Ruslan L. Stratonovich

Theory of Information and its Value

Edited by Roman V. Belavkin
Panos M. Pardalos · Jose C. Principe

Springer

Theory of Information and its Value

Roman V. Belavkin • Panos M. Pardalos
Jose C. Principe
Editors

Theory of Information and its Value

Springer

Editors
Roman V. Belavkin
Faculty of Science and Technology
Middlesex University
London, UK

Panos M. Pardalos
Industrial and Systems Engineering
University of Florida
Gainesville, FL, USA

Jose C. Principe
Electrical & Computer Engineering
University of Florida
Gainesville, FL, USA

Author
Ruslan L. Stratonovich (Deceased)

ISBN 978-3-030-22832-3 ISBN 978-3-030-22833-0 (eBook)
https://doi.org/10.1007/978-3-030-22833-0

Mathematics Subject Classification: 94A17, 94A05, 60G35

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

It would be impossible for us to start this book without mentioning the main achievements of its remarkable author, Professor Ruslan Leontievich Stratonovich (RLS or Ruslan later). He was a brilliant mathematician, probabilist and theoretical physicist, best known for the development of the symmetrized version of stochastic calculus (an alternative to Itô calculus), with stochastic differential equations and integrals now bearing his name. His unique and beautiful approach to stochastic processes was invented in the 1950s during the time of his doctoral work on the solution to the notorious nonlinear filtering problem. The importance of this work was immediately recognized by the great Andrei Kolmogorov, who invited Ruslan, then a graduate student, for a discussion of his first papers.

This work was so much ahead of its time that its initial reception in the Soviet mathematical community was mixed, mainly due to misunderstandings of the differences between the Itô and Stratonovich approaches. These and perhaps other factors related to the cold war had obscured some of the achievements of Stratonovich's early papers on optimal nonlinear filtering, which, apart from the general solution to the nonlinear filtering problem, contained also the equation for the Kalman–Bucy filter as a special (linear) case as well as the forward–backward procedures for computing posterior probabilities, which were later rediscovered in the hidden Markov models theory. Nonetheless, the main papers were quickly translated into English, and thanks to the remarkable hard work of RLS, by 1966 he had already published two monographs—the *Topics in the Theory of Random Noise* (see [54] for a recent reprint) and *Conditional Markov Processes* [52]. These books were also promptly translated and had a better reception in the West, with the latter book being edited by Richard Bellman. In 1968, Professor W. Murray Wonham wrote in a letter to RLS '*Perhaps you are the prophet, who is honored everywhere except in his own land*'.

Despite the difficulties, the following ten years in the 1960s were very productive for Ruslan, who quickly became recognized as one of the top scientists in the World in his field. He became a Professor by 1969 (in the post he remained at the Department of Physics all his life). At that time in the 1960s, he managed to form a group of young and talented graduate students including Grishanin, B. A., Sosulin, Yu. G., Kulman, N. K., Kolosov, G. E., Mamayev, D. D., Platonov, A. A. and Belavkin, V. P.

Fig. 1 Stratonovich R. L. (second left) with his group in front of Physics Department, Moscow State University, 1971. From left to right: Kulman, N. K., Stratonovich, R. L., Mamayev, D. D., Sosulin, Yu. G., Kolosov, G. E., Grishanin, B. A. Picture taken by Slava (V. P.) Belavkin

(Figure 1). These students began working under his supervision in completely new areas of information theory, stochastic and adaptive optimal control, cybernetics and quantum information. Ruslan was young and had a somewhat legendary reputation among students and colleagues at the university, so that many students aspired to work with him, even though he was known to be a very hard working and demanding supervisor. At the same time, he treated his students as equals and gave them a lot of freedom. Many of his former students recall that Ruslan had an amazing gift to predict the solution before it was derived. Sometimes, he surprised his colleagues by giving an answer to some difficult problem, and when they asked him how he obtained it, his answer was '*just verify it yourself, and you will see it is correct*'. Ruslan and his students spent a lot of their spare time together, either playing tennis during the summer time or skiing in winters, such that they developed long-lasting friendships.

In the mid-1960s, Ruslan read several specialist courses on various topics, including adaptive Bayesian inference, logic and information theory. The latter course included a lot of original material, which emphasized the connection between information theory with statistical thermodynamics and introduced the Value of Information, which he pioneered in his 1965 paper [47] and later developed together with his students (mainly Grishanin, B. A.). This course motivated Ruslan to write his third book called '*Theory of Information*'. Its first draft was ready in 1967, and it is remarkable to note that RLS even had some negotiations with Springer to pub-

lish the book in English, which unfortunately did not happen, perhaps due to some bureaucratic difficulties in the Soviet Union. The monograph was quite large and included parts on quantum information theory, which he developed together with his new student Slava (V. P.) Belavkin. Although there was an agreement to publish the book with a leading Soviet scientific publisher, the publication was delayed for some unexplained reasons. In the end, an abridged version of the book was published by a different publisher (Soviet Radio) in 1975 [53], which did not include the quantum information parts!

Nonetheless, the book had become a classic even without having been translated into English. Several anecdotal stories exist regarding the book that it was used in the West and discussed at seminars with the help of Russian-speaking graduate students. For example, Professor Richard Bucy used translated parts of the book in his seminars on information theory, and in the 1990s he even suggested that the book be published in English. In fact, in 1994 and 1995 Ruslan visited his former student and collaborator Slava Belavkin in the University of Nottingham, United Kingdom, who worked there at the Department of Mathematics (Figure 2). They had a plan to publish a second edition of the book together in English and to include the parts on quantum information. Because quantum information theory had progressed in the 1970s and 1980s, it was necessary to update the quantum parts of the manuscript, and this had become the Achilles's heel to their plan. During Ruslan's visit, they spent more time working on joint papers, which seemed as a more urgent matter.

Recollections of Roman Belavkin

I also visited my father (V.P. Belavkin) in Nottingham during the summer of 1994, and I remember very clearly how happy Ruslan was during that visit (Figure 3), especially for the ability to mow the lawn in his backyard—a completely new experience for someone who lived in a small Moscow flat all his life. Two years later, in January 1997, Ruslan died after catching a flu during the winter examinations at the Moscow State University. I went to his funeral at the Department of Physics, from which I too had already graduated. It was a very big and sad event attended by a crowd of students and colleagues. In the next couple of years, my father collaborated with Valentina, Ruslans wife, on an English translation of the book, the first version of which was in fact finished. Valentina too passed away two years later, and my father never finished this project.

Before telling the reader how the translation of this book eventually came about, I would like to write a few words about this book from my personal experience, how it became one of my favourite texts on information theory, and why I believe it is so relevant today.

Having witnessed first-hand the development of quantum information and filtering theory in the 1980s (my father's study in our small Moscow flat was also my bedroom), I decided that my career could do without non-commutative probability and stochastics. So, although I graduated from the same department as my father

Fig. 2 Stratonovich with his wife during their visit to Nottingham, England, 1994. From left to right: Robin Hudson, Slava Belavkin, Ruslan Stratonovich, Valentina Stratonovich, Nadezda Belavkina

and Ruslan, I became interested in Artificial Intelligence (AI), and a couple of years later I managed to get a scholarship to do a PhD in cognitive modelling of human learning at the University of Nottingham. I was fortunate enough to be in the same city with my parents, which allowed me to take a cycle ride through Wollaton Park and visit them either for lunch or dinner. Although we often had scientific discussions, I felt comfortable that my area of research was far away and independent of my father's territory. That, however, turned out to be a false impression.

During that time at the end of 1990s, I came across many heuristics and learning algorithms using randomization in the form of the so-called soft-max rule, where decisions were sampled from a Boltzmann distribution with a *temperature* parameter controlling how much randomization was necessary. And although using these heuristics had clearly improved the performance of the algorithms and cognitive models, I was puzzled by these links with statistical physics and thermodynamics.

Fig. 3 Stratonovich with his wife at Belavkins' home in Nottingham, England, 1994. From left to right: Nadezda Belavkina, Slava Belavkin, Roman Belavkin, Ruslan Stratonovich and Valentina Stratonovich

The fact that it was more than just a coincidence became clear when I saw that performance of cognitive models could be improved by relating the temperature parameter dynamically to entropy. Of course, I could not help sharing these naive ideas with my father, and to my surprise he did not criticize them. Instead, he went to his study and brought an old copy of Ruslan's *Theory of Information*. I spent the next few days going through various chapters of the book, and I was immediately impressed by the self-contained, and at the same time, very detailed and deep style of the presentation. Ruslan managed to start each chapter with basic and fundamental ideas, supported by very understandable examples, and then developed the material to such depth and detail that no questions seemed to remain unanswered. However, the main value of the book was in the ideas unifying theories of information, optimization and statistical physics.

My main focus was on Chapters 3 and 9, which covered variational problems leading to optimal solutions in the form of exponential family distributions (the 'soft-max'), defined and developed the value of information theory and explored many interesting examples. The value of information is an amalgamation of theories of optimal statistical decisions and information, and its applications go far beyond problems of information transmission. For example, the relation to machine learning and cognitive modelling was very immediately clear—learning from mathematical point of view was simply an optimization problem with information constraints (otherwise, there is nothing to learn), and a solution to such a problem could only be a randomized policy, where randomization was the consequence of incomplete information. Furthermore, the temperature parameter was simply the Lagrange multiplier defined by the constraint, which also meant that an optimal temperature could be derived (at least in theory) giving the solution to the notorious 'exploration-exploitation' dilemma in reinforcement learning theory. A few years later, I applied

these ideas to evolutionary systems and derived optimal control strategies for mutation rates in genetic algorithms (controlling randomization of DNA sequences). Similar applications can be developed to control learning rates in artificial neural networks and other data analysis algorithms.

History of this translation

This publication is a new translation of the 1975 book, which incorporates some parts from the original translation by Ruslan's wife, Valentina Stratonovich. The publication has become possible, thanks to the initiatives of Professors Panos Pardalos and Jose Principe. The collaboration was initiated at the '*First International Conference on Dynamics of Information Systems*' organized by Panos in the University of Florida in 2009 [22]. It is fair to say that at that time it was the only conference dedicated to more than traditional information-theoretic aspects of data and systems analysis, but also to the importance of analysing and understanding the value of information. Another very important achievement of this conference was the first attempt to develop a geometric approach to the value of information, which is why one of the invited speakers to the conference was Professor Shun-Ichi Amari. It was at this conference that the editors of this book first met together. Panos, who by that time was the author and editor of dozens of books on global optimization and data science, expressed his amazement at the unfortunate fact that this book had still not been available in English. Jose Principe, known for his pioneering work on information-theoretic learning, had already recognized the importance and relevance of this book to modern applications and was planning the translation of specific chapters. It was clear that there was a huge interest in the topic of value of information, and we began discussing the possibility of making the new English translation of this classic book, and finishing the project, which unfortunately was never completed by Ruslan Stratonovich and Slava Belavkin.

Panos suggested that Vladimir Stozhkov, one of his Russian-speaking PhD students, should do the initial translation. Vladimir took on the bulk of this work. The equations for each chapter were coordinated by Matt Emigh and entered in LaTeX by students and visitors in the Department of Computational Neuro-Engineering Laboratory (CNEL), University of Florida, during the Summer and Fall of 2016 as follows: Sections 1.1–1.5 by Carlos Loza, 1.6–1.7 by Ryan Burt, 2.1–3.4 by Ying Ma, 3.5–4.3 by Zheng Cao, 4.4–5.3, 6.1–6.5 and 8.6–8.8 by Isaac Sledge, 5.4–5.7 by Catia Silva, 5.8–5.11 by John Henning, 6.6–6.7 by Eder Santana, 7.3–7.5 by Paulo Scalassara, 7.6–8.5 by Shulian Yu and Chapters 9–12 by Matt Emigh. This translation and equations were then edited by Roman Belavkin, who also combined it with the translation by Valentina Stratonovich in order to achieve a better reflection of the original text and terminology. In particular, the introductory paragraphs of each chapter are largely based on Valentina's translation.

We would like to take the opportunity to thank Springer, and specifically Razia Amzad and Elizabeth Loew, for making the publication of this book possible. We

also acknowledge the help of the Laboratory of Algorithms and Technologies for Network Analysis (LATNA) at the Higher School of Economics in Nizhny Novgorod, Russia, for facilitating meetings and collaborations among the editors and accommodating many fruitful discussions on the topic in this beautiful Russian city.

With the emergence of data-driven economy, progress in machine learning and AI algorithms and increased computational resources, the need for a better understanding of information, its value and limitations is greater than ever. This is why we believe this book is even more relevant today than when it was first published. The vast amount of examples pertaining to all kinds of stochastic processes and problems makes it a treasure trove of information for any researcher working in the areas of data science or machine learning. It is a great pleasure to be able to contribute a little to this project and see that finally this amazing book will be open to the rest of the World.

London, UK — Roman Belavkin
Gainesville, FL, USA — Panos Pardalos
Gainesville, FL, USA — Jose Principe

Preface

This book was inspired by the author's lectures on information theory in the Department of Physics at Moscow State University, 1963–1965. Initially, the book was written in accordance with the content of those lectures. The plan was to organize the book in order to reflect all the paramount achievements of Shannon's information theory. However, while working on the book the author 'lost his way' and used a more familiar style, in which the development of own ideas dominated over a thorough recollection of existing results. That led to an inclusion of the original material into the book and to an original interpretation of many central constructs of the theory. The original material extruded a part of steady results, which were about to be included into the book. For instance, the chapter devoted to known steady methods of encoding and decoding in noisy channels was discarded.

The material included in the book is organized in three stages: the first, second, third variational problems and the first, second, third asymptotical theorems, respectively. This creates a clear panorama of the most fundamental content of Shannon's information theory.

Every writing style has its own benefits and disadvantages. The disadvantage of the chosen style is that the work of many scientists in the field of interest remains non-reflected (or insufficiently reflected). This fact should not be regarded as an indication of insufficient respect to them. As a rule, an assessment of the material's originality and the attribution of the author's ownership of the results are not given. The only exception is made for a few explicit facts.

The book adopts 'the principle of increasing complexity of material', which means that simpler and easily understood material is placed in the beginning of a book (as well as in the beginning of a chapter). The reader is not required to be familiar with more difficult and specific material situated towards the end of a chapter/the book. This principle allows the inclusion of complicated material into the book, while not making it difficult for a fairly wide range of readers. The hope is that many readers will gain useful knowledge for themselves from the book.

While considering general questions, the author tried to lead statements with the most generality possible. To achieve this, he often used the language of measure theory. For example, he utilized the notation $P(dx)$ for probability distribution. This

should not scare off those who did not master the specified language. The point is that, omitting the details, a reader can always use a simple dictionary which converts those 'intimidating' terms into those more familiar. For instance, 'probability measure $P(dx)$' can be treated as probability $P(x)$ in the case of a discrete random variable or as product $p(x)\,dx$ in the case of a continuous random variable, where $p(x)$ is a probability density function and $dx = dx_1 \dots dx_r$ is a differential corresponding to a space of dimensionality r.

Focusing on various readers, the author did not attach significance to consistency of terminology. He thought that it did not matter whether we say 'expectation' or 'mean value', 'probabilistic measure' or 'distribution'. If we employ the apparatus of generalized functions, then there always exists a probability density and it can be used. Often there is no need to distinguish between minimization signs min. and inf. By this we mean that if infimum is not attained within a considered region then we can always 'turn on' an absolutely standard ε-procedure and, as a matter of fact, nothing essential will change as a result. In the book, we pass from a discrete probabilistic space to a continuous probabilistic space in a free manner. The author tried to spare the reader any concern of inessential details and distractions from the main ideas.

The general constructs of the theory are illustrated in the book by numerous cases and examples. Due to a common importance of the theory and the examples, their statement will not require a special radiophysics terminology. If a reader is interested in application of the stated material to the problem of messages transmission through radiochannels, he/she should fill abstract concepts with radiophysical content. For instance, when considering noisy channels (Chapter 7), an input stochastic process $x = \{x_t\}$ should be treated as a variable informational parameter of signal $s(t, x_t)$ emitted by a radiotransmitter. Also, an output process $y = \{y_t\}$ should be treated as a signal at a receiver input. A proper concretization of concepts is needed for application of any mathematical theory.

The author expresses acknowledgements to the first reader of this book B.A. Grishanin, who rendered significant assistance while the book was being prepared to print, and professor B.I. Tikhonov for discussions involving a variety of subjects and a number of precious comments.

Moscow, Russia Ruslan L. Stratonovich

Introduction

The term ‘information’ mentioned in the title of the book is understood here not in the broad sense in which the word is understood by people working in the press, radio, media, but in the narrow scientific sense of Claude Shannon’s theory. In other words, the subject of this book is the special mathematical discipline, Shannon’s information theory, which can solve its own quite specific problems.

This discipline consists of abstractly formulated theorems and results, which can have different specializations in various branches of knowledge. Information theory has numerous applications in the theory of message transmission in the presence of noise, the theory of recording and registering devices, mathematical linguistics and other sciences including genetics.

Information theory, together with other mathematical disciplines, such as the theory of optimal statistical decisions, the theory of optimal control, the theory of algorithms and automata, game theory and so on, is a part of theoretical cybernetics—a discipline dealing with problems of control. Each of the above disciplines is an independent field of science. However, this does not mean that they are completely separated from each other and cannot be bridged. Undoubtedly, the emergence of complex theories is possible and probable, where concepts and results from different theories are combined and which interconnect distinct disciplines. The picture resembles trees in a forest: their trunks stand apart, but their crowns intertwine. At first, they grow independently, but then their twigs and branches intertwine making their new common crown.

Of course, generally speaking, the statement about uniting different disciplines is just an assertion, but, in fact, the merging of some initially disconnected fields of science is now an actual fact. As is evident from a number of works and from this book, the following three disciplines are inosculating:

1. statistical thermodynamics as a mathematical theory
2. Shannon’s information theory
3. the theory of optimal statistical decisions (together with its multi-step or sequential variations, such as optimal filtering and dynamic programming).

This book will demonstrate that the three disciplines mentioned above are cemented by 'thermodynamic' methods with typical attributes such as 'thermodynamic' parameters and potentials, Legendre transforms, extremum distributions, and asymptotic nature of the most important theorems.

Statistical thermodynamics can be referred to as cybernetic disciplines only conditionally. However, in some problems of statistical thermodynamics, its cybernetic nature manifests itself quite clearly. It is sufficient to recall the second law of thermodynamics and 'Maxwell's demon', which is a typical automaton converting information into physical entropy. Information is 'fuel' for perpetual motion of the second kind. These points will be discussed in Chapter 12.

If we consider statistical thermodynamics as a cybernetic discipline, then L. Boltzmann and J. C. Maxwell should be called the first outstanding cyberneticists. It is important to bear in mind that the formula expressing entropy in terms of probabilities was introduced by L. Boltzmann, who also introduced the probability distribution that was the solution to the first variational problem (of course, it does not matter how we call the functions in this formula—energy or cost function).

During the emergence of Shannon's information theory, the appearance of a well-known notion of thermodynamics, namely entropy, was regarded by some scientists as a curious coincidence, and little attention was given to the fact. It was thought that this entropy had nothing to do with physical entropy (despite the work of Maxwell's demon). In this connection, we can recall a countless number of quotation marks around the word 'entropy' in the first edition of the collection of Shannon's papers translated into Russian (the collection under the editorship of A.N. Zheleznov, Foreign Literature, 1953). I believe that now even terms such as 'temperature' in information theory can be written without quotation marks and understood merely as a parameter incorporated in the expression for the extreme distribution. Similar laws are valid both in information theory and statistical physics, and we can conditionally call them 'thermodynamics'.

At the beginning (from 1948 until 1959), only one 'thermodynamic' notion appeared in Shannon's information theory—entropy. There seemed to be no room for energy and other analogous thermodynamic potentials in it. In that regard, the theory looked feeble in comparison with statistical thermodynamics. This, however, was short-lived. The situation changed when scientists realized that in applied information theory, regarded as the theory of signal transmission, the cost function was the analogue of energy, and risk or average cost was the analogue of average energy. It became evident that a number of main concepts and relations between them are similar in two disciplines. In particular, if the first variational problem is considered, then we can speak about resemblance, 'isomorphism' of the two theories. Mathematical relationships between the corresponding notions of both disciplines are the same, and they are the contents of a mathematical theory that is considered in this book.

The content of information theory is not limited to the specified relations. Besides entropy the theory contains other notions such as the Shannon's amount of information. In addition to the first variational problem, related to an extremum of entropy under fixed risk (i.e. energy), there are also possible variational problems, in

which entropy is replaced by the Shannon's amount of information. Therefore, the content of information theory is broader than the mathematical content of statistical thermodynamics.

New variational problems reveal a remarkable analogy with the first variational problem. They contain the same ensemble of conjugate parameters and potentials, the same formulae linking potentials via the Legendre transform. And this is no surprise. It is possible to show that all those phenomena emerge when considering any non-singular variational problem.

Let us consider this subject more thoroughly. Suppose that at least two functionals $\Phi_1[P]$, $\Phi_2[P]$ of the argument ('distribution') P are given, and it is required to find an extremum of one functional subject to a fixed value constraint of the second one, say,

$$\Phi_1[P] = \mathrm{extr}_P \quad \text{subject to} \quad \Phi_2[P] = A.$$

Introducing the Lagrange multiplier α, which serves as a 'thermodynamic' parameter canonically conjugate with parameter A, we study an extremum of the expression

$$K = \Phi_1[P] + \alpha\Phi_2[P] = \mathrm{extr}_P \tag{1}$$

under a fixed value of α. The extreme value

$$K(\alpha) = \Phi_1[P^{\mathrm{extr}}] + \alpha\Phi_2[P^{\mathrm{extr}}] = \Phi_1 + \alpha A$$

serves as a 'thermodynamic' potential since $dK = d\Phi_1 + \alpha d\Phi_2 + \Phi_2 d\alpha = \Phi_2 d\alpha$, i.e.

$$dK/d\alpha = A. \tag{2}$$

The relation used

$$d\Phi_1 + \alpha d\Phi_2 \equiv [\Phi_1(P^{\mathrm{extr}} + \delta P) - \Phi_1(P^{\mathrm{extr}})] + \alpha[\Phi_2(P^{\mathrm{extr}} + \delta P) - \Phi_2(P^{\mathrm{extr}})] = 0$$

follows from (1). Here, we take a partial variation $\partial P = P^{\mathrm{extr}}(\alpha + d\alpha) - P^{\mathrm{extr}}(\alpha)$ due to the increment of parameter A, i.e. parameter α. Function $L(A) = K - \alpha A = K - \alpha(dK/d\alpha)$ is a potential conjugate by Legendre with $K(\alpha)$. Meanwhile, it follows from (2) that

$$dL(A)/dA = -\alpha.$$

These relations, which are common in thermodynamics, are apparently valid regardless of the nature of the functionals Φ_1, Φ_2 and the argument P.

The results of information theory discussed in this book are related to the three variational problems whose solution yields a number of relations, parameters and potentials. Variational problems play an important role in theoretical cybernetics because it concerns mainly *optimal* constructions and procedures.

As can be seen from the book contents, these variational problems are related also to the major laws (theorems) having asymptotic nature (i.e. valid for large composite systems). The first variational problem corresponds to stability of the canonical distribution, which is essential in statistical physics, as well as to asymptotic equivalence of constraints imposed on specific and mean values of a function.

The second variational problem is related to the famous result of Shannon about asymptotic zero probability of error for the transmission of messages through noisy channels. The third variational problem is connected with asymptotic equivalence of the values of information of Shannon's and Hartley's type. The latter results are a splendid example of unity of discrete and continuous worlds. They are also a good example that when the complexity of a discrete system grows, it is convenient to describe it by continuous mathematical objects. Finally, they are an example of how a complex continuous system behaves asymptotically similarly to a complex discrete system. It would be tempting to observe something similar, say, in a future asymptotical theory of dynamical systems and automata.

Contents

1 Definition of information and entropy in the absence of noise 1
- 1.1 Definition of entropy in the case of equiprobable outcomes 3
- 1.2 Entropy and its properties in the case of non-equiprobable outcomes 5
- 1.3 Conditional entropy. Hierarchical additivity 8
- 1.4 Asymptotic equivalence of non-equiprobable and equiprobable outcomes 12
- 1.5 Asymptotic equiprobability and entropic stability 16
- 1.6 Definition of entropy of a continuous random variable 22
- 1.7 Properties of entropy in the generalized version. Conditional entropy 28

2 Encoding of discrete information in the absence of noise and penalties 35
- 2.1 Main principles of encoding discrete information 36
- 2.2 Main theorems for encoding without noise. Independent identically distributed messages 40
- 2.3 Optimal encoding by Huffman. Examples 44
- 2.4 Errors of encoding without noise in the case of a finite code sequence length 48

3 Encoding in the presence of penalties. First variational problem 53
- 3.1 Direct method of computing information capacity of a message for one example 54
- 3.2 Discrete channel without noise and its capacity 56
- 3.3 Solution of the first variational problem. Thermodynamic parameters and potentials 58
- 3.4 Examples of application of general methods for computation of channel capacity 65
- 3.5 Methods of potentials in the case of a large number of parameters 70
- 3.6 Capacity of a noiseless channel with penalties in a generalized version 74

4 First asymptotic theorem and related results 77
4.1 Potential Γ or the cumulant generating function 78
4.2 Some asymptotic results of statistical thermodynamics. Stability of the canonical distribution 82
4.3 Asymptotic equivalence of two types of constraints 89
4.4 Some theorems about the characteristic potential 94

5 Computation of entropy for special cases. Entropy of stochastic processes 103
5.1 Entropy of a segment of a stationary discrete process and entropy rate 104
5.2 Entropy of a Markov chain 107
5.3 Entropy rate of part of the components of a discrete Markov process and of a conditional Markov process 113
5.4 Entropy of Gaussian random variables 123
5.5 Entropy of a stationary sequence. Gaussian sequence 128
5.6 Entropy of stochastic processes in continuous time. General concepts and relations 134
5.7 Entropy of a Gaussian process in continuous time 137
5.8 Entropy of a stochastic point process 144
5.9 Entropy of a discrete Markov process in continuous time 153
5.10 Entropy of diffusion Markov processes 157
5.11 Entropy of a composite Markov process, a conditional process, and some components of a Markov process 161

6 Information in the presence of noise. Shannon's amount of information 173
6.1 Information losses under degenerate transformations and simple noise 173
6.2 Mutual information for discrete random variables 178
6.3 Conditional mutual information. Hierarchical additivity of information 181
6.4 Mutual information in the general case 187
6.5 Mutual information for Gaussian variables 189
6.6 Information rate of stationary and stationary-connected processes. Gaussian processes 196
6.7 Mutual information of components of a Markov process 202

7 Message transmission in the presence of noise. Second asymptotic theorem and its various formulations 217
7.1 Principles of information transmission and information reception in the presence of noise 218
7.2 Random code and the mean probability of error 221
7.3 Asymptotic zero probability of decoding error. Shannon's theorem (second asymptotic theorem) 225
7.4 Asymptotic formula for the probability of error 228

7.5 Enhanced estimators for optimal decoding 232
7.6 Some general relations between entropies and mutual informations for encoding and decoding 243

8 Channel capacity. Important particular cases of channels 249
8.1 Definition of channel capacity 249
8.2 Solution of the second variational problem. Relations for channel capacity and potential 252
8.3 The type of optimal distribution and the partition function 259
8.4 Symmetric channels 262
8.5 Binary channels 264
8.6 Gaussian channels 267
8.7 Stationary Gaussian channels 277
8.8 Additive channels 284

9 Definition of the value of information 289
9.1 Reduction of average cost under uncertainty reduction 290
9.2 Value of Hartley's information amount. An example 294
9.3 Definition of the value of Shannon's information amount and α-information 300
9.4 Solution of the third variational problem. The corresponding potentials. 304
9.5 Solution of a variational problem under several additional assumptions. 313
9.6 Value of Boltzmann's information amount 318
9.7 Another approach to defining the value of Shannon's information 321

10 Value of Shannon's information for the most important Bayesian systems 327
10.1 Two-state system 327
10.2 Systems with translation invariant cost function 331
10.3 Gaussian Bayesian systems 338
10.4 Stationary Gaussian systems 346

11 Asymptotic results about the value of information. Third asymptotic theorem 353
11.1 On the distinction between the value functions of different types of information. Preliminary forms 354
11.2 Theorem about asymptotic equivalence of the value functions of different types of information 357
11.3 Rate of convergence between the values of Shannon's and Hartley's information 369
11.4 Alternative forms of the main result. Generalizations and special cases 379
11.5 Generalized Shannon's theorem 385

12 Information theory and the second law of thermodynamics 391
12.1 Information about a physical system being in thermodynamic equilibrium. The generalized second law of thermodynamics 392
12.2 Influx of Shannon's information and transformation of heat into work 395
12.3 Energy costs of creating and recording information. An example . 399
12.4 Energy costs of creating and recording information. General formulation 403
12.5 Energy costs in physical channels 405

Correction to: Theory of Information and its Value C1

A Some matrix (operator) identities 409
A.1 Rules for operator transfer from left to right 409
A.2 Determinant of a block matrix 410

References 413

Index 417

The original version of this book was revised: Author name has been included in all chapters source line. The correction to this book is available at https://doi.org/10.1007/978-3-030-22833-0_13

Chapter 1
Definition of information and entropy in the absence of noise

In modern science, engineering and public life, a big role is played by information and operations associated with it: information reception, information transmission, information processing, storing information and so on. The significance of information has seemingly outgrown the significance of the other important factor, which used to play a dominant role in the previous century, namely, energy.

In the future, in view of a complexification of science, engineering, economics and other fields, the significance of correct control in these areas will grow and, therefore, the importance of information will increase as well.

What is information? Is a theory of information possible? Are there any general laws for information independent of its content that can be quite diverse? Answers to these questions are far from obvious. Information appears to be a more difficult concept to formalize than, say, energy, which has a certain, long established place in physics.

There are two sides of information: quantitative and qualitative. Sometimes it is the total amount of information that is important, while other times it is its quality, its specific content. Besides, a transformation of information from one format into another is technically a more difficult problem than, say, transformation of energy from one form into another. All this complicates the development of information theory and its usage. It is quite possible that the general information theory will not bring any benefit to some practical problems, and they have to be tackled by independent engineering methods.

Nevertheless, general information theory exists, and so do standard situations and problems, in which the laws of general information theory play the main role. Therefore, information theory is important from a practical standpoint, as well as in fundamental science, philosophy and expanding the horizons of a researcher.

From this introduction one can gauge how difficult it was to discover the laws of information theory. In this regard, the most important milestone was the work of Claude Shannon [44, 45] published in 1948–1949 (the respective English originals are [38, 39]). His formulation of the problem and results were both perceived as a surprise. However, on closer investigation one can see that the new theory extends and develops former ideas, specifically, the ideas of statistical thermodynamics due

R. V. Belavkin et al. (eds.), R. L. Stratonovich (Author), *Theory of Information and its Value*, https://doi.org/10.1007/978-3-030-22833-0_1

to Boltzmann. The deep mathematical similarities between these two directions are not accidental. It is evidenced in the use of the same formulae (for instance, for entropy of a discrete random variable). Besides that, a logarithmic measure for the amount of information, which is fundamental in Shannon's theory, was proposed for problems of communication as early as 1928 in the work of R. Hartley [19] (the English original is [18]).

In the present chapter, we introduce the logarithmic measure of the amount of information and state a number of important properties of information, which follow from that measure, such as the additivity property.

The notion of the amount of information is closely related to the notion of entropy, which is a measure of uncertainty. Acquisition of information is accompanied by a decrease in uncertainty, so that the amount of information can be measured by the amount of uncertainty or entropy that has disappeared.

In the case of a discrete message, i.e. a discrete random variable, entropy is defined by the Boltzmann formula

$$H_\xi = -\sum_\xi P(\xi)\ln P(\xi),$$

where ξ is a random variable, and $P(\xi)$ is its probability distribution.

In this chapter, we emphasize the fact that this formula is a corollary (in the asymptotic sense) of the simpler Hartley's formula $H = \ln M$.

The fact that the set of realizations (the set of all possible values of a random variable) of an entropically stable (this property is defined in Section 1.5) random variable can be partitioned into two subsets, is an essential result of information theory. The first subset has infinitesimal probability, and therefore it can be discarded. The second subset contains approximately e^{H_ξ} realizations (i.e. variants of representation), and it is often substantially smaller than the total number of realizations. Realizations of that second subset can be approximately treated as equiprobable. According to the terminology introduced by Boltzmann, those equiprobable realizations can be called 'micro states'.

Information theory is a mathematical theory that employs definitions and methods of probability theory. In these discussions, there is no particular need to hold to a special 'informational' terminology, which is usually used in applications. Without detriment to the theory, the notion of 'message' can be substituted by the notion of 'random variable', and the notion of 'sequence of messages' by 'stochastic process', etc. Then in order to apply the general theory, certainly, we require a proper formalization of the applicable concepts in the language of probability theory. We will not pay extra attention here to such a translation.

The main results of Sections 1.1–1.3, related to discrete random variables (or processes) having a finite or countable number of states, can be generalized to the case of continuous (and even arbitrary) random variables assuming values from a multidimensional real space. For such a generalization, one should overcome certain difficulties and put up with certain complications of the most important concepts and formulae. The main complication consists in the following. In contrast to the case of

discrete random variables, where it is sufficient to define entropy using one measure or one probability distribution, in the general case it is necessary to introduce two measures to do so. Therefore, entropy is now related to two measures instead of one, and thus it characterizes the relationship between these measures. In our presentation, the general version of the formula for entropy is derived from the Boltzmann formula by using as an example the condensation of the points representing random variables.

There are several books on information theory, in particular, Goldman [16], Feinstein [12] and Fano [10] (the corresponding English originals are [11, 15] and [9]). These books conveniently introduce readers to the most basic notions (also see the book by Kullback [30] or its English original [29]).

1.1 Definition of entropy in the case of equiprobable outcomes

Suppose we have M equiprobable outcomes of an experiment. For example, when we roll a standard die, $M = 6$. Of course, we cannot always perform the formalization of conditions so easily and accurately as in the case of a die. We assume though that the formalization has been performed and, indeed, one of M outcomes is realized, and they are equivalent in probabilistic terms. Then there is a priori uncertainty directly connected with M (i.e. the greater the M is, the higher the uncertainty is). The quantity measuring the above uncertainty is called *entropy* and is denoted by H:

$$H = f(M), \tag{1.1.1}$$

where $f(\cdot)$ is some increasing non-negative function defined at least for natural numbers.

When rolling a dice and observing the outcome number, we obtain information whose amount is denoted by I. After that (i.e. a posteriori) there is no uncertainty left: the a posteriori number of outcomes is $M = 1$ and we must have $H_{\text{ps}} = f(1) = 0$. It is natural to measure the amount of information received by the value of disappeared uncertainty:

$$I = H_{\text{pr}} - H_{\text{ps}}. \tag{1.1.2}$$

Here, the subscript 'pr' means 'a priori', whereas 'ps' means 'a posteriori'.

We see that the amount of received information I coincides with the initial entropy. In other cases (in particular, for formula (1.2.3) given below) a message having entropy H can also transmit the amount of information I equal to H.

In order to determine the form of function $f(\cdot)$ in (1.1.1) we employ very natural *additivity principle*. In the case of a die it reads: the entropy of two throws of a die is twice as large as the entropy of one throw, the entropy of three throws of a die is three times as large as the entropy of one throw, etc. Applying the additivity principle to other cases means that the entropy of several independent systems is equal to the sum of the entropies of individual systems. However, the number M of outcomes for a complex system is equal to the product of the numbers m of outcomes for each

one of the 'simple' (relative to the total system) subsystems. For two throws of dice, the number of various pairs (ξ_1, ξ_2) (where ξ_1 and ξ_2 both take one out of six values) equals to $36 = 6^2$. Generally, for n throws the number of equivalent outcomes is 6^n. Applying formula (1.1.1) for this number, we obtain entropy $f(6^n)$. According to the additivity principle, we find that

$$f(6^n) = nf(6).$$

For other $m > 1$ the latter formula would take the form

$$f(m^n) = nf(m). \tag{1.1.3}$$

Denoting $x = m^n$, we have $n = \ln x / \ln m$. Then it follows from (1.1.3) that

$$f(x) = K \ln x, \tag{1.1.4}$$

where $K = f(m)/\ln m$ is a positive constant, which is independent of x. It is related to the choice of the units of information. Thus, a representation of function $f(\cdot)$ has been determined up to a choice of the measurement units. It is easy to verify that the aforementioned constraint $f(1) = 0$ is satisfied, indeed.

Hartley was the first [19] (the English original is [18]) who introduced the logarithmic measure of information. That is why the quantity $H = K \ln M$ is called the *Hartley's amount of information*.

Let us indicate three main choices of the measurement units for information:

1. If we set $K = 1$ in (1.1.4), then the entropy will be measured in natural units (nats) [1]:
$$H_{\text{nat}} = \ln M; \tag{1.1.5}$$
2. If we set $K = 1/\ln 2$, then we will have the entropy expressed in binary units (bits) [2]:
$$H_{\text{bit}} = \frac{1}{\ln 2} \ln M = \log_2 M; \tag{1.1.6}$$
3. Finally, we will have a physical scale of K if we consider the Boltzmann constant $k = 1.38 \cdot 10^{-23}$ J/K. The entropy measured in those units will be equal to
$$S = H_{\text{ph}} = k \ln M. \tag{1.1.7}$$

It is easy to see from the comparison of (1.1.5) and (1.1.6) that 1 nat is greater than 1 bit in $\log_2 e = 1/\ln 2 \approx 1.44$ times.

In what follows, we shall use natural units (formula (1.1.5)), dropping subscript 'nat', unless otherwise stipulated.

Suppose the random variable ξ takes any of the M equiprobable values, say, $1, \ldots, M$. Then the probability of each individual value is equal to $P(\xi) = 1/M$, $\xi = 1, \ldots, M$. Consequently, formula (1.1.5) can be rewritten as

[1] 'nat' refers to *natural digit* that means natural unit.

[2] 'bit' refers to *binary digit* that means binary unit (sign).

$$H = -\ln P(\xi). \tag{1.1.8}$$

1.2 Entropy and its properties in the case of non-equiprobable outcomes

1. Suppose now the probabilities of different outcomes are unequal. If, as earlier, the number of outcomes equals to M, then we can consider a random variable ξ, which takes one of M values. Considering an index of the corresponding outcome as ξ, we obtain that those values are nothing else but $1,\dots,M$. Probabilities $P(\xi)$ of those values are non-negative and satisfy the normalization constraint: $\sum_{\xi} P(\xi) = 1$.

If we formally apply equality (1.1.8) to this case, then each ξ should have its own entropy

$$H(\xi) = -\ln P(\xi). \tag{1.2.1}$$

Thus, we attribute a certain value of entropy to each realization of the variable ξ. Since ξ is a random variable, we can also regard this entropy as a random variable.

As in Section 1.1, the a posteriori entropy, which remains after the realization of ξ becomes known, is equal to zero. That is why the information we obtain once the realization is known is numerically equal to the initial entropy

$$I(\xi) = H(\xi) = -\ln P(\xi). \tag{1.2.2}$$

Similar to entropy $H(\xi)$, information I depends on the actual realization (on the value of ξ), i.e., it is a random variable. One can see from the latter formula that information and entropy are both large when a posteriori probability of the given realization is small and vice versa. This observation is quite consistent with intuitive ideas.

Example 1.1. Suppose we would like to know whether a certain student has passed an exam or not. Let the probabilities of these two events be

$$P(\text{pass}) = 7/8, \qquad P(\text{fail}) = 1/8.$$

One can see from these probabilities that the student is quite strong. If we were informed that the student had passed the exam, then we could say: 'Your message has not given me a lot of information. I have already expected that the student passed the exam'. According to formula (1.2.2) the information of this message is quantitatively equal to

$$I(\text{pass}) = \log_2(8/7) = 0.193 \text{ bits}.$$

If we were informed that the student had failed, then we would say 'Really?' and would feel that we have improved our knowledge to a greater extent. The amount of information of such a message is equal to

$$I(\text{fail}) = \log_2(8) = 3 \text{ bits}.$$

In theory, however, a greater role is not played by random entropy (random information, respectively) (1.2.1), (1.2.2), but by the average entropy defined by the formula

$$H_\xi = \mathbb{E}[H(\xi)] = -\sum_\xi P(\xi)\ln P(\xi). \tag{1.2.3}$$

We shall call it the *Boltzmann's entropy* or the Boltzmann's amount of information.[3] In the aforementioned example averaging over both messages yields

$$I_\xi = H_\xi = \frac{7}{8}0.193 + \frac{3}{8} = 0.544 \text{ bits}.$$

The subscript ξ of symbol H_ξ (as opposed to the argument of $H(\xi)$) is a dummy, i.e., the average entropy H_ξ describes the random variable ξ (depending on the probabilities P_ξ), but it does not depend on the actual value of ξ, i.e., on the realization of ξ. Further, we shall use this notation system in its extended form including conditional entropies.

Uncertainty of the type $0\ln 0$ occurring in (1.2.3), for vanishing probabilities is always understood in the sense $0\ln 0 = 0$. Consequently, a set of M outcomes can be always complemented by any outcomes having zero probability. Also, deterministic (non-random) quantities can be added to the entropy index. For instance, the equality $H_\xi = H_{\xi,7}$ is valid for every random variable ξ.

2. Properties of entropy.

Theorem 1.1. *Both random and average entropies are always non-negative.*

This property is connected with the fact that probability cannot exceed one and that the constant K in (1.1.4) is necessarily positive. Since $P(\xi) \leqslant 1$, we have $-\ln P(\xi) = H(\xi) \geqslant 0$. Certainly, this inequality remains valid after averaging as well.

Theorem 1.2. *Entropy attains the maximum value of* $\ln M$ *when the outcomes (realizations) are equiprobable, i.e. when* $P(\xi) = 1/M$.

Proof. This property is the consequence of the Jensen's inequality (for instance, see Rao [35] or its English original [34])

$$\mathbb{E}[f(\zeta)] \leqslant f(\mathbb{E}[\zeta]) \tag{1.2.4}$$

that is valid for every concave function $f(x)$. (Function $f(x) = \ln x$ is concave for $x > 0$, because $f''(x) = -x^{-2} < 0$). Indeed, denoting $\zeta = 1/P(\xi)$ we have

$$\mathbb{E}[\zeta] = \mathbb{E}\left[\frac{1}{P(\xi)}\right] = \sum_{\xi=1}^{M} \frac{P(\xi)}{P(\xi)} = M, \tag{1.2.5}$$

$$\mathbb{E}[f(\zeta)] = \mathbb{E}\left[\ln \frac{1}{P(\xi)}\right] = \mathbb{E}[H(\xi)] = H_\xi. \tag{1.2.6}$$

[3] Boltzmann's entropy is commonly referred to as 'Shannon's entropy', or just 'entropy' within the field of information theory.

Substituting (1.2.5), (1.2.6) to (1.2.4), we obtain

$$H_\xi \leqslant \ln M.$$

For a particular form of function $f(\zeta) = \ln \zeta$ it is easy to verify inequality (1.2.4) directly. Averaging the obvious inequality

$$\ln \frac{\zeta}{\mathbb{E}[\zeta]} \leqslant \frac{\zeta}{\mathbb{E}[\zeta]} - 1, \tag{1.2.7}$$

we obtain (1.2.4).

In the general case, it is convenient to consider the tangent line $f(\mathbb{E}[\zeta]) + f'(\mathbb{E}[\zeta])(\zeta - \mathbb{E}[\zeta])$ for function $f(\zeta)$ at point $\zeta = \mathbb{E}[\zeta]$ in order to prove (1.2.4). Concavity implies

$$f(\zeta) \leqslant f(\mathbb{E}[\zeta]) + f'(\mathbb{E}[\zeta])(\zeta - \mathbb{E}[\zeta]).$$

Averaging the latter inequality, we derive (1.2.4).

As is seen from the provided proof, Theorem 1.2 would remain valid if we replaced the logarithmic function by any other concave function in the definition of entropy. Let us now consider the properties of entropy, which are specific for the logarithmic function, namely, properties related to the additivity of entropy.

Theorem 1.3. *If random variables ξ_1, ξ_2 are independent, then the full (joint) entropy $H_{\xi_1\xi_2}$ is decomposed into the sum of entropies:*

$$H_{\xi_1\xi_2} = H_{\xi_1} + H_{\xi_2}. \tag{1.2.8}$$

Proof. Suppose that ξ_1, ξ_2 are two random variables such that the first one assumes values $1,\dots,m_1$ and the second one—values $1,\dots,m_2$. There are $m_1 m_2$ pairs of $\xi = (\xi_1, \xi_2)$ with probabilities $P(\xi_1, \xi_2)$. Numbering the pairs in an arbitrary order by indices $\xi = 1,\dots,m_1 m_2$ we have

$$H_\xi = -\mathbb{E}[\ln P(\xi)] = -\mathbb{E}[\ln P(\xi_1, \xi_2)] = H_{\xi_1\xi_2}.$$

In view of independence, we have

$$P(\xi_1, \xi_2) = P(\xi_1)P(\xi_2).$$

Therefore, $\ln P(\xi_1, \xi_2) = \ln P(\xi_1) + \ln P(\xi_2)$; $H(\xi_1\xi_2) = H(\xi_1) + H(\xi_2)$. Averaging of the latter equality yields $H_\xi = H_{\xi_1\xi_2} = H_{\xi_1} + H\xi_2$, i.e. (1.2.7). The proof is complete.

If, instead of two independent random variables, there are two independent groups $(\eta_1,\dots,\eta_r$ and $\zeta_1,\dots,\zeta_s)$ of random variables $(P(\eta_1,\dots,\eta_r,\zeta_1,\dots,\zeta_s) = P(\eta_1,\dots,\eta_r)P(\zeta_1,\dots,\zeta_s))$, then the provided reasoning is still applicable when denoting ensembles $(\eta_1,\dots,\eta_r)$ and $(\zeta_1,\dots,\zeta_s)$ as ξ_1 and ξ_2, respectively.

The property mentioned in Theorem 1.3 is a manifestation of the additivity principle, which was taken as a base principle in Section 1.1 and led to the logarithmic function (1.1.1). This property can be generalized to the case of several independent random variables $\xi_1, \ldots, \xi_n$, which yields,

$$H_{\xi_1 \ldots \xi_n} = \sum_{j=1}^{n} H_{\xi_j}, \tag{1.2.9}$$

and is easy to prove by an analogous method.

1.3 Conditional entropy. Hierarchical additivity

Let us generalize formulae (1.2.1), (1.2.3) to the case of conditional probabilities. Let $\xi_1, \ldots, \xi_n$ be random variables described by the joint distribution $P(\xi_1, \ldots, \xi_n)$. The conditional probabilities

$$P(\xi_k, \ldots, \xi_n \mid \xi_1, \ldots, \xi_{k-1}) = \frac{P(\xi_1, \ldots, \xi_n)}{P(\xi_1, \ldots, \xi_{k-1})} \quad (k \leqslant n)$$

are associated with the random conditional entropy

$$H(\xi_k, \ldots, \xi_n \mid \xi_1, \ldots, \xi_{k-1}) = -\ln P(\xi_k, \ldots, \xi_n \mid \xi_1, \ldots, \xi_{k-1}). \tag{1.3.1}$$

Let us introduce a special notation for the result of averaging (1.3.1) over $\xi_k, \ldots, \xi_n$:

$$H_{\xi_k \ldots \xi_n}(\mid \xi_1, \ldots \xi_{k-1}) = -\sum_{\xi_k \ldots \xi_n} P(\xi_k, \ldots, \xi_n \mid \xi_1, \ldots, \xi_{k-1}) \times \\ \times \ln P(\xi_k, \ldots, \xi_n \mid \xi_1, \ldots, \xi_{k-1}), \tag{1.3.2}$$

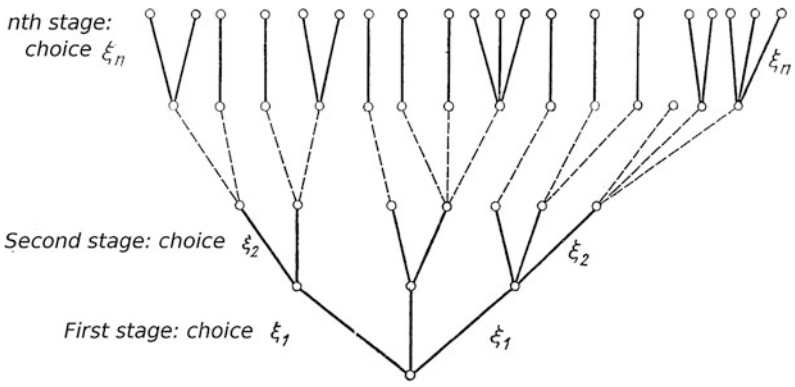

Fig. 1.1 Partitioning stages and the decision tree in the general case

and also for the result of total averaging:

$$H_{\xi_k,\dots,\xi_n|\xi_1,\dots,\xi_{k-1}} = \mathbb{E}[H(\xi_k,\dots,\xi_n \mid \xi_1,\dots,\xi_{k-1})]$$
$$= -\sum_{\xi_1\dots\xi_n} P(\xi_1\dots\xi_n)\ln P(\xi_k,\dots,\xi_n \mid \xi_1,\dots,\xi_{k-1}). \qquad (1.3.3)$$

If, in addition, we vary k and n, then we will form a large number of different entropies, conditional and non-conditional, random and non-random. They are related by identities that will be considered below.

Before we formulate the main hierarchical equality (1.3.4), we show how to introduce a hierarchical set of random variables $\xi_1,\dots,\xi_n$, even if there was just one random variable ξ initially.

Let ξ take one of M values with probabilities $P(\xi)$. The choice of one realization will be made in several stages. At the first stage, we indicate which subset (from a full ensemble of non-overlapping subsets $E_1,\dots,E_{m_1}$) the realization belongs to. Let ξ_1 be the index of such a subset. At the second stage, each subset is partitioned into smaller subsets $E_{\xi_1\xi_2}$. The second random variable ξ_2 points to which smaller subset the realization of the random variable belongs to. In turn, those smaller subsets are further partitioned until we obtain subsets consisting of a single element. Apparently, the number of nontrivial partitioning stages n cannot exceed $M-1$. We can juxtapose a fixed partitioning scheme with a 'decision tree' depicted on Figure 1.1. Further considerations will be associated with a particular selected 'tree'.

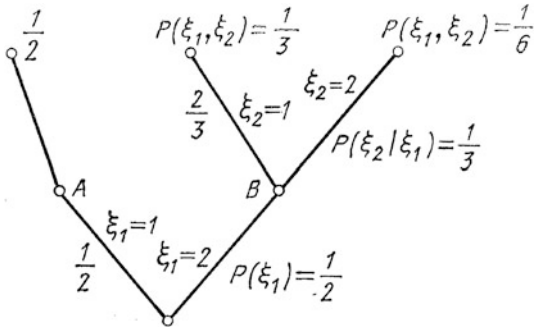

Fig. 1.2 A 'decision tree' for one particular example

In order to indicate a realization of ξ, it is necessary and sufficient to fix the assembly of realizations $(\xi_1,\xi_2,\dots,\xi_n)$. In order to indicate the 'node' of $(k+1)$-th stage, we need to specify values $\xi_1,\dots,\xi_k$. Then the value of ξ_{k+1} points to the branch we use to get out from this node.

As an example of a 'decision tree' we can give the simple 'tree' represented on Figure 1.2, which was considered by Shannon.

At each stage, the choice is associated with some uncertainty. Consider the entropy corresponding to this uncertainty. The first stage has one node, and the entropy of the choice is equal to H_{ξ_1}. Fixing ξ_1, we determine the node of the second stage.