Information Theory and Coding by Example

MARK KELBERT
YURI SUHOV

CAMBRIDGE

INFORMATION THEORY AND CODING BY EXAMPLE

MARK KELBERT
Swansea University, and Universidade de São Paulo

YURI SUHOV
University of Cambridge, and Universidade de São Paulo

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

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

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521769358

First published 2013

Printed in the United Kingdom by CPI Group Ltd. Croydon CR0 4YY

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

ISBN 978-0-521-76935-8 Hardback
ISBN 978-0-521-13988-5 Paperback

INFORMATION THEORY AND CODING BY EXAMPLE

This fundamental monograph introduces both the probabilistic and the algebraic aspects of information theory and coding. It has evolved from the authors' years of experience teaching at the undergraduate level, including several Cambridge Mathematical Tripos courses. The book provides relevant background material, a wide range of worked examples and clear solutions to problems from real exam papers. It is a valuable teaching aid for undergraduate and graduate students, or for researchers and engineers who want to grasp the basic principles.

MARK KELBERT is a Reader in Statistics in the Department of Mathematics at Swansea University. For many years he has also been associated with the Moscow Institute of Information Transmission Problems and the International Institute of Earthquake Prediction Theory and Mathematical Geophysics (Moscow).

YURI SUHOV is a Professor of Applied Probability in the Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge (Emeritus). He is also affiliated to the University of São Paulo in Brazil and to the Moscow Institute of Information Transmission Problems.

Preface

This book is partially based on the material covered in several Cambridge Mathematical Tripos courses: the third-year undergraduate courses *Information Theory* (which existed and evolved over the last four decades under slightly varied titles) and *Coding and Cryptography* (a much younger and simplified course avoiding cumbersome technicalities), and a number of more advanced Part III courses (Part III is a Cambridge equivalent to an MSc in Mathematics). The presentation revolves, essentially, around the following core concepts: (a) the entropy of a probability distribution as a measure of 'uncertainty' (and the entropy rate of a random process as a measure of 'variability' of its sample trajectories), and (b) coding as a means to measure and use redundancy in information generated by the process.

Thus, the contents of this book includes a more or less standard package of information-theoretical material which can be found nowadays in courses taught across the world, mainly at Computer Science and Electrical Engineering Departments and sometimes at Probability and/or Statistics Departments. What makes this book different is, first of all, a wide range of examples (a pattern that we followed from the onset of the series of textbooks *Probability and Statistics by Example* by the present authors, published by Cambridge University Press). Most of these examples are of a particular level adopted in Cambridge Mathematical Tripos exams. Therefore, our readers can make their own judgement about what level they have reached or want to reach.

The second difference between this book and the majority of other books on information theory or coding theory is that it covers both possible directions: probabilistic and algebraic. Typically, these lines of inquiry are presented in different monographs, textbooks and courses, often by people who work in different departments. It helped that the present authors had a long-time association with the Institute for Information Transmission Problems, a section of the Russian Academy of Sciences, Moscow, where the tradition of embracing a broad spectrum of problems was strongly encouraged. It suffices to list, among others,

the names of Roland Dobrushin, Raphail Khas'minsky, Mark Pinsker, Vladimir Blinovsky, Vyacheslav Prelov, Boris Tsybakov, Kamil Zigangirov (probability and statistics), Valentin Afanasiev, Leonid Bassalygo, Serguei Gelfand, Valery Goppa, Inna Grushko, Grigorii Kabatyansky, Grigorii Margulis, Yuri Sagalovich, Alexei Skorobogatov, Mikhail Tsfasman, Victor Zinov'yev, Victor Zyablov (algebra, combinatorics, geometry, number theory), who worked or continue to work there (at one time, all these were placed in a five-room floor of a converted building in the centre of Moscow). Importantly, the Cambridge mathematical tradition of teaching information-theoretical and coding-theoretical topics was developed along similar lines, initially by Peter Whittle (Probability and Optimisation) and later on by Charles Goldie (Probability), Richard Pinch (Algebra and Geometry), Tom Körner and Keith Carne (Analysis) and Tom Fisher (Number Theory).

We also would like to add that this book has been written by authors trained as mathematicians (and who remain still mathematicians to their bones), who nevertheless have a strong background in applications, with all the frustration that comes with such work: vagueness, imprecision, disputability (involving, inevitably, personal factors) and last – but by no means least – the costs of putting any mathematical idea – however beautiful – into practice. Still, they firmly believe that mathematisation is the mainstream road to survival and perfection in the modern competitive world, and therefore that Mathematics should be taken and studied seriously (but perhaps not beyond reason).

Both aforementioned concepts (entropy and codes) forming the base of the information-theoretical approach to random processes were introduced by Shannon in the 1940s, in a rather accomplished form, in his publications [139], [141]. Of course, entropy already existed in thermodynamics and was understood pretty well by Boltzmann and Gibbs more than a century ago, and codes have been in practical (and efficient) use for a very long time. But it was Shannon who fully recognised the role of these concepts and put them into a modern mathematical framework, although, not having the training of a professional mathematician, he did not always provide complete proofs of his constructions. [Maybe he did not bother.] In relevant sections we comment on some rather bizarre moments in the development of Shannon's relations with the mathematical community. Fortunately, it seems that this did not bother him much. [Unlike Boltzmann, who was particularly sensitive to outside comments and took them perhaps too close to his heart.] Shannon definitely understood the full value of his discoveries; in our view it puts him on equal footing with such towering figures in mathematics as Wiener and von Neumann.

It is fair to say that Shannon's name still dominates both the probabilistic and the algebraic direction in contemporary information and coding theory. This is quite extraordinary, given that we are talking of the contribution made by a person who

was active in this area more than 40 years ago. [Although on several advanced topics Shannon, probably, could have thought, re-phrasing Einstein's words: "Since mathematicians have invaded the theory of communication, I do not understand it myself anymore."]

During the years that passed after Shannon's inceptions and inventions, mathematics changed drastically, and so did electrical engineering, let alone computer science. Who could have foreseen such a development back in the 1940s and 1950s, as the great rivalry between Shannon's information-theoretical and Wiener's cybernetical approaches was emerging? In fact, the latter promised huge (even fantastic) benefits for the whole of humanity while the former only asserted that a modest goal of correcting transmission errors could be achieved within certain limits. Wiener's book [171] captivated the minds of 1950s and 1960s thinkers in practically all domains of intellectual activity. In particular, cybernetics became a serious political issue in the Soviet Union and its satellite countries: first it was declared "a bourgeois anti-scientific theory", then it was over-enthusiastically embraced. [A quotation from a 1953 critical review of cybernetics in a leading Soviet ideology journal *Problems of Philosophy* reads: "Imperialists are unable to resolve the controversies destroying the capitalist society. They can't prevent the imminent economical crisis. And so they try to find a solution not only in the frenzied arms race but also in ideological warfare. In their profound despair they resort to the help of pseudo-sciences that give them some glimmer of hope to prolong their survival." The 1954 edition of the Soviet *Concise Dictionary of Philosophy* printed in hundreds of thousands of copies defined cybernetics as a "reactionary pseudo-science which appeared in the USA after World War II and later spread across other capitalist countries: a kind of modern mechanicism." However, under pressure from top Soviet physicists who gained authority after successes of the Soviet nuclear programme, the same journal, *Problems of Philosophy*, had to print in 1955 an article proclaiming positive views on cybernetics. The authors of this article included Alexei Lyapunov and Sergei Sobolev, prominent Soviet mathematicians.]

Curiously, as was discovered in a recent biography on Wiener [35], there exist "secret [US] government documents that show how the FBI and the CIA pursued Wiener at the height of the Cold War to thwart his social activism and the growing influence of cybernetics at home and abroad." Interesting comparisons can be found in [65].

However, history went its own way. As Freeman Dyson put it in his review [41] of [35]: "[Shannon's theory] was mathematically elegant, clear, and easy to apply to practical problems of communication. It was far more user-friendly than cybernetics. It became the basis of a new discipline called 'information theory' ... [In modern times] electronic engineers learned information theory, the gospel according to Shannon, as part of their basic training, and cybernetics was forgotten."

Not quite forgotten, however: in the former Soviet Union there still exist at least seven functioning institutes or departments named after cybernetics: two in Moscow and two in Minsk, and one in each of Tallinn, Tbilisi, Tashkent and Kiev (the latter being a renowned centre of computer science in the whole of the former USSR). In the UK there are at least four departments, at the Universities of Bolton, Bradford, Hull and Reading, not counting various associations and societies. Across the world, cybernetics-related societies seem to flourish, displaying an assortment of names, from concise ones such as the Institute of the Method (Switzerland) or the Cybernetics Academy (Italy) to the Argentinian Association of the General Theory of Systems and Cybernetics, Buenos Aires. And we were delighted to discover the existence of the Cambridge Cybernetics Society (Belmont, CA, USA). By contrast, information theory figures only in a handful of institutions' names. Apparently, the old Shannon *vs.* Wiener dispute may not be over yet.

In any case, Wiener's personal reputation in mathematics remains rock solid: it suffices to name a few gems such as the Paley–Wiener theorem (created on Wiener's numerous visits to Cambridge), the Wiener–Hopf method and, of course, the Wiener process, particularly close to our hearts, to understand his true role in scientific research and applications. However, existing recollections of this giant of science depict an image of a complex and often troubled personality. (The title of the biography [35] is quite revealing but such views are disputed, e.g., in the review [107]. In this book we attempt to adopt a more tempered tone from the chapter on Wiener in [75], pp. 386–391.) On the other hand, available accounts of Shannon's life (as well as other fathers of information and coding theory, notably, Richard Hamming) give a consistent picture of a quiet, intelligent and humorous person. It is our hope that this fact will not present a hindrance for writing Shannon's biographies and that in future we will see as many books on Shannon as we see on Wiener.

As was said before, the purpose of this book is twofold: to provide a synthetic introduction both to probabilistic and algebraic aspects of the theory supported by a significant number of problems and examples, and to discuss a number of topics rarely presented in most mainstream books. Chapters 1–3 give an introduction into the basics of information theory and coding with some discussion spilling over to more modern topics. We concentrate on typical problems and examples [many of them originated in Cambridge courses] more than on providing a detailed presentation of the theory behind them. Chapter 4 gives a brief introduction into a variety of topics from information theory. Here the presentation is more concise and some important results are given without proofs.

Because the large part of the text stemmed from lecture notes and various solutions to class and exam problems, there are inevitable repetitions, multitudes of

notation and examples of pigeon English. We left many of them deliberately, feeling that they convey a live atmosphere during the teaching and examination process.

Two excellent books [52] and [36] had a particularly strong impact on our presentation. We feel that our long-term friendship with Charles Goldie played a role here, as well as YS's amicable acquaintance with Tom Cover. We also benefited from reading (and borrowing from) the books [18], [110], [130] and [98]. The warm hospitality at a number of programmes at the Isaac Newton Institute, University of Cambridge, in 2002–2010 should be acknowledged, particularly Stochastic Processes in Communication Sciences (January–July 2010). Various parts of the material have been discussed with colleagues in various institutions, first and foremost, the Institute for Information Transmission Problems and the Institute of Mathematical Geophysics and Earthquake Predictions, Moscow (where the authors have been loyal staff members for a long time). We would like to thank James Lawrence, from Statslab, University of Cambridge, for his kind help with figures.

References to PSE I and PSE II mean the books by the present authors *Probability and Statistics by Example*, Cambridge University Press, Volumes I and II. We adopted the style used in PSE II, presenting a large portion of the material through 'Worked Examples'. Most of these Worked Examples are stated as problems (and many of them originated from Cambridge Tripos Exam papers and keep their specific style and spirit).

Contents

1

Essentials of Information Theory

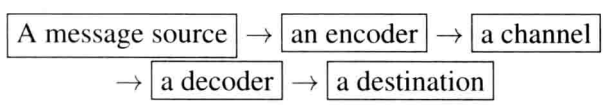
Throughout the book, the symbol $\mathbb{P}$ denotes various probability distributions. In particular, in Chapter 1, $\mathbb{P}$ refers to the probabilities for sequences of random variables characterising sources of information. As a rule, these are sequences of independent and identically distributed random variables or discrete-time Markov chains; namely, $\mathbb{P}(U_1 = u_1, \ldots, U_n = u_n)$ is the joint probability that random variables $U_1, \ldots, U_n$ take values $u_1, \ldots, u_n$, and $\mathbb{P}(V = v | U = u, W = w)$ is the conditional probability that a random variable V takes value v, given that random variables U and W take values u and w, respectively. Likewise, $\mathbb{E}$ denotes the expectation with respect to $\mathbb{P}$.

The symbols p and P are used to denote various probabilities (and probability-related objects) loosely. The symbol $\sharp A$ denotes the cardinality of a finite set A. The symbol $\mathbf{1}$ stands for an indicator function. We adopt the following notation and formal rules for logarithms: $\ln = \log_e$, $\log = \log_2$, and for all $b > 1$: $0 \cdot \log_b 0 = 0 \cdot \log_b \infty = 0$. Next, given $x > 0$, $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the maximal integer that is no larger than x and the minimal integer that is no less than x, respectively. Thus, $\lfloor x \rfloor \leq x \leq \lceil x \rceil$; equalities hold here when x is a positive integer ($\lfloor x \rfloor$ is called the integer part of x.)

The abbreviations LHS and RHS stand, respectively, for the left-hand side and the right-hand side of an equation.

1.1 Basic concepts. The Kraft inequality. Huffman's encoding

A typical scheme used in information transmission is as follows:



Example 1.1.1 (a) A message source: a Cambridge college choir.
(b) An encoder: a BBC recording unit. It translates the sound to a binary array and writes it to a CD track. The CD is then produced and put on the market.
(c) A channel: a customer buying a CD in England and mailing it to Australia. The channel is subject to 'noise': possible damage (mechanical, electrical, chemical, etc.) incurred during transmission (transportation).
(d) A decoder: a CD player in Australia.
(e) A destination: an audience in Australia.
(f) The goal: to ensure a high-quality sound despite damage.

In fact, a CD can sustain damage done by a needle while making a neat hole in it, or by a tiny drop of acid (you are not encouraged to make such an experiment!). In technical terms, typical goals of information transmission are:

(i) fast encoding of information,
(ii) easy transmission of encoded messages,
(iii) effective use of the channel available (i.e. maximum transfer of information per unit time),
(iv) fast decoding,
(v) correcting errors (as many as possible) introduced by noise in the channel.

As usual, these goals contradict each other, and one has to find an optimal solution. This is what the chapter is about. However, do not expect perfect solutions: the theory that follows aims mainly at providing knowledge of the basic principles. A final decision is always up to the individual (or group) responsible.

A large part of this section (and the whole of Chapter 1) will deal with *encoding* problems. The aims of encoding are:

(1) compressing data to reduce redundant information contained in a message,
(2) protecting the text from unauthorised users,
(3) enabling errors to be corrected.

We start by studying *sources* and *encoders*. A source emits a sequence of letters (or symbols),

$$u_1\, u_2\, \ldots\, u_n \ldots, \tag{1.1.1}$$

where $u_j \in I$, and $I(= I_m)$ is an m-element set often identified as $\{1, \ldots, m\}$ (a source alphabet). In the case of literary English, $m = 26 + 7$, 26 letters plus 7 punctuation symbols: . , : ; – ( ). (Sometimes one adds ? ! ‘ ’ and ”). Telegraph English corresponds to $m = 27$.

A common approach is to consider (1.1.1) as a *sample* from a random source, i.e. a sequence of random variables

$$U_1, U_2, \ldots, U_n, \ldots \tag{1.1.2}$$

and try to develop a theory for a reasonable class of such sequences.

Example 1.1.2 (a) The simplest example of a random source is a sequence of independent and identically distributed random variables (IID random variables):

$$\mathbb{P}(U_1 = u_1,\, U_2 = u_2, \ldots, U_k = u_k) = \prod_{j=1}^{k} p(u_j), \tag{1.1.3a}$$

where $p(u) = \mathbb{P}(U_j = u)$, $u \in I$, is the marginal distribution of a single variable. A random source with IID symbols is often called a *Bernoulli source*.

A particular case where $p(u)$ does not depend on $u \in U$ (and hence equals $1/m$) corresponds to the equiprobable Bernoulli source.

(b) A more general example is a Markov source where the symbols form a discrete-time Markov chain (DTMC):

$$\mathbb{P}(U_1 = u_1,\, U_2 = u_2, \ldots,\, U_k = u_k) = \lambda(u_1) \prod_{j=1}^{k-1} P(u_j, u_{j+1}), \tag{1.1.3b}$$

where $\lambda(u) = \mathbb{P}(U_1 = u)$, $u \in I$, are the initial probabilities and $P(u,u') = \mathbb{P}(U_{j+1} = u' | U_j = u)$, $u, u' \in I$, are transition probabilities. A Markov source is called *stationary* if $\mathbb{P}(U_j = u) = \lambda(u)$, $j \geq 1$, i.e. $\lambda = \{\lambda(u), u = 1, \ldots, m\}$ is an invariant row-vector for matrix $P = \{P(u,v)\}$: $\sum\limits_{u \in I} \lambda(u) P(u,v) = \lambda(v)$, $v \in I$, or, shortly, $\lambda P = \lambda$.

(c) A 'degenerated' example of a Markov source is where a source emits repeated symbols. Here,

$$\begin{gathered} \mathbb{P}(U_1 = U_2 = \cdots = U_k = u) = p(u), \;\; u \in I, \\ \mathbb{P}(U_k \neq U_{k'}) = 0, \;\; 1 \leq k < k', \end{gathered} \tag{1.1.3c}$$

where $0 \leq p(u) \leq 1$ and $\sum\limits_{u \in I} p(u) = 1$.

An initial piece of sequence (1.1.1)

$$\mathbf{u}^{(n)} = (u_1, u_2, \ldots, u_n) \;\; \text{or, more briefly,} \;\; \mathbf{u}^{(n)} = u_1 u_2 \ldots u_n$$

is called a (source) sample *n-string*, or *n-word* (in short, a string or a word), with digits from I, and is treated as a 'message'. Correspondingly, one considers a random n-string (a random message)

$$\mathbf{U}^{(n)} = (U_1, U_2, \ldots, U_n) \;\; \text{or, briefly,} \;\; \mathbf{U}^{(n)} = U_1 U_2 \ldots U_n.$$

An encoder (or coder) uses an alphabet $J (= J_q)$ which we typically write as $\{0, 1, \ldots, q-1\}$; usually the number of encoding symbols $q < m$ (or even $q \ll m$); in many cases $q = 2$ with $J = \{0,1\}$ (a binary coder). A *code* (also *coding*, or

encoding) is a map, f, that takes a symbol $u \in I$ into a finite string, $f(u) = x_1 \dots x_s$, with digits from J. In other words, f maps I into the set J^* of all possible strings:

$$f : I \to J^* = \bigcup_{s \geq 1} \big(J \times \cdots (s \text{ times}) \times J\big).$$

Strings $f(u)$ that are images, under f, of symbols $u \in I$ are called *codewords* (in code f). A code has (constant) length N if the value s (the length of a codeword) equals N for all codewords. A message $\mathbf{u}^{(n)} = u_1 u_2 \dots u_n$ is represented as a concatenation of codewords

$$f(\mathbf{u}^{(n)}) = f(u_1) f(u_2) \dots f(u_n);$$

it is again a string from J^*.

Definition 1.1.3 We say that a code is *lossless* if $u \neq u'$ implies that $f(u) \neq f(u')$. (That is, the map $f : I \to J^*$ is one-to-one.) A code is called *decipherable* if any string from J^* is the image of at most one message. A string x is a *prefix* in another string y if $y = xz$, i.e. y may be represented as a result of a concatenation of x and z. A code is *prefix-free* if no codeword is a prefix in any other codeword (e.g. a code of constant length is prefix-free).

A prefix-free code is decipherable, but not vice versa:

Example 1.1.4 A code with three source letters $1, 2, 3$ and the binary encoder alphabet $J = \{0, 1\}$ given by

$$f(1) = 0, \;\; f(2) = 01, \;\; f(3) = 011$$

is decipherable, but not prefix-free.

Theorem 1.1.5 (The Kraft inequality) *Given positive integers $s_1, \dots, s_m$, there exists a decipherable code $f : I \to J^*$, with codewords of lengths $s_1, \dots, s_m$, iff*

$$\sum_{i=1}^{m} q^{-s_i} \leq 1. \tag{1.1.4}$$

Furthermore, under condition (1.1.4) *there exists a prefix-free code with codewords of lengths $s_1, \dots, s_m$.*

Proof (I) Sufficiency. Let (1.1.4) hold. Our goal is to construct a prefix-free code with codewords of lengths $s_1, \dots, s_m$. Rewrite (1.1.4) as

$$\sum_{l=1}^{s} n_l q^{-l} \leq 1, \tag{1.1.5}$$

or

$$n_s q^{-s} \leq 1 - \sum_{l=1}^{s-1} n_l q^{-l},$$

where n_l is the number of codewords of length l and $s = \max s_i$. Equivalently,

$$n_s \leq q^s - n_1 q^{s-1} - \cdots - n_{s-1} q. \tag{1.1.6a}$$

Since $n_s \geq 0$, deduce that

$$n_{s-1} q \leq q^s - n_1 q^{s-1} - \cdots - n_{s-2} q^2,$$

or

$$n_{s-1} \leq q^{s-1} - n_1 q^{s-2} - \cdots - n_{s-2} q. \tag{1.1.6b}$$

Repeating this argument yields subsequently

$$\begin{array}{ccccc} n_{s-2} & \leq & q^{s-2} - n_1 q^{s-3} & - \ \ldots & -n_{s-3} q \\ \vdots & & \vdots & & \vdots \\ n_2 & \leq & q^2 - n_1 q & & \end{array} \tag{1.1.6.s$-$1}$$

$$n_1 \leq q. \tag{1.1.6.s}$$

Observe that actually either $n_{i+1} = 0$ or n_i is less than the RHS of the inequality, for all $i = 1, \ldots, s-1$ (by definition, $n_s \geq 1$ so that for $i = s-1$ the second possibility occurs). We can perform the following construction. First choose n_1 words of length 1, using distinct symbols from J: this is possible in view of (1.1.6.s). It leaves $(q - n_1)$ symbols unused; we can form $(q - n_1)q$ words of length 2 by appending a symbol to each. Choose n_2 codewords from these: we can do so in view of (1.1.6.s$-$1). We still have $q^2 - n_1 q - n_2$ words unused: form n_3 codewords, etc. In the course of the construction, no new word contains a previous codeword as a prefix. Hence, the code constructed is prefix-free.

(II) Necessity. Suppose there exists a decipherable code in J^* with codeword lengths $s_1, \ldots, s_m$. Set $s = \max s_i$ and observe that for any positive integer r

$$\left(q^{-s_1} + \cdots + q^{-s_m}\right)^r = \sum_{l=1}^{rs} b_l q^{-l}$$

where b_l is the number of ways r codewords can be put together to form a string of length l.

Because of decipherability, these strings must be distinct. Hence, we must have $b_l \leq q^l$, as q^l is the total number of l-strings. Then

$$\left(q^{-s_1} + \cdots + q^{-s_m}\right)^r \leq rs,$$

and

$$q^{-s_1} + \cdots + q^{-s_m} \leq r^{1/r} s^{1/r} = \exp\left[\frac{1}{r}(\log r + \log s)\right].$$

This is true for any r, so take $r \to \infty$. The RHS goes to 1. □

Remark 1.1.6 A given code obeying (1.1.4) is not necessarily decipherable.

Leon G. Kraft introduced inequality (1.1.4) in his MIT PhD thesis in 1949.

One of the principal aims of the theory is to find the 'best' (that is, the shortest) decipherable (or prefix-free) code. We now adopt a probabilistic point of view and assume that symbol $u \in I$ is emitted by a source with probability $p(u)$:

$$\mathbb{P}(U_k = u) = p(u).$$

[At this point, there is no need to specify a joint probability of more than one subsequently emitted symbol.]

Recall, given a code $f : I \mapsto J^*$, we encode a letter $i \in I$ by a prescribed codeword $f(i) = x_1 \ldots x_{s(i)}$ of length $s(i)$. For a random symbol, the generated codeword becomes a random string from J^*. When f is lossless, the probability of generating a given string as a codeword for a symbol is precisely $p(i)$ if the string coincides with $f(i)$ and 0 if there is no letter $i \in I$ with this property. If f is not one-to-one, the probability of a string equals the sum of terms $p(i)$ for which the codeword $f(i)$ equals this string. Then the length of a codeword becomes a *random variable*, S, with the probability distribution

$$\mathbb{P}(S = s) = \sum_{1 \leq i \leq m} \mathbf{1}(s(i) = s) p(i). \tag{1.1.7}$$

We are looking for a decipherable code that minimises the expected word-length:

$$\mathbb{E}S = \sum_{s \geq 1} s\mathbb{P}(S = s) = \sum_{i=1}^{m} s(i) p(i).$$

The following problem therefore arises:

$$\begin{aligned} &\text{minimise } g(s(1), \ldots, s(m)) = \mathbb{E}S \\ &\text{subject to } \sum_i q^{-s(i)} \leq 1 \text{ (Kraft)} \\ &\text{with } s(i) \text{ positive integers.} \end{aligned} \tag{1.1.8}$$