new mathematical monographs: 18

Entropy in Dynamical Systems

Tomasz Downarowicz

CAMBRIDGE

CAMBRIDGE

more information - www.cambridge.org/9780521888851

Entropy in Dynamical Systems

This comprehensive text on entropy covers three major types of dynamics: measure-preserving transformations; continuous maps on compact spaces; and operators on function spaces.

Part I contains proofs of the Shannon–McMillan–Breiman Theorem, the Ornstein–Weiss Return Time Theorem, the Krieger Generator Theorem, the Sinai and Ornstein Theorems, and among the newest developments, the Ergodic Law of Series. In Part II, after an expanded exposition of classical topological entropy, the book addresses Symbolic Extension Entropy. It offers deep insight into the theory of entropy structure and explains the role of zero-dimensional dynamics as a bridge between measurable and topological dynamics. Part III explains how both measure-theoretic and topological entropy can be extended to operators on relevant function spaces.

Intuitive explanations, examples, exercises and open problems make this an ideal text for a graduate course on entropy theory. More experienced researchers can also find inspiration for further research.

Tomasz Downarowicz is Full Professor in Mathematics at Wroclaw University of Technology, Poland.

NEW MATHEMATICAL MONOGRAPHS

All the titles listed below can be obtained from good booksellers or from Cambridge University Press. For a complete series listing visit www.cambridge.org/mathematics.

1 M. Cabanes and M. Enguehard *Representation Theory of Finite Reductive Groups*
2 J. B. Garnett and D. E. Marshall *Harmonic Measure*
3 P. Cohn *Free Ideal Rings and Localization in General Rings*
4 E. Bombieri and W. Gubler *Heights in Diophantine Geometry*
5 Y. J. Ionin and M. S. Shrikhande *Combinatorics of Symmetric Designs*
6 S. Berhanu, P. D. Cordaro and J. Hounie *An Introduction to Involutive Structures*
7 A. Shlapentokh *Hilbert's Tenth Problem*
8 G. Michler *Theory of Finite Simple Groups I*
9 A. Baker and G. Wüstholz *Logarithmic Forms and Diophantine Geometry*
10 P. Kronheimer and T. Mrowka *Monopoles and Three-Manifolds*
11 B. Bekka, P. de la Harpe and A. Valette *Kazhdan's Property (T)*
12 J. Neisendorfer *Algebraic Methods in Unstable Homotopy Theory*
13 M. Grandis *Directed Algebraic Topology*
14 G. Michler *Theory of Finite Simple Groups II*
15 R. Schertz *Complex Multiplication*
16 S. Bloch *Lectures on Algebraic Cycles (2nd Edition)*
17 B. Conrad, O. Gabber and G. Prasad *Pseudo-reductive Groups*

Entropy in Dynamical Systems

TOMASZ DOWNAROWICZ
Wroclaw University of Technology, Poland

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521888851

First published 2011

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data

Downarowicz, Tomasz, 1956–
Entropy in Dynamical Systems / Tomasz Downarowicz.
p. cm. – (New Mathematical Monographs ; 18)
Includes bibliographical references and index.
ISBN 978-0-521-88885-1 (Hardback)
1. Topological entropy–Textbooks. 2. Topological dynamics–Textbooks. I. Title.
QA611.5.D685 2011
515′.39–dc22

2010050336

ISBN 978-0-521-88885-1 Hardback

To my Parents

Contents

Preface

This book is designed as a comprehensive lecture on entropy in three major types of dynamics: measure-theoretic, topological and operator. In each case the study is restricted to the most classical case of the action of iterates of a single transformation (or operator) on either a standard probability space or on a compact metric space. We do not venture into studying actions of more general groups, dynamical systems on noncompact spaces or equipped with infinite measures. On the other hand, we do not restrict the generality by adding more structure to our spaces. The most structured systems addressed here in detail are smooth transformations of the compact interval. The primary intention is to create a self-contained course, from the basics through more advanced material to the newest developments. Very few theorems are quoted without a proof, mainly in the chapters or sections marked with an asterisk. These are treated as "nonmandatory" for the understanding of the rest of the book, and can be skipped if the reader chooses. Our facts are stated as generally as possible within the assumed scope, and wherever possible our proofs of classical theorems are different from those found in the most popular textbooks. Several chapters contain very recent results for which this is a textbook debut.

We assume familiarity of the reader with basics of ergodic theory, measure theory, topology and functional analysis. Nevertheless, the most useful facts are recalled either in the main text or in the appendix.

Some elementary statements and minor passages are left without a proof, as an exercise for the reader. Such statements are collected at the end of each chapter, together with other exercises of independent interest. It is planned that solutions to selected exercises will be made available shortly after the book has occurred in print, at the publisher's website www.cambridge.org/9780521888851.

Acknowledgments

First of all, I wish to express my gratitude to those who helped me with the mathematical issues, especially to Mike Boyle, Dan Rudolph, Jean-Paul Thouvenot and Benjy Weiss. I thank all those who read the preliminary version and helped me fix the mistakes. Large portions were proofread by Bartosz Frej, Jacek Serafin and David Burguet. I am grateful to the authorities of my Institute and Faculty for their understanding and a reduction on other university duties allowing me to better focus on writing the book. My warmest thanks are directed to Sylwia for her love and priceless help in organizing my work and everyday life during this busy time.

The research of the author was partially supported by the Polish Ministry of Education Grant Number N N201 394537.

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.

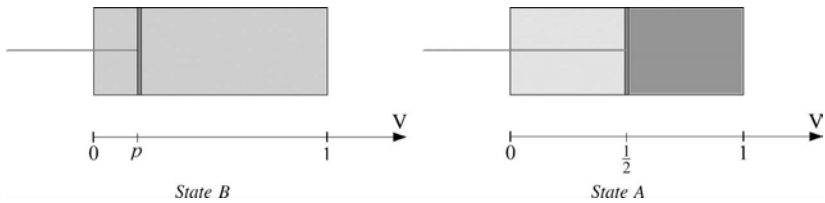
Ludwig Boltzmann gave another, probabilistic meaning to entropy. For each state A the (negative) difference between the entropy of A and the entropy of the "maximal state" B is nearly proportional to the logarithm of the probability that the system spontaneously assumes state A,

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

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 non-negative 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

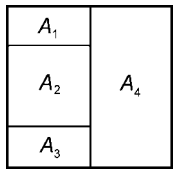
defined on the space Ω and it assumes the constant value $I(A_i)$ at all points ω belonging to the set A_i. Formally,

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

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 $I_{\mathcal{P}}(\omega) \leq N(\omega) \leq I_{\mathcal{P}}(\omega) + 1$ 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 $H_\mu(\mathcal{P})$.

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

Ω is infinite and so 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 $I(A) = -\log_2(\mu(A)) = -\log_2\left(\frac{\#A}{\#\Omega}\right)$. By Equation (0.3.1) we get

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

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 interpretation above is a subject of much discussion, 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.

0.3.4 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

$$h_\mu(T, \mathcal{P}) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) \tag{0.3.2}$$

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" the entire future evolution. Usually, however, the observer cannot fully determine the "identity" of the initial state. The observer knows only the values of a few measurements, which give only a rough information, and the future of the system is, from his/her standpoint, random. In particular, the values of 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/she 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 ω belongs.*

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

0.3.5 Dynamical entropy as data compression rate

The interpretation of entropy given in this subsection 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).

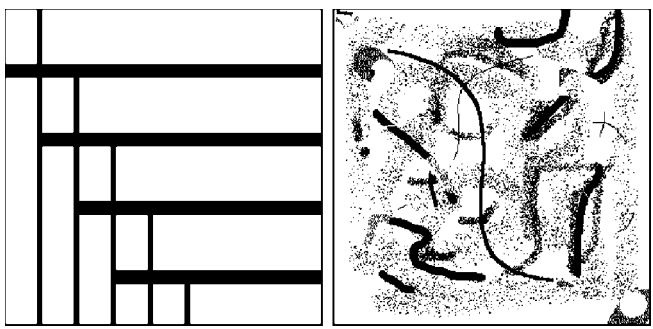
First we discuss the connection between entropy and the horizontal data compression. Consider a collection of computer files, each in form of a long string B (we will call it a *block*) of symbols belonging to some finite alphabet Λ. For simplicity let us assume that all files are binary, i.e., that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

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 us 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 us 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 simply!

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 tediously 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 random it looks, the more information must be passed over to the recipient, and the less it can be compressed no matter how intelligent a zipping algorithm is used.

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* $s(B)/m(B)$ 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 (0.3.2)), 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

$$\frac{s(B)}{m(B)} \approx \frac{1}{n}H_{\mu_{(B)}}(\Lambda^n). \tag{0.3.3}$$

As we will prove later, this idea works perfectly well; in most cases the combinatorial entropy estimates the compression rate very accurately. For now 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 (appearing in B) of some carefully chosen length n 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 $B = A_1 A_2 \dots A_k$ 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 (0.3.3) 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 (in this book it is Theorem 4.2.3). 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 (see Theorem 4.2.3).

0.3.6 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." Entropy equals lack of diversity.

0.4 Conventions

In the main body of the book (Parts I – III) we are using a consistent notational system. Every symbol has an assigned fixed meaning throughout the book. If a letter is multiply used, the meanings are distinguished by font types. The complete list of symbols is provided at the end.

The main conventions include:

- The capital letters X, Y, Z (sometimes with primes or subscripts) are reserved to denote phase spaces of dynamical systems, lowercase x, y, z are their elements. The lowercase Greek letters μ, ν, ξ denote probability measures, while Gothic capitals $\mathfrak{A}, \mathfrak{B}$, etc. stand for sigma-algebras. The letters T, S, R are used for transformations of the phase space that govern the dynamical system. Boldface $\boldsymbol{T}$ represents an operator on a function space. Factor maps and other auxiliary maps between spaces are π, ϕ, ψ. Dual maps on relevant spaces of measures are denoted by the same letter as the map on points (exception: $\boldsymbol{T}^*$ denotes the dual to a Markov operator). The images by major maps of elements of their domains are written (whenever possible) without parentheses, for example Tx, $T\mu$, $\pi\mu$, $\boldsymbol{T}f$.
- The script capitals $\mathcal{P}, \mathcal{Q}, \mathcal{R}$ stand for measurable partitions with elements (cells) denoted A, B, C, etc. The letters B and C are also used to denote finite blocks and their associated cylinders (which in fact are cells of certain partitions of appropriate symbolic spaces). The alphabet in a symbolic system is Λ (rarely Δ). If we need to distinguish between the alphabet and the associated zero-coordinate partition of the symbolic space, we use $\mathcal{P}_\Lambda$ for the latter. A special meaning is reserved to the Gothic capital $\mathfrak{P}$ (with subscripts); it is used for various spaces whose elements are partitions.
- The letters $\mathcal{U}, \mathcal{V}$ represent open covers and their cells are U, V, while $\mathcal{F}, \mathcal{G}, \mathcal{H}$ represent finite families of functions (measurable or continuous) on X.
- The symbols $\mathbb{Z}$, $\mathbb{N}$, $\mathbb{N}_0$ and $\mathbb{R}$ denote the sets of all integers, positive integers (natural numbers), nonnegative integers and real numbers, respectively. The letter $\mathbb{S}$ is used as either $\mathbb{Z}$ or $\mathbb{N}_0$. We try to consistently reserve n for integers representing the time; whereas k indexes refining sequences of partitions or covers, while i, j, l, m (sometimes also p, q, r, s, t) are integer indices of all kinds.
- The letters H and $\mathbf{H}$ are reserved to denote various notions of static entropy, with the boldface version used for topological notions. Similarly, h and $\mathbf{h}$ will be used for dynamical entropy, respectively, measure-theoretic and topological. Calligraphic $\mathcal{H}$ is used for a net or sequence of functions such as an entropy structure.

Some other conventions:

- From now on we choose to use only logarithms to base 2. We write just $\log$.
- A sequence will be written as $(a_i)_{i\geq 1}$ or (a_i), or just "the sequence a_i," when this is not ambiguous.
- Throughout this book, in order to avoid confusingly sounding words we use "decreasing" and "increasing" in the meaning of "nonincreasing" and "nondecreasing," with the adverb "strictly" when the monotonicity is sharp.

Part I

Entropy in ergodic theory

1

Shannon information and entropy

1.1 Information and entropy of probability vectors

We agree (applying the continuous extension) that the real function

$$\eta(t) = -t \log t \tag{1.1.1}$$

assumes the value 0 at $t = 0$. It is strictly concave, i.e., $\eta(pt + qs) > p\eta(t) + q\eta(s)$ for every $t, s \in [0, 1]$, where $p \in (0, 1),\ q = 1 - p$. Like every concave nonnegative function on $[0, 1]$, η satisfies the *subadditivity condition*

$$\eta(t + s) \leq \eta(t) + \eta(s),$$

whenever $t, s, t + s \in [0, 1]$ (Exercise 1.1). By iterating and by continuity, we also obtain *countable subadditivity*

$$\eta\Big(\sum_{i=1}^{\infty} t_i\Big) \leq \sum_{i=1}^{\infty} \eta(t_i),$$

whenever all above arguments of η belong to $[0, 1]$.

Let $\mathbf{P}$ and $\mathbf{S}$ denote the set of all countable probability vectors (i.e., nonnegative, with sum equal to 1) and subprobability vectors (likewise, but with sum in $[0, 1]$), respectively. Both sets are contained in the space ℓ^1 of all absolutely summable sequences, and we will regard them with the ℓ^1 topology. It is an elementary exercise to check that relatively on $\mathbf{P}$ this topology coincides with the topology of the pointwise convergence (Exercise 1.2), but on $\mathbf{S}$ this is no longer true. For instance $\mathbf{P}$ is closed in ℓ^1, while it is dense in $\mathbf{S}$ in the topology of the pointwise convergence. Of course, we are mainly interested in probability vectors. Subprobabilistic vectors will be technically useful in one place in the proof of Fact 1.1.11, so until then we are forced to check all statements for them as well.

Below, we define the key notions of entropy theory.

Definition 1.1.2 If $\mathbf{p} = (p_i)_{i\in\mathbb{N}}$ is a probability vector, its associated *information function* $I_{\mathbf{p}} : \mathbb{N} \to [0, \infty]$ is defined by

$$I_{\mathbf{p}}(i) = -\log p_i.$$

The *entropy* of $\mathbf{p}$ is defined as

$$H(\mathbf{p}) = \sum_{i=1}^{\infty} p_i I_{\mathbf{p}}(i) = -\sum_{i=1}^{\infty} p_i \log(p_i) = \sum_{i=1}^{\infty} \eta(p_i).$$

This nonnegative value can be infinite but it is certainly finite for vectors with at most finitely many nonzero terms and vectors tending to zero sufficiently fast (see Fact 1.1.4 below). The function H can be applied to any countable sequence with values in $[0, 1]$ (in particular to subprobabilistic vectors) and here it satisfies the following:

Fact 1.1.3 *The function H is concave and on the set where H is finite the concavity is strict.*

Proof Let $\mathbf{p} = (p_1, p_2, \dots)$, $\mathbf{q} = (q_1, q_2, \dots)$ and $\mathbf{r} = (r_1, r_2, \dots)$ belong to $[0, 1]^{\mathbb{N}}$, and suppose that $\mathbf{r} = p\mathbf{p} + q\mathbf{q}$ where $p \in (0, 1),\ q = 1 - p$. Then by concavity of the function η

$$H(\mathbf{r}) = \sum_{i=1}^{\infty} \eta(pp_i + qq_i) \geq \sum_{i=1}^{\infty} \big(p\eta(p_i) + q\eta(q_i)\big) = pH(\mathbf{p}) + qH(\mathbf{q}),$$

and since η is strictly concave and all terms of the above sums are nonnegative, equality holds when either $p_i = q_i$ for all i, or both sides are infinite. □

We note the following criterion for finiteness of the function H on probability vectors:

Fact 1.1.4 *If a probability vector $\mathbf{p} = (p_i)$ satisfies $\sum_{i=1}^{\infty} ip_i < \infty$, then $H(\mathbf{p}) < \infty$.*

Proof Because the function $-\log t$ is decreasing, while $-t \log t$ is increasing (certainly for values below $1/4$), we have

$$H(\mathbf{p}) = -\sum_i p_i \log(p_i) =$$

$$p_1 \log p_1 + \sum_{i\geq 2: p_i > 2^{-i}} p_i(-\log(p_i)) + \sum_{i\geq 2: p_i \leq 2^{-i}} (-p_i \log(p_i)) \leq$$

$$p_1 \log p_1 + \sum_{i\geq 2: p_i > 2^{-i}} p_i(-\log(2^{-i})) + \sum_{i\geq 2: p_i \leq 2^{-i}} (-2^{-i} \log(2^{-i})) \leq$$

$$p_1 \log p_1 + \sum_i i p_i + \sum_i i 2^{-i} < \infty.$$

□

Moreover, for vectors as above the following holds: If we let $p = 1/\sum_i i p_i$ (clearly, $p \in (0, 1]$), then $H(\mathbf{p}) \leq \frac{1}{p} H(p, 1-p)$, and equality is attained if and only if $\mathbf{p}$ is the geometric distribution $p_i = p(1-p)^{i-1}$. Although this fact can be proved using analysis (constrained maximum), we will prove it using dynamical methods much later, in Section 4.3 (Fact 4.3.7).

Let $\mathbf{P}_m$ (respectively, $\mathbf{S}_m$) denote the subset of $\mathbf{P}$ (respectively, of $\mathbf{S}$) consisting of all m-dimensional probability (respectively, subprobability) vectors, i.e., satisfying $p_i = 0$ for all $i > m$. Obviously, $\mathbf{P}_m$ (and $\mathbf{S}_m$) are compact, and the function H is continuous (hence uniformly continuous) on these sets, and assumes the maximal value equal to $\log m$ at the probability vector $\mathbf{p} = (\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m}, 0, 0, 0, \dots)$.

Below we provide a tool very useful for handling countable vectors (and later countable partitions):

Definition 1.1.5 For $\mathbf{p} \in \mathbf{P}$ we let $\mathbf{p}_{(m)} \in \mathbf{P}_m$ denote the vector obtained from $\mathbf{p}$ by taking its $m-1$ largest terms and, as the mth term, the sum of the rest, and ordering the resulting m terms decreasingly. For $\mathbf{p} \in \mathbf{S}$, $\mathbf{p}_{(m)}$ is defined identically, and it belongs to $\mathbf{S}_m$.

It is not hard to see that the map $\mathbf{p} \mapsto \mathbf{p}_{(m)}$ is uniformly continuous in ℓ^1. Moreover, we have

Fact 1.1.6

$$H(\mathbf{p}) = \lim_m \uparrow H(\mathbf{p}_{(m)}).$$

Proof By the finite and countable subadditivity of η we have

$$H(\mathbf{p}_{(m)}) \leq H(\mathbf{p}_{(m+1)}) \text{ and } H(\mathbf{p}_{(m)}) \leq H(\mathbf{p}).$$

On the other hand, ordering the terms p_i of $\mathbf{p}$ decreasingly, we can write

$$H(\mathbf{p}) = \lim_m \sum_{i=1}^{m-1} \eta(p_i) \leq \lim_m H(\mathbf{p}_{(m)}). \tag{1.1.7}$$

□

Combining the above fact with the uniform continuity of the map $\mathbf{p} \mapsto \mathbf{p}_{(m)}$ and that of H on $\mathbf{P}_m$ (and on $\mathbf{S}_m$), we conclude the following

Fact 1.1.8 *The functions $\mathbf{p} \mapsto H(\mathbf{p}_{(m)})$ are ℓ^1-uniformly continuous and $\mathbf{p} \mapsto H(\mathbf{p})$ is ℓ^1-lower semicontinuous on $\mathbf{P}$ (and on $\mathbf{S}$) (see Appendix A.1.4 for the definition of lower semicontinuity).* □

We shall be needing another observation:

Fact 1.1.9 *For each $0 \leq M < \infty$ the set of all decreasingly ordered countable probability vectors $\mathbf{p}$ with $H(\mathbf{p}) \leq M$ is compact in ℓ^1. The same holds for subprobability vectors.*

Before the proof we note that the statement does not hold without the ordering. Indeed, if $\mathbf{p}_n$ is the probability vector whose all terms are 0 except the nth term which is 1, then $H(\mathbf{p}_n) = 0$, and the set $\{\mathbf{p}_n : n \geq 1\}$ is 2-separated in ℓ^1.

Proof of Fact 1.1.9 Let $\mathbf{p}$ be a decreasingly ordered probability vector. If $H(\mathbf{p}) \leq M$, then for every $\varepsilon > 0$ the joint mass of the terms p_i smaller than $2^{-\frac{M}{\varepsilon}}$ is at most ε, for otherwise already the sum of $-p_i \log p_i$ over these terms would exceed $\varepsilon \cdot \frac{M}{\varepsilon} = M$. The cardinality of the terms larger than or equal to $2^{-\frac{M}{\varepsilon}}$ is clearly bounded by $K(\varepsilon) = 2^{\frac{M}{\varepsilon}}$. Thus, $\mathbf{p}$ has the following property:

- For every $\varepsilon > 0$ the sum of the terms above index $K(\varepsilon)$ is at most ε.

The set of all probability vectors with this property is totally bounded in ℓ^1. Indeed, every such vector can be, up to ε, approximated by its restriction to the initial $K(\varepsilon)$ terms, while the set of all subprobability vectors of dimension $K(\varepsilon)$ obviously has a finite ε-net. This net becomes a 2ε-net in the set in question. On the other hand, by lower semicontinuity of H, the set of probability vectors with $H(\mathbf{p}) \leq M$ is closed in ℓ^1, and its subset of decreasing vectors is also closed. We have shown that the set of decrasingly ordered probability vectors $\mathbf{p}$ with $H(\mathbf{p}) \leq M$ is closed in ℓ^1 and contained in a totally bounded set. By completeness of the space ℓ^1, such a set is compact. The proof for subprobability vectors is identical. □

Before we continue we need some more notation. Let ξ be a probability distribution on $[0,1]^{\mathbb{N}}$. The *barycenter* of ξ is the sequence $x^\xi = (x_1^\xi, x_2^\xi, \dots)$ such that for each natural i, $x_i^\xi = \int x_i \, d\xi(x)$ (here $x = (x_1, x_2, \dots)$). This notion generalizes convex combinations of vectors, which correspond to barycenters of finitely supported probability distributions ξ. Let $\mathbf{p}^\xi = (p_1^\xi, p_2^\xi, \dots)$ be the barycenter of a probability distribution ξ supported on $\mathbf{P}$. We claim that then $\mathbf{p}^\xi \in \mathbf{P}$. Indeed,

$$\sum_{i=1}^{\infty} p_i^\xi = \sum_{i=1}^{\infty} \int p_i \, d\xi = \int \sum_{i=1}^{\infty} p_i \, d\xi = 1,$$

where the central equality follows from monotone convergence of the finite sums to the infinite sum and linearity of the integral. By the same argument, the barycenter of a distribution supported by $\mathbf{S}$ belongs to $\mathbf{S}$.

A real function f on $\mathbf{P}$ (respectively on $\mathbf{S}$) is *supharmonic* if for every probability distribution ξ on $\mathbf{P}$ (respectively on $\mathbf{S}$), we have $f(\mathbf{p}^\xi) \geq \int f(\mathbf{p}) \, d\xi$. (The notions of barycenter and of supharmonic function are discussed in a more general context in Appendix A.2.3.) The following holds.

Fact 1.1.10 *As a concave lower semicontinuous function, the entropy H is supharmonic on $\mathbf{P}$ and on $\mathbf{S}$ (see Fact A.2.10).* □

The next fact will become important in Section 3.1. It says that on the set of probability vectors $\mathbf{p}$ such that $H(\mathbf{p}) \leq M$, the supharmonic property of H is ℓ^1*-uniformly strict*, in the following sense:

Fact 1.1.11 *Fix some positive number M. For every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever ξ is a probability distribution on $\mathbf{P}$ with barycenter $\mathbf{p}^\xi$ such that $H(\mathbf{p}^\xi) \leq M$ and $\int H(\mathbf{p}) \, d\xi > H(\mathbf{p}^\xi) - \delta$, then*

$$\int \|\mathbf{p}^\xi - \mathbf{p}\|_1 \, d\xi < \varepsilon,$$

where $\|\cdot\|_1$ denotes the norm in ℓ^1.

Proof The ℓ^1-uniform strictness of the concavity of H is obvious on the interval $[0,1]$ because this set is compact, as is the set of all probability measures supported by this set, and H (which is equal to η) is uniformly continuous and strictly concave. This property easily passes to any finite-dimensional cube $[0,1]^m$ ($m \in \mathbb{N}$) and thus to $\mathbf{S}_m$.

Let us proceed to countable probability vectors, as in the assertion. We can change the order of coordinates so that $\mathbf{p}^\xi$ becomes decreasingly ordered.