcup_{\mathbf{n}=\mathbf{n}(\mathbf{i})} \Gamma_{\mathbf{i}}$.

Since the molecules are all of the same kind, the dynamics, and therefore the measure L will certainly be invariant with respect to the exchange of molecules. Therefore if $\mathbf{n}(\mathbf{i}) = \mathbf{n}(\mathbf{i}')$ then $L(\Gamma_{\mathbf{i}}) = L(\Gamma_{\mathbf{i}}')$. Let us consider first the molecules distinguishable. Then $L(\Gamma_{\mathbf{n}}) = N(\mathbf{n})p_{\mathbf{n}}$ where $p_{\mathbf{n}}$ is the common volume of all $\Gamma_{\mathbf{i}}$ with $\mathbf{n}(\mathbf{i}) = \mathbf{n}$, and

$$N(\mathbf{n}) = \frac{n!}{n_1! \cdots n_m!}.$$

If we also assume that all the compartment have the same shape and size then we get that $p_{\mathbf{n}}$ does not depend on $\mathbf{n}$. Indeed, $p_{\mathbf{n}} = V_m^n$, the n-the power of the volume of a single compartment since after $\mathbf{i}$ is given what is left to determine is only the exact position of each molecule in its compartment. Boltzmann entropy of state $\mathbf{n}$ is equal to

$$H(\mathbf{n}) = \log N(\mathbf{n}) \approx n(-\sum_i f_i \log f_i + \log V_m)$$

where $f_i = n_i/n$. Again, if the molecules are indistinguishable it is necessary to subtract $\log n!$. □

Let us call a system **volume-defined** (this is the one given preference in textbooks) if the two macroscopic variables V (volume) and E (energy) determine the rest. Here, V can be given exactly, and E is given with some precision. The entropy

$$S(V, E)$$

is given as the volume of the set of phase points corresponding to coordinates in volume V and energy between $E - \Delta E$ and E. The previous example shows that this is approximately the volume of the set with the simple limitation that the energy is $< E$ (the volumes of an n-dimensional ball and a shell of it are close).

Consider a system consisting of two parts 1, 2, such that both parts have separate descriptions, both microscopic and macroscopic. Then a cell corresponding to the joint system is the Cartesian product of its projection cells in systems 1 and 2, and its volume is also the product of the corresponding volumes. This way, the entropy of the joint system is the sum of the two entropies, so entropy is an additive function of the state.

Entropy was first defined in a so-called "phenomenological" model, using only the notions of temperature, heat and reversibility, without reference to atoms. We will point out the connection more closely a little later, but the last example already gives some indication.

The growth of entropy Since entropy measures the amount of phase space occupied by a macrostate it sounds plausible that a system tends to be in states with high entropy but not obvious without further discussion. Anyway, this is exactly how classical thermodynamical systems are known to behave: according to one formulation of the second law of thermodynamics, entropy in isolated systems cannot decrease. There is also a "phenomenological" formulation, saying that heat cannot be transformed into work (leaving everything else unchanged). There are some other related entropy increase properties, so let us list all of those interesting us:

1. in an isolated system, entropy cannot decrease;
2. an isolated system undergoes an irreversible transformation exactly when entropy actually increases;
3. an equilibrium state of an isolated system is a maximum (or at least a strongly pronounced local maximum) of entropy.

Just as the precise notion of reversibility had to be formulated statistically, the same can be expected for these entropy increase properties for Boltzmann entropy.

Formulation and proof are, of course, different and rigorous proof is unavailable for most realistic systems. We will analyze the heuristic arguments for the increase of Boltzmann entropy later but first we want to understand the ordinary physical consequences.

Temperature and pressure Suppose that two volume-defined systems are brought into contact and are allowed to come into thermodynamic equilibrium without a change to their volume. Then according to the maximum entropy principle, the joint entropy is maximal, so a small reversible energy exchange between the two containers should not change it. If at this point therefore the amount of energy dE was transmitted reversibly from system 2 to system 1 and the entropy of system S_i increased by dS_i then $dS_1 + dS_2 = 0$; in other words,

$$\frac{dS_1}{dE_1} = \frac{dS_2}{dE_2}.$$

In the Example 2.3, if our gas is ideal gas, all its energy is kinetic and therefore we have $dS/dE = dS/dK = n/K$. The quantity K/n is the average kinetic energy of a molecule, and is also called the temperature T of a gas. This and the above equilibrium property allow us to define temperature in general for volume-defined systems as follows. First we express energy via entropy and volume as $E(S, V)$. Then we define

$$T = \frac{\partial E(V, S)}{\partial S}.$$

We can also introduce pressure for a volume-defined system by

$$p = \frac{-\partial E(V, S)}{\partial V}$$

getting the equation

$$dE = TdS - pdV$$

for reversible changes in volume-defined systems. Here, the first term is the heat part dQ of the energy received by the system, and the second one is the work part. The equation turns into inequality for irreversible changes.

Conditions of entropy increase Let

$$M = \log L(\Omega)$$

be the maximal value of the entropy. Let $C(d)$ be the union of all cells Γ with

$$B(\Gamma) < M - d$$

(it is, of course, a tail of this sequence). We are looking for an entropy increase property in the following form. (It is still somewhat loosely formulated but hopefully gives the idea.)

(2.5) Boltzmann entropy increase property
There is a function $f(t, d, e)$ with $\lim_{t\to\infty} f(t, d, e) = \infty$ such that if d is large enough (but still small compared to M), $d < e$ and $\Gamma \not\subset C(e)$ then

$$L(U^t\Gamma \cap C(d)) < 2^{-f(t,d,e)}L(\Gamma).$$

□

This says that if we start from a cell with entropy at least $M - e$ then within time t, the the proportion of those of its elements that do not end up in a cell with entropy $M - d$ is at most $2^{-f(t,d,e)}$.

Two properties will provide a satisfactory condition. The first one says the union $C(d)$ of all "small" cells taken together is small. We call it the Noncompensation Condition to suggest that the smallness of small cells is not compensated by their quantity. The condition holds for typical systems simply because the total number of cells is small.

(2.6) Noncompensation Condition There is a constant $c > 0$ such that for d large enough (but still small compared to M) we have

$$L(C(d)) \leq 2^{-cd}.$$

□

In Example 2.4, the total number of cells is the total number of partitions of n into $\sum_{i=1}^{m} n_i$, which is less than n^m. Therefore $L(C(d)) \leq 2^{-(d-m\log n)}$, so the condition holds with $c = 0.5$ for all $d > 2m\log n$, while M is of the order of n.

(2.7) Example The enormous differences in cell sizes seem to be typical for statistical mechanical systems even without reference to the small number of cells. Suppose, for example, that about a container of a mole of gas, our macroscopic variable is just the binary number $0.0\omega_1,\omega_2$ giving approximately the relative quantity of gas in the left half of the container. Then the cells $\Gamma_{0.000}$ and $\Gamma_{0.011}$ are absolutely negligible in volume compared to the cells $\Gamma_{0.001}$ and $\Gamma_{0.010}$. □

Our proposed replacement of the Boltzmann entropy satisfies the Noncompensation Condition automatically, with $c = 1$.

The second 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.

(2.8) Weak Mixing Condition
There is a function $h(t,d,e)$ with $\lim_{t\to\infty} h(t,d,e) = \infty$ such that if $d < e$, further $\Gamma \not\subset C(e)$ and D is a union of cells with $L(D) \leq 2^{-d}$

$$L(U^t\Gamma \cap D) < 2^{-h(t,d,e)}L(\Gamma).$$

□

This property is mostly very difficult to prove but is at the same time very plausible in typical physical systems. It says that as t grows the transformation U^t, while preserving the volume of the cell Γ, will distribute its content "thinly" over a large area, so that only a small fraction of it can intersect the small "compact" set D (it is compact since it is a union of cells). In other words, after a while, the cells will be "mixed".

The combination of the two conditions gives, of course, the Boltzmann entropy increase property with $f(t,d,e) = h(t,cd,e)$.

Entropy of Gibbs ensembles A distribution over the state space, given with the help of a density $p(\omega)$ with respect to volume, is another possible approach to formalize the notion of a macrostate. The interpretation of an ensemble is somewhat debatable but let us accept, for a moment, that at some time t_0, we have reason to assume that the probability density to find the system in state ω is $p(\omega)$. The **entropy** of the ensemble is defined as

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

In the special case when p is the macrostate-ensemble defined in (2.1) we have $G(p_\Gamma) = \log L(\Gamma)$, *i.e.* the Gibbs entropy is the same as the Boltzmann entropy.

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. We recall Liouville's Theorem which says that the volume L is invariant under the transformation U^t. This implies that $p(U^t\omega) = p(\omega)$ from which one can imply that $G(p^t) = G(p)$, *i.e.* 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, especially in case of a nonisolated system. Such a system does not have a deterministic evolution describable by a transformation U^t since the evolution of the system ω is only the projection of the (possibly deterministic) evolution of a larger system (ω,ξ). The most popular ensemble for such cases is the Gibbs **canonical ensemble** (a generalization of the so-called Boltzmann distribution) defined by a density proportional to $e^{-E(\omega)/(kT)}$, with the energy $E(\omega)$, the Boltzmann constant k and the temperature T. It can be shown invariant if the system ω is part of a large system (ω,ξ) where ξ is a heat reservoir of temperature T.

Maxwell's demon Imagine a container of gas divided into parts A and B by a wall with a tiny door on the wall. Assume that the door is infinitely easy to open or close, and next to it sits a being called the demon, with the powers to measure the speeds of the molecules passing by and to operate the door. These are the only interactions of the demon with the system: otherwise, it belongs to an altogether different "dimension". Assume that it is very unlikely that ever more than one molecule passes by the door. The demon's strategy is that whenever a molecule from part A would hit the door the door is opened for a moment to let this molecule pass into part B. Suppose that parts A and B have the same size and had originally about the same density and temperature of gas. As a result of the activity of the demon, eventually all molecules pass from part A to part B. If we now consider the

system as a whole then it seems that its entropy strongly decreases in time, contradicting the second law. Since we have some trust in the second law we suspect something physically impossible about the existence of such a demon.

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. The problem with these explanations is that they introduce additional physical assumptions which are somehow alien from the general mathematical nature of the second law (increase of disorder). Also, several such explanations are refuted by more refined models (see Bennett's article in the Scientific American).

Modern explanations A convincing modern explanation emerged in a principle announced by Landauer, who relied on some ideas of Szilard (see the nice exposition by Bennett in the Scientific American).

After the demon acts on the information obtained from her observation, she is stuck with this useless piece of data. So, unless she gets rid of it the decrease of the entropy of the container is not the only result of the transformation, since the information content of the demon increases. Since, however, the demon was defined abstractly and not in terms of some macroscopic variables, it is difficult to translate the increase of this information content into an increase of Boltzmann entropy. Landauer's solution was to introduce a new 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.

Zurek [3] saw the need to be able to talk about the increase of a certain quantitiy associated with this system without worrying about how information later will be erased. He restricted attention to a system consisting of two parts: a classical one, known to satisfy the second law, and a "computerized demon", whose state is determined by the content of its memory but which has no other physical characteristics. He created a special macroscopic variable d (without actually distinguishing macroscopic and microscopic), whose value is equal to the demon's whole memory state.

He defined then a quantity associated with such a system that is essentially

$$Z(a,d) = B(a,d) + K(a,d) \tag{2.9}$$

though he actually identified (a,d) with d. The quantity $K(x)$ is the description complexity of the value x: this will be discussed extensively later but its intuitive meaning is "information content". Zurek called the quantity "physical entropy", but we will not use this terminology.

Notice that $B(a,d) = B(a)$ 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. Therefore

$$Z(a,d) \approx B(a) + K(d).$$

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 always be neatly divided 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 implies the upper bound $T(B(a_2) - B(a_1))$ on the first kind of work and one can derive from Landauer's principle the bound $T(K(d_2) - K(d_1))$ for the second kind of work, which gives Zurek's bound.

Formally, the quantity we recommend looks similar to Zurek's but is defined more generally, and will be shown to make reasonable 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, and in a general model.

3. Complexity

From recursive function theory, I will assume that the notion of a **partial recursive function** of natural numbers is known.

These are the partial functions $f(x)$ for whom there is a computer program computing $f(x)$ from x wherever $f(x)$ is defined. An everywhere defined partial recursive function is called a **recursive function**, or a **computable function**. A set of numbers is **recursively enumerable** if it is the range of values of some recursive function, *i.e.* it can indeed be enumerated by a program. If there is some standard correspondence between natural numbers and some other countable set of objects then this also defines the notion of partial recursive function, etc. for these objects. Such a correspondence exists, in particular, between numbers and pairs of numbers, and between numbers and finite binary strings.

Let $F(p, y)$ be a partial recursive function whose first argument is a binary string and its output is a natural number. (The notion of such a function is naturally defined over the space $\mathbf{B} \times Y$ where $\mathbf{B}$ is the set of binary strings.) Such a funtion will be called a **self-delimiting machine** if whenever p is a prefix of q an $F(p, y)$ is defined then $F(q, y) = F(p, y)$. Let

$$K_F(x \mid y)$$

be the length of the shortest string p such that $F(p, y) = x$. This quantity is called the complexity of x with respect to y, on machine F.

(3.1) Complexity Invariance Theorem *There is a machine G on which the complexity function K_G is optimal within an additive constant. That is, for every other machine F there is a constant c_F such that for all x, y we have*

$$K_G(x \mid y) \leq K_F(x \mid y) + c_F.$$

The invariance theorem was first proved by Solomonoff, then by Kolmogorov, for a slightly different version of complexity. The use of self-delimiting machines originates from Levin.

Let us introduce the notation

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

This will mean that for some constant c and for all x we have $f(x) \leq g(x) + c$. The notation $\dot{<}$ will mean the same with a multiplicative constant. The notation $\stackrel{\pm}{=}$ means that both $\stackrel{+}{<}$ and $\stackrel{+}{>}$ hold. With this notation, the theorem's formula can be written as $K_G(x \mid y) \stackrel{+}{<} K_F(x \mid y)$.

The function $K(x \mid y) = K_G(x \mid y)$ is called the **complexity** of the natural number x conditional on the information y. We write $K(x) = K(x \mid 0)$ when Y is the one-element set $\{0\}$. In general, when there are some objects in the condition, we always assume tacitly that the complexity function over the appropriate space Y is considered.

It is easy to see that

$$K(n) \stackrel{+}{<} \log n + 2 \log\log n,$$

and

$$K(f(x) \mid y) \stackrel{+}{<} K(x \mid g(y)) \stackrel{+}{<} K(x). \tag{3.2}$$

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 the set

$$\{(x, r) : r \in \mathbf{Q},\ f(x) < r\}$$

is recursively enumerable. These are the functions for which 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. Thus, we can compute arbitrarily exact upper bounds on $K(x)$ but we will not know how close we are to the limit. It is known that no nontrivial lower bounds can be computed for $K(x)$. But there are strong statistical lower bound results. There is a single property from which these can be derived:

$$\sum_x 2^{-K(x|y)} \leq 1.$$

It is easy to prove this inequality by considering programs to our standard self-delimiting machine that are obtained by coin-tossing.

A nice property of the function $K(x \mid y)$ is the following.

(3.3) 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 $2^{-K(x|y)} \stackrel{+}{>} f(x,y)$.*

This property was proved by Levin and independently, by Chaitin.

4. Randomness

In what follows we will often use spaces Ω, X, Y in which computability is defined (see the Appendix) as a set of parameters. In the Appendix, we fixed an increasing sequence Ω_n of subsets of Ω whose union is the whole space and defined 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 $\Omega \times \mathcal{M}(\Omega) \times Y$ 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. Suppose that a certain casino claims that it draws elements from Ω according to the distribution μ. Then it must accept the following deal: I prove that $f_\mu(\omega)$ is a test of randomness. I offer two dollars for a game, and ask for ω. 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 $f_\mu(\omega)$ as large as possible.

(4.1) Universal Test Theorem *Among all randomness tests, there is a certain one, denoted by $t_\mu(\omega \mid y)$, 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*

$$f_\mu(\omega, y) \stackrel{\cdot}{<} t_\mu(\omega \mid y).$$

This theorem corresponds to the theorems of Martin-Löf and Levin for universal tests. The property that t takes only values 2^n is only for convenience. The theorem can be interpreted as follows. Suppose that I decide to catch the casino owner if she cheats in a certain way. For example, if ω is supposed to be a sequence of coin tosses then I could make my test function greater for those ω's in which the frequency of 1's is at least 60%. The theorem says that $t_\mu(\omega)$ anticipates, in some sense, all these attempts since it is "almost" as large as any other test. The theorem, as well as the notion of a test, are due to Martin-Löf: however, not exactly in this form. This form originates essentially from Levin.

We can write the equation of the theorem then as $f_\mu(\omega, y) \stackrel{\cdot}{<} t_\mu(\omega \mid y)$.

The **algorithmic 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. Thus, $H_\mu(\omega)$ measures the randomness of ω with respect to μ. The higher it is the more random is ω. It is easy to see that H can take arbitrarily large negative values, even the value $-\infty$. For example, if μ is computable and $\mu(\alpha) = 0$ for an open cell then $H_\mu(\omega) = -\infty$ everywhere on α. In other words, an object can be infinitely nonrandom. The measure of such objects has, of course, 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 ∞.

We define $H_\mu(\omega) = H_\mu(\omega \mid 0)$ where Y is chosen as the one-element set $\{0\}$. From now on, we will often state properties for $H_\mu(\omega)$ that are also true more generally for $H_\mu(\omega \mid y)$. 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. The following theorem states some easy properties of $H(\omega)$.

(4.2) Theorem

$$\mu\{\omega : H_\mu(\omega) < m\} \quad < \quad 2^m \ (-\infty < m < \infty), \tag{4.3}$$

$$H_\nu(y \mid \mu) \quad \overset{+}{<} \quad -\log \int 2^{-H_{\mu,\nu}(\omega,y)} \mu(d\omega). \tag{4.4}$$

If f is a computable function then

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

Proof The first inequality is an application of Markov's inequality.

The second one comes from the fact that the right-hand side is a test for ν with parameter μ. The last statement is easy. ∎

To give some idea of how $H_\mu(\omega)$ depends on ω and μ we give an upper bound. For $\omega \in \Omega_n \setminus \Omega_{n-1}$, let $m(\omega) = \mu(\Omega_n)$. Then

$$H_\mu(\omega) \overset{+}{<} \log(m(\omega) + 1) + 2\log(\log(m(\omega) + 1) + 1).$$

5. Randomness in terms of complexity

Additivity of information Note that the Coding Theorem 3.3 can be interpreted as the equality

$$K(x \mid y) \overset{\pm}{=} H_\#(x \mid y)$$

where $\#$ is the counting measure. In other words, $2^{-K(x|y)}$ is the universal randomness test for the counting measure.

The following additivity property is used often. It is stated here in a very general form, which makes it look complex.

(5.1) Addition Theorem *We have*

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

More generally,

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

The earliest form of this theorem is due to Kolmogorov and Levin. In its present form, it is due to Gács and Levin and independently, Chaitin. The general form is new here. Notice that it differs only by having μ, ν everywhere in the appropriate conditions and subscripts. (When μ is in the subscript it does not have to be in the condition.) The proof is in the Appendix. A simple corollary is

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

The function $H_\mu(\omega)$ behaves quite differently for different measures μ. For example, (3.2) implies

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

In contrast, if μ is a probability measure then

$$H_\nu(y) \stackrel{+}{>} H_{\mu,\nu}(\omega, y).$$

This comes from the fact that $2^{-H_\nu(y)}$ is a test for $\mu \times \nu$. It comes from these considerations that the following relation does not follow easily from the addition property (it would be if $H_\mu(\omega) \stackrel{+}{<} H_{\mu\times\#}(\omega,z)$ would hold): For z from a countable set Y, we have

$$H_\mu(\omega) \stackrel{+}{<} H_\mu(\omega \mid z) + K(z). \tag{5.2}$$

Since we will need it for the proof of the Addition Theorem, we spell out this in a lemma in a more general form.

(5.3) Lemma
Relation (5.2) *holds; moreover, for a computable function* $f(y,z)$ *on a countable set* Y, *we have*

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

Proof The function

$$g_\mu(\omega, y) = \sum 2^{-H_\mu(\omega \mid f(y,z)) - K(z)}$$

is lower semicomputable in μ, and $\int g_\mu(\omega,y)\mu(d\omega) \le \sum 2^{-K(z)} \le 1$. Hence $g_\mu(\omega,y) < 2^{-H_\mu(\omega|y)}$. The left-hand side is a sum, hence the inequality holds for each element of the sum: just what we had to prove. ∎

The quantity

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

is called the **mutual information** between the objects x, y. Its general form is

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

(with respect to the measures μ, ν). We can introduce the notation

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

and call this 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 relation between the two kinds of information is clarified by the complexity addition theorem:

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

The right-hand side can be interpreted as the information that the pair $(x, K(x))$ carries about x. More generally,

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

On a countable set Y, Lemma 5.3 is equivalent to

$$I_\mu(y:x) \stackrel{+}{<} K(y). \tag{5.4}$$

The meaning of this inequality is simple: an object cannot carry more information about a string than its own complexity.

Canonical cells For the following, we want to represent each element of $\omega \in \Omega$ of our space simply as a string $\omega_1, \omega_2, \ldots$ of bits. The bits $\omega_1, \omega_2, \ldots$ would contain information about the point ω in decreasing order of importance from a macroscopic point of view. For example, if Ω is the phase space of a container of gas, then 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. It turns out that this is possible for those measures in which the boundary of each cell has measure 0.

The notion of open cell is introduced in the Appendix. Open cells are a fixed sequence of open neighborhoods. A typical example would be, in a 2-dimensional space, to take all open rectangles whose corners have rational coordinates. 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 $\mu(\Omega^0) = 0$. Then it can be shown that the function $\mu(\alpha)$ is computable on the set $\mathcal{M}^0(\Omega)$ considered as a subspace of $\mathcal{M}(\Omega)$. Let us give the correspondence between bits and elements of Ω^0.

Let $\mathbf{B}$ be the set of finite binary strings $s = s_1 \cdots s_n$ $(s = 0, 1)$. For a binary string of length s, let

$$l(s)$$

denote its length. We introduce a special object Λ called the "empty string", a binary string of length 0, and put it also into $\mathbf{B}$.

We will assume that there is a computable function Γ_s assigning to each element of $\mathbf{B}$ an open cell with the following properties.

(a) $\Gamma_\Lambda = \Omega$.

(b) For each $s \in \mathbf{B}$, the interiors of the sets Γ_{s0} and Γ_{s1} are disjoint and the closure of their union contains Γ_s.

(c) For $\omega \in \Omega^0$, the set $\{\Gamma_s : \omega \in \Gamma_s\}$ forms a basis of its neighborhoods.

If s has length n then Γ_s will be called a canonical n-cell, or simply canonical cell, or n-cell. From now on, whenever Γ denotes a subset of Ω, it means a canonical cell. We will also use the notation

$$l(\Gamma_s) = l(s).$$

The three properties say that if we restrict ourselves to the set Ω^0 then the canonical cells behave just like binary subintervals: they divide Ω^0 in half, then each half again in half, etc. The last requirement says that around each point, these canonical cells become arbitrarily small. It is easy to see that if $\Gamma_{s_1}, \Gamma_{s_2}$ are two canonical cells then they either are disjoint or one of them contains the other. If $\Gamma_{s_1} \subset \Gamma_{s_2}$ then s_2 is a prefix of s_1. For each $\omega \in \Omega^0$ and each n, there is a unique canonical n-cell Γ_s containing ω. We will write $s = \omega_1 \ldots \omega_n$. If, for a moment, we write $\Gamma_s^0 = \Gamma_s \cap \Omega^0$ then we have the disjoint union

$$\Gamma_s^0 = \Gamma_{s0}^0 \cup \Gamma_{s1}^0.$$

Thus, for elements of Ω^0, we can talk about the n-th bit ω_n of the description of ω: it is uniquely determined. Also, the 2^n cells of the form Γ_s for $l(s) = n$ form a partition

$$\mathcal{P}_n$$

of Ω^0. Let us denote

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

The characterization of tests The Coding Theorem 3.3 implies that if x runs on a discrete space then $H_{\#}(x) \stackrel{+}{=} K(x)$. It also implies easily the following generalization.

(5.5) Discrete Test Characterization *If X is a discrete space then*

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

We would like to see a similar characterization of tests over arbitrary spaces in terms of complexity. Let us denote

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

for canonical cells Γ. The following theorem is known in complexity theory for randomness tests: the proof is the same for tests over arbitrary measures. It is in a class of similar theorems proved by Martin-Löf, Levin, Schnorr and the author.

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

$$H_\mu(\omega) \stackrel{+}{=} \inf_{\omega \in \Gamma} H_\mu(\Gamma). \tag{5.7}$$

This can also be written as

$$H_\mu(\omega) \stackrel{+}{=} \inf_n K(\omega^n \mid \mu) + \log \mu(\Gamma_{\omega^n}).$$

The constant in $\stackrel{+}{=}$ here depends on μ. Since we assumed μ computable, we can actually delete it from the condition in the complexity. We left it there only since we would really like to prove that the equation (5.7) holds uniformly over $\mathcal{M}(\Omega^0)$ and not only for computable measures. For some annoying technical reasons, we can only prove a slightly weaker statement. This, and its proof, is mainly of technical interest, and are postponed to the Appendix. Note that there is an analogous theorem whose proof is very easy, and which connects complexity and fine-grained algorithmic entropy in a similar way on a discrete set.

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

(5.8) Stability Theorem

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

Proof Let $f(\Gamma) = 2^{-K(l(\Gamma))} \int_\Gamma 2^{-H(\omega)}\mu(d\omega)$. This function is semicomputable and $\sum_\Gamma f(\Gamma) \leq 1$. Therefore $f(\Gamma) \stackrel{\cdot}{<} 2^{-K(\Gamma)}$. Rearranged, this gives:

$$\mu(\Gamma)^{-1} \int_\Gamma 2^{-H(\omega)}\mu(d\omega) \stackrel{\cdot}{<} 2^{-H(\Gamma)+K(l(\Gamma))}.$$

From here, we conclude with Markov's inequality. ∎

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

Algorithmic entropy Consider an isolated physical system with state space Ω whose development is described by a transformation group U^t. We will assume that $U^t\omega$ is computable as a function of t and ω. We will assume the existence of a computable invariant measure L (the "Liouville measure"): it has the property that $L(U^tA) = L(A)$ for all t and all measurable sets A. Under suitable conditions, the existence, uniqueness and computability of L can also be proven. Our physical system is thus determined by the space Ω, the canonical cells Γ_s, the transformation group U^t and the measure L.

In general, the partitions given by the canonical cells have no connection with the dynamics U^t. In particular, they are not "generated" from the first partition into Γ_0 and Γ_1 by U^t. An important exception is the following system.

(6.1) 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 are, of course, cells of the form Γ_{ω^n}. For volume L, we can choose the measure arising from the tossing of an unbiased coin. Then all have the same volume 2^{-n}. Or, we can choose some other stationary measure, like the tossing of a coin with probability $p < 1/2$ of 1. Then the volume of Γ_{ω^n} is $p^k(1-p)^{n-k}$ where k is the frequency of 1's in ω^n. □

We defined $H(\omega) = H_L(\omega)$ as the fine-grained algorithmic entropy of a state ω. Let

$$E_m = \{\omega : H(\omega) < m\}. \tag{6.2}$$

In (4.3), we have shown $L(E_m) < 2^m$ for all m. For finite space volume $L(\Omega)$, this implies that

$$L(E_{\log L(\Omega)-m}) < 2^{-m}L(\Omega).$$

i.e. the proportial size of the set of those points where algorithmic entropy is not close to its maximum is very small.

Algorithmic Boltzmann entropy According to (5.7) we have for all $\omega \in \Omega^0$:

$$H_L(\omega) \stackrel{+}{=} \inf_n K(\omega^n) + \log L(\Gamma_{\omega^n}). \tag{6.3}$$

This formula says that we should include (the program for) more and more bits of ω^n into the macroscopic description as long as the complexity increase buys us an even greater decrease in the Boltzmann entropy

$$B(\omega_n) = \log L(\Gamma_{\omega^n})$$

which is our *a priori* uncertainty about ω. For the systems of interest in physics, and for those precisions in the macroscopic parameters, for which Boltzmann's entropy is generally considered, we expect the additive term $K(\omega^n) \stackrel{+}{<} 2n$ to be negligible compared to the other one since, as pointed out in connection with the Noncompensation Condition 2.6, the total number of macroscopic cells is typically small compared to the volume of the large cells.

Let us define

$$H^n(\omega) = \min_{i\le n} H(\Gamma_{\omega^i}). \tag{6.4}$$

Let $i_n(\omega)$ be the i where the minimum is achieved, and let us whatch its growth as a function of n. For certain ω, it may remain small until some very large n when it suddenly jumps to n. But there is another way of looking at the question. In general, what is given is a canonical cell Γ, and we have $H(\omega) = \inf_{\omega\in\Gamma} H(\Gamma)$. The Stability Theorem 5.8 implies

$$L\{\omega \in \Gamma : H(\omega) < H(\Gamma) - m - K(l(\Gamma))\} < 2^{-m}L(\Gamma).$$

In other words, for most elements ω of the canonical cell Γ, the value $H(\omega)$ is not much less than $H(\Gamma)$. In a sense, a string ω^n describing a canonical cell never has too few bits since it is always a nearly optimal description for most states in it. But it may very well have too many bits in the sense that deleting some of the last ones results in a significant decrease of the entropy of the canonical cell. In other words, the deletion may result in a great decrease of complexity accompanied with only a smaller amount of increase in $\log L$.

Gibbs entropy Let μ be a measure and $p(\omega) \in C_B(\Omega)$ a computable (hence continuous) nonnegative function (this condition could be relaxed somewhat but the result given here is sufficient to see the ideas). Then

$$\nu(f) = \int f(\omega)p(\omega)\mu(d\omega)$$

defines a finite measure. We have

$$1 \geq \int 2^{-H_\mu(\omega)}\mu(d\omega) = \int 2^{-\log p(\omega) - H_\mu(\omega)}\nu(d\omega).$$

Therefore $H_\mu(\omega) + \log p(\omega)$ is a test of randomness for ν. The following inequality follows from Markov's inequality. It holds for all integers m, positive or negative.

$$\nu(\Omega)^{-1}\nu\{\omega : H(\omega) < \log \nu(\Omega) - \log p(\omega) - m\} < 2^{-m}.$$

The following also holds.

(6.5) Theorem

$$H_\nu(\omega) \stackrel{+}{=} H_\mu(\omega) + \log p(\omega)$$

where the constant in $\stackrel{+}{=}$ depends on the definition of the function $p(\omega)$.

Proof For an arbitrary lower semicomputable function $f(\omega, \nu)$ we have

$$\int f(\omega, \nu)\nu(d\omega) = \int f(\omega, \nu)p(\omega)\mu(d\omega).$$

Therefore f is a test for ν if and only if fp is a test for μ. The maximal f will therefore be $\doteq 2^{-H_\mu(\omega)+\log p(\omega)}$. ■

Notice that the theorem does not allow to compute $H_\mu(\omega)$ from $H_\nu(\omega)$ in the places where $p(\omega) = 0$. The uniform bound (9.6) on $H_\nu(\omega)$ implies

$$H_\nu(\omega) \stackrel{+}{<} \log \nu(\Omega) + K(\lfloor \log \nu(\Omega) \rfloor).$$

Combining these, we see that $\log \nu(\Omega) - \log p(\omega)$ is nearly an upper bound to $H_\mu(\Omega)$: more exactly,

$$H_\mu(\omega) \stackrel{+}{<} \log \nu(\Omega) - \log p(\omega) + K(\lfloor \log \nu(\Omega) \rfloor).$$

The last term disappears, of course, if ν is a probability measure. On the other hand, Markov's inequality implies that $H_\mu(\omega)$ is, with overwhelming ν-probability, close to this upper bound. From this, it is easy to see the following relation to Gibbs entropy:

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

This is what we mean by saying that the Gibbs entropy is the average of algorithmic entropy.

7. Entropy increase properties

Fine-grain nondecrease Let us fix a finite time interval $[0, T]$ and investigate the dependence of the function $H(U^t\omega)$ on t. We assumed that the function $U^t\omega$ is a computable one. Since U^t is measure-preserving, the function $2^{-H(U^t\omega)}$ is a parametrized randomness test. From the theorem introducing $H(\omega \mid t)$, we obtain therefore

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

This is, in essence, our entropy nondecrease formula. It can also be regarded as a special case of a more general randomness-conservation property formulated by L.A. Levin in several ways, see its latest form in [1]. To disappoint those who rejoice in a fast increase of entropy, it is also an entropy nonincrease formula. Indeed, the same inequality between $H(\omega)$ and $H(U^t\omega)$, can also be used between $H(U^t\omega)$ and $H(\omega)$. We get therefore

$$-I(t : \omega) \stackrel{+}{<} H(U^t\omega) - H(\omega) \stackrel{+}{<} I(t : U^t\omega). \tag{7.1}$$

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. But the amount of increase is only due to the information that t may have on $U^t\omega$. The latter can be greater but not too much, as we will see. Let us first explore the nondecrease property.

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

$$\lambda\{t : H(U^t\omega) < H(\omega) - K(T) - m\} \stackrel{.}{<} 2^{-m}T.$$

This theorem follows, by an application of Markov's Inequality, from the following lemma.

(7.3) Lemma

$$T^{-1}\int_0^T 2^{I(t:\omega)}dt \stackrel{.}{<} 2^{K(T)}.$$

Proof The function $f(\omega) = T^{-1}\int_0^T 2^{-H(t,\omega)}dt$ is a randomness test and therefore

$$-\log f(\omega) \stackrel{+}{>} H(\omega \mid T) \stackrel{+}{>} H(\omega) - K(T)$$

(using Lemma 5.3). Hence

$$T^{-1}\int_0^T 2^{H(\omega|t)}dt \stackrel{.}{<} 2^{-H(\omega)+K(T)}$$

which by rearrangement gives just what we want. ∎

The theorem says that no matter what interval of time we consider, if the maximum entropy during that interval is H then the proportion of time in which it can be $H - m - K(T)$ will be only about 2^{-m}. The decrease $K(T)$ can be made very small by choosing a simple value of T. If we choose it to be of the form $T = 2^n$ for a positive or negative integer n then $K(T)$ is at most $2\log\log T$ which is generally negligible even if a unit of T is of the duration of some atomic event. We could have chosen an arbitrary interval $[s, s+T]$ since $U^{s+t}\omega = U^t(U^s\omega)$. It is remarkable that the theorem is true on all time scales, not only on the time scale on which the ergodic theorem works (if it is applicable). Thus, if ω has an orbit on which, *e.g.* due to the ergodic theorem, the entropy reaches near its upper bound $\log L(\Omega)$ at some t_0 then the entropy will stay near this upper bound on the largest proportion of each interval around t_0. And the decrease of entropy always happens in an "accelerated" fashion.

The lower the decrease the more "urgent" it is since the relative time spent in the lower regions is decreased by a factor of 2 for every step of decrease, on all time scales. On the interval $[1, T]$ the "steepest" (measured from T) monotonically increasing function of t which has the behavior that is proved here is, in some sense,

$$H(\omega_t) = \log t.$$

Coarse-grain increase 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 $\log L(\Omega)$. Let

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

This set is contained in E_m defined in (6.2) above and therefor just like E_m, has volume at most 2^{-m}. This is exactly the Noncompensation Condition 2.6 with $c = 1$ for the algorithmic Boltzmann entropy; remember that we used such a condition in the argument for the increase of Boltzmann entropy. The rest of the argument about why $H^n(\omega)$ should increase is based on the same mixing condition, but with the cell set $C(d)$ defined with the help of $H^n(\omega^n)$ rather $B(\omega^n)$.

Let us argue that the coarse-grained algorithmic entropy is preferable to Boltzmann entropy. We give a (rather frivolous) example system in which Boltzmann entropy could actually slightly decrease.

(7.4) Example Take a large container filled with ideal gas and a few large balloons. For a while, we constrain the balloons to be fixed. Then we release them. The balloons begin to fly around, gaining energy from collisions with the gas molecules. Eventually, 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. But no matter how we fix their positions and velocities, it is easy to see that the volume of the cell will be essentially determined by the energy of the system consisting of the gas alone. After the balloons are released the energy of the gas becomes smaller by the amount transferred to the balls and hence the log volume of the cell, which is the Boltzmann entropy, becomes correspondingly smaller. □

If this example is ridiculous it becomes less so 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: it is given by specifying the "global" charges and magnetizations of all the tiny areas on the disks and the silicon memories. However, as the capacity of the memory increases (and the size of a site storing an individual bit decreases), effects like the one above become less and less negligible and the gulf between the microscopic and the macroscopic quantities becomes much smaller than in classical statistical physics. There may come a point where it is not reasonable to consider the memory state as part of the macroscopical description. But it is desirable that when we convert from one representation to another one our entropy measure do not suffer a large jump. In terms of our scheme, we are talking about increasing n. 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.

(7.5) Example An example of a chaotic transformation in which Boltzmann entropy does not increase is the "baker's map" of example 6.1. Let us choose the unbiased coin-tossing measure. 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 $H^n(U^t z)$ will, however, increase fast. Let us consider a typical example of an infinite sequence z which has $z^n = 0 \cdots 0$ and whose other bits are random (this has now a precise meaning but let us just use the informal understanding). Then $K(z^n) \leq 2 \log n$, $\log \Gamma_{z^n} = -n$, therefore

$$H^n(z) \leq 2 \log n - n.$$

Now, for $t > n$, the string $(U^t z)^n$ consists of random bits, so essentially, $H^n(U^t z) \geq 0$. Between time 0 and n, this algorithmic coarse-grained entropy increases linearly from $-n$ to 0. The fine-grained algorithmic entropy $H(U^t z)$ will, on the other hand, only increase slowly since in our choice of the precision n, we can follow the increase of t. We have

$$\begin{aligned} H(U^t z) &\stackrel{+}{<} H^{t+n}(U^t z) \leq 2\log n + 2\log t + t - (t+n) \\ &= 2\log n + 2\log t - n. \end{aligned}$$

On the right-hand side of the first equation, the first two terms upper-bound the complexity of n and t, the second term is the length of the nonzero part of the string $(U^t z)^{t+n}$. Therefore $K((U^t z)^{n+t})$ is bounded by the sum of the first three terms, while the last term gives the logvolume of the cell. This inequality shows that $H(U^t z)$ can only grow as slowly as $\log t$. The general inequality (7.1) gives $H(U^t\omega) - H(\omega) \stackrel{+}{<} I(t : (U^t\omega))$. For rational values of t, the inequality (5.4) gives $I(t : (U^t\omega)) \stackrel{+}{<} K(t)$. Hence

$$H(U^t\omega) - H(\omega) \stackrel{+}{<} K(t).$$

This means that for some simple rational values of t, the algorithmic entropy hardly increases at all. For example, if t is an integer of the form 2^n then the increase is at most $2\log\log t$.

We can also use independent biased coin tossings for the measure, where the probability of 1 is $p = 0.3$. If $\omega^n = 0\cdots 0$ then $B(\omega^n) = n\log 0.7$. But for a random ω^n, this value will be $\approx n(0.3\log 0.3 + 0.7\log 0.7)$ which is considerably smaller. Therefore Boltzmann entropy decreases strongly. The algorithmic Boltzmann entropy is, however, negative for the all 0 starting cell and grows to ≈ 0 as expected. This, however, merely shows that we should not expect the Boltzmann entropy to rise above $\log L(\Omega)$ even if it starts higher. If we do not want that then the biased coin-tossing measure is about as bad as the unbiased one. □

The reason that Boltzmann entropy does not increase to $\log L(\Omega)$ in the baker's map with any stationary measure is just that, in case of the uniform distribution, the n-cells have the same volume, and in the more general case still, 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"). Therefore the Noncompensation Condition (the more trivial of the two conditions) is not satisfied. If for this same map we use n-cells of different enough volumes to satisfy this condition then Boltzmann entropy will increase. In fact, when $H^n(\omega)$ is high this can be treated as a concise expression of the fact that in every "simple" partition with extremely different cell sizes, the point ω would end up in a large cell.

The example of the shift transformation suggests that actually, in typical chaotic systems, the parameter n can be made a function of t as long as it grows slower than linearly with t. Thus, if $n(t)$ is a function of t such that $\lim_{t\to\infty} n(t)/t = 0$ 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 speed with which $n(t)$ can grow is connected with the speed of mixing of n-cells under U^t, and seems an interesting measure of the chaoticity of U^t in general.

In conclusion, we suggest that the new quantity $H^n_L(\omega)$ extends the idea of entropy increase to a wider class of chaotic systems than the one in which it has originally worked, and can also serve as a useful tool for formulating conjectures concerning the nature of chaoticity and its extent.

8. Maxwell's demon

Entropy balance This section handles a problem for which the algorithmic entropy $H(\omega)$ is sufficient: it is not necessary to cut off the process of its finite approximations.

Consider two systems, $\mathcal{X}$ and $\mathcal{Y}$, where $\mathcal{Y}$ is considered the environment from which $\mathcal{X}$ is temporarily isolated. Due to isolation, suppose that these systems are nearly independent at time 0, when we want to "do something" to $\xi \in \mathcal{X}$. Now we somehow couple the systems, giving rise to a joint Hamiltonian. 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 $(\xi_t, \eta_t) = U^t(\xi, \eta)$. Notice now that for our measure,

$$H(\xi) + H(\eta) = H(\xi, \eta) + I(\xi, \eta).$$

Let

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

(8.1) Entropy Balance Theorem

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

Proof According to (7.1) applied to the joint system we have $\Delta H(\xi, \eta) \stackrel{+}{>} -I(t : \xi, \eta)$. Note that this formula cannot be applied now to the parts of the system since they do not have their own transformations now. Using this, we have

$$\begin{aligned} H(\xi_t) + H(\eta_t) &= H(\xi_t, \eta_t) + I(\xi_t, \eta_t) \\ &\stackrel{+}{>} H(\xi, \eta) - I(t : \xi, \eta) + I(\xi_t, \eta_t) \\ &= H(\xi) + H(\eta) + I(\xi_t, \eta_t) - I(\xi, \eta) - I(t : \xi, \eta), \end{aligned}$$

which gives the statement by rearrangement. ∎

This theorem says that if the two systems were originally independent (this means $I(\xi, \eta) \approx 0$) then, apart from those rare times which contain information about (ξ, η), a decrease in the entropy of ξ must be accompanied by an increase in the entropy of η.

Maxwell's demon The entropy balance theorem is not new, of course, for Boltzmann entropy. But its new form is useful for our discussion of Maxwell's demon in Section 2. If, for example, η is Maxwell's demon, and ξ is the gas whose entropy she is trying to decrease then our theorem provides an explanation, in a neat inequality, why is it that she will, after a while, have trouble concentrating.

Let us, as in Zurek's setup, count the whole state η of the demon into its 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 above 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 $H(\eta)$, *i.e.* the information content of the demon's memory.

Landauer's thesis Landauer advanced a conjecture (or "thesis") that was considered the modern contribution to the solution of the Maxwell demon paradox. The thesis says that erasing one bit of information from a system requires "heat dissipation" of size $kT \ln 2$, *i.e.* as much as would the decrease of (binary) entropy by 1. The thesis requires interpretation for our model. What does "erasing a bit

of information" mean, and what does "heat dissipation" mean? We will interpret the act of "doing something" to ξ by the fact that we couple it with a formerly independent η which can be considered the environment. Now, if "erasing a bit" means decreasing the entropy of ξ by 1 and "dissipating heat by $kT\ln 2$" means increasing the entropy of η by 1 then our entropy balance theorem confirms the thesis.

For our concept of entropy, and the systems (computer parts) typically considered by Landauer, it is fine to interpret "erasing a bit" as decreasing the entropy by 1. Indeed, consider a system that is used as a computer memory in an optimal way. Suppose that the memory contains the string s. Then probably $H(\xi) = K(\Gamma) + \log L(\Gamma)$ where $\xi \in \Gamma$ and Γ can be described by saying what bits are in the memory, hence $K(\Gamma) = K(s)$. Erasing the information means putting 0's everywhere into the memory. Arguably, we will have $H(\xi_t) = K(\Gamma') + \log L(\Gamma')$ where $\xi_t \in \Gamma'$. Here, Γ' just says that the memory contains all 0's, hence $K(\Gamma') \stackrel{+}{=} 0$. We can require that the memory have approximately the same amount of Boltzmann entropy $\log L(\Gamma)$ no matter what bits it holds since these bits belong to its macroscopic state. In other words, $\log L(\Gamma') = \log L(\Gamma)$. In this interpretation, the erasure indeed means $\Delta H(\xi) = K(s)$.

In order to argue, on the basis of Boltzmann entropy alone, that turning all these bits to 0 results in heat dissipation, we need to consider this operation of turning all bits to 0 as some kind of general operation that does something similar with all possible memory contents, *i.e.* decreases the volume of the whole phase space. This artificiality is avoided here.

It is more controversial to interpret the increase of $H(\eta)$ as heat dissipation. To see this, consider a memory consisting of a row of pendulums swinging transversally. The pendulums all look the same, and the bit 1 means that the pendulum swings while the bit 0 means that it hangs motionless. Let the string stored in ξ this way be s. Let η be an identical row of pendulums, each of which hangs motionless. Now we can use η to erase the memory ξ by moving it next to ξ in the right moment. No question that we decreased $H(\xi)$ by $K(s)$ and increased $H(\eta)$ by $K(s)$. But where is the "heat dissipation"? The change in η is obviously reversible.

There is a way to bring in heat dissipation into the picture but for us, it is not essential here. We will be satisfied with the result that erasing a bit of information in a system results in an equal amount of entropy increase in the environment.

9. Appendix

9.1 Computability

For a set E, we denote the function taking the value 1 in E and 0 outside by $1_E(\omega)$, and call it the **indicator function** of E.

We will need to use some concepts of topology; however, hardly any more knowledge is assumed than the basic definitions and properties of basis, open and closed sets, closure, interior and compactness.

The kind of set in which we can talk about computability will be a topological space X with the properties that allow it to turn into a separable metric space. But we are not interested in the metric itself. For us, a topological space will be defined with the help of a sequence

$$\mathbf{A} = \alpha_1, \alpha_2, \ldots$$

of sets called its **open basis**. The basis must satisfy some simple axioms that we assume known. Basis elements are called **open cells**. A set is defined as open when it is a union of open cells, and closed when it is the complement of an open set. The closure of a set is the smallest closed set containing it, its interior is the largest open set contained in it. Let $\mathrm{Clo}(A)$ denote the closure of a set $A \subset X$, and $\mathrm{Int}(A)$ its interior. For a set E, let us call the interior of its complement the **open complement** of E. We assume that the set of open cells is closed under the operation of finite union and intersection, and open complement.

For sets E_1, E_2, let

$$E_1 < E_2$$

denote the relation $\mathrm{Clo}(E_1) \subset \mathrm{Int}(E_2)$. Our space will be assumed to satisfy the following usual separation property. First, points are closed sets. Second, given open cells $\alpha_1 < \alpha_2$ there is an open cell α_3 with $\alpha_1 < \alpha_3 < \alpha_2$. This property implies that each open set G is the union of basis elements $\alpha < G$. Let us call an open set G **constructive** if the set of basis elements $\alpha < G$ is recursively enumerable. We assume that the set

$$\{(\alpha_1, \alpha_2) : \alpha_1 < \alpha_2\}$$

is recursively enumerable. This implies that all open cells are constructive but is a little stronger. What it says is that every strong inclusion $\alpha_1 < \alpha_2$ between open cells will be "eventually established". (The procedure establishing such facts is the enumeration of all such pairs.) The property is equivalent with assuming that $\{(\alpha_1, \alpha_2) : \mathrm{Clo}(\alpha_1) \cap \mathrm{Clo}(\alpha_2) = \emptyset\}$ is recursively enumerable. This means that whenever two open cells are "well separated" this fact will be "eventually established".

Let us assume that there is a fixed sequence of open sets $\Omega_1 < \Omega_2 < \ldots$ of open cells whose union is Ω. A set is called **bounded** if it is in Ω_n for some n.

A space Ω together with the above enumerations and satisfying the above properties will be called a **computable space**. We will denote it by

$$(\Omega, \mathbf{A}, <)$$

since what is given is Ω, the sequence of cells α_i and the enumeration of the relation $<$.

(9.1) Examples

- The set $\mathbf{N}$ of natural numbers with the so-called discrete topology, in which each number is an open set by itself. The basis is the set of all finite and co-finite sets. In what follows whenever we speak about computability over a countable space the discrete topology is assumed.
- The set $\mathbf{R}$ of real numbers. Open cells are all finite unions of open rational intervals.
- The set $\overline{\mathbf{R}}$ of real numbers with $-\infty$, ∞ added to it. Open cells are the same.
- A finite-dimensional Euclidean space. The open cells are finite unions of open cubes with rational boundaries.
- The set of infinite 0-1 sequences. For a finite binary string x let Γ_x be the set of all infinite continuations of x. Open cells are all finite unions of sets of the form Γ_x.

□

Let X, Y be two spaces with corresponding bases. Intuitively, a function $f : X \to Y$ is computable if there exists an algorithm that from arbitrarily close approximations of x will compute arbitrarily close approximations of $f(x)$. This has the consequence (maybe unexpected for some) that all computable functions are continuous, and that *e.g.* there are no nonconstant computable functions from the set of real numbers to the integers, *i.e.* that the function taking value 0 for $x < 0$ and 1 otherwise is not computable. On reflection, one must accept that such a function is indeed not computable. Let us turn to the formal definition. Let us recall that f is **continuous** if the inverse image $f^{-1}(G)$ of each open set is open, or equivalently, if the inverse image of each open cell is open. Computability is the constructive equivalent of this definition. We could require that the inverse image of each constructively open cell is constructively open, *i.e.* for each open cell α_2 the set of all $\alpha_1 < f^{-1}(\alpha_2)$ is recursively enumerable. The actual definition requires a little more: namely that there is a recursive enumeration of all these sets simultaneously. Thus, f is **computable** if it is continuous and the set

$$\{(\alpha_1, \alpha_2) : \alpha_1 < f^{-1}(\alpha_2)\}$$

is recursively enumerable. An **index** of f is a program for the enumeration of this set. This definition turns out to be the natural one in all simple examples.

It is easy to see that the expected properties of computable functions hold, *e.g.* if f, g are computable then so is the superposition $f(g(\omega))$. It follows that maxima, minima and rational linear combinations of computable real-valued functions are computable. An element ω of Ω is **computable** if the constant function (on any space) with value ω is computable.

Given two computable spaces X, Y, there is a simple canonical definition of the computable space

$$X \times Y.$$

A function f over Ω taking real values and ∞ as values is called **lower semicomputable** if the subset

$$\{(\omega, r) : f(\omega) > r\}$$

of $\Omega \times \mathbf{R}$ is constructively open. An example of a lower semicomputable function is any function that takes the value 1 on a cell and 0 elsewhere. A real function is upper semicomputable if its negative is lower semicomputable. It can be proven that it is computable if and only if it is both upper and lower semicomputable. (The analogy with semicontinuity is more than an analogy.) One can prove that lower semicomputable functions are the ones obtained as the limit of a monotonically increasing computable sequence of computable functions.

As an exercise with these notions, one can show that if f is upper semicomputable and g is computable then $f(g(\omega))$ is also upper semicomputable.

The following theorem can be proved by a routine technique.

(9.2) Theorem *Let $(\Omega, \mathbf{A}, <)$ be a computable space. There is a partial recursive function assigning to each $\alpha_1 < \alpha_2$ an index of a computable function $h(\omega) = h_{\alpha_1,\alpha_2}(\omega)$ such that $h(\omega)$ is 0 if $\omega \in \alpha_1$, 1 if $\omega \notin \alpha_2$ and is in $[0, 1)$ otherwise.*

Let us call functions of the form h_{α_1,α_2} **elementary separator functions**. Let us call a function an **elementary function** if it is obtained from elementary separator functions by finitely many applications of rational linear combination, maximum and minimum. All the elementary functions can be, of course, enumerated.

Let us recall that a closed set F in a topological space is **compact** if it has the following property. If F_i is a (countable or uncountable) set of closed subsets of F with the property that any finitely many of them have a nonempty intersection then the intersection of all of the F_i is nonempty. Let us call the computable space Ω **locally compact** if all the sets $\mathrm{Clo}(\Omega_n)$ are compact.

For a subset Y of Ω, let $C(Y)$ be the set of bounded continuous functions over Ω that are 0 outside Y. Let

$$C_B(\Omega) = \bigcup_n C(\Omega_n).$$

For any function f, let us define, as usual,

$$\|f\|_\infty = \sup_\omega |f(\omega)|.$$

The metric defined by the distance $\|f - g\|_\infty$ turns $C_B(\Omega)$ into a topological space. For each function f and positive number r, we can define the ball $B(f, r) = \{g : \|g - f\|_\infty < r\}$. Let us define open cells in $C_B(\Omega)$ by taking the balls $B(f, r)$ for all elementary functions f and all rational numbers r, and adding all sets obtained by taking finite unions, intersections and open complements. The following theorem can be proved by routine techniques.

(9.3) Theorem *The above definition turns $C_B(\Omega)$ into a computable space. The function $(f, \omega) \to f(\omega)$ on the space $C_B(\Omega) \times \Omega$ is computable.*

Notice that we have now two different definitions of a computable function f: first, f is computable if it is computable as a function from Ω to $\mathbf{R}$; second, it is computable if it is a computable element of $C_B(\Omega)$. It can be shown that the two definitions are equivalent. From now on, whenever we speak of $C(\Omega)$ or $C_B(\Omega)$ we assume that Ω is (locally) compact.

A (nonnegative, σ-finite) measure μ over Ω is a linear functional over $C_B(\Omega)$ that is bounded over each $C(\Omega_n)$ and is nonnegative on all nonnegative functions. We write $\mu(f)$ also as

$$\int f(\omega)\mu(d\omega) = \int f d\mu.$$

A measure is determined by its values over the set of elementary functions. For an open set G, we define, in accordance with the usual procedure, $\mu(G) = \sup_{f \subset G, f \leq 1} \mu(f)$. Also according to this procedure, the set function $\mu(G)$ can be extended to a σ-additive set function over the Borel sets of Ω.

Let $\mathcal{M}(\Omega)$ be the set of all measures over Ω. Let us define $C(f,r)$ as the set of measures μ with the property that $\mu(f) < r$. Let us define open cells in $\mathcal{M}(\Omega)$ by taking the sets $C(f,r)$ for all elementary functions f and all rational numbers r, and adding all sets obtained by taking finite unions, intersections and open complements. The following theorem can be proved by routine techniques.

(9.4) Theorem *The above definition turns $\mathcal{M}(\Omega)$ into a computable space. The function $(\mu, f) \to \mu(f)$ on the space $\mathcal{M} \times C_B(\Omega)$ is computable. The topology of $\mathcal{M}(\Omega)$ is the usual weak topology, so if Ω is compact then so is $\mathcal{M}$.*

Notice that we have now two different definitions of a computable measure μ: first, μ is computable if it is computable as a function from $C_B(\Omega)$ to $\mathbf{R}$; second, it is computable if it is a computable element of $\mathcal{M}(\Omega)$. It can be shown that the two definitions are equivalent. Moreover, it can be shown that this is equivalent to the property that $\mu(f)$, as a real function defined just on the countable set of elementary functions is computable. It can also be shown that function $(\mu, \alpha) \to \mu(\alpha)$ is lower semicomputable over $\mathcal{M} \times \mathbf{A}$.

A measure is called **lower semicomputable** when it is lower semicomputable as a function over $C_B(\Omega)$. It can be shown that this is equivalent to the property that it is lower semicomputable as a function over the countable set of elementary functions.

9.2 Some proofs

The next results are preparation for the Addition Theorem. The theorem below is a generalization of the universal test theorem.

(9.5) Theorem

Let $f_\mu(x,y)$ be an arbitrary lower semicomputable function with $F_\mu(y) = -\log \int 2^{f_\mu(x,y)}\mu(dx)$. Then for all y with finite $F_\mu(y)$ we have

$$-\log f_\mu(x,y) \stackrel{+}{>} F_\mu(y) + H_\mu(x \mid y, \lceil F(y) \rceil).$$

Proof Let us construct a lower semicomputable function $g_\mu(x,y,m)$ for integers m with the property that $\int g_\mu(x,y,m)\mu(dx) \leq 2^{-m}$ and for all y with $F_\mu(y) \geq m$ we have $g_\mu(x,y,m) = f_\mu(x,y)$. Such a g can be constructed by simply watching the approximation of f grow and cutting it off as soon as it would give $F_\mu(y) < m$. Now $h_\mu(x,y,m) = 2^m g_\mu(x,y,m)$ is a test of x and hence it is $\dot{<} 2^{-H_\mu(x|y,m)}$. Substituting $\lceil F_\mu(y) \rceil$ for m gives the result. ∎

If we take $f_\mu(x,y) = 1$ in the above lemma, we have

$$H_\mu(x \mid \lceil \mu(X) \rceil) \stackrel{+}{<} \log \mu(X).$$

This and Lemma 5.3 gives an explicit upper bound

$$H_\mu(x) \stackrel{+}{<} \log \mu(X) + K(\lfloor \log \mu(X) \rfloor) \tag{9.6}$$

on $H_\mu(x)$ for finite measures.

(9.7) Lemma *For $i < j$ we have*

$$i + H_\mu(x \mid i) \leq j + H_\mu(x \mid j).$$

Proof From the above lemma, with $f(i, n) = i + n$ we have

$$H_\mu(x \mid i) - H_\mu(x \mid j) \stackrel{+}{<} K(j - i) \stackrel{+}{<} j - i.$$

Rearrangement gives the result. ∎

Proof of Theorem 5.1 To prove the inequality $\stackrel{+}{<}$, let us define the function

$$G_{\mu,\nu}(x, y, m) = \min_{i \geq m} i + H_\mu(x \mid y, i, \nu).$$

This function is upper semicomputable and is a decreasing function of n. Therefore $G_{\mu,\nu}(x, y) = G_{\mu,\nu}(x, y, H_\nu(y \mid \mu))$ is also upper semicomputable since it is obtained by substituting an upper semicomputable function into m.

According to the above lemma,

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

Now, we have

$$\int 2^{-G_{\mu,\nu}(x,y)} \mu(dx) \leq 2^{-H_\nu(y|\mu)}.$$

Therefore $\int 2^{-G} d\mu d\nu \leq 1$. This implies that $H_{\mu,\nu}(x, y) \stackrel{+}{<} G_{\mu,\nu}(x, y)$ which was to be proved.

To prove the inequality $\stackrel{+}{>}$, let $f_\mu(x, y, \nu) = 2^{-H_{\mu,\nu}(x,y)}$. Let $F_\mu(y, \nu) = -\log \int f_\mu(x, y, \nu) \mu(dx)$. According to Theorem 9.5,

$$-\log f_\mu(x, y, \nu) \stackrel{+}{>} F + H_\mu(x \mid y, \lceil F \rceil, \nu).$$

Inequality (4.4) implies $F_\mu(y, \nu) \stackrel{+}{>} H_\nu(y \mid \mu)$. This, with Lemma 9.7 implies the same inequality, with $H_\nu(y \mid \mu)$ in place of $F_\mu(y, \nu)$ which was to be proved. ∎

Here is the proof of the Test Characterization Theorem—in a slightly generalized form.

(9.8) Theorem *We have*

$$H_\mu(\omega) \stackrel{+}{<} \inf_{\omega \in \Gamma} H_\mu(\Gamma)$$

uniformly in ω and the measure $\mu \in \mathcal{M}^0(\Omega)$.

Let $\mathcal{G}$ be a decreasing sequence of cells containing the measure $\mu \in \mathcal{M}^0(\Omega)$, and forming a neighborhood basis of μ. Let

$$H'_\mathcal{G}(\Gamma) = K(\Gamma \mid \mathcal{G}) + \log \mu(\Gamma).$$

Then

$$H_\mu(\omega) \stackrel{+}{>} \inf_{\omega \in \Gamma} H'_\mathcal{G}(\Gamma).$$

Proof To prove $\stackrel{+}{<}$ let $f_\mu(\Gamma) = 2^{-H_\mu(\Gamma)}$. Then it is enough to show that $\sup_{\Gamma:\omega\in\Gamma} f_\mu(\Gamma)$ is a test. It is lower semicomputable over $\mathcal{M}^0(\Omega)$, so only the integral condition needs proof. We will show that even the sum, not only the supremum, satisfies this condition. Let $1_\Gamma(\omega)$ be the indicator function of the cell Γ. Then

$$\int \sum_{\Gamma:\omega\in\Gamma} f_\mu(\omega)\mu(d\omega) = \sum_\Gamma 2^{-K(\Gamma|\mu)} \int \frac{1_\Gamma(\omega)}{\mu(\Gamma)}\mu(d\omega) = \sum_\Gamma 2^{-K(\Gamma|\mu)} \leq 1$$

which was to be proved.

To prove $\stackrel{+}{>}$ let $g_\mu(\omega) = 2^{-H_\mu(\omega)}$. By the definition of the semicomputability of $g_\mu(\omega)$, there is a recursively enumerable set T of triples (Γ, m, γ) where γ is a cell in $\mathcal{M}$, Γ is a canonical cell in Ω and m is an integer such that for an arbitrary μ,

$$g_\mu(\omega) = \sup\{\, 2^m : \omega \in \Gamma,\ \mu \in \gamma,\ (\Gamma, m, \gamma) \in T \,\}.$$

Now, with the help of the sequence $\mathcal{G}$, we can enumerate the set T_m of those Γ for which there is a triple (Γ, m, γ) in T with $\delta \in \Gamma$. Let us create a subsequence T'_m of T_m whose elements are disjoint but still, their union is the same. We take the elements of T_m one-by-one and "process" them, putting some elements of T_m into T'_m. Suppose that the first n elements Γ'_i of T' are defined. Take the first unprocessed element Γ of T_m. Break up the set $\Gamma \setminus \bigcup_i \Gamma'_i$ into a finite number of canonical cells (this is always possible) and put them into T'_m. This finishes the processing of Γ. We have

$$g_\mu(\omega) = \sup\{\, 2^m : \exists \Gamma \in T'_m\ \omega \in \Gamma \,\}.$$

Due to the disjointness of the set T'_m, each 2^m occurs at most once in this supremum, and therefore it increases at most by a factor of 2 if we replace it with a sum. Let $g_\mathcal{G}(\Gamma) = \sum_{\Gamma\in T'_m} 2^m$. It follows that

$$\begin{aligned} 0.5 g_\mu(\omega) &\geq \sum\{\, 2^m : \exists \Gamma \in T'_m\ \omega \in \Gamma \,\} \\ &= \sum_{\omega\in\Gamma} g_\mathcal{G}(\Gamma) = \sum_\Gamma g_\mathcal{G}(\Gamma) 1_\Gamma(\omega). \end{aligned}$$

Hence

$$1 \geq \int g_\mu(\omega) \geq 0.5 \sum_\Gamma g_\mathcal{G}(\Gamma)\mu(\Gamma).$$

Since the function $g_\mathcal{G}(\Gamma)\mu(\Gamma)$ is lower semicomputable we have $g_\mathcal{G}(\Gamma)\mu(\Gamma) \stackrel{+}{<} 2^{-K(\Gamma|\mathcal{G})}$. In other words,

$$-\log g_\mathcal{G}(\Gamma) \stackrel{+}{>} K(\Gamma \mid \mathcal{G}) + \mu(\Gamma).$$

We are done since the definition of $g_\mathcal{G}(\Gamma)$ implies $H_\mu(\omega) \geq \inf_{\omega\in\Gamma} -\log g_\mathcal{G}(\Gamma)$. ∎

REFERENCES

1. L. A. Levin. Randomness conservation inequalities: Information and independence in mathematical theories. *Information and Control*, 61(1):15–37, 1984.
2. M. Li and P. Vitanyi. Two decades of applied Kolmogorov complexity. *Russian Math. Surveys*, 43:129–164, 1988.
3. W.H. Zurek. Algorithmic randomness and physical entropy. *Physical Review A*, 40(8):4731–4751, October 1989.