Introduction

0.1 The leitmotiv

Nowadays, nearly every kind of information is turned into 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 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. Sometimes (but not always), we can imagine that the observation has been conducted since forever in the past as well. In this manner, the history of our recording takes on the form of a unilateral or bilateral sequence of symbols from some finite alphabet. 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 of all sequences of symbols, and the action on this space is the familiar shift transformation advancing the enumeration.

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 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 book 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 book. 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 in the book.

We also study entropy (both measure-theoretic and topological) for operators on function spaces, which generalize classical dynamical systems. The reference to data compression is not as clear here and we concentrate more on technical properties that carry over from dynamical systems, leaving the precise connection with information theory open for further investigation.

0.2 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 book. 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 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, so-called 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 results, from the point of view of data compression, 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, 1970a].

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 book 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, 2005a].

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 was 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."

0.3 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; this interpretation plays a central role in the approach to entropy in dynamical systems presented in the book. The notation used in this section is temporary.

0.3.1 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, is

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

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 us 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 $-dQ/T_1$, while that of the other rises by 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. By 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.

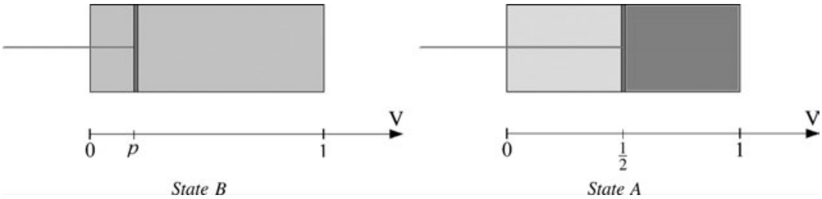
The first and second laws of thermodynamics together imply that an isolated system will tend to the state of maximal entropy among all states of the same energy. The energy distributed in this state is incapable of any further activity. The state of maximal entropy is often called the “thermodynamical 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 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 $1/T_0$ 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 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).

0.3.2 Shannon entropy

In probability theory, a *probability vector* $\mathbf{p}$ is a sequence of finitely many nonnegative numbers $\{p_1, p_2, \dots, p_n\}$ 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 $\{A_1, A_2, \dots, A_n\}$ 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 $\mathcal{P}$:

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

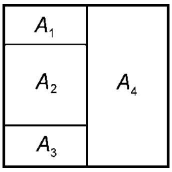
In this setup entropy can be linked with *information*. Given a measurable set A, the information $I(A)$ associated with A is defined as $-\log_2(\mu(A))$. The *information function* $I_{\mathcal{P}}$ associated with a partition $\mathcal{P} = \{A_1, A_2, \dots, A_n\}$ is defined on the space Ω and it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

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

where $\mathbb{1}_{A_i}$ 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 an 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 $I_{\mathcal{P}}(\omega) \le N(\omega) \le I_{\mathcal{P}}(\omega) + 1$ 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.

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

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 $H_\mu(\mathcal{P})$.

0.3.3 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

$$S_{max} - S(A) = k \cdot I(A), \tag{0.3.1}$$

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 (0.3.1), 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 a quantum approach, the space