Entropy

Tomasz Downarowicz
Institute of Mathematics and Computer Science
Wroclaw University of Technology

Extracted from
Entropy in Dynamical Systems
Cambridge University Press
New Mathematical Monographs 18
Cambridge 2011

1 Introduction

Nowadays, nearly every kind of information is turned into the digital form. Digital cameras turn every image into a computer file. The same happens to musical recordings or movies. Even our mathematical work is registered mainly as computer files. Analog information is nearly extinct.

While studying dynamical systems (in any understanding of this term) sooner or later one is forced to face the following question: How can the information about the evolution of a given dynamical system be most precisely turned into a digital form? Researchers specializing in dynamical systems are responsible for providing the theoretical background for such a transition.

So suppose that we do observe a dynamical system, and that we indeed turn our observation into the digital form. That means, from time to time, we produce a digital "report", a computer file, containing all our observations since the last report. Assume for simplicity that such reports are produced at equal time distances, say, at integer times. Of course, due to bounded capacity of our recording devices and limited time between the reports, our files have bounded size (in bits). Because the variety of digital files of bounded size is finite, we can say that at every integer moment of time we produce just one *symbol*, where the collection of all possible symbols, i.e. the *alphabet*, is finite.

An illustrative example is filming a scene using a digital camera. Every unit of time, the camera registers an image, which is in fact a bitmap of some fixed size (camera resolution). The camera turns the live scene into a sequence of bitmaps. We can treat every such bitmap as a single symbol in the alphabet of the "language" of the camera.

The sequence of symbols is produced as long as the observation is being conducted. We have no reason to restrict the global observation time, and we can agree that it goes on forever and we usually also assume that the observation has been conducted since ever in the past as well. In this manner, the history of our recording takes on the form

of a bilateral sequence of symbols from some finite alphabet, say Λ. Advancing in time by a unit corresponds, on one hand, to the unit-time evolution of the dynamical system, on the other, to shifting the enumeration of our sequence of symbols. This way we have come to the conclusion that the digital form of the observation is nothing else but an element of the space $\Lambda^{\mathbb{Z}}$ of all bilateral sequences of symbols, and the action on this space is the familiar shift transformation σ advancing the enumeration: for $x = (x(n))_{n\in\mathbb{Z}}$,

$$\sigma(x)(n) = x(n+1).$$

Now, in most situations, such a "digitalization" of the dynamical system will be *lossy*, i.e., it will capture only some aspects of the observed dynamical system, and much of the information will be lost. For example, the digital camera will not be able to register objects hidden behind other objects, moreover, it will not see objects smaller than one pixel or their movements until they pass from one pixel to another. However, it may happen that, after a while, each object will eventually become detectable, and that we will be able to reconstruct its trajectory from the recorded information.

Of course, lossy digitalization is always possible and hence presents a lesser kind of challenge. We will be much more interested in *lossless* digitalization. When and how is it possible to digitalize a dynamical system so that no information is lost, i.e., in such a way that after viewing the entire sequence of symbols we can completely reconstruct the evolution of the system?

In this note the task of encoding a system with possibly smallest alphabet is refereed to as "data compression". The reader will find answers to the above question at two major levels: measure-theoretic, and topological. In the first case the digitalization is governed by the *Kolmogorov-Sinai entropy* of the dynamical system, the first major subject of this note. In the topological setup the situation is more complicated. Topological entropy, our second most important notion, turns out to be insufficient to decide about digitalization that respects the topological structure. Thus another parameter, called *symbolic extension entropy*, emerges as the third main object discussed here.

2 A few words about the history of entropy

Below we review very briefly the development of the notion of entropy focusing on the achievements crucial for the genesis of the basic concepts of entropy discussed in this note. For a more complete survey we refer to the expository article [Katok, 2007].

The term "entropy" was coined by a German physicist Rudolf Clausius from Greek "en-" = in + "trope" = a turning [Clausius, 1850]. The word reveals analogy to "energy" and was designed to mean the form of energy that any energy eventually and inevitably "turns into" – a useless heat. The idea was inspired by an earlier formulation by French physicist and mathematician Nicolas Léonard Sadi Carnot [Carnot, 1824] of what is now known as the *second Law of Thermodynamics*: entropy represents the energy no longer capable to perform work, and in any isolated system it can only grow.

Austrian physicist Ludwig Boltzmann put entropy into the probabilistic setup of statistical mechanics [Boltzmann, 1877]. Entropy has also been generalized around 1932 to quantum mechanics by John von Neumann [see von Neumann, 1968].

Later this led to the invention of entropy as a term in probability and information theory by an American electronic engineer and mathematician Claude Elwood Shannon, now recognized as the father of the information theory. Many of the notions have not changed much since they first occurred in Shannon's seminal paper *A Mathematical Theory of Communication* [Shannon, 1948]. Dynamical entropy in dynamical systems was created by one of the most influential mathematicians of modern times, Andrei Nikolaevich Kolmogorov, [Kolmogorov, 1958, 1959] and improved by his student Yakov Grigorevich Sinai who practically brought it to the contemporary form [Sinai, 1959].

The most important theorem about the dynamical entropy, the Shannon-McMillan-Breiman Theorem gives this notion a very deep meaning. The theorem was conceived by Shannon [Shannon, 1948], and proved in increasing strength by Brockway McMillan [McMillan, 1953] (L^1-convergence), Leo Breiman [Breiman, 1957] (almost everywhere convergence), and Kai Lai Chung [Chung, 1961] (for countable partitions). In 1970 Wolfgang Krieger obtained one of the most important, from the point of view of data compression, results about the existence (and cardinality) of finite generators for automorphisms with finite entropy [Krieger, 1970].

In 1970 Donald Ornstein proved that Kolmogorov-Sinai entropy was a *a complete invariant* in the class of *Bernoulli systems*, a fact considered one of the most important features of entropy (alternatively of Bernoulli systems) [Ornstein, 1970].

In 1965, Roy L. Adler, Alan G. Konheim and M. Harry McAndrew carried the concept of dynamical entropy over to topological dynamics [Adler et al., 1965] and in 1970 Efim I. Dinaburg and (independently) in 1971 Rufus Bowen redefined it in the language of metric spaces [Dinaburg, 1970; Bowen, 1971]. With regard to entropy in topological systems, probably the most important theorem is the Variational Principle proved by L. Wayne Goodwyn (the "easy" direction) and Timothy Goodman (the "hard" direction), which connects the notions of topological and Kolmogorov-Sinai entropy [Goodwyn, 1971; Goodman, 1971] (earlier Dinaburg proved both directions for finite dimensional spaces [Dinaburg, 1970]).

The theory of symbolic extensions of topological systems was initiated by Mike Boyle around 1990 [Boyle, 1991]. The outcome of this early work is published in [Boyle et al., 2002]. The author of this note contributed to establishing that invariant measures and their entropies play a crucial role in computing the so-called symbolic extension entropy [Downarowicz, 2001; Boyle and Downarowicz, 2004; Downarowicz, 2005].

Dynamical entropy generalizing the Kolmogorov-Sinai dynamical entropy to noncommutative dynamics occurred as an adaptation of von Neumann's quantum entropy in a work of Robert Alicki, Johan Andries, Mark Fannes and Pim Tuyls [Alicki et al., 1996] and then applied to doubly stochastic operators by Igor I. Makarov [Makarov, 2000]. The axiomatic approach to entropy of doubly stochastic operators, as well as topological entropy of Markov operators have been developed in [Downarowicz and Frej, 2005].

The term "entropy" is used in many other branches of science, sometimes distant from physics or mathematics (such as sociology), where it no longer maintains its rigorous quantitative character. Usually, it roughly means "disorder", "chaos", "decay

of diversity" or "tendency toward uniform distribution of kinds".

3 Multiple meanings of entropy

In the following paragraphs we review some of the various meanings of the word "entropy" and try to explain how they are connected. We devote a few pages to explain how dynamical entropy corresponds to data compression rate.

3.1 Entropy in physics

In classical physics, a physical system is a collection of objects (bodies) whose *state* is parametrized by several characteristics such as the distribution of density, pressure, temperature, velocity, chemical potential, etc. The change of entropy of a physical system, as it passes from one state to another, equals

$$\Delta S = \int \frac{dQ}{T},$$

where dQ denotes an element of heat being absorbed (or emitted; then it has the negative sign) by a body, T is the absolute temperature of that body at that moment, and the integration is over all elements of heat active in the passage. The above formula allows to compare entropies of different states of a system, or to compute entropy of each state up to an additive constant (this is satisfactory in most cases). Notice that when an element dQ of heat is transmitted from a warmer body of temperature T_1 to a cooler one of temperature T_2 then the entropy of the first body changes by $-\frac{dQ}{T_1}$, while that of the other rises by $\frac{dQ}{T_2}$. Since $T_2 < T_1$, the absolute value of the latter fraction is larger and jointly the entropy of the two-body system increases (while the global energy remains the same).

A system is *isolated* if it does not exchange energy or matter (or even information) with its surroundings. In virtue of the first Law of Thermodynamics, the conservation of energy principle, an isolated system can pass only between states of the same global energy. The second Law of Thermodynamics introduces irreversibility of the evolution: an isolated system cannot pass from a state of higher entropy to a state of lower entropy. Equivalently, it says that it is impossible to perform a process whose only final effect is the transmission of heat from a cooler medium to a warmer one. Any such transmission must involve an outside work, the elements participating in the work will also change their states and the overall entropy will rise.

The first and second laws of thermodynamics together imply that an isolated system will tend to the state of maximal entropy among all states of the same energy. The energy distributed in this state is uncapable of any further activity. The state of maximal entropy is often called the "thermodynamical death" of the system.

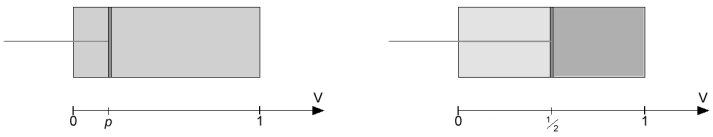
Ludwig Boltzmann gave another, probabilistic meaning to entropy. For each state A the (negative) difference between the entropy of A and the entropy of the "maximal state" B is nearly proportional to the logarithm of the probability that the system

spontaneously assumes state A,

$$S(A) - S_{max} \approx k \log_2(\mathsf{Prob}(A)).$$

The proportionality factor k is known as the Boltzmann constant. In this approach the probability of the maximal state is almost equal to 1, while the probabilities of states of lower entropy are exponentially small. This provides another interpretation of the second Law of Thermodynamics: the system spontaneously assumes the state of maximal entropy simply because all other states are extremely unlikely.

Example Consider a physical system consisting of an ideal gas enclosed in a cylindrical container of volume 1. The state B of maximal entropy is clearly the one where both pressure



State B *State A*

and temperature are constant (P_0 and T_0, respectively) throughout the container. Any other state can be achieved only with help from outside. Suppose one places a piston at a position $p < \frac{1}{2}$ in the cylinder (the left figure; thermodynamically, this is still the state B) and then slowly moves the piston to the center of the cylinder (position $\frac{1}{2}$), allowing the heat to flow between the cylinder and its environment, where the temperature is T_0, which stabilizes the temperature at T_0 all the time. Let A be the final state (the right figure). Note that both states A and B have the same energy level inside the system.

To compute the jump of entropy one needs to examine what exactly happens during the passage. The force acting on the piston at position x is proportional to the difference between the pressures:

$$F = c\left(P_0 \frac{1-p}{1-x} - P_0 \frac{p}{x}\right).$$

Thus, the work done while moving the piston equals:

$$W = \int_p^{\frac{1}{2}} F\,dx = cP_0\big((1-p)\ln(1-p) + p\ln p + \ln 2\big).$$

The function

$$p \;\mapsto\; (1-p)\ln(1-p) + p\ln p$$

is negative and assumes its minimal value $-\ln 2$ at $p = \frac{1}{2}$.

Thus the above work W is positive and represents the amount of energy delivered to the system from outside. During the process the compressed gas on the right emits heat, while the depressed gas on the left absorbs heat. By conservation of energy (applied to the enhanced system including the outside world), the gas altogether will emit heat to the environment equivalent to the delivered work $\Delta Q = -W$. Since the temperature is constant

all the time, the change in entropy between states B and A of the gas is simply $\frac{1}{T_0}$ times ΔQ, i.e.,

$$\Delta S = \frac{1}{T_0} \cdot cP_0\big(-(1-p)\ln(1-p) - p\ln p - \ln 2\big).$$

Clearly ΔS is negative. This confirms, what was already expected, that the outside intervention has lowered the entropy of the gas.

This example illustrates very clearly Boltzmann's interpretation of entropy. Assume that there are N particles of the gas independently wandering inside the container. For each particle the probability of falling in the left or right half of the container is $\frac{1}{2}$. The state A of the gas occurs spontaneously if pN and $(1-p)N$ particles fall in the left and right halves of the container, respectively. By elementary combinatorics formulae, the probability of such an event equals

$$\mathsf{Prob}(A) = \frac{N!}{(pN)!((1-p)N)!}2^{-N}.$$

By Stirling's formula ($\ln n! \approx n\ln n - n$ for large n), the logarithm of $\mathsf{Prob}(A)$ equals approximately

$$N\big(-(1-p)\ln(1-p) - p\ln p - \ln 2\big),$$

which is indeed proportional to the drop ΔS of entropy between the states B and A (see above).

3.2 Shannon entropy

In probability theory, a *probability vector* $\mathbf{p}$ is a sequence of finitely many nonnegative numbers $\{p_1, p_2, \ldots, p_n\}$ whose sum equals 1. The Shannon entropy of a probability vector $\mathbf{p}$ is defined as

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

(where $0\log_2 0 = 0$). Probability vectors occur naturally in connection with finite partitions of a probability space. Consider an abstract space Ω equipped with a probability measure μ assigning probabilities to measurable subsets of Ω. A finite partition $\mathcal{P}$ of Ω is a collection of pairwise disjoint measurable sets $\{A_1, A_2, \ldots, A_n\}$ whose union is Ω. Then the probabilities $p_i = \mu(A_i)$ form a probability vector $\mathbf{p}_{\mathcal{P}}$. One associates the entropy of this vector with the (ordered) partition $\mathcal{P}$:

$$H_\mu(\mathcal{P}) = H(\mathbf{p}_{\mathcal{P}}).$$

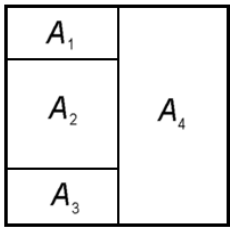
In this setup entropy can be linked with *information*. Given a measurable set A, the information $I(A)$ associated with A is defined as $-\log_2(\mu(A))$. The *information function* $I_{\mathcal{P}}$ associated with a partition $\mathcal{P} = \{A_1, A_2, \ldots, A_n\}$ is defined on the space Ω and it assumes the constant value $I(A_i)$ at all points ω belonging to the set A_i. Formally,

$$I_{\mathcal{P}}(\omega) = \sum_{i=1}^{n} -\log_2(\mu(A_i))\mathbb{I}_{A_i}(\omega),$$

where $\mathbb{I}_{A_i}$ is the characteristic function of A_i. One easily verifies that the expected value of this function with respect to μ coincides with the entropy $H_\mu(\mathcal{P})$.

We shall now give an interpretation of the information function and entropy, the key notions in entropy theory. The partition $\mathcal{P}$ of the space Ω associates with each element $\omega \in \Omega$ the "information" that gives answer to the question "in which A_i are you?". That is the best knowledge we can acquire about the points, based solely on the partition. One bit of information is equivalent to acquiring an answer to a binary question, i.e., a question of a choice between two possibilities. Unless the partition has two elements, the question "in which A_i are you?" is not binary. But it can be replaced by a series of binary questions and one is free to use any arrangement (tree) of such questions. In such an arrangement, the number of questions $N(\omega)$ (i.e., the amount of information in bits) needed to determine the location of the point ω within the partition may vary from point to point (see the Example below). The smaller the expected value of $N(\omega)$ the better the arrangement. It turns out that the best arrangement satisfies $I_{\mathcal{P}}(\omega) \leq N(\omega) \leq I_{\mathcal{P}}(\omega) + 1$ for μ-almost every ω. The difference between $I_{\mathcal{P}}(\omega)$ and $N(\omega)$ follows from the crudeness of the measurement of information by counting binary questions; the outcome is always a positive integer. The real number $I_{\mathcal{P}}(\omega)$ can be interpreted as the precise value. Entropy is the expected amount of information needed to locate a point in the partition.

Example Consider the unit square representing the space Ω, where the probability is the Lebesgue measure (i.e., the surface area), and the partition $\mathcal{P}$ of Ω into four sets A_i of probabilities $\frac{1}{8}, \frac{1}{4}, \frac{1}{8}, \frac{1}{2}$, respectively, as shown in the figure. The information function equals



$-\log_2(\frac{1}{8}) = 3$ on A_1 and A_3, $-\log_2(\frac{1}{4}) = 2$ on A_2 and $-\log_2(\frac{1}{2}) = 1$ on A_4. The entropy of $\mathcal{P}$ equals

$$H(\mathcal{P}) = \frac{1}{8} \cdot 3 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{2} \cdot 1 = \frac{7}{4}.$$

The arrangement of questions that optimizes the expected value of the number of questions asked is the following:

1. *Are you in the left half?*

The answer "no", locates ω in A_4 using one bit. Otherwise the next question is:

2. *Are you in the central square of the left half?*

The "yes" answer locates ω in A_2 using two bits. If not, the last question is:

3. *Are you in the top half of the whole square?*

Now "yes" or "no" locate ω in A_1 or A_3, respectively. This takes three bits.

$$\text{Question 1} \begin{cases} yes \rightarrow \text{Question 2} \begin{cases} yes \rightarrow A_2 \text{ (2 bits)} \\ no \rightarrow \text{Question 3} \begin{cases} yes \rightarrow A_1 \text{ (3 bits)} \\ no \rightarrow A_3 \text{ (3 bits)} \end{cases} \end{cases} \\ no \rightarrow A_4 \text{ (1 bit)} \end{cases}$$

In this example the number of questions equals exactly the information function at every point and the expected number of question equals the entropy $\frac{7}{4}$. There does not exist a better arrangement of questions. Of course, such an accuracy is possible only when the probabilities of the sets A_i are integer powers of 2; in general the information is not integer valued.

Another interpretation of Shannon entropy deals with the notion of *uncertainty*. Let X be a random variable defined on the probability space Ω and assuming values in a finite set $\{x_1, x_2, \ldots, x_n\}$. The variable X generates a partition $\mathcal{P}$ of Ω into the sets $A_i = \{\omega \in \Omega : \mathrm{X}(\omega) = x_i\}$ (called the preimage partition). The probabilities $p_i = \mu(A_i) = \mathsf{Prob}\{\mathrm{X} = x_i\}$ form a probability vector called the *distribution* of X. Suppose an experimenter knows the distribution of X and tries to guess the outcome of X before performing the experiment, i.e., before picking some $\omega \in \Omega$ and reading the value $\mathrm{X}(\omega)$. His/her *uncertainty* about the outcome is the expected value of the information he/she is *missing* to be certain. As explained above that is exactly the entropy $H_\mu(\mathcal{P})$.

3.3 Connection between Shannon and Boltzmann entropy

Both notions in the title of this subsection refer to probability and there is an evident similarity in the formulae. But the analogy fails to be obvious. In the literature many different attempts toward understanding the relation can be found. In simple words, the interpretation relies on the distinction between the macroscopic state considered in classical thermodynamics and the microscopic states of statistical mechanics. A thermodynamical state A (a distribution of pressure, temperature, etc.) can be realized in many different ways ω at the microscopic level, where one distinguishes all individual particles, their positions and velocity vectors. As explained above, the difference of Boltzmann entropies $S(A) - S_{max}$ is proportional to $\log_2(\mathsf{Prob}(A))$, the logarithm of the probability of the macroscopic state A in the probability space Ω of all microscopic states ω. This leads to the equation

$$S_{max} - S(A) = k \cdot I(A), \tag{3.1}$$

where $I(A)$ is the probabilistic information associated with the set $A \subset \Omega$. So, Boltzmann entropy seems to be closer to Shannon information rather than Shannon entropy. This interpretation causes additional confusion, because $S(A)$ appears in this equation with negative sign, which reverses the direction of monotonicity; the more information is "associated" with a macrostate A the smaller its Boltzmann entropy. This is usually explained by interpreting what it means to "associate" information with a state. Namely, the information about the state of the system is an information available to an outside observer. Thus it is reasonable to assume that this information acually "escapes" from the system, and hence it should receive the negative sign. Indeed, it is the knowledge about the system possessed by an outside observer that increases the usefulness of the energy contained in that system to do physical work, i.e., it decreases the system's entropy.

The interpretation goes further: each microstate in a system appearing to the observer as being in macrostate A still "hides" the information about its "identity". Let

$I_h(A)$ denote the joint information still hiding in the system if its state is identified as A. This entropy is clearly maximal at the maximal state, and then it equals S_{max}/k. In a state A it is diminished by $I(A)$, the information already "stolen" by the observer. So, one has

$$I_h(A) = \frac{S_{max}}{k} - I(A).$$

This, together with (3.1), yields

$$S(A) = k \cdot I_h(A),$$

which provides a new interpretation to the Boltzmann entropy: it is proportional to the information still "hiding" in the system provided the macrostate A has been detected.

So far the entropy was determined up to an additive constant. We can compute the *change* of entropy when the system passes from one state to another. It is very hard to determine the proper additive constant of the Boltzmann entropy, because the entropy of the maximal state depends on the level of precision of identifying the microstates. Without quantum approach, the space Ω is infinite and such is the maximal entropy. However, if the space of states is assumed finite, the absolute entropy obtains a new interpretation, already in terms of the Shannon entropy (not just of the information function). Namely, in such case, the highest possible Shannon entropy $H_\mu(\mathcal{P})$ is achieved when $\mathcal{P} = \xi$ is the partition of the space Ω into single states ω and μ is the uniform measure on Ω, i.e., such that each state has probability $(\#\Omega)^{-1}$. It is thus natural to set

$$S_{max} = k \cdot H_\mu(\xi) = k \log_2 \#\Omega.$$

The detection that the system is in state A is equivalent to acquiring the information $I(A) = -\log_2(\mu(A)) = -\log_2\big(\frac{\#A}{\#\Omega}\big)$. By the equation (3.1) we get

$$S(A) = k(-\log_2 \#\Omega + \log_2\big(\tfrac{\#A}{\#\Omega}\big)) = k \log_2 \#A.$$

The latter equals (k times) the Shannon entropy of μ_A, the normalized uniform measure restricted to A. In this manner we have compared the Boltzmann entropy directly with the Shannon entropy and we have gotten rid of the unknown additive constant.

The whole above interpretation is a subject of many discussions, as it makes entropy of a system depend on the seemingly nonphysical notion of "knowledge" of a mysterious observer. The classical *Maxwell's paradox* [Maxwell, 1871] is based on the assumption that it is possible to acquire information about the parameters of individual particles without any expense of heat or work. To avoid such paradoxes, one must agree that every bit of acquired information has its physical entropy equivalent (equal to the Boltzmann constant k), by which the entropy of the memory of the observer increases. In consequence, erasing one bit of information from a memory (say, of a computer) at temperature T, results in the emission of heat in amount kT to the environment. Such calculations set limits on the theoretical maximal speed of computers, because the heat can be driven away with a limited speed only.

3.4 Dynamical entropy

This is the key entropy notion in ergodic theory; a version of the Kolmogorov-Sinai entropy for one partition. It refers to Shannon entropy, but it differs significantly as it makes sense only in the context of a measure-preserving transformation. Let T be a measurable transformation of the space Ω, which preserves the probability measure μ, i.e., such that $\mu(T^{-1}(A)) = \mu(A)$ for every measurable set $A \subset \Omega$. Let $\mathcal{P}$ be a finite measurable partition of Ω and let $\mathcal{P}^n$ denote the partition $\mathcal{P} \vee T^{-1}(\mathcal{P}) \vee \cdots \vee T^{-n+1}(\mathcal{P})$ (the least common refinement of n preimages of $\mathcal{P}$). By a subadditivity argument, the sequence of Shannon entropies $\frac{1}{n} H_\mu(\mathcal{P}^n)$ converges to its infimum. The limit

$$h_\mu(T, \mathcal{P}) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) \tag{3.2}$$

is called *the dynamical entropy of the process generated by $\mathcal{P}$ under the action of T*. This notion has a very important physical interpretation, which we now try to capture.

First of all, one should understand that in the passage from a physical system to its mathematical model (a dynamical system) (Ω, μ, T), the points $\omega \in \Omega$ should not be interpreted as particles nor the transformation T as the way the particles move around the system. Such an interpretation is sometimes possible, but has a rather restricted range of applications. Usually a point ω (later we will use the letter x) represents the physical state of the entire physical system. The space Ω is hence called the *phase space*. The transformation T is interpreted as the *set of physical rules* causing the system that is currently at some state ω to assume in the following instant of time (for simplicity we consider models with discrete time) the state $T\omega$. Such a model is *deterministic* in the sense that the initial state has "imprinted" entire future evolution. Usually, however, the observer cannot fully determine the "identity" of the initial state. He knows only the values of a few measurements, which give only a rough information, and the future of the system is, from his standpoint, random. In particular, the values of his future measurements are random variables. As time passes, the observer learns more and more about the evolution (by repeating his measurements) through which, in fact, he learns about the initial state ω. A finite-valued random variable X imposes a finite partition $\mathcal{P}$ of the phase space Ω. After time n, the observer has learned the values $\mathrm{X}(\omega), \mathrm{X}(T\omega), \ldots, \mathrm{X}(T^n\omega)$ i.e., he/she has learned which element of the partition $\mathcal{P}^n$ contains ω. His/her acquired *information* about the "identity" of ω equals $I_{\mathcal{P}^n}(\omega)$, the expected value of which is $H_\mu(\mathcal{P}^n)$. It is now seen directly from the definition that:

- *The dynamical entropy equals the average (over time and the phase space) gain in one step of information about the initial state*.

Notice that it does not matter whether in the end (at time infinity) the observer determines the initial state completely, or not. What matters is the "gain of information in one step".

If the transformation T is invertible, we can also assume that the evolution of the system runs from time $-\infty$, i.e., it has an infinite past. In such case ω should be called the *current state* rather than initial state (in a process that runs from time $-\infty$, there is no initial state). Then the entropy $h_\mu(T, \mathcal{P})$ can be computed alternatively using

conditional entropy:

$$h_\mu(T,\mathcal{P}) = \lim_n H(\mathcal{P}|T(\mathcal{P}) \vee T^2(\mathcal{P}) \cdots \vee T^{n-1}(\mathcal{P})) = H(\mathcal{P}|\mathcal{P}^-),$$

where $\mathcal{P}^-$ is the sigma-algebra generated by all partitions $T^n(\mathcal{P})$ ($n \geq 0$) and is called *the past*. This formula provides another interpretation:

- *The dynamical entropy equals the expected amount of information about the current state ω acquired, in addition to was already known from the infinite past, by learning the element of the partition $\mathcal{P}$ to which belongs ω.*

Notice that in this last formulation the averaging over time is absent.

3.5 Dynamical entropy as data compression rate

The interpretation of entropy given in this paragraph is going to be fundamental for our understanding of dynamical entropy, in fact, we will also refer to a similar interpretation when discussing topological dynamics.

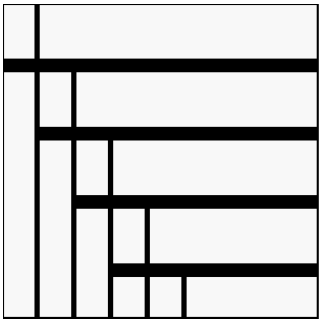
We will distinguish two kinds of data compression: "horizontal" and "vertical". In horizontal data compression we are interested in replacing computer files by other files, as short as possible. We want to "shrink them horizontally". Vertical data compression concerns infinite sequences of symbols interpreted as *signals*. Such signals occur for instance in any "everlasting" data transmission, such as television or radio broadcasting. Vertical data compression attempts to losslessly translate the signal maintaining the same speed of transmission (average lengths of incoming files) but using a smaller alphabet. We call it "vertical" simply by contrast to "horizontal". One can imagine that the symbols of a large alphabet, say of cardinality 2^k, are binary columns of k zeros or ones, and then the vertical data compression will reduce not the length but the "height" of the signal. This kind of compression is useful for data transmission "in real time"; a compression device translates the incoming signal into the optimized alphabet and sends it out at the same speed as the signal arrives (perhaps with some delay).

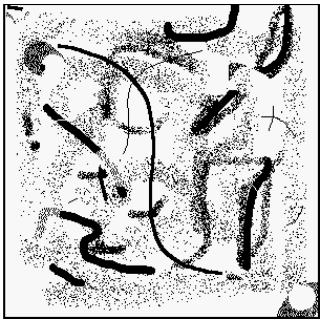
First we discuss the connection between entropy and the horizontal data compression. Consider a collection of computer files, each in form of a long string B (we will call it a *block*) of symbols belonging to some finite alphabet Λ. For simplicity let us assume that all files are binary, i.e., that $\Lambda = \{0, 1\}$.

Suppose we want to compress them to save the disk space. To do it, we must establish a coding algorithm ϕ which replaces our files B by some other (preferably shorter) files $\phi(B)$ so that no information is lost, i.e., we must also have a decoding algorithm ϕ^{-1} allowing to reconstruct the original files when needed. Of course, we assume that our algorithm is efficient, that is, it compresses the files as much as possible. Such an algorithm allows to measure the effective information content of every file: a file carries s bits of information (regardless of its original size) if it can be compressed to a binary file of length $s(B) = s$. This complies with our previous interpretation of information: each symbol in the compressed file is an answer to a binary question, and $s(B)$ is the optimized number of answers needed to identify the original file B.

Somewhat surprisingly, the amount of information $s(B)$ depends not only on the initial size $m = m(B)$ of the original file B but also on subtle properties of its structure. Evidently $s(B)$ is not the simple-minded Shannon information function. There are 2^m binary blocks of a given length m, all of them are "equally likely" so that each has "probability" 2^{-m}, and hence each should carry the same "amount of information" equal to $m \log_2 2 = m$. But $s(B)$ does not behave that simple!

Example Consider the two bitmaps shown in this figure. They have the same dimensions





and the same "density", i.e., the same amount of black pixels. As uncompressed computer files, they occupy exactly the same amount of disk space. However, if we compress them, using nearly any available "zipping" program, the sizes of the zipped files will differ significantly. The left hand side picture will shrink nearly 40 times, while the right hand side one only 8 times. Why? To quickly get an intuitive understanding of this phenomenon imagine that you try to pass these pictures over the phone to another person, so that he/she can literally copy it based on your verbal description. The left picture can be precisely described in a few sentences containing the precise coordinates of only two points, while the second picture, if we want it precisely copied, requires tedious dictating the coordinates of nearly all black pixels. Evidently, the right hand side picture carries more information. A file can be strongly compressed if it reveals some regularity or predictability, which can be used to shorten its description. The more it looks random, the more information must be passed over to the recipient, and the less it can be compressed using no matter how intelligent zipping algorithm.

How can we *a priori*, i.e., without experimenting with compression algorithms, just by looking at the file's internal structure, predict the *compression rate* $\frac{s(B)}{m(B)}$ of a given block B? Here is an idea: The compression rate should be interpreted as the *average information content per symbol*. Recall that the dynamical entropy was interpreted similarly, as the expected gain of information per step. If we treat our long block as a portion of the orbit of some point ω representing a shift-invariant measure μ on the symbolic space $\Lambda^{\mathbb{N}\cup\{0\}}$ of all sequences over Λ, then the global information carried by this block should be approximately equal to its length (number of steps in the shift map) times the dynamical entropy of μ. It will be only an approximation, but it should work. The alphabet Λ plays the role of the finite partition $\mathcal{P}$ of the symbolic space, and the partition $\mathcal{P}^n$ used in the definition of the dynamical entropy can be identified with Λ^n – the collection of all blocks over Λ of length n. Any shift-invariant measure on $\Lambda^{\mathbb{N}\cup\{0\}}$ assigns values to all blocks $A \in \Lambda^n$ ($n \in \mathbb{N}$) following some rules of

consistency; we skip discussing them now. It is enough to say that a long block B (of a very large length m) nearly determines a shift-invariant measure: for subblocks A of lengths n much smaller than m (but still very large) it determines their *frequencies*:

$$\mu_{(B)}(A) = \frac{\#\{1 \le i \le m-n+1 : B[i, i+n-1] = A\}}{m-n+1},$$

i.e., it associates with A the probability of seeing A in B at a randomly chosen "window" of length n. Of course, this measure is not completely defined (values on longer blocks are not determined), so we cannot perform the full computation of the dynamical entropy. But instead, we can use the approximate value $\frac{1}{n}H_{\mu_{(B)}}(\Lambda^n)$ (see (3.2)), which is defined and practically computable for some reasonable length n. We call it *the combinatorial entropy of the block* B. In other words, we decide that the compression rate should be approximately

$$\frac{s(B)}{m(B)} \approx \frac{1}{n}H_{\mu_{(B)}}(\Lambda^n). \tag{3.3}$$

This idea works perfectly well; in most cases the combinatorial entropy estimates the compression rate very accurately. We replace a rigorous proof with a simple example.

Example We will construct a lossless compression algorithm and apply it to a file B of a finite length m. The compressed file will consist of a *decoding instruction* followed by the coded image $\phi(B)$ of B. To save on the output length, the decoding instruction must be relatively short compared to m. This is easily achieved in codes which refer to relatively short components of the block B. For example, the instruction of the code may consist of the complete list of subblocks A of some carefully chosen length n (appearing in B) followed by the list of their images $\Phi(A)$. The images may have different lengths (as short as possible). The assignment $A \mapsto \Phi(A)$ will depend on B, therefore it must be included in the output file. The coded image $\phi(B)$ is obtained by cutting B into subblocks $B = A_1A_2 \dots A_k$ of length n and concatenating the images of these subblocks: $\phi(B) = \Phi(A_1)\Phi(A_2)\cdots\Phi(A_k)$. There are additional issues here: in order for such a code to be invertible, the images $\Phi(A)$ must form a *prefix free* family (i.e., no block in this family is a prefix of another). Then there is always a unique way of cutting $\phi(B)$ back into the images $\Phi(A_i)$. But this does not affect essentially the computations. For best compression results, it is reasonable to assign shortest images to the subblocks appearing in B with highest frequencies. For instance, consider a long binary block

$$B \;=\; 010001111001111...110 \;=\; 010, 001, 111, 001, 111, ..., 110$$

On the right, B is shown divided into subblocks of length $n = 3$. Suppose that the frequencies of the subblocks in this division are:

$$\begin{array}{cccc} 000-0\% & 001-40\% & 010-10\% & 011-10\% \\ 100-0\% & 101-0\% & 110-10\% & 111-30\% \end{array}$$

The theoretical value of the compression rate (obtained using the formula (3.3) for $n = 3$) is

$$\big(-0.4\log_2(0.4) - 0.3\log_2(0.3) - 3\cdot 0.1\log_2(0.1)\big)/3 \approx 68.2\%.$$

A binary prefix free code giving shortest images to most frequent subblocks is

$$\begin{aligned} 001 &\mapsto 0, \\ 111 &\mapsto 10, \\ 010 &\mapsto 110, \\ 011 &\mapsto 1110, \\ 110 &\mapsto 1111. \end{aligned}$$

The compression rate achieved on B using this code equals

$$(0.4 \times 1 + 0.3 \times 2 + 0.1 \times 3 + 0.1 \times 4 + 0.1 \times 4)/3 = 70\%$$

(ignoring the finite length of the decoding instruction, which is simply a recording of the above code). This code is nearly optimal (at least for this file).

We now focus on the vertical data compression. Its connection with the dynamical entropy is easier to describe but requires a more advanced apparatus. Since we are dealing with an infinite sequence (the signal), we can assume it represents some genuine (not only approximate as it was for a long but finite block) shift-invariant probability measure μ on the symbolic space $\Lambda^{\mathbb{Z}}$. Recall that the dynamical entropy $h = h_\mu(\sigma, \Lambda)$ (where σ denotes the shift map) is the expected amount of new information per step (i.e., per incoming symbol of the signal). We intend to replace the alphabet by a possibly small one. It is obvious that if we manage to losslessly replace the alphabet by another, say Λ_0, then the entropy h cannot exceed $\log_2 \#\Lambda_0$. Conversely, it turns out that any alphabet of cardinality $\#\Lambda_0 > 2^h$ is sufficient to encode the signal. This is a consequence of the famous Krieger Generator Theorem. Thus we have the following connection:

$$\log_2 \#\Lambda_0 - 1 \leq h \leq \log_2 \#\Lambda_0,$$

where Λ_0 is the smallest alphabet allowing to encode the signal. In this manner the cardinality of the optimal alphabet is completely determined by the entropy. If 2^h happens to be an integer we seem to have two choices, but there is an easy way to decide which one to choose.

3.6 Topological entropy

By a *topological dynamical system* we understand the pair (X, T), where X is a compact metric space, $T : X \to X$ is continuous.

Just like in the measure-theoretic case, we are interested in a notion of entropy that captures the complexity of the dynamics, interpreted as the amount of information transmitted by the system per unit of time. Again, the initial state carries complete information about the evolution (both forward and backward in time or just forward, depending on whether T is invertible or not), but the observer cannot "read" all this information immediately. Since we do not fix any particular measure, we want to use the metric (or, more generally, the topology) to describe the "amount of information" about the initial state, acquired by the observer in one step (one measurement). A reasonable interpretation relies on the notion of *topological resolution*. Intuitively, resolution is a parameter measuring the ability of the observer to distinguish between points. A

resolution is topological, when this ability agrees with the topological structure of the space. The simplest such resolution is based on the metric and a positive number ε: two points are "indistinguishable" if they are less than ε apart. Another way to define a topological resolution (applicable in all topological spaces) refers to an open cover of X. Points cannot be distinguished when they belong to a common cell of the cover.

By compactness, the observer is able to "see" only a finite number N of "classes of indistinguishability" and classify the current state of the system to one of them. The logarithm to base 2 of N roughly corresponds to the number of binary questions answering which is equivalent to what he has learned, i.e., the amount of acquired information. The static entropy, instead of an expectation (which requires a measure), will now be replaced by the supremum, over the space, of this information. The rest is done just like in the measure-theoretic case; we define the topological dynamical entropy with respect to a resolution as the average (along the time) information acquired per step. Finally we pass to the supremum as the resolution refines.

Notice that "indistinguishability" is not an equivalence relation; the "classes" often overlap without being equal. This makes the interpretation of a topological resolution a bit fuzzy and its usage in rigorous computations – rather complicated.

We will describe in more detail topological entropy in the sense of Dinaburg and Bowen, using the metric [comp. Dinaburg, 1970; Bowen, 1971]. Let X be endowed with a metric d. For $n \in \mathbb{N}$, by d^n we will mean the metric

$$d^n(x,y) = \max\{d(T^i x, T^i y) : i = 0, \ldots, n-1\}.$$

Of course $d^1 = d$, $d^{n+1} \geq d^n$ for each natural n, and, by compactness of X, all these metrics are pairwise uniformly equivalent.

Following the concept of indistinguishability in the resolution determined by a distance $\varepsilon > 0$, a set $F \subset X$ is said to be (n,ε)-*separated* if the distances between distinct points of F in the metric d^n are at least ε:

$$\forall_{x,y \in F}\ d_n(x,y) \geq \varepsilon.$$

By compactness, the cardinalities of (n,ε)-separated sets in X are finite and bounded. By $s(n,\varepsilon)$ we will denote the maximal cardinality of an (n,ε)-separated set:

$$s(n,\varepsilon) = \max\{\#F : F \text{ is } (n,\varepsilon)\text{-separated}\}.$$

It is clear that $s(n,\varepsilon)$ (hence also the first two terms defined below) depends decreasingly on ε. So, we can apply the general scheme:

$$\begin{aligned}
\mathbf{H}(n,\varepsilon) &= \log s(n,\varepsilon),\\
\mathbf{h}(T,\varepsilon) &= \limsup_{n\to\infty} \tfrac{1}{n}\mathbf{H}(n,\varepsilon),\\
\mathbf{h}(T) &= \lim_{\varepsilon\to 0} \uparrow \mathbf{h}(T,\varepsilon).
\end{aligned}$$

For the interpretation, suppose we observe the system through a device whose resolution is determined by the distance ε. Then we can distinguish between two

n-orbits $(x, Tx, \ldots, T^{n-1}x)$ and $(y, Ty, \ldots, T^{n-1}y)$ if and only if for at least one $i \in \{0, \ldots, n-1\}$ the points $T^i x, T^i y$ can be distinguished (which means their distance is at least ε), i.e., when the points x, y are (n, ε)-separated. Thus, $s(n, \varepsilon)$ is the *maximal* number of pairwise distinguishable n-orbits that exists in the system. The term $\mathbf{h}(T, \varepsilon)$ is hence the *rate of the exponential growth of the number of ε-distinguishable n-orbits* (and $\mathbf{h}(T)$ is the limit of such rates as the resolutions refine).

The connection between topological entropy and Kolmogorov-Sinai entropy is established by the celebrated *Variational Principle* [Goodwyn, 1971; Goodman, 1971]. By $\mathcal{M}_T(X)$ we will denote the set of all T-invariant Borel probability measures on X (which is nonempty by an appropriate fixpoint theorem).

Theorem 3.4 (Variational Principle)**.**

$$\mathbf{h}(T) = \sup\{h_\mu(T) : \mu \in \mathcal{M}_T(X)\}.$$

3.7 Symbolic extension entropy

Given a topological dynamical system (X, T) of finite entropy, we are interested in computing the "amount of information" per unit of time transferred in this system. Suppose we want to compute *verbatim* the vertical data compression, as described in Section 3.5: the logarithm of the minimal cardinality of the alphabet allowing to losslessly encode the system in real time. Since we work with topological dynamical systems, we want the coding to respect not only the measurable, but also the topological, structure. There is nothing like Krieger Generator Theorem in topological dynamics. A topological dynamical system of finite entropy (even invertible) need not be conjugate to a subshift over a finite alphabet. Thus, we must first create its *lossless digitalization*, which has only one possible form: a subshift, in which the original system occurs as a factor. In other words, we seek for a *symbolic extension*. Only then we can try to optimize the alphabet. It turns out that such a vertical data compression is not governed by the topological entropy only by a different (possibly much larger) parameter.

Precisely, what are symbolic extensions? First of all, we do not accept extensions in form of subshifts over infinite alphabets, even countable, like $\mathbb{N}_0 \cup \infty$. Such extensions are useless in terms of vertical data compression. So, in all we are about to say, a symbolic extension always means a subshift over a finite alphabet.

Notice that if a system has infinite topological entropy, it simply does not have any symbolic extensions, as these have finite topological entropy and the entropy of a factor is smaller than or equal to that of the extension. Thus we will focus exclusively on systems (X, T) with finite topological entropy.

Next, we want to explain that, no matter whether T is invertible or not, we are always going to seek for a symbolic extension in form of a *bilateral* subshift, i.e., a subset of $\Lambda^{\mathbb{Z}}$, never of $\Lambda^{\mathbb{N}_0}$. Without any assumptions, the system may happen to posses an invertible factor of positive topological entropy, which eliminates the existence of unilateral symbolic extensions. So, if we want a unified theory of symbolic extensions, we should agree on the definition of a symbolic extension given below:

Definition 3.5. *Let (X,T) be a topological dynamical system. By a* symbolic extension *of (X,T) we understand a bilateral subshift (Y,σ), where $Y \subset \Lambda^{\mathbb{Z}}$ (Λ - finite), together with a topological factor map $\phi : Y \to X$.*

The construction of symbolic extensions with minimized topological entropy relies on controlling the measure-theoretic entropies of the measures in the extension. This leads to the following notion:

Definition 3.6. *Let (Y,σ) be an extension of (X,T) via a map $\phi : Y \to X$. The* extension entropy function *is defined on the collection of all T-invariant measures $\mathcal{M}_T(X)$ on X as*

$$h^{\phi}_{\mathsf{ext}}(\mu) = \sup\{h(\nu,S) : \nu \in \mathcal{M}_S(Y),\ \phi\nu = \mu\},$$

where h denotes the entropy function on $\mathcal{M}_S(Y)$.

We now introduce the key notions of this section:

Definition 3.7. *Let (X,T) be a topological dynamical system. The* symbolic extension entropy function *is defined on $\mathcal{M}_T(X)$ as*

$$h_{\mathsf{sex}}(\mu) = h_{\mathsf{sex}}(\mu,T) = \inf\{h^{\phi}_{\mathsf{ext}}(\mu) :\ \phi \textit{ is a symbolic extension of } (X,T)\}.$$

The topological symbolic extension entropy *of (X,T) is*

$$\mathbf{h}_{\mathsf{sex}}(T) = \inf\{\mathbf{h}(S) :\ (Y,\sigma) \textit{ is a symbolic extension of } (X,T)\}.$$

In both cases, the supremum over the empty set is ∞.

It turns out that the following equality holds (see [Boyle and Downarowicz, 2004]):

Theorem 3.8 (Symbolic Extension Entropy Variational Principle)**.**

$$\mathbf{h}_{\mathsf{sex}}(T) = \sup\{h_{\mathsf{sex}}(\mu) : \mu \in \mathcal{M}_T(X)\}.$$

A system has no symbolic extensions if and only if $\mathbf{h}_{\mathsf{sex}}(T) = \infty$ if and only if $h_{\mathsf{sex}}(\mu,T) = \infty$ for all (equivalently, some) invariant measures.

If $\mathbf{h}_{\mathsf{sex}}(T)$ is finite, we can create a symbolic extension whose topological entropy is only a bit larger, for example, smaller than $\log(\lfloor 2^{\mathbf{h}_{\mathsf{sex}}(T)} \rfloor + 1)$. The alphabet in this symbolic extension can be optimized to contain $\lfloor 2^{\mathbf{h}_{\mathsf{sex}}(T)} \rfloor + 1$ elements. On the other hand, there is no way to do better than that. In this manner the symbolic extension entropy controls the vertical data compression in the topological sense.

An effective method allowing to establish (or at least estimate) the symbolic extension entropy function (and hence the topological symbolic extension entropy) is quite complicated and refers to the notion of an *entropy structure*, which describes how entropy emerges for invariant measures as the topological resolution refines. In most dynamical systems the topological symbolic extension entropy is essentially larger than topological entropy; it can even be infinite (i.e., a system has no symbolic extension)

in systems with finite topological entropy. The details are described in [Boyle and Downarowicz, 2004].

Let us mention that historically, in the first attempt to capture the "entropy jump" when passing a to symbolic extension Mike Boyle defined *topological residual entropy* which, in our notation, is the difference

$$\mathbf{h}_{\mathsf{res}}(T) = \mathbf{h}_{\mathsf{sex}}(T) - \mathbf{h}(T).$$

[see Boyle, 1991; Boyle et al., 2002].

3.8 Entropy as disorder

The interplay between Shannon and Boltzmann entropy has led to associating with the word "entropy" some colloquial understanding. In all its strict meanings (described above), entropy can be viewed as a measure of disorder and chaos, as long as by "order" one understands that "things are segregated by their kind" (e.g. by similar properties or parameter values). Chaos is the state of a system (physical or dynamical) in which elements of all "kinds" are mixed evenly throughout the space. For example, a container with gas is in its state of maximal entropy when the temperature and pressure are constant. That means there is approximately the same amount of particles in every unit of the volume, and the proportion between slow and fast particles is everywhere the same. States of lower entropy occur when particles are "organized": slower ones in one area, faster ones in another. A signal (an infinite sequence of symbols) has large entropy (i.e., compression rate) when all subblocks of a given length n appear with equal frequencies in all sufficiently long blocks. Any trace of "organization" and "logic" in the structure of the file allows for its compression and hence lowers its entropy. These observations generated a colloquial meaning of entropy. To have order in the house, means to have food separated from utensils and plates, clothing arranged in the closet by type, trash segregated and deposited in appropriate recycling containers, etc. When these things get mixed together "entropy" increases causing disorder and chaos. Entropy is a term in social sciences, too. In a social system, order is associated with classification of the individuals by some criteria (stratification, education, skills, etc.) and assigning to them appropriate positions and roles in the system. Law and other mechanisms are enforced to keep such order. When this classification and assignment fails, the system falls into chaos called "entropy". Individuals loose their specialization. Everybody must do all kinds of things in order to survive. Ergo,

Entropy equals lack of diversity.

References

Adler, R. L., Konheim, A. G., and McAndrew, M. H. 1965. Topological entropy. *Trans. Amer. Math. Soc.*, **114**, 309–319.

Alicki, R., Andries, J., Fannes, M., and Tuyls, P. 1996. An algebraic approach to the Kolmogorov-Sinai entropy. *Rev. Math. Phys.*, **8**(2), 167–184.

Boltzmann, L. 1877. Über die beziehung dem zweiten Haubtsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. *Wiener Berichte*, **76**, 373–435.

Bowen, R. 1971. Entropy for group endomorphisms and homogeneous spaces. *Trans. Amer. Math. Soc.*, **153**, 401–414.

Boyle, M. 1991. Quotients of subshifts. Adler conference lecture. Unpublished.

Boyle, M., and Downarowicz, T. 2004. The entropy theory of symbolic extensions. *Invent. Math.*, **156**(1), 119–161.

Boyle, M., Fiebig, D., and Fiebig, U.-R. 2002. Residual entropy, conditional entropy and subshift covers. *Forum Math.*, **14**(5), 713–757.

Breiman, L. 1957. The individual ergodic theorem of information theory. *Ann. Math. Statist.*, **28**, 809–811.

Carnot, S. 1824. *Reflections on the Motive Power of Fire and on Machines Fitted to Develop that Power*. Paris: Bachelier. French title: *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*.

Chung, K. L. 1961. A note on the ergodic theorem of information theory. *Ann. Math. Statist.*, **32**, 612–614.

Clausius, R. 1850. Über die bewegende Kraft der Wärme, Part I, Part II. *Annalen der Physik*, **79**, 368–397, 500–524. English translation "On the Moving Force of Heat, and the Laws regarding the Nature of Heat itself which are deducible therefrom" Phil. Mag. (1851), 2, 1–21, 102–119.

Dinaburg, E. I. 1970. A correlation between topological entropy and metric entropy. *Dokl. Akad. Nauk SSSR*, **190**, 19–22.

Downarowicz, T. 2001. Entropy of a symbolic extension of a dynamical system. *Ergodic Theory Dynam. Systems*, **21**(4), 1051–1070.

Downarowicz, T. 2005. Entropy structure. *J. Anal. Math.*, **96**, 57–116.

Downarowicz, T., and Frej, B. 2005. Measure-theoretic and topological entropy of operators on function spaces. *Ergodic Theory Dynam. Systems*, **25**(2), 455–481.

Goodman, T. N. T. 1971. Relating topological entropy and measure entropy. *Bull. London Math. Soc.*, **3**, 176–180.

Goodwyn, L. W. 1971. Topological entropy bounds measure-theoretic entropy. 69–84. Lecture Notes in Math., Vol. 235.

Katok, A. 2007. Fifty years of entropy in dynamics: 1958–2007. *J. Mod. Dyn.*, **1**(4), 545–596.

Kolmogorov, A. N. 1958. A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk SSSR (N.S.)*, **119**, 861–864.

Kolmogorov, A. N. 1959. Entropy per unit time as a metric invariant of automorphisms. *Dokl. Akad. Nauk SSSR*, **124**, 754–755.

Krieger, W. 1970. On entropy and generators of measure-preserving transformations. *Trans. Amer. Math. Soc.*, **149**, 453–464.

Makarov, I. I. 2000. Dynamical entropy for Markov operators. *J. Dynam. Control Systems*, **6**(1), 1–11.

Maxwell, J. C. 1871. *Theory of Heat*. Unveränderter Nachdruck der ersten Auflage von 1932. Die Grundlehren der mathematischen Wissenschaften, Band 38. Dover Publications, Inc.

McMillan, B. 1953. The basic theorems of information theory. *Ann. Math. Statistics*, **24**, 196–219.

Ornstein, D. S. 1970. Bernoulli shifts with the same entropy are isomorphic. *Advances in Math.*, **4**, 337–352 (1970).

Shannon, C. E. 1948. A Mathematical theory of communication. *Bell System Tech.*, 379–423, 623–656.

Sinai, Y. G. 1959. On the concept of entropy for a dynamic system. *Dokl. Akad. Nauk SSSR*, **124**, 768–771.

von Neumann, J. 1968. *Mathematische Grundlagen der Quantenmechanik*. Unveränderter Nachdruck der ersten Auflage von 1932. Die Grundlehren der mathematischen Wissenschaften, Band 38. Berlin: Springer-Verlag.

Institute of Mathematics and Computer Science
Wroclaw University of Technology
Wybrzee Wyspiaskiego 27, 50-370 Wroclaw, POLAND
e-mail: downar@pwr.wroc.pl