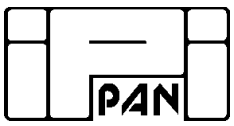
POLISH ACADEMY OF SCIENCES
INSTITUTE OF COMPUTER SCIENCE

Łukasz Dębowski

A Short Course in Universal Coding

Draft dated July 26, 2024



Contents

Preface

Information theory is a useful tool in the analysis of notions such as randomness, data compression, and prediction. It is intimately linked with mathematical foundations of computer science, probability, and statistics. It has many practical applications in statistical language modeling, machine learning, artificial intelligence, cryptography, bioinformatics, and also physics. In an introductory course, however, it may be good to focus on some setting of a moderate generality that both motivates the most important concepts and does not overwhelm with too many radiating developments.

In this textbook, I sketch basic results of information theory as regarded through the lens of universal coding. Universal coding consists in compressing and predicting an unknown random process. An example of such a process is an infinite sequence of randomly typed letters and spaces. This setting inspires excursions to foundations of computer science, probability, and statistics—understood as branches of mathematics. Consequently, the readers are faced with questions what computation is, what randomness is, and what learning is. These investigations have a practical aspect since we are interested in the best prediction of quite arbitrary strings of symbols. A particular case of such data are texts in natural language. Predicting them is a fundamental problem of artificial intelligence and computational linguistics, as witnessed by the advance of large statistical language models.

This textbook constitutes an updated one-semester course for STEM master students that introduces the most important topics and forks off my previous attempts to attack the subject. The main body of the book consists of fourteen lectures that assume little prior knowledge of probability calculus or theory of computation. The course progresses fast to fill in almost all knowledge gaps with rigorous reasonings. Only a few theorems are not followed by proofs—to sweep more complicated issues under the carpet. The exposition of many advanced subjects has been simplified down to the necessary essence while sacrificing stronger or more general results. These conceptual shortcuts concern measure theory, martingales, ergodic properties, and algorithmic information theory, in particular. Each lecture is accompanied with

the reading section that proposes starting pointers to the literature, where the missing demonstrations can be found as well. As duly expected, each lecture is followed by a list of traditional pen-and-paper exercises. What is less usual, to make a link with practical applications, I added a collection of programming tasks at the end of the book.

While writing this textbook, I adapted some material from my prior open access textbook [34]. That book has been quite popular on ResearchGate but I have not been satisfied with the text since I was clinically depressed while writing it. Moreover, I wanted to craft a simple-minded introduction to the more advanced monograph [36], which resumes some of my research and was composed for doctoral students in mathematics. There are also a few new topics here, drawn from my recent papers on universal prediction and language modeling. Let me hope that the readers will forgive me revamping previous works. *Repetitio est mater studiorum*, as my high school mathematics professor Olga Stande used to say.

One more remark. It was Andrey Kolmogorov who expressed an unusual opinion that information theory is more fundamental than the measure-theoretic approach to probability:

> Information theory must precede probability theory, and not be based on it. By the very essence of this discipline, the foundations of information theory have a finite combinatorial character. [...] The concepts of information theory as applied to infinite sequences give rise to very interesting investigations, which, without being indispensable as a basis of probability theory, can acquire a certain value in the investigation of the algorithmic side of mathematics as a whole. [27]

This statement should be particularly appreciated since Andrey Kolmogorov founded both the modern measure-theoretic probability calculus [83] and the algorithmic information theory [84]. In this textbook, I exercise Kolmogorov's idea that it is coding that motivates all later probabilistic and algorithmic constructions. The notions of information, probability, and computation are intertwined and it is hard to speak of one without mentioning the other.

Notations

We list some basic notations that we apply further.

- $\mathbb{N} := \{1, 2, 3, ...\}$ is the set of natural numbers without zero, $\mathbb{Z} := \{..., -1, 0, 1, ...\}$ is the set of integers, $\mathbb{Q} := \{p/q : p \in \mathbb{Z}, q \in \mathbb{N}\}$ is the set of rational numbers, and $\mathbb{R}$ is the set of real numbers. A set is called countable if its elements can be mapped one-to-one to a subset of natural numbers. Sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are countable but $\mathbb{R}$ is not.

- For a countable set $\mathbb{X}$, called an alphabet, $\mathbb{X}^n$ is the set of sequences of length n and $\mathbb{X}^+ := \bigcup_{n=1}^{\infty} \mathbb{X}^n$, called the Kleene plus, is the set of non-empty finite sequences. Symbol λ denotes the empty sequence, whereas $\mathbb{X}^0 := \{\lambda\}$. Then $\mathbb{X}^* := \mathbb{X}^0 \cup \mathbb{X}^+$, called the Kleene star, is the set of all finite sequences, also called strings. Set $\mathbb{X}^*$ is countable. Strings consisting of individually listed symbols are abbreviated as $x_j^k := (x_j, x_{j+1}, ..., x_k)$, where $j \leq k$ and $x_i \in \mathbb{X}$.

- Relation $A \subset B$ is the inclusion of sets. Symbols $A \cup B$, $A \cap B$, and $A \setminus B$ denote the union, the intersection, and the difference of sets A and B, respectively. Notation 2^A stands for the set of subsets of set A, called the power set of A. The cardinality of set A is denoted as $\# A$. Notations $[a, b] := \{r \in \mathbb{R} : a \leq r \leq b\}$ and $(a, b) := \{r \in \mathbb{R} : a < r < b\}$ stand for closed and open intervals of real numbers.

- We denote the length $|w| := k$ of a string $w \in \mathbb{X}^k$. The same notation for a real number denotes the absolute value, $|x| := x$ if $x \geq 0$ and $|x| := -x$ if $x < 0$. The floor and the ceiling functions are

$$\lceil x \rceil := \min \{y \in \mathbb{Z} : y \geq x\}, \qquad \lfloor x \rfloor := \max \{y \in \mathbb{Z} : y \leq x\}.$$

  The binary and the natural logarithm are

$$y = \log x \iff 2^y = x, \qquad y = \ln x \iff \exp y = x.$$

  For a proposition ϕ, the indicator function is $\mathbf{1}\{\phi\} := 1$ if ϕ is true and $\mathbf{1}\{\phi\} := 0$ if ϕ is false.

Chapter 1

Codes

General codes. Non-singular codes. Uniquely decodable or instantaneous codes. Comma-separated codes. Fixed-length codes. Prefix-free and suffix-free codes. Binary codes for natural numbers. Turning non-singular codes into prefix-free codes. Expected code length. Huffman codes.

The founding concept of information theory is as simple as the notion of codes. The readers may be familiar with many practical codes such as the Morse code, the Braille alphabet, the ISBN numbers for books, the PESEL numbers for citizens of Poland, the DNA code for aminoacids, etc. In general, a code is a mapping between a set of discrete objects and the set of strings of symbols from a fixed alphabet. The most popular choice is to take strings of digits, binary digits in particular. The coded objects can be also strings such as sequences of letters from the Roman alphabet. From this point of view, translations of texts between two different human languages can be also considered codes. In fact, any object that is processed by modern computers, be it an image or a tune, takes form of a binary code word at some stage of information processing.

Let us repeat that the general idea of coding consists in representing arbitrary objects from a countable set of distinct possibilities—such as letters, words, sentences, or whole finite texts—as unique finite sequences of binary digits. With the advent of personal computers and mobile devices, this idea seems as transparent as the idea of the alphabet but it took some effort to discover its profound implications for foundations of mathematics. Without coding, computer science and statistics would not be either thinkable or feasible. We dare to say that the results of the invention of coding are comparable to the implications of the invention of the alphabet.

In physics of complex systems, a spontaneous transformation of a continuous system into a system governed by discrete signals was called ritualization. Quite a few such ritualizations have been recognized: genes and proteins, human language, written numbers, and coined money. It is a good question how ritualization emerges in general. There can be different degrees of ritualization of a given system. In particular, codified law and formal mathematics can be considered more ritualized subsystems within the previously partly ritualized system of human language. Seen from this perspective, the abstract idea of coding comes as a relatively late sort of ritualization.

Let us make a more systematic exposition of the concept of coding. The object of interest of coding theory are functions, called codes, that map elements of a countable set $\mathbb{X}$ into finite sequences over a countable set $\mathbb{Y}$, called strings. The set of strings is denoted as $\mathbb{Y}^* := \mathbb{Y}^0 \cup \mathbb{Y}^+$, where $\mathbb{Y}^+ := \bigcup_{n=1}^{\infty} \mathbb{Y}^n$, $\mathbb{Y}^0 := \{\lambda\}$, and λ is the empty string. Set $\mathbb{Y}$ is called an alphabet. Of a special interest are binary codes, i.e., codes for which the output alphabet $\mathbb{Y}$ is the set the binary digits $\{0, 1\}$, succinctly called bits. In information theory, we often restrict ourselves to binary codes for two reasons. The first one is the ease of computer processing. The second one is to have some simple fixed unit of the amount of information. On the other hand, the input set $\mathbb{X}$ consists typically of letters, digits, or even strings of symbols, such as words in natural language.

Let us proceed to formal definitions. The first definition formalizes what we have said so far.

Definition 1.1 (code) *For two countable sets $\mathbb{X}$ and $\mathbb{Y}$, a function $B : \mathbb{X} \to \mathbb{Y}^*$ is called a code.*

Strings $B(x)$ will be called code words.

Codes are primarily used to represent individual entities being elements of the input set as preferably distinct strings over the output alphabet. Therefore, the following property is desired in the first step.

Definition 1.2 (non-singular code) *A code $B : \mathbb{X} \to \mathbb{Y}^*$ is called non-singular if for $B(x) = B(x')$ for $x, x' \in \mathbb{X}$ implies $x = x'$.*

To recall, more generally, an injection is a function $f : \mathbb{X} \to \mathbb{Y}$ such that $f(x) \neq f(x')$ for any $x, x' \in \mathbb{X}$ where $x \neq x'$. Thus, a non-singular code is simply a code that is an injection.

Example 1.3 *An example of a non-singular code:*

symbol x:	*code word $B(x)$:*
a	*0*
b	*1*
c	*10*
d	*11*

Example 1.4 (Morse code) *The international Morse code, applied in telegraphs, is a non-singular binary code for letters of the Latin alphabet that consists of long signals (dahs) and short signals (dits).*

letter:	*Morse code:*
A	$\cdot\ -$
B	$-\ \cdot\ \cdot\ \cdot$
C	$-\ \cdot\ -\ \cdot$
D	$-\ \cdot\ \cdot$
E	$\cdot$
...	...

The main practical purpose of coding is to transmit some representations of strings written with symbols from an input alphabet through a digital device which processes only strings consisting of symbols from a smaller output alphabet. Thus the idea of a particularly good code is that we should be able to reconstruct coded symbols x_i from the concatenation of their code words $B(x_i)$. The concatenation of code words is formally called the code extension.

Definition 1.5 (code extension) *For a code $B : \mathbb{X} \to \mathbb{Y}^*$ we define its extension $B^* : \mathbb{X}^* \to \mathbb{Y}^*$ as concatenation*

$$B^*(x_1, x_2, ..., x_n) := B(x_1)B(x_2)...B(x_n), \tag{1.1}$$

where $x_i \in \mathbb{X}$.

The following condition of unique decodability is desired.

Definition 1.6 (uniquely decodable code) *A code $B : \mathbb{X} \to \mathbb{Y}^*$ is called uniquely decodable if its extension $B^* : \mathbb{X}^* \to \mathbb{Y}^*$ is non-singular.*

Uniquely decodable codes are important both in theory and applications. The condition of unique decodability is stronger than non-singularity.

Example 1.7 *The non-singular code given in Example 1.3 is not uniquely decodable because $B(ba) = 10 = B(c)$.*

Example 1.8 *However, this code is uniquely decodable:*

symbol x*:*	*code word* $B(x)$*:*
a	$0c$
b	$1c$
c	$10c$
d	$11c$

The above code is a special case of a more general construction called a comma-separated code, which is a certain general recipe for a uniquely decodable code.

Definition 1.9 (comma-separated code) *Let* $c \notin \mathbb{Y}$*. Code* $B : \mathbb{X} \to (\mathbb{Y} \cup \{c\})^*$ *is called comma-separated if for each* $x \in \mathbb{X}$ *there exists a string* $w \in \mathbb{Y}^*$ *such that* $B(x) = wc$*. Symbol* c *is called the comma.*

Theorem 1.10 *Non-singular comma-separated codes are uniquely decodable.*

Proof: For a non-singular comma-separated code B, let us decompose $B(x) = \phi(x)c$. We first observe that $B(x_1)...B(x_n) = B(y_1)...B(y_m)$ holds only if $n = m$ (the same number of c's on both sides of equality) and $\phi(x_i) = \phi(y_i)$ for $i \in \{1, 2, ..., n\}$. Next, we observe that function ϕ is a non-singular code. Hence string $B(x_1)...B(x_n)$ may be only the image of $(x_1, ..., x_n)$ under the mapping B^*. This means that code B is uniquely decodable. □

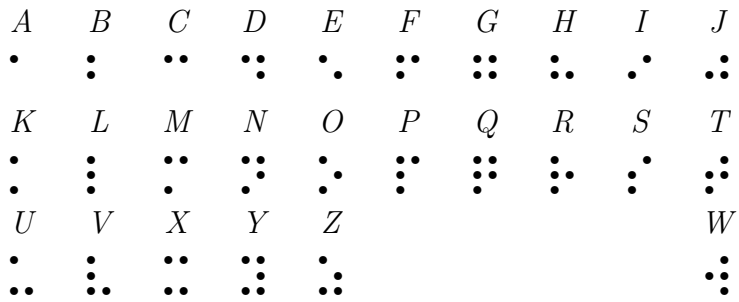
Another recipe for producing a uniquely decodable code is to restrict the length of code words.

Definition 1.11 (fixed-length code) *Let* n *be a fixed natural number. Code* $B : \mathbb{X} \to \mathbb{Y}^n$ *is called a fixed-length code.*

Example 1.12 *An example of a fixed-length code:*

symbol x*:*	*code word* $B(x)$*:*
a	00
b	01
c	10
d	11

Example 1.13 (Braille alphabet) *The Braille alphabet, used by the blind, is a fixed-length binary code for letters of the Latin alphabet. Each letter code consists of six slots which are filled with a raised dot or left blank.*



Example 1.14 (ASCII code) *The ASCII code, applied in computers, is a fixed-length binary code for letters of the Latin alphabet, Arabic digits, punctuation, and some additional symbols (128 symbols in total).*

symbol:	*ASCII code:*
...	...
A	*100 0001*
B	*100 0010*
C	*100 0011*
D	*100 0100*
E	*100 0101*
...	...

Theorem 1.15 *Non-singular fixed-length codes are uniquely decodable.*

Proof: Consider a fixed-length code B. We observe that $B(x_1)...B(x_n) = B(y_1)...B(y_m)$ holds only if $n = m$ (the same length of strings on both sides of equality) and $B(x_i) = B(y_i)$ for $i \in \{1, 2, ..., n\}$. Because B is a non-singular code, string $B(x_1)...B(x_n)$ may be only the image of $(x_1, ..., x_n)$ under the mapping B^*. Hence, code B is uniquely decodable. □

Yet another way to produce a uniquely decodable code is to require that no code word is a prefix or a suffix of another code word. There are two mirror-like definitions.

Definition 1.16 (prefix-free code) *A code B is called prefix-free if no code word $B(x)$ is a prefix of another code word $B(y)$, i.e., $B(y) = B(x)u$ for a string $u \in \mathbb{Y}^*$ implies $x = y$.*

Definition 1.17 (suffix-free code) *A code B is called suffix-free if no code word $B(x)$ is a suffix of another code word $B(y)$, i.e., $B(y) = uB(x)$ for a string $u \in \mathbb{Y}^*$ implies $x = y$.*

Example 1.18 *Codes in Examples 1.8 and 1.12 are prefix-free. Moreover, the code in Example 1.12 is also suffix-free.*

Example 1.19 *A code which is prefix-free but not suffix-free:*

symbol x:	*code word $B(x)$:*
a	*10*
b	*0*
c	*110*
d	*111*

Example 1.20 *A code which is suffix-free but not prefix-free:*

symbol x:	*code word $B(x)$:*
a	*01*
b	*0*
c	*011*
d	*111*

Example 1.21 (Unicode) *The Unicode Standard is a standard for encoding texts written down according to all presently used and many historical writing systems of the world. The Unicode Standard defines 144 697 characters and several encoding systems. The most popular one, UTF-8, is a prefix-free code which applies one byte (i.e., eight bit) code words for ASCII symbols and up to four bytes for other symbols. Any ASCII text is also a UTF-8 text.*

Theorem 1.22 *Prefix-free and suffix-free codes are uniquely decodable.*

Proof: Without loss of generality we restrict ourselves to prefix-free codes. The proof for suffix-free codes is mirror-like. Let B be a prefix-free code and assume that $B(x_1)...B(x_n) = B(y_1)...B(y_m)$. By the prefix-free property the initial segments $B(x_1)$ and $B(y_1)$ must match exactly and $x_1 = y_1$. The analogous argument applied by induction yields $x_i = y_i$ for $i \in \{2, .., n\}$ and $n = m$. Thus code B is uniquely decodable. □

In the proof of the above theorem, we have shown that when we read the concatenation of code words from left to right, we immediately know where are the boundaries between the code words. For this reason, prefix-free codes are also called instantaneous codes.

An important property of a code that we may like to optimize is its length. Let $|w|$ denote the length of a string $w \in \mathbb{Y}^*$, measured in symbols of alphabet $\mathbb{Y}$, i.e.,

$$|w| = n \iff w \in \mathbb{Y}^n. \tag{1.2}$$

To conveniently operate with lengths of code words, let us also denote the floor and the ceiling functions of real numbers as

$$\lceil x \rceil := \min \{y \in \mathbb{Z} : y \geq x\}, \tag{1.3}$$
$$\lfloor x \rfloor := \max \{y \in \mathbb{Z} : y \leq x\}. \tag{1.4}$$

The above notation is easier to remember than the old-fashioned notation $[x] := \lfloor x \rfloor$ for the floor function.

As an example, we will inspect several useful codes for natural numbers and we will evaluate their lengths.

Example 1.23 (military code) *To produce the military order of binary strings, we first sort strings according to their length and then alphabetically:*

$$\lambda, 0, 1, 00, 01, 10, 11, 000, ... \tag{1.5}$$

Subsequently, we assign consecutive strings to consecutive natural numbers to obtain the military code for natural numbers:

number n*:*	*code word* $\mathrm{mil}(n)$*:*
1	λ
2	*0*
3	*1*
4	*00*
5	*01*
6	*10*
7	*11*
8	*000*
...	...

The military code $\mathrm{mil} : \mathbb{N} \to \{0,1\}^*$ *is non-singular but not uniquely decodable. Its length is* $|\mathrm{mil}(n)| = \lfloor \log n \rfloor$*, where* $\log n$ *is the binary logarithm of number* n*.*

Example 1.24 (binary expansion) *The standard binary expansion of a natural number* n *will be written as* $\mathrm{bin}(n) := 1\,\mathrm{mil}(n)$*:*

number n:	*code word* $\mathrm{bin}(n)$*:*
1	*1*
2	*10*
3	*11*
4	*100*
...	...

Code $\mathrm{bin} : \mathbb{N} \to \{0,1\}^*$ *is non-singular but not uniquely decodable. Its length is* $|\mathrm{bin}(n)| = \lfloor \log n \rfloor + 1$*. Code* bin *is also called the Elias beta code.*

Example 1.25 (leading zeros) *Let us define a fixed-length code* $bin_n(k) : \{0,1,...,n-1\} \to \{0,1\}^*$ *as*

$$\mathrm{bin}_n(k-1) := 0^{|\mathrm{mil}(n)|-|\mathrm{mil}(k)|}1\,\mathrm{mil}(k), \tag{1.6}$$

i.e., we add leading zeros to obtain a code word of length $\mathrm{mil}(n)+1$*. For example:*

number n:	*code word* $\mathrm{bin}_4(n)$*:*
0	*00*
1	*01*
2	*10*
3	*11*

A simple example of a prefix-free code for arbitrarily large natural numbers is as follows

Example 1.26 (unary code) *The unary prefix-free code for natural numbers is* $\mathrm{una}(n) := 0^{n-1}1$*:*

number n:	*code word* $\mathrm{una}(n)$*:*
1	*1*
2	*01*
3	*001*
4	*0001*
...	...

Code $\mathrm{una} : \mathbb{N} \to \{0,1\}^*$ *is prefix-free. Its length is* $|\mathrm{una}(n)| = n$*. The unary code is also called the Elias alpha code.*

Now we will analyze how to turn a non-singular code into a prefix-free code and how much this operation costs.

Theorem 1.27 *Consider a non-singular code $B : \mathbb{X} \to \{0,1\}^*$ and a prefix-free code $D : \mathbb{N} \to \{0,1\}^*$. The code $E : \mathbb{X} \to \{0,1\}^*$ given as*

$$E(x) := D(|B(x)| + 1)B(x) \tag{1.7}$$

is prefix-free.

Proof: Suppose that $E(x)$ is a prefix of $E(y)$. Since code D is prefix-free then $D(|B(x)|+1) = D(|B(y)|+1)$ and consequently $|B(x)| = |B(y)|$. Hence $E(x) = E(y)$ and consequently $B(x) = B(y)$. Since code B is non-singular, we deduce $x = y$, i.e., code E is prefix-free. □

In this way, we can improve the non-singular code $\mathrm{mil} : \mathbb{N} \to \{0,1\}^*$ as the prefix-free codes

$$\mathrm{una}'(n) := \mathrm{una}(|\mathrm{mil}(n)| + 1)\,\mathrm{mil}(n), \tag{1.8}$$
$$\mathrm{una}''(n) := \mathrm{una}'(|\mathrm{mil}(n)| + 1)\,\mathrm{mil}(n). \tag{1.9}$$

Code una' is called the Elias gamma code. Code una'' is called the Elias delta code. The lengths of these codes are:

$$|\mathrm{una}'(n)| = 2\,\lfloor \log n \rfloor + 1, \tag{1.10}$$
$$|\mathrm{una}''(n)| = \lfloor \log n \rfloor + 2\,\lfloor \log(\lfloor \log n \rfloor + 1) \rfloor + 2. \tag{1.11}$$

Thus the cost of turning a non-singular code into a prefix-free code becomes negligible for very long code words:

$$\lim_{n\to\infty} \frac{|\mathrm{una}''(n)|}{|\mathrm{mil}(n)|} = 1. \tag{1.12}$$

In the second turn we may ask what is the shortest code to encode a given set of symbols, where the symbols appear with given probabilities. Formally, we introduce discrete probability distributions.

Definition 1.28 (probability distribution) *For a countable set $\mathbb{X}$, a discrete probability distribution is a function $p : \mathbb{X} \to [0,1]$ such that $p(x) \geq 0$ and $\sum_{x\in\mathbb{X}} p(x) = 1$.*

If $p(x)$ is the relative frequency of symbol x in a plain text then the following expected code length is the average length of a code word in the encoded text.

Definition 1.29 (expected code length) *The expected code length is*

$$\ell(p|B) := \sum_{x\in\mathbb{X}} p(x)\,|B(x)|\,. \tag{1.13}$$

Example 1.30 *Consider the following distribution and a code:*

symbol x*:*	$p(x)$*:*	*code word* $B(x)$*:*
a	$1/2$	$0c$
b	$1/6$	$1c$
c	$1/6$	$10c$
d	$1/6$	$11c$

We have $\ell(p|B) = 2 \cdot \frac{1}{2} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} = 2\frac{1}{3}$.

We are interested in codes that minimize the expected code length for a given probability distribution. In this regard, both comma-separated codes and fixed-length codes have advantages and drawbacks. If certain symbols appear more often than others then comma-separated codes allow to code them as shorter strings and thus to spare space. On the other hand, if all symbols are equally probable then a fixed-length code without a comma occupies less space than the same code with a comma.

The drawbacks of prefix-free codes cannot be pointed out easily. In the following, we will show that for a finite alphabet $\mathbb{X}$ and a distribution $p : \mathbb{X} \to [0, 1]$, there is a specific code, called the Huffman code, which is prefix-free and minimizes the expected length $\ell(p|B)$ among all prefix-free codes. To introduce this code we need first to uncover a relationship between prefix-free codes and binary trees.

Definition 1.31 (binary tree) *A binary tree is a directed acyclic connected graph where each node has at most two children nodes (left and/or right one) and at most one parent node. The node which has no parents is called the root node. The nodes which have no children are called leaf nodes. We assume that edges to the left children are labeled with* 0*'s whereas edges to the right children are labeled with* 1*'s. Moreover, some nodes may be labeled with some symbols as well.*

Definition 1.32 (path) *We say that a binary tree contains a path* $w \in \{0, 1\}^*$ *if there is a sequence of edges starting from the root node and labeled with the consecutive symbols of* w*. We say that the path ends with symbol* $a \in \mathbb{X}$ *if the last edge of the sequence ends in a node labeled with symbol* a*.*

Definition 1.33 (code tree) *The code tree for a code* $B : \mathbb{X} \to \{0, 1\}^*$ *is a labeled binary tree which contains a path* w *if and only if* $B(a) = w$ *for some* $a \in \mathbb{X}$*, and in that case we require that path* w *ends with symbol* a*.*

Example 1.34 *Consider codes:*

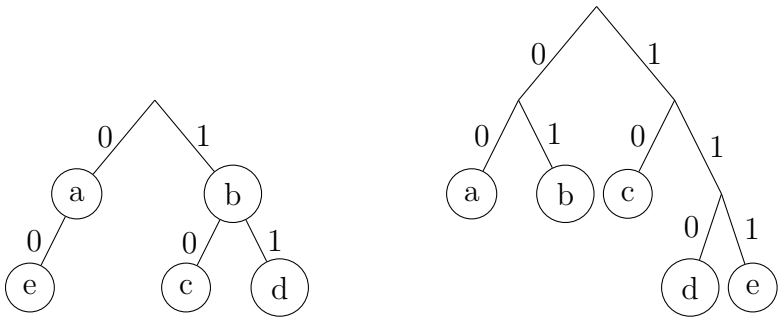


Figure 1.1: The code trees for the codes from Example 1.34.

symbol x:	*code word $B(x)$:*	*code word $D(x)$:*
a	*0*	*00*
b	*1*	*01*
c	*10*	*10*
d	*11*	*110*
e	*00*	*111*

The code trees for these codes are depicted in Figure 1.1.

It is easy to observe the following fact.

Theorem 1.35 *There is a one-to-one correspondence between binary codes and code trees. Moreover, a code is prefix-free if and only if the leaf nodes are the only nodes labeled.*

In the next step, we will add some weights to the code trees, which stem from the distribution of symbols.

Definition 1.36 (weighted code tree) *The weighted code tree for a prefix code $B : \mathbb{X} \to \{0,1\}^*$ and a probability distribution $p : \mathbb{X} \to [0,1]$ is the code tree for code B where the nodes are enhanced with the following weights: (1) for a leaf node with symbol a, we add weight $p(a)$, (2) to other (internal) nodes, we ascribe weights equal to the sum of weights of their children.*

Example 1.37 *Consider this distribution and the code D from Example 1.34:*

symbol x:	*$p(x)$:*	*code word $D(x)$:*
a	0.2	*00*
b	0.3	*01*
c	0.1	*10*
d	0.2	*110*
e	0.2	*111*

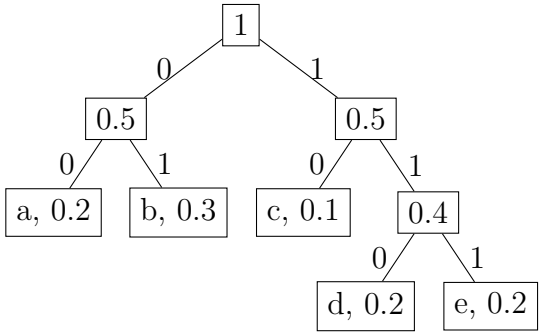


Figure 1.2: The weighted code tree for Example 1.37.

The weighted code tree is depicted in Figure 1.2.

Now we can define the Huffman code.

Definition 1.38 (Huffman code) *Let $\mathbb{X}$ be a finite set. The Huffman code for a probability distribution $p : \mathbb{X} \to [0, 1]$ is a prefix-free code whose weighted code tree is constructed by the following algorithm:*

1. *Create a leaf node with weight $p(x)$ for each symbol x and make a list of these nodes.*
2. *While there is more than one node in the list:*
   (a) *Remove two nodes of the lowest weight from the list.*
   (b) *Create a new internal node with these two nodes as children and with weight equal to the sum of the two nodes' weights.*
   (c) *Add the new node to the list.*
3. *The remaining node is the root node and the tree is complete.*

Example 1.39 *The Huffman code for the distribution from Example 1.37 is:*

symbol x:	*$p(x)$:*	*Huffman code $B(x)$:*
a	0.2	*00*
b	0.3	*10*
c	0.1	*110*
d	0.2	*111*
e	0.2	*01*

The corresponding Huffman code tree is depicted in Figure 1.3.

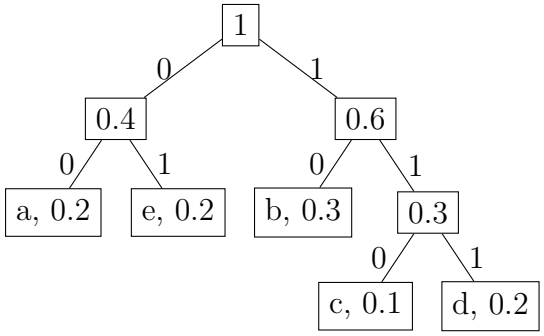


Figure 1.3: The Huffman code tree for Example 1.39.

It can be proved that no prefix-free code fares better than the Huffman code if the probability distribution is fixed.

Theorem 1.40 *For a fixed probability distribution $p : \mathbb{X} \to [0,1]$, the Huffman code achieves the minimum expected length $\ell(p|B)$ across prefix-free codes.*

Proof: Let us fix distribution $p : \mathbb{X} \to [0,1]$. A prefix-free code B and its corresponding tree will be called optimal if $\ell(p|B)$ achieves the minimum across prefix-free codes. We will use the this fact:

(H) Consider two symbols x and y with the smallest probabilities. Then there is an optimal code D such that these two symbols are sibling leaves in the lowest level of D's code tree.

To prove fact (H), we observe the following. Every internal node in a code tree for an optimal code must have two children. (Surely, if some internal node had only a single child, we might discard this node.) Then let B be an optimal code and let symbols a and b be two siblings at the maximal depth of B's code tree. Assume without loss of generality that $p(x) \leq p(y)$ and $p(a) \leq p(b)$. We have $p(x) \leq p(a)$, $p(y) \leq p(b)$, $|B(a)| \geq |B(x)|$, and $|B(b)| \geq |B(y)|$. Now let D's code tree differ from the B's code tree by switching $a \leftrightarrow x$ and $b \leftrightarrow y$. Then we obtain

$$\begin{aligned}\ell(p|D) - \ell(p|B) &= -p(x)\,|B(x)| - p(a)\,|B(a)| + p(a)\,|B(x)| + p(x)\,|B(a)| \\ &\quad - p(y)\,|B(y)| - p(b)\,|B(b)| + p(b)\,|B(y)| + p(y)\,|B(b)| \\ &= (p(a) - p(x))(|B(x)| - |B(a)|) \\ &\quad + (p(b) - p(y))(|B(y)| - |B(b)|) \leq 0. \qquad (1.14)\end{aligned}$$

Hence code D is also optimal.

Now we will proceed by induction on the number of symbols in the alphabet $\mathbb{X}$. If $\mathbb{X}$ contains only two symbols, then Huffman code is optimal. In the second step, we assume that Huffman code is optimal for $n-1$ symbols and we prove its optimality for n symbols. Let D be an optimal code for n symbols. By fact (H), without loss of generality, we may assume that symbols x and y having the smallest probabilities occupy two sibling leaves in the lowest level of D's code tree. Then from the weighted code tree of D we construct a code D' for $n-1$ symbols by removing nodes with symbols x and y and ascribing a symbol z to its parent node. Hence we have

$$\ell(p'|D') = \ell(p|D) - p(x) - p(y), \tag{1.15}$$

where $p'(z) := p(x) + p(y)$ and $p'(u) := p(u)$ if $u \notin \{x, y\}$. On the other hand, let B' be the Huffman code for p' and let B be the code constructed from B' by adding leaves with symbols x and y to the node with symbol z. By construction, code B is the Huffman code for p. We have

$$\ell(p'|B') = \ell(p|B) - p(x) - p(y). \tag{1.16}$$

Because $\ell(p'|B') \leq \ell(p'|D')$ by optimality of Huffman code B', we obtain $\ell(p|B) \leq \ell(p|D)$. Hence Huffman code B is also optimal. □

***

To recapitulate, this chapter concerned the idea of coding. In particular, we have learned about various non-singular and prefix-free codes. We have also minimized the average length of the prefix-free code, which yields the Huffman code. In Chapter 2, we will see that uniquely decodable and prefix-free codes satisfy some important inequalities and can be connected to a functional of a probability distribution called the Shannon entropy.

Further reading

The idea of coding was discovered gradually, with important examples preceding formal general definitions. The binary expansion for natural numbers was invented by Gottfried Leibniz in 1689 [89], who was inspired by binary hexagrams of the Chinese divination text *Yijing*. Some famous binary codes for letters of the Latin alphabet were proposed later by Samuel Morse and Louis Braille, both around 1837. Modern uniquely decodable codes bear from the seminal works by Claude Shannon [114, 115], who laid foundations of information theory. The Huffman code was invented by David Huffman

[71]. August Sardinas and George Patterson [113] constructed an algorithm for checking whether a given code is uniquely decodable. Various codes for natural numbers were studied by Peter Elias [44]. The most popular contemporary textbooks on information theory, which contain also an exposition of coding theory, are by Thomas Cover and Joy Thomas [26] and by Imre Csiszár and János Körner [30]. The idea of ritualization, the phase transition of a continuous system into a system governed by discrete signals, has been proposed by Rainer Feistel and Werner Ebeling [49].

Thinking exercises

1. Which of the following codes are prefix-free? Which of these codes cannot be Huffman codes for any probability distribution?

   (a) $\{0, 01, 1\}$,
   (b) $\{01, 101, 11\}$,
   (c) $\{0, 10, 110\}$,
   (d) $\{001, 01, 1\}$,
   (e) $\{01, 10, 11\}$,
   (f) $\{0, 10, 11\}$,
   (g) $\{00, 010, 110, 11\}$,
   (h) $\{00, 010, 10, 11\}$.

2. Find the Huffman codes for these distributions:

   (a)

symbol x:	$p(x)$:
a	1/12
b	1/6
c	1/4
d	1/3
e	1/6

   (b)

symbol x:	$p(x)$:
a	1/11
b	3/11
c	2/11
d	1/11
e	3/11
f	1/11

(c)

symbol x:	$p(x)$:
a	1/7
b	1/2
c	1/6
d	3/42
e	5/42

3. *Elias omega code:* As we have discussed the standard binary expansion is an example of a code for natural numbers which is not prefix-free. We can correct this code to make it prefix-free by prepending the binary expansion with a recursive representation of its length. If we repeat this procedure until its natural limit, we obtain the Elias omega code [44].

   The algorithm for the Elias omega encoding is as follows:

   (a) Put 0 at the end of the code.

   (b) If the coded number n is 1 then stop. Else, write the binary representation $\text{bin}(n)$ of the coded number n before the code.

   (c) Repeat the previous step with the coded number n equal to the number of digits written in the previous step minus 1.

   In this way we obtain the following correspondence:

number n:	code word:
1	0
2	10 0
3	11 0
4	10 100 0
5	10 101 0
6	10 110 0
7	10 111 0
8	11 1000 0
...	...

   Find the algorithm for decoding the Elias omega code.

4. *Complete code:* A uniquely decodable code $B : \mathbb{X} \to \{0,1\}^*$ is called complete if

$$\sum_{x \in \mathbb{X}} 2^{-|B(x)|} = 1. \tag{1.17}$$

Show that Huffman codes are complete. Are codes una(n), una$'(n)$, and una$''(n)$ complete? Is the Elias omega code complete?

5. Assume that $\mathbb{X}$ is finite and code $B : \mathbb{X} \to \{0,1\}^*$ is prefix-free and complete. Show that for any sequence of bits $(b_i)_{i\in\mathbb{N}}$ where $b_i \in \{0,1\}$ there exists a unique sequence $(x_i)_{i\in\mathbb{N}}$ where $x_i \in \mathbb{X}$ such that

$$B(x_1)B(x_2)B(x_3)... = b_1b_2b_3... \tag{1.18}$$

Is it true as well if alphabet $\mathbb{X}$ is infinite?

6. *Fix-free code:* A code is called is called fix-free if it is both suffix-free and prefix-free. Here is a code which is fix-free and complete [56]:

symbol x:	code word $B(x)$:
a	01
b	000
c	100
d	110
e	111
f	0010
g	0011
h	1010
i	1011

Find a few other examples of codes that are fix-free and complete.

Chapter 2

Inequalities

Kraft inequality. Kraft converse. Shannon-Fano code. Convex and concave functions. Jensen inequality. Shannon entropy. Kullback-Leibler divergence. Source coding inequality. Markov inequality. Barron inequality.

In a mathematical theory, one-sided inequalities play an important role by establishing some impossibility results. Sometimes, such inequalities can be chained into sandwich bounds. These often state asymptotic equivalence of the compared quantities. This chapter is focused on showing several simple but important inequalities that arise for uniquely decodable codes and their expected lengths. We will show that links between codes and probabilities are pretty close and lay foundations to a certain common area of computer science and probability. This area is called information theory.

Three such inequalities called the Kraft inequality, the Jensen inequality, and the Markov inequality are central to further developments. These three inequalities should be remembered. They bridge codes with probability distributions, motivating the ideas of information measures such as Shannon entropy and universal codes discussed later. They also yield a data-compression interpretation of seemingly purely probabilistic statements. Various corollaries of these inequalities such as the source coding inequality and the Barron inequality will reappear throughout the course.

To begin, so called incomplete distributions are a prerequisite. Incomplete distributions are non-negative functions of discrete symbols $x \in \mathbb{X}$ that add up to less than 1. In the case when they add up to 1 exactly, we already called them probability distributions in Chapter 1.

Definition 2.1 (incomplete distribution) *An incomplete distribution or a semi-distribution is a function* $p : \mathbb{X} \to [0, 1]$ *such that* $p(x) \geq 0$ *and* $\sum_{x \in \mathbb{X}} p(x) \leq 1$.

The reason for considering such defective distributions is the observation that for a prefix-free code B, quantity $p(x) = 2^{-|B(x)|}$ defines an incomplete distribution. This fact is called the Kraft inequality.

Theorem 2.2 (Kraft inequality) *For any prefix-free code $B : \mathbb{X} \to \{0,1\}^*$ we have inequality*

$$\sum_{x \in \mathbb{X}} 2^{-|B(x)|} \leq 1. \tag{2.1}$$

Proof: Let string u be the k-th element of set $\{0,1\}^l$ enumerated in the lexicographic order. We define interval

$$s(u) := [k2^{-l}, (k+1)2^{-l}) \subset \mathbb{R} \tag{2.2}$$

as the set of all real numbers whose binary expansions begin with string $0.u$. We observe that code B is prefix-free if and only if intervals $s(B(x))$ and $s(B(y))$ are disjoint for $x \neq y$.

Let us denote the length of an interval $[a, b)$ as

$$\Lambda([a, b)) := b - a. \tag{2.3}$$

Subsequently, we observe that the length of $s(u)$ is $\Lambda(s(u)) = 2^{-|u|}$. By disjointness of intervals $s(B(x))$ and inclusion

$$\bigcup_{x \in \mathbb{X}} s(B(x)) \subset [0, 1), \tag{2.4}$$

we obtain

$$\sum_{x \in \mathbb{X}} 2^{-|B(x)|} = \sum_{x \in \mathbb{X}} \Lambda(s(B(x))) \leq \Lambda([0, 1)) = 1. \tag{2.5}$$

□

What is somewhat surprising, the Kraft inequality can be generalized to uniquely decodable codes as well.

Theorem 2.3 (Kraft inequality) *For any uniquely decodable code $B : \mathbb{X} \to \{0,1\}^*$ we have inequality (2.1).*

Proof: Consider an arbitrary L. Let $a(m, n, L)$ denote the number of sequences $(x_1, ..., x_n)$ such that $|B(x_i)| \leq L$ and the length of $B^*(x_1, ..., x_n)$ equals m. We have

$$\left(\sum_{x : |B(x)| \leq L} 2^{-|B(x)|} \right)^n = \sum_{m=1}^{nL} a(m, n, L) \cdot 2^{-m}. \tag{2.6}$$

Because the code is uniquely decodable, we have $a(m,n,L) \leq 2^m$. Therefore

$$\sum_{x:|B(x)|\leq L} 2^{-|B(x)|} \leq (nL)^{1/n} \xrightarrow{n\to\infty} 1. \tag{2.7}$$

Letting $L \to \infty$, we obtain (2.1). □

There exists also a theorem converse to the Kraft inequality. Namely, if we have a length function that satisfies the Kraft inequality then there exists a prefix-free code of the same length. Thus, if we seek for the shortest uniquely decodable code, it suffices to look for it in the class of prefix-free codes. In particular, the Huffman code, which is optimal in the class of prefix-free codes, is also optimal in the class of uniquely decodable codes.

The exact theorem is as follows.

Theorem 2.4 (Kraft converse) *Let $\mathbb{X}$ be a countable set. If function $l : \mathbb{X} \to \mathbb{N}$ satisfies inequality*

$$\sum_{x\in\mathbb{X}} 2^{-l(x)} \leq 1 \tag{2.8}$$

then there exists a prefix-free code $B : \mathbb{X} \to \{0,1\}^*$ such that $|B(x)| = l(x)$.

Proof: Because the code domain $\mathbb{X}$ is countable, we may assume without loss of generality that $\mathbb{X} = \{1,2,...,n\}$ or $\mathbb{X} = \mathbb{N}$. Then we define B by iteration as follows. Let the interval $s(u)$ for a string $u \in \{0,1\}^*$ be defined as in (2.2). First, we denote sets of intervals $s(B(y))$ excluded before the x-th iteration as $N(1) := \emptyset$ and $N(x) := \bigcup_{y=1}^{x-1} s(B(y))$ for $x > 1$. Next, we define $B(x) := u$, where u is the *first* element of set $\{0,1\}^{l(x)}$ in the lexicographic order such that sets $s(u)$ and $N(x)$ are disjoint. It is obvious that B defined in this way is prefix-free and satisfies $|B(x)| = l(x)$, as long as strings u with the requested property exist.

Now we will show that strings u with the requested property exist if inequality (2.8) is satisfied. The proof of existence rests on this fact, which can be shown easily by induction: Set $[0,1) \setminus N(x)$ can be represented as a sum of finitely many intervals $[k2^{-l}, (k+1)2^{-l})$ of *different* l, which appear in $[0,1)$ in order of decreasing l. Let 2^{-m} be the length of the largest available of these intervals. By the mentioned fact, we have

$$2^{-m+1} > 1 - \sum_{y=1}^{x-1} 2^{-l(y)} \geq 2^{-m}. \tag{2.9}$$

The requested string u exists if and only if $2^{-l(x)} \leq 2^{-m}$. In view of (2.9), the latter condition holds if and only if

$$1 - \sum_{y=1}^{x-1} 2^{-l(y)} \geq 2^{-l(x)}. \tag{2.10}$$

But this condition is satisfied by (2.8). □

Let us define the prefix-free Shannon-Fano code, which is an alternative to the prefix-free Huffman code introduced in Chapter 1.

Definition 2.5 (Shannon-Fano code) *A prefix-free code $B : \mathbb{X} \to \{0,1\}^*$ is called a Shannon-Fano code for a probability distribution $p : \mathbb{X} \to [0,1]$ if*

$$|B(x)| = \lceil -\log p(x) \rceil . \tag{2.11}$$

Theorem 2.6 *Shannon-Fano codes exist for any probability distribution.*

Proof: We have

$$\sum_{x \in \mathbb{X}} 2^{-\lceil -\log p(x) \rceil} \leq \sum_{x \in \mathbb{X}} 2^{\log p(x)} = \sum_{x \in \mathbb{X}} p(x) = 1. \tag{2.12}$$

Hence Shannon-Fano codes exist by Theorem 2.4. □

The Shannon-Fano code is not necessarily the shortest prefix-free code since it can be sometimes outperformed by the Huffman code.

Example 2.7 *Consider the following distribution and codes:*

symbol x:	*$p(x)$:*	*code word $B(x)$:*	*code word $D(x)$*
a	$1 - 2^{-5}$	*0*	*0*
b	2^{-6}	*100000*	*10*
c	2^{-6}	*100001*	*11*

Code B is a Shannon-Fano code, whereas code D is the Huffman code. For no symbol code D is worse than code B, whereas for less probable symbols code D is much better.

Subsequently, we may ask about bounds for the expected length $\ell(p|B)$ of various uniquely decodable codes. To answer this question we will introduce the Jensen inequality for convex functions and the Kullback-Leibler divergence.

Definition 2.8 (convex and concave functions) *A real function* $f : (a, b) \to \mathbb{R}$ *is called convex if*

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) \tag{2.13}$$

for $p_i \geq 0$, $i = 1, 2$, *and* $p_1 + p_2 = 1$. *Moreover,* f *is called strictly convex if*

$$p_1 f(x_1) + p_2 f(x_2) > f(p_1 x_1 + p_2 x_2) \tag{2.14}$$

for $x_1 \neq x_2$, $p_i > 0$, $i = 1, 2$, *and* $p_1 + p_2 = 1$. *We say that function* f *is concave if* $-f$ *is convex, whereas* f *is strictly concave if* $-f$ *is strictly convex.*

A practical criterion of convexity is as follows.

Theorem 2.9 *A twice differentiable function* $f : (a, b) \to \mathbb{R}$ *is convex if its second derivative is positive,* $f''(x) \geq 0$ *for all* $x \in (a, b)$, *and strictly convex if its second derivative is strictly positive,* $f(x) > 0$ *for all* $x \in (a, b)$.

Proof: Let f' be the first derivative of f. By the mean value theorem for any $a < x_1 < x_2 < b$ there exists such an $x \in [x_1, x_2]$ that

$$f'(x_2) - f'(x_1) = (x_2 - x_1) f''(x). \tag{2.15}$$

Moreover, for any $a < x_1 < x_2 < b$ there also exists such an $x \in [x_1, x_2]$ that

$$f(x_2) - f(x_1) = (x_2 - x_1) f'(x). \tag{2.16}$$

For any $x_1 < x < x_2$, we have $x = p_1 x_1 + p_2 x_2$, where $p_1 = (x_2 - x)/(x_2 - x_1)$ and $p_2 = (x - x_1)/(x_2 - x_1)$. Moreover, by the above two displayed inequalities, there exist $x_1 \leq \tilde{x}_1 \leq \tilde{x} \leq \tilde{x}_2 \leq x_2$ such that

$$\begin{aligned} p_1 f(x_1) + p_2 f(x_2) - f(x) &= \frac{x_2 - x}{x_2 - x_1}(f(x_1) - f(x)) + \frac{x - x_1}{x_2 - x_1}(f(x_2) - f(x)) \\ &= \frac{(x_2 - x)(x - x_1)}{x_2 - x_1}\left(f'(\tilde{x}_2) - f'(\tilde{x}_1)\right) \\ &= \frac{(x_2 - x)(x - x_1)}{x_2 - x_1}(\tilde{x}_2 - \tilde{x}_1) f''(\tilde{x}) \end{aligned} \tag{2.17}$$

Hence if $f''(x) \geq 0$ for all $x \in (a, b)$ then f is convex, whereas if $f''(x) > 0$ for all $x \in (a, b)$ then f is strictly convex. □

According to the above criterion, some examples of strictly convex functions are: $f(x) = x^2$, $f(x) = \exp(x)$, and $f(x) = -\log x$.

The following inequality, called the Jensen inequality, states that the expectation of a convex function is greater than the function of the expected argument.

Theorem 2.10 (Jensen inequality) *If* : $f(a,b) \to \mathbb{R}$ *is a convex function and* p *is a discrete probability distribution over real values then*

$$\sum_x p(x)f(x) \geq f\left(\sum_x p(x)\cdot x\right). \tag{2.18}$$

Moreover, if f *is strictly convex then*

$$\sum_x p(x)f(x) = f\left(\sum_x p(x)\cdot x\right) \tag{2.19}$$

holds if and only if distribution p *is concentrated on a single value.*

Proof: We use the fact that if function f is convex then for any $y \in (a,b)$ there exists a linear function $h(x) = cx + d$ such that $h(x) \leq f(x)$ and $h(y) = f(y)$. In particular if we fix $y = \sum_x p(x)\cdot x$ then we obtain

$$\sum_x p(x)f(x) \geq \sum_x p(x)h(x) = c\sum_x p(x)\cdot x + d = h(y) = (y). \tag{2.20}$$

Additionally, we use the fact that if function f is strictly convex then the linear function h satisfies $h(x) = f(x)$ if and only if $x = y$. Consequently, $\sum_x p(x)f(x) = f(y)$ implies

$$\sum_x p(x)[f(x) - h(x)] = \sum_x p(x)f(x) - f(y) = 0. \tag{2.21}$$

Since $f(x) - h(x)$ is positive for $x \neq y$, hence $p(y) = 1$. □

Now we define a functional of discrete probability distributions called the Shannon entropy.

Definition 2.11 (Shannon entropy) *The Shannon entropy of a probability distribution* p *is denoted as*

$$H(p) := -\sum_{x:p(x)>0} p(x)\log p(x). \tag{2.22}$$

By definition, Shannon entropy is non-negative, $H(p) \geq 0$ since $-\log p(x) \geq 0$. Quantity $-\log p(x)$ is called the pointwise entropy. Shannon entropy is the expectation of the pointwise entropy.

Let us observe that if B is the Shannon-Fano code then there holds a symmetric bound for the expected length of the code

$$H(p) \leq \ell(p|B) = \sum_{x\in\mathbb{X}} p(x)\,|B(x)| \leq H(p) + 1 \tag{2.23}$$

since $-\log p(x) \leq |B(x)| \leq -\log p(x) + 1$. We may ask whether for other codes, such as the Huffman code, we have a similar inequality. The upper bound $\ell(p|D) \leq H(p)+1$ for the Huffman code D follows by inequality (2.23) since $\ell(p|D) \leq \ell(p|B)$ by the optimality of D. Hence it is rather interesting to ask whether also $H(p) \leq \ell(p|D)$. If such an inequality holds for any uniquely decodable code then the Shannon entropy sets an absolute lower bound for lossless data compression.

To answer this question, we will consider another functional, called the Kullback-Leibler divergence.

Definition 2.12 (Kullback-Leibler divergence) *The Kullback-Leibler (KL) divergence or relative entropy of an incomplete distribution q given a probability distribution p is*

$$D(p\|q) := \sum_{x:p(x)>0} p(x) \log \frac{p(x)}{q(x)}. \tag{2.24}$$

Quantity $H(p\|q) := -\sum_{x:p(x)>0} p(x) \log q(x) = H(p) + D(p\|q)$ is called the cross entropy of q with respect to p.

From the Jensen inequality, we can prove that also Kullback-Leibler divergence is non-negative.

Theorem 2.13 *For an incomplete distribution q and a probability distribution p, we have*

$$D(p\|q) \geq 0, \tag{2.25}$$

where the equality holds if and only if $p = q$.

Proof: By the Jensen inequality for the strictly convex function $h(x) = -\log x$, we have

$$\begin{aligned} D(p\|q) &= -\sum_{x:p(x)>0} p(x) \log \frac{q(x)}{p(x)} \geq -\log \left(\sum_{x:p(x)>0} p(x) \frac{q(x)}{p(x)} \right) \\ &= -\log \left(\sum_{x:p(x)>0} q(x) \right) \geq -\log 1 = 0, \end{aligned} \tag{2.26}$$

with the equality if and only if $p = q$. □

In fact, the probability that $\log \frac{p(x)}{q(x)}$ is negative is very small. In the following, for a proposition ϕ we write the indicator function

$$\mathbf{1}\{\phi\} := \begin{cases} 1, & \text{if proposition } \phi \text{ is true,} \\ 0, & \text{if proposition } \phi \text{ is false.} \end{cases} \tag{2.27}$$

Let us notice that $\mathbf{1}\{y \geq \epsilon\} \leq y/\epsilon$ for $y \geq 0$ and $\epsilon > 0$. This fact is called the Markov inequality and is often applied in probability calculus, as we will see in Chapter 4. Now we will use it as follows.

Theorem 2.14 (Barron inequality) *For an incomplete distribution q and a probability distribution p, we have*

$$\sum_{x:p(x)>0} p(x)\mathbf{1}\left\{\log \frac{p(x)}{q(x)} \leq -m\right\} \leq 2^{-m}. \tag{2.28}$$

Proof: By the Markov inequality, we may write

$$\begin{aligned} \sum_{x:p(x)>0} p(x)\mathbf{1}\left\{\log \frac{p(x)}{q(x)} \leq -m\right\} &= \sum_{x:p(x)>0} p(x)\mathbf{1}\left\{\frac{q(x)}{p(x)} \geq 2^{m}\right\} \\ &\leq \sum_{x:p(x)>0} p(x) \cdot \frac{q(x)}{p(x)} \cdot 2^{-m} \\ &= \sum_{x:p(x)>0} q(x) \cdot 2^{-m} \leq 2^{-m}. \end{aligned} \tag{2.29}$$

□

The non-negativity of KL divergence and the Barron inequality imply two desired theorems which link coding with the Shannon entropy.

Theorem 2.15 (source coding inequality) *For any uniquely decodable code $B : \mathbb{X} \to \{0,1\}^*$ and any probability distribution $p : \mathbb{X} \to [0,1]$, we have*

$$\ell(p|B) - H(p) = \sum_{x:p(x)>0} p(x)\left[|B(x)| + \log p(x)\right] \geq 0. \tag{2.30}$$

Proof: Introduce function $q(x) = 2^{-|B(x)|}$, which is an incomplete distribution by the Kraft inequality. By non-negativity of KL divergence, we obtain

$$\sum_{x:p(x)>0} p(x)\left[|B(x)| + \log p(x)\right] = \sum_{x:p(x)>0} p(x) \log \frac{p(x)}{q(x)} = D(p\|q) \geq 0. \tag{2.31}$$

□

Theorem 2.16 (Barron inequality) *For any uniquely decodable code* $B : \mathbb{X} \to \{0,1\}^*$ *and any probability distribution* $p : \mathbb{X} \to [0,1]$*, we have*

$$\sum_{x:p(x)>0} p(x)\mathbf{1}\left\{|B(x)| + \log p(x) \leq -m\right\} \leq 2^{-m}. \tag{2.32}$$

Proof: Introduce function $q(x) = 2^{-|B(x)|}$, which is an incomplete distribution by the Kraft inequality. Theorem 2.14 yields

$$\begin{aligned}\sum_{x:p(x)>0} p(x)\mathbf{1}\left\{|B(x)| + \log p(x) \leq -m\right\} &= \sum_{x:p(x)>0} p(x)\mathbf{1}\left\{\log \frac{p(x)}{q(x)} \leq -m\right\} \\ &\leq 2^{-m}. \end{aligned} \tag{2.33}$$

□

***

To recapitulate, in this chapter, we have obtained that the Shannon entropy and the pointwise entropy set an unexcelled lower bound for the length of any uniquely decodable code in a probabilistic sense. For the Shannon-Fano and the Huffman codes, these lengths are close to the lower bound. In the following chapters, we will use the approximate equivalence of optimal code lengths and entropies. Sometimes it is more convenient to think of the amount of information as a code length and sometimes it is more convenient to think about it as the pointwise entropy. These two perspectives complement each other.

Further reading

The Kraft inequality for prefix-free codes was proved by Leon Kraft [85] and for uniquely decodable codes—by Brockway McMillan [95]. The Jensen inequality was discovered by Johan Jensen [76]. The Kullback-Leibler divergence was introduced by Solomon Kullback and Richard Leibler [87]. The Shannon-Fano code is an invention of Claude Shannon [114] and Robert Fano [46]. The fact called the Barron inequality was a part of information-theoretic folklore until it was formally stated by Andrew Barron in his PhD thesis [5]. It may be also helpful to look into the textbooks on information theory by Thomas Cover and Joy Thomas [26] and by Imre Csiszár and János Körner [30]. A modern textbook about convex functions is by Stephen Boyd and Lieven Vandenberghe [11].

Thinking exercises

1. Consider a probability distribution:

symbol x:	$p(x)$:
a	1/3
b	1/3
c	4/15
d	1/15

   Show that there are two Huffman codes for this distribution: one has lengths $(1, 2, 3, 3)$ and the other has lengths $(2, 2, 2, 2)$. Use this result to demonstrate that the length of some Huffman code word can be greater for some symbol than the length of the Shannon-Fano code.

2. We say that p is dyadic distribution if for each element x in the domain of p there exists an integer k such that $p(x) = 2^{-k}$. Show that the length of the Huffman code B for a dyadic distribution p is the same as the length of the Shannon-Fano code and satisfies $\ell(p|B) = H(p)$.

3. *Rényi entropy:* Consider a distribution $p : \mathbb{X} \to [0, 1]$. We define

   (a) the Hartley entropy

   $$H_0(p) := \log \# \{x \in \mathbb{X} : p(x) > 0\}, \tag{2.34}$$

   (b) the Shannon entropy

   $$H_1(p) := H(p) = -\sum_{x \in \mathbb{X}} p(x) \log p(x), \tag{2.35}$$

   (c) the collision entropy

   $$H_2(p) := -\log \sum_{x \in \mathbb{X}} p(x)^2, \tag{2.36}$$

   (d) the min-entropy

   $$H_\infty(p) := -\log \max_{x \in \mathbb{X}} p(x). \tag{2.37}$$

   Show that the above functionals are special cases of the Rényi entropy

   $$H_\gamma(p) := -\frac{1}{\gamma - 1} \log \sum_{x \in \mathbb{X}} p(x)^\gamma, \quad \gamma \in (0, 1) \cup (1, \infty), \tag{2.38}$$

   where $H_\delta(p) := \lim_{\gamma \to \delta} H_\gamma(p)$ for $\delta \in \{0, 1, \infty\}$. Show also the general inequality $H_\delta(p) \leq H_\gamma(p)$ for $\delta \geq \gamma$ [106, 48].

4. Define $[x]_+ := x$ if $x > 0$ and $[x]_+ := 0$ if $x \leq 0$. Similarly, we put $[x]_- := [-x]_+$. Show that $x = [x]_+ - [x]_-$. For a uniquely decodable code $B : \mathbb{X} \to \{0,1\}^*$ and a probability distribution $p : \mathbb{X} \to [0,1]$ define redundancy $R(x) = |B(x)| + \log p(x)$. Demonstrate that for any $m > 0$, we have

$$\sum_{x:p(x)>0} p(x) \left[R(x)\right]_- \leq 4. \tag{2.39}$$

$$\sum_{x:p(x)>0} p(x) \mathbf{1}\left\{R(x) \geq m\right\} \leq \frac{\sum_{x:p(x)>0} p(x)R(x) + 4}{m}. \tag{2.40}$$

5. *Bregman divergence:* Let ϕ be a differentiable and strictly convex function of a vector $x = (x_1, x_2, ..., x_k)$. Bregman divergence is defined as

$$d_\phi(x,y) = \phi(x) - \phi(y) - \sum_i (x_i - y_i) \frac{\partial \phi(y)}{\partial y_i}.$$

Show that for $\phi(p) = -H(p) := \sum_i p_i \log p_i$, Bregman divergence equals Kullback-Leibler divergence, $d_\phi(p,q) = D(p\|q) := \sum_i p_i \log \frac{p_i}{q_i}$. What is the Bregman divergence for $\phi(x) = \sum_i x_i^2$? Show that $d_\phi(x,y) \geq 0$ and equality holds if and only if $x = y$.

6. *Generalized Pythagoras theorem:* Define $\arg\min_{x \in S} f(x)$ as the argument $x \in S$ for which function f attains the minimal value. Let S be a convex set of points, let $x_1 \in S$ and let $x_2 = \arg\min_{x \in S} d_\phi(x, x_3)$. Show that

$$d_\phi(x_1, x_2) + d_\phi(x_2, x_3) \leq d_\phi(x_1, x_3).$$

Hint: Let $x_\lambda = \lambda x_1 + (1-\lambda) x_2$. Show that

$$0 \leq \left. \frac{\partial d_\phi(x_\lambda, x_3)}{\partial \lambda} \right|_{\lambda=0} = d_\phi(x_1, x_3) - d_\phi(x_1, x_2) - d_\phi(x_2, x_3).$$

Chapter 3

Entropy

Finite probability spaces. Discrete random variables. Expectation. Probability as a random variable. Independence. Shannon entropy. Conditional entropy. Mutual information. Conditional mutual information. Chain rules. Venn diagrams. Triple information.

In this chapter, we will discuss some information measures for discrete random variables that are collectively called Shannon information measures. Shannon information measures are four entities: entropy, conditional entropy, mutual information, and conditional mutual information. Shannon entropy, as a functional of a probability distribution, has been already encountered in Chapter 2. In order to generalize this concept for a random variable, we need to formally define probability measures and discrete random variables on a finite probability space. We will also introduce the notions of expectation and conditional independence.

The concept of probability has many philosophically competing interpretations. Probability can be:

- the relative frequency of a certain event in a repeatable experiment,
- a learner's degree of belief in the propensity of a phenomenon,
- a convenient generalization of weights in the arithmetic mean,
- the relative volume of a figure when related to another figure.

All these interpretations are important and useful. None of them should be perceived as the only correct interpretation. The following formal definition generalizes and abstract from all these particular interpretations. It simply

defines a probability measure as a normalized and additive function of events defined on a suitable domain.

Definition 3.1 (finite probability space) *A finite probability space $(\Omega, \mathcal{J}, P)$ is a triple where:*

- Ω*, called an event space, is a certain set.*
- $\mathcal{J} \subset 2^{\Omega}$*, called a finite field, is a finite subset of subsets of* Ω *which satisfies*
  1. $\Omega \in \mathcal{J}$*,*
  2. $A \in \mathcal{J}$ *implies* $A^c \in \mathcal{J}$*, where* $A^c := \Omega \setminus A$*,*
  3. $A_1, A_2, ..., A_n \in \mathcal{J}$ *implies* $\bigcup_{i=1}^{n} A_i \in \mathcal{J}$*.*
- $P : \mathcal{J} \to [0, 1]$*, called a probability measure, is a function that satisfies*
  1. $P(\Omega) = 1$*,*
  2. $P(A) \geq 0$ *for* $A \in \mathcal{J}$*,*
  3. $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$ *for* $A_i \in \mathcal{J}$ *where* $A_i \cap A_j = \emptyset$*.*

The elements of $\mathcal{J}$ are called events, whereas the elements of Ω are called elementary events.

If the event space Ω is finite and $\mathcal{J} = 2^{\Omega}$, that is, $\mathcal{J}$ is the set of all subsets of Ω, then we may define a probability measure $P : \mathcal{J} \to [0, 1]$ by setting its values $P(\{\omega\}) \geq 0$ for all elementary events $\omega \in \Omega$ in such a way that

$$\sum_{\omega \in \Omega} P(\{\omega\}) = 1. \tag{3.1}$$

For an event $A \in \mathcal{J}$, we also have

$$P(A) = \sum_{\omega \in A} P(\{\omega\}). \tag{3.2}$$

Here are some examples:

Example 3.2 (fair coin) *The elementary outcomes of one coin toss are $\Omega = \{H, T\}$ (head and tail). Assuming that the outcomes of tossing are equally likely, we have $P(\{\omega\}) = 1/2$ so that $P(\Omega) = 1$.*

Example 3.3 (cubic die) *The elementary outcomes of one cubic die toss are* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *Assuming that the outcomes of tossing are equally likely, we have* $P(\{\omega\}) = 1/6$.

Example 3.4 (three cubic dice) *The elementary outcomes of three cubic die tosses are* $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$. *Assuming that the outcomes of tossing are equally likely, we have* $P(\{\omega\}) = 1/6^3$.

Let $\#\,\Omega$ be the number of elements in Ω. If the event space Ω is a finite set, probability measure P such that $P(\{\omega\}) = 1/\#\,\Omega$ is called the uniform measure. The above examples are uniform measures. If the event space Ω is a countably infinite set then the uniform measure does not exist.

Having a probability space, we can define discrete random variables.

Definition 3.5 (discrete random variable) *Let* $(\Omega, \mathcal{J}, P)$ *be a finite probability space. Function* $X : \Omega \to \mathbb{X}$ *is called a discrete random variable if set* $\mathbb{X}$ *is countable and for all* $x \in \mathbb{X}$ *we have*

$$(X = x) := \{\omega \in \Omega : X(\omega) = x\} \in \mathcal{J}. \tag{3.3}$$

The discrete random variable $Y : \Omega \to \mathbb{R} \bigcup \{-\infty, \infty\}$ *is called a discrete extended real random variable.*

We recall that we write $\mathbf{1}\{\phi\} = 1$ if proposition ϕ is true and $\mathbf{1}\{\phi\} = 0$ if proposition ϕ is false. For a discrete random variable $\Phi : \Omega \to \{\text{true}, \text{false}\}$ taking values in propositions, we generalize notation (3.3) as

$$(\Phi) := \{\omega \in \Omega : \Phi(\omega) \text{ is true}\}. \tag{3.4}$$

For a discrete extended real random variable, we define its expectation as the weighted average of the random variable's values, where the weights are given as the probabilities of particular values.

Definition 3.6 (expectation) *For a discrete extended real random variable* $Y : \Omega \to [0, \infty]$ *the expectation is defined as*

$$\mathbf{E}\, Y := \sum_{y : P(Y=y) > 0} y P(Y = y). \tag{3.5}$$

For discrete extended real random variables $Y_1, Y_2 : \Omega \to [0, \infty]$, *the expectation of random variable* $Y_1 - Y_2$ *is defined as*

$$\mathbf{E}(Y_1 - Y_2) := \mathbf{E}\, Y_1 - \mathbf{E}\, Y_2 \tag{3.6}$$

if $\mathbf{E}\, Y_1 < \infty$ *or* $\mathbf{E}\, Y_2 < \infty$.

In particular $\mathbf{E}(Y_1 - Y_2)$ is not defined if both $\mathbf{E}\,Y_1 = \infty$ and $\mathbf{E}\,Y_2 = \infty$. It can be demonstrated that $\mathbf{E}(X+Y) = \mathbf{E}\,X + \mathbf{E}\,Y$ and $\mathbf{E}(aY+b) = a\,\mathbf{E}\,Y + b$ if all the expectations are defined.

In information theory the following real variables play an important role.

Definition 3.7 *Let X and Y be discrete variables and A be an event on a probability space $(\Omega, \mathcal{J}, P)$. We define $P(X)$ as a discrete variable such that*

$$P(X)(\omega) = P(X = x) \iff X(\omega) = x. \tag{3.7}$$

Analogously we define $P(X|Y)$ and $P(X|A)$ as

$$P(X|Y)(\omega) = P(X = x|Y = y) \iff X(\omega) = x \text{ and } Y(\omega) = y, \tag{3.8}$$

$$P(X|A)(\omega) = P(X = x|A) \iff X(\omega) = x, \tag{3.9}$$

where the conditional probability is $P(B|A) := P(B \cap A)/P(A)$ for $P(A) > 0$.

Definition 3.8 (independence) *Random variables $X_1, X_2, ..., X_n$ are called independent if*

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i). \tag{3.10}$$

We also say that variables $X_1, X_2, X_3, ...$ are independent if $X_1, X_2, ..., X_n$ are independent for any n.

In this formalism, now we will cast the concept of Shannon entropy introduced in Chapter 2. Some interpretation of this quantity is the average uncertainty carried by a random variable or a tuple of random variables, regardless of their particular values. We expect that uncertainty adds for independent variables. Thus entropy $H(X)$ of a random variable X should be a functional of random variable $P(X)$ which is additive for independent random variables. Formally, for $P(X, Y) = P(X)P(Y)$, we postulate $H(X, Y) = H(X)+H(Y)$. Because $\log(xy) = \log x + \log y$ for the logarithm function, the following definition comes as a very natural idea.

Definition 3.9 (Shannon entropy) *The Shannon entropy of a discrete variable X is defined as*

$$H(X) := \mathbf{E}\,[-\log P(X)]. \tag{3.11}$$

To remain consistent with the Definition 2.11, symbol log *denotes the binary logarithm: $y = \log x \iff 2^y = x$.*

Because $\log P(X) \leq 0$, we put the minus sign in the definition (3.11) so that the Shannon entropy be positive. We notice that

$$H(X) = H(p) = -\sum_{x:p(x)>0} p(x) \log p(x), \tag{3.12}$$

where $p(x) = P(X = x)$ is the discrete distribution of variable X. Thus the entropy of a random variable is the entropy of its distribution. Moreover, we can verify that for $P(X, Y) = P(X)P(Y)$,

$$H(X, Y) = \mathbf{E}\left[-\log P(X, Y)\right] = \mathbf{E}\left[-\log P(X) - \log P(X)\right] \tag{3.13}$$

$$= \mathbf{E}\left[-\log P(X)\right] + \mathbf{E}\left[-\log P(X)\right] = H(X) + H(Y). \tag{3.14}$$

Example 3.10 *Let $P(X = 0) = 1/3$ and $P(X = 1) = 2/3$. Then*

$$H(X) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = \log 3 - 2/3 = 0.918.... \tag{3.15}$$

We obtain the same value for $P(X = 0) = 2/3$ and $P(X = 1) = 1/3$ because entropy depends on distribution $P(X)$ rather than on particular values of X. On the other hand, for $P(X = 0) = 1/2$ and $P(X = 1) = 1/2$, we have

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2 = 1. \tag{3.16}$$

The plot of function $H(X)$ for a binary variable (see Figure 3.1) shows that $H(X)$ achieves maximum 1 when the variable values are equally probable, whereas $H(X)$ achieves minimum 0 when the probability is concentrated on a single value.

What is the range of function $H(X)$ in general? Because function $f(p) = -p \log p$ is strictly positive for $p \in (0, 1)$ and equals 0 for $p = 1$, it can be easily seen that:

Theorem 3.11 *$H(X) \geq 0$, whereas $H(X) = 0$ if and only if X assumes only a single value.*

This fact agrees intuitively with the idea that constants carry no uncertainty.

On the other hand, assume that X takes values $x \in \{1, 2, ..., n\}$ with equal probabilities $P(X = x) = 1/n$. Then we have

$$H(X) = -\sum_{x=1}^{n} \frac{1}{n}\log\frac{1}{n} = \sum_{x=1}^{n} \frac{1}{n}\log n = \log n. \tag{3.17}$$

As we will see now, $\log n$ is the maximal value of $H(X)$ when variable X assumes values in $\{1, 2, ..., n\}$. That fact agrees with the intuition that the highest uncertainty occurs for uniformly distributed variables.

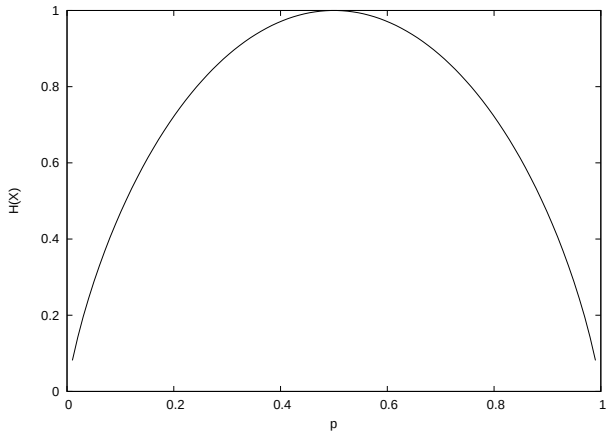


Figure 3.1: Shannon entropy $H(X) = -p\log p - (1-p)\log(1-p)$ for $P(X=0) = p$ and $P(X=1) = 1-p$.

Theorem 3.12 *Let X assume values in $\{1, 2, ..., n\}$. We have $H(X) \leq \log n$, whereas $H(X) = \log n$ if and only if $P(X = x) = 1/n$.*

Remark: If the range of variable X is infinite then entropy $H(X)$ may be infinite.

Proof: Let $p(x) = P(X = x)$ and $q(x) = 1/n$. Then

$$0 \leq D(p\|q) = \sum_{x:p(x)>0} p(x)\log\frac{p(x)}{1/n} = \log n - H(X), \tag{3.18}$$

where the equality occurs if and only if $p = q$. □

The next important problem is the behavior of the Shannon entropy under conditioning. The intuition is that given additional information, the uncertainty should decrease. So should the Shannon entropy. There are, however, two distinct ways of defining conditional entropy.

Definition 3.13 (conditional entropy) *The Shannon conditional entropy of a discrete variable X given event A is*

$$H(X|A) := H(p) \text{ for } p(x) = P(X = x|A). \tag{3.19}$$

The Shannon conditional entropy of X given a discrete variable Y is defined as

$$H(X|Y) := \sum_{y:P(Y=y)>0} P(Y = y)H(X|Y = y). \tag{3.20}$$

Both $H(X|A)$ and $H(X|Y)$ are non-negative.

Theorem 3.14 *$H(X|Y) = 0$ holds if and only if $X = f(Y)$ for a certain function f except for a set of probability 0.*

Proof: Observe that $H(X|Y) = 0$ if and only if $H(X|Y = y) = 0$ for all y such that $P(Y = y) > 0$. This holds if and only if given $(Y = y)$ with $P(Y = y) > 0$, variable X is concentrated on a single value. Denoting this value as $f(y)$, we obtain $X = f(Y)$, except for the union of those sets $(Y = y)$ which have probability 0. □

Let us note that inequality $H(X|A) \leq H(X)$ need not hold.

Example 3.15 *Let $P(X = 0|A) = P(X = 1|A) = 1/2$, whereas $P(X = 0|A^c) = 1$ and $P(X = 1|A^c) = 0$. Assuming $P(A) = 1/2$, we have $P(X = 0) = (1/2) \cdot (1/2) + (1/2) = 3/4$ and $P(X = 0) = (1/2) \cdot (1/2) = 1/4$ so*

$$H(X) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = \log 4 - \frac{3}{4}\log 3 = 0.811.... \tag{3.21}$$

On the other hand, we have $H(X|A) = \log 2 = 1$.

Despite that fact, it is true that $H(X|Y) \leq H(X)$ holds in general. Thus entropy decreases given additional information on average. Before we prove it, let us observe:

Theorem 3.16 *We have*

$$H(X|Y) = \mathbf{E}\left[-\log P(X|Y)\right]. \tag{3.22}$$

Proof: Observe

$$\begin{aligned} H(X|Y) &= \sum_{y:P(Y=y)>0} P(Y = y)H(X|Y = y) && (3.23)\\ &= -\sum_{x,y:P(X=x,Y=y)>0} P(Y = y)P(X = x|Y = y)\log P(X = x|Y = y) && (3.24)\\ &= -\sum_{x,y:P(X=x,Y=y)>0} P(X = x, Y = y)\log P(X = x|Y = y) && (3.25)\\ &= \mathbf{E}\left[-\log P(X|Y)\right]. && (3.26) \end{aligned}$$

□

Because $P(Y)P(X|Y) = P(X,Y)$, by Theorem 3.16 we obtain

$$H(Y) + H(X|Y) = H(X,Y). \tag{3.27}$$

Hence

$$H(X,Y) \geq H(Y). \tag{3.28}$$

To show that $H(X)$ is greater than $H(X|Y)$, it is convenient to introduce another important concept.

Definition 3.17 (mutual information) *The Shannon mutual information between discrete variables X and Y is defined as*

$$I(X;Y) := \mathbf{E}\left[\log \frac{P(X,Y)}{P(X)P(Y)}\right]. \tag{3.29}$$

Let us observe that $I(X;X) = H(X)$. Hence entropy is sometimes called self-information.

Mutual information is non-negative because it is a special instance of the KL divergence.

Theorem 3.18 *We have*

$$I(X;Y) \geq 0, \tag{3.30}$$

where the equality holds if and only if X and Y are independent.

Proof: Let $p(x,y) = P(X = x, Y = y)$ and $q(x,y) = P(X = x)P(Y = y)$. Then we have

$$I(X;Y) = \sum_{(x,y):p(x,y)>0} p(x,y) \log \frac{p(x,y)}{q(x,y)} = D(p\|q) \geq 0 \tag{3.31}$$

with the equality exactly for $p = q$, by Theorem 2.13. □

By the definition of mutual information and by Theorem 3.16,

$$H(X,Y) + I(X;Y) = H(X) + H(Y), \tag{3.32}$$

$$H(X|Y) + I(X;Y) = H(X). \tag{3.33}$$

Hence by Theorem 3.18, we have

$$H(X) + H(Y) \geq H(X,Y), \tag{3.34}$$

$$H(X) \geq H(X|Y),\ I(X;Y). \tag{3.35}$$

Moreover, we have $H(X|Y) = H(Y)$ if X and Y are independent, which also agrees with intuition.

In a similar fashion as for entropy, we may introduce conditional mutual information.

Definition 3.19 (conditional mutual information) *The Shannon conditional mutual information between discrete variables X and Y given event A is*

$$I(X;Y|A) := D(p\|q) \text{ for } p(x,y) = P(X=x, Y=y|A) \\ \text{and } q(x,y) = P(X=x|A)P(Y=y|A). \quad (3.36)$$

The Shannon conditional mutual information between discrete variables X and Y given variable Z is defined as

$$I(X;Y|Z) := \sum_{z:P(Z=z)>0} P(Z=z)I(X;Y|Z=z). \quad (3.37)$$

Both $I(X;Y|A)$ and $I(X;Y|Z)$ are non-negative. As in the case of conditional entropy, the following proposition is true:

Theorem 3.20 *We have*

$$I(X;Y|Z) := \mathbf{E}\left[\log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right]. \quad (3.38)$$

The notion of conditional information is useful when analyzing conditional independence.

Definition 3.21 (conditional independence) *Random variables $X_1, X_2, ..., X_n$ are called conditionally independent given a random variable Z if*

$$P(X_1, X_2, ..., X_n|Z) = \prod_{i=1}^{n} P(X_i|Z). \quad (3.39)$$

We also say that variables $X_1, X_2, X_3, ...$ are conditionally independent given Z if $X_1, X_2, ..., X_n$ are conditionally independent given Z for any n.

Example 3.22 *Let $Y = f(Z)$ be a function of variable Z, whereas X be an arbitrary variable. Variables X and Y are conditionally independent given Z. Indeed, we have*

$$P(X=x, Y=y|Z=z) = P(X=x|Z=z)\mathbf{1}\{y=f(z)\} \quad (3.40)$$
$$= P(X=x|Z=z)P(Y=y|Z=z). \quad (3.41)$$

Example 3.23 *Let variables X, Y, and Z be independent assuming with equal probability values 0 and 1. Variables $U = X + Z$ and $W = Y + Z$ are conditionally independent given Z. Indeed, we have*

$$P(U=u, W=w|Z=z) = P(X=u-z, Y=w-z) \\ = P(X=u-z)P(Y=w-z) = P(U=u|Z=z)P(W=w|Z=z). \quad (3.42)$$

It can be checked, however, that U and V are not independent.

As in the case of plain mutual information the following fact is true:

Theorem 3.24 *We have*

$$I(X;Y|Z) \geq 0, \tag{3.43}$$

where the equality holds if and only if X and Y are conditionally independent given Z.

Of particular interest is this generalization of formula (3.33):

Theorem 3.25 (chain rule) *We have*

$$I(X;Y|Z) + I(X;Z) = I(X;Y,Z). \tag{3.44}$$

Remark: Hence, variables X and (Y,Z) are independent if and only if X and Z are independent and X and Y are independent given Z.

Proof:

$$I(X;Y|Z) + I(X;Z) = \mathbf{E}\left[\log \frac{P(X,Y,Z)P(Z)}{P(X,Z)P(Y,Z)}\right] + \mathbf{E}\left[\log \frac{P(X,Z)}{P(X)P(Z)}\right] \tag{3.45}$$

$$= \mathbf{E}\left[\log \frac{P(X,Y,Z)}{P(X)P(Y,Z)}\right] = I(X;Y,Z). \tag{3.46}$$

□

Finally, one can ask whether conditional entropy and mutual information may be expressed by entropies of tuples of variables. The answer is positive if the entropies are finite.

Theorem 3.26 *If entropies $H(X)$, $H(Y)$, and $H(Z)$ are finite, we observe these identities:*

$$H(X|Y) = H(X,Y) - H(Y), \tag{3.47}$$

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y), \tag{3.48}$$

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \tag{3.49}$$

$$= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z), \tag{3.50}$$

where all terms are finite and non-negative.

Proof: (Left as an exercise.) □

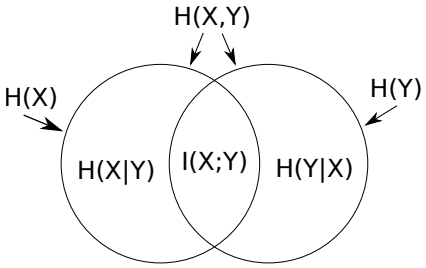


Figure 3.2: Venn diagram for two random variables.

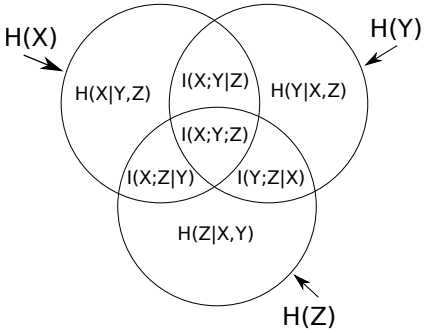


Figure 3.3: Venn diagram for three random variables.

The dependence between the Shannon entropy, conditional entropy and mutual information can be depicted by Venn diagrams. The diagram for two variables is given in Figure 3.2, whereas the diagram for three variables is presented in Figure 3.3. Quantity $I(X;Y;Z)$, appearing in Figure 3.3, is called triple information or interaction information. It can be defined as

$$I(X;Y;Z) := I(X;Y) - I(X;Y|Z). \tag{3.51}$$

We can easily see that $I(X;Y;Z) = I(Y;X;Z) = I(Y;Z;X)$. For some random variables X, Y, Z, we have $I(X;Y;Z) > 0$, whereas for some other we have $I(X;Y;Z) < 0$. We leave constructing such examples as an exercise.

***

Recapitulating this chapter, we have learned about the algebraic properties of Shannon entropy for discrete random variables. This establishes the first link between information theory and probability. In Chapter 5, we will establish some further link in the instance of universal codes that learn the underlying probability distribution. Prior to this, in Chapter 4, we will recall the law of large numbers that establishes a frequency interpretation of probability and the base for a mathematical theory of learning.

Further reading

The algebra of Shannon information measures such as entropy and mutual information was discovered by Claude Shannon [114]. Many years later Zhen Zhang and Raymond Yeung [127] discovered inequalities for information measures that cannot be reduced to the inequalities discussed in this chapter. See also Raymond Yeung's textbook for an overview [126]. It may be also helpful to look into the textbooks by Thomas Cover and Joy Thomas [26] and by Imre Csiszár and János Körner [30]. Shannon information measures in the non-discrete setting were studied by Izrail Gelfand, Andrey Kolmogorov, and Akiva Yaglom [55], by Roland Dobrushin [41], by Aaron Wyner [125], and by Łukasz Dębowski [32, 35]. Those generalized information measures also satisfy the familiar inequalities and additionally enjoy a certain continuity.

Thinking exercises

1. We toss a coin until the first tail is obtained. The outcome of the experiment is the number of tosses. Compute its entropy.

2. Show that function $d(X,Y) = H(X|Y) + H(Y|X)$ is almost a metric on random variables, i.e., it satisfies:

   - $d(X,Y) = 0$ if there is a one-to-one mapping between X and Y;
   - $d(X,Y) = d(Y,X)$;
   - $d(X,Z) \leq d(X,Y) + d(Y,Z)$.

3. For a function g show that

$$H(g(X)) \leq H(X), \tag{3.52}$$
$$H(X|g(Y)) \geq H(X|Y), \tag{3.53}$$
$$I(X;g(Y)) \leq I(X;Y). \tag{3.54}$$

4. Prove Theorem 3.26.

5. *Data-processing inequality:* Let X and Z be conditionally independent given Y. Show that

$$I(X;Y) \geq I(X;Z). \tag{3.55}$$

6. *Chain rule:* Prove the chain rule

$$H(X_1, ..., X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_1, ...X_{i-1}). \tag{3.56}$$

Chapter 4

Independence

Prequential probability spaces. Prequential distributions. Stochastic processes. Consistency conditions. IID processes. Uniform measure. Bernoulli process. Variance. Markov inequality. Weak law of large numbers. Convergence in probability. Limits of sequences. Countably additive probability spaces. Kolmogorov process theorem. Real random variables. Borel-Cantelli lemma. Almost sure convergence. Expectation. Convergence of expectations. Riesz theorem. Strong law of large numbers.

In Chapter 3, we mentioned that probability can be regarded as a limiting relative frequency of an event in a series of repeatable experiments. Given this interpretation, we would like to discuss arbitrarily large collections of random variables. Assuming no prior knowledge of probability for infinite spaces, in this chapter, we will make first steps towards stochastic processes. Stochastic processes are exactly infinite sequences of random variables. We will present a general condition for their existence.

Subsequently, we will analyze the simplest of stochastic processes called IID processes, which are sequences of independent identically distributed random variables. In this setting, we will observe an important fact that frequencies of events tend to their probabilities. This mathematical fact is called the law of large numbers and is a foundation of statistical inference. The law of large numbers comes in two flavors. The first kind is the weak law, which is conceptually simpler but complicated to use, whereas the second kind is the strong law which requires some mathematical imagination but leads to simpler application. Let us recall how these two laws differ.

Weak law

First we will discuss the weak law of large numbers, which is more elementary. To construct a probability space that supports an infinite collection of arbitrary random variables, it is convenient to apply an event space whose elements are simply infinite sequences of values of these random variables. In the following, for a finite set $\mathbb{X}$, let us write strings $x_j^k := (x_j, x_{j+1}, ..., x_k)$, where $x_i \in \mathbb{X}$.

Definition 4.1 (prequential probability space) *For a finite set $\mathbb{X}$, consider the event space consisting of infinite sequences*

$$\Omega = \{x = (x_1, x_2, x_3, ...) : x_i \in \mathbb{X}\}. \tag{4.1}$$

A set $A \subset \Omega$ is called an event of depth n if there exists a set $B \subset \mathbb{X}^n$ such that

$$A = [B] := \{(x_1, x_2, x_3, ...) : x_1^n \in B, x_i \in \mathbb{X} \text{ for } i > n\}. \tag{4.2}$$

Let $\mathcal{J}_n$ be the set of events of depth n and let $\mathcal{J} = \bigcup_{n=1}^{\infty} \mathcal{J}_n$ be the set of events of any depth. For these Ω and $\mathcal{J}$ and a certain function $P : \mathcal{J} \to [0, 1]$, triple $(\Omega, \mathcal{J}, P)$ is called a prequential probability space if each $(\Omega, \mathcal{J}_n, P_n)$ is a finite probability space with measure $P_n(A) = P(A)$ for all $A \in \mathcal{J}_n$. We also say that $(\Omega, \mathcal{J})$ is a prequential measurable space and P is a probability measure on this space.

Definition 4.2 (prequential distribution) *Consider a function $Q : \mathbb{X}^* \to [0, 1]$. It is called a prequential distribution if*

$$\sum_{x_1 \in \mathbb{X}} Q(x_1) = 1, \tag{4.3}$$

$$\sum_{x_{n+1} \in \mathbb{X}} Q(x_1^{n+1}) = Q(x_1^n), \quad n \geq 1. \tag{4.4}$$

Theorem 4.3 *Function P is a probability measure on the prequential measurable space $(\Omega, \mathcal{J})$ if and only if there exists a prequential distribution Q such that for each $A \in \mathcal{J}_n$ with $A = [B]$ we have*

$$P(A) = \sum_{x_1^n \in B} Q(x_1^n). \tag{4.5}$$

Proof: It suffices to observe that equality (4.5) holds if and only if $P([\{x_1^n\}]) = Q(x_1^n)$, whereas function Q defined by this constraint is a prequential distribution if and only if P is a probability measure on $(\Omega, \mathcal{J})$. □

Now we can discuss stochastic processes.

Definition 4.4 (stochastic process) *A stochastic process is a sequence*

$$(X_i)_{i\in\mathbb{N}} := (X_1, X_2, X_3, ...) \tag{4.6}$$

of random variables $X_i : \Omega \to \mathbb{X}$ sharing the same image $\mathbb{X}$.

Example 4.5 *Let Ω be given by (4.1) and let $X_i(x) := x_i$. So that process $(X_i)_{i\in\mathbb{N}}$ were supported on the prequential probability space $(\Omega, \mathcal{J}, P)$, it is necessary and sufficient that probability measure P satisfies so called consistency conditions*

$$\sum_{x_1\in\mathbb{X}} P(X_1 = x_1) = 1, \tag{4.7}$$

$$\sum_{x_{n+1}\in\mathbb{X}} P(X_1^{n+1} = x_1^{n+1}) = P(X_1^n = x_1^n), \quad n \geq 1. \tag{4.8}$$

The second consistency condition (4.8) can be rewritten for $P(X_1^n = x_1^n) > 0$ using the concept of conditional probability as

$$\sum_{x_{n+1}\in\mathbb{X}} P(X_{n+1} = x_{n+1} | X_1^n = x_1^n) = 1. \tag{4.9}$$

In particular, the consistency conditions are obviously satisfied by a sequence of independent identically distributed random variables, which is briefly called an IID process.

Definition 4.6 (IID process) *A stochastic process $(X_i)_{i\in\mathbb{N}}$ is called an IID process if for all $n \in \mathbb{N}$, we have*

$$P(X_1^n = x_1^n) = \prod_{i=1}^{n} \pi(x_i) \tag{4.10}$$

for a certain function $\pi : \mathbb{X} \to [0, 1]$.

It is enlightening to observe that uniform measures on a product space induce some IID processes. If we have the event space

$$\Omega = \{x = (x_1, x_2, ..., x_n) : \omega_i \in \mathbb{X}\}, \tag{4.11}$$

random variables $X_i(x) := x_i$, and a uniform measure $P(\{x\}) = 1/\#\Omega$ then process X_1^n is an IID process. Thus independence arises naturally on a prequential space when all outcomes are equally probable. For the event

space (4.1) and an IID process $(X_i)_{i\in\mathbb{N}}$ such that $X_i(x) := x_i$ and $P(X_i = x_i) = 1/\#\mathbb{X}$, measure P is also called the uniform measure or the Lebesgue measure.

Another important example of an IID process is the Bernoulli process, being a sequence of binary random variables.

Example 4.7 (Bernoulli process) *The Bernoulli(θ) process $(Y_i)_{i\in\mathbb{N}}$ is an IID process where $P(Y_i = 1) = \theta$ and $P(Y_i = 0) = 1 - \theta$. Events $(Y_i = 1)$ are called successes, whereas events $(Y_i = 0)$ are called failures. The total number of successes is $S_n := \sum_{i=1}^n Y_i$. Its distribution is*

$$P(S_n = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{4.12}$$

where $k \in \{0, 1, ..., n\}$ and $k! := 1 \cdot 2 \cdot ... \cdot k$ is the factorial of k with $0! := 1$. In spite of this complicated expression, the expectation of S_n can be simply computed as

$$\mathbf{E}\, S_n = \mathbf{E} \sum_{i=1}^n Y_i = \sum_{i=1}^n \mathbf{E}\, Y_i = n\theta. \tag{4.13}$$

In the following, let us compute the variance of the number of successes for the Bernoulli process. The variance is defined as follows:

Definition 4.8 (variance) *For a real random variable Y, the variance is defined as*

$$\operatorname{Var} Y := \mathbf{E}(Y - \mathbf{E}\, Y)^2 \tag{4.14}$$

if the expectations are defined.

We can write equivalently

$$\operatorname{Var} Y = \mathbf{E}(Y - \mathbf{E}\, Y)^2 = \mathbf{E}\, Y^2 - 2\,\mathbf{E}\, Y\,\mathbf{E}\, Y + (\mathbf{E}\, Y)^2 = \mathbf{E}\, Y^2 - (\mathbf{E}\, Y)^2. \tag{4.15}$$

In particular, we obtain $\operatorname{Var}(aY) = a^2 \operatorname{Var} Y$.

Let us note that if real random variables X and Y are independent then they are uncorrelated, i.e., $\mathbf{E}(XY) = \mathbf{E}\, X\,\mathbf{E}\, Y$. Consequently, for uncorrelated random variables, we obtain

$$\begin{aligned}\operatorname{Var}(X+Y) &= \mathbf{E}(X+Y)^2 - (\mathbf{E}\, X + \mathbf{E}\, Y)^2 \\ &= \mathbf{E}\, X^2 - 2\,\mathbf{E}(XY) + \mathbf{E}\, Y^2 - (\mathbf{E}\, X)^2 + 2\,\mathbf{E}\, X\,\mathbf{E}\, Y - (\mathbf{E}\, Y)^2 \\ &= \mathbf{E}\, X^2 - (\mathbf{E}\, X)^2 + \mathbf{E}\, Y^2 - (\mathbf{E}\, Y)^2 = \operatorname{Var} X + \operatorname{Var} Y. \end{aligned} \tag{4.16}$$

The above observation can be generalized to a sequence of independent random variables $(Y_i)_{i\in\mathbb{N}}$ and $S_n := \sum_{i=1}^n Y_i$ as $\operatorname{Var} S_n = \sum_{i=1}^n \operatorname{Var} Y_i$.

Example 4.9 (Bernoulli process) *For the Bernoulli(θ) process $(Y_i)_{i\in\mathbb{N}}$ and the total number of successes $S_n := \sum_{i=1}^n Y_i$, by the independence of Y_i, we have*

$$\operatorname{Var} S_n = \operatorname{Var} \sum_{i=1}^n Y_i = \sum_{i=1}^n \operatorname{Var} Y_i = n\theta(1-\theta). \tag{4.17}$$

Let $(Y_i)_{i\in\mathbb{N}}$ be an arbitrary real IID process with expectation $\mathbf{E}\, Y_i = \mu$ and variance $\operatorname{Var} Y_i = \sigma^2$. Let us consider the empirical average $\frac{1}{n}\sum_{i=1}^n Y_i$. Its expectation equals $\mathbf{E}\,\frac{1}{n}\sum_{i=1}^n Y_i = \mu$. Is it true that the probability of a fixed deviation $\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mu\right| \geq \epsilon > 0$ tends to zero for $n \to \infty$? To show this fact, let us make a simple but an important observation called the Markov inequality.

Theorem 4.10 (Markov inequality) *Let $Z : \Omega \to [0,\infty]$ be a non-negative real random variable. For any $\epsilon > 0$ we have*

$$P(Z \geq \epsilon) \leq \frac{\mathbf{E}\, Z}{\epsilon}. \tag{4.18}$$

Proof: Since $\mathbf{1}\{z \geq \epsilon\} \leq z/\epsilon$ for $z \geq 0$, we have

$$\begin{aligned} P(Z \geq \epsilon) &= \sum_{z:P(Z=z)>0} \mathbf{1}\{z \geq \epsilon\}\, P(Z=z) \\ &\leq \sum_{z:P(Z=z)>0} \frac{z}{\epsilon} P(Z=z) = \frac{\mathbf{E}\, Z}{\epsilon}. \end{aligned} \tag{4.19}$$

□

Hence we derive an important fact called the weak law of large numbers.

Theorem 4.11 (weak law of large numbers) *Let $(Y_i)_{i\in\mathbb{N}}$ be a real IID process with expectation $\mathbf{E}\, Y_i = \mu$ and variance $\operatorname{Var} Y_i = \sigma^2$, where $|\mu|, \sigma^2 < \infty$. Then for any $\epsilon > 0$, we have*

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mu\right| \geq \epsilon\right) = 0. \tag{4.20}$$

Proof: By the Markov inequality and independence of Y_i, we have

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mu\right| \geq \epsilon\right) &= P\left(\left(\frac{1}{n}\sum_{i=1}^n Y_i - \mu\right)^2 \geq \epsilon^2\right) \\ &\leq \frac{\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right)}{\epsilon^2} \\ &= \frac{\sum_{i=1}^n \operatorname{Var} Y_i}{n^2\epsilon^2} = \frac{\sigma}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0. \end{aligned} \tag{4.21}$$

□

A more general property applies more often, so we will give it a name.

Definition 4.12 (convergence in probability) *For a real stochastic process $(Y_n)_{n\in\mathbb{N}}$, we say that Y_n converge in probability to a real random variable Y, written $\lim_{n\to\infty} Y_n = Y$ i.p., if for any $\epsilon > 0$, we have*

$$\lim_{n\to\infty} P(|Y_n - Y| \geq \epsilon) = 0. \tag{4.22}$$

In particular, the weak law of large numbers states that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_n = \mu \text{ i.p.} \tag{4.23}$$

In other words, the empirical average converges in probability to the expectation if the expectation and the variance of Y_i are finite.

Strong law

So far, we tried to use only elementary probability calculus, which deals with sequences of nested finite probability spaces and convergence in probability. For more advanced constructions in universal coding, this formalism is too weak. We want to discuss random variables being limits of sequences of discrete random variables, such as general real random variables. By the way, we will introduce a stronger concept of probabilistic convergence, called the almost sure convergence. In turn, our consideration will lead to the strong law of large numbers.

Let us observe that the values of random processes usually fluctuate or oscillate. For this reason, in order to discuss a probabilistic convergence of irregular sequences, it is advisable to define first the upper and the lower limit of a sequence of real numbers.

Definition 4.13 (limits of a sequence) *Let $(a_n)_{n\in\mathbb{N}} = (a_1, a_2, a_3, ...)$ be a sequence of extended real numbers, $a_n \in \mathbb{R} \cup \{-\infty, \infty\}$. The supremum $\sup_{m\geq n} a_m$ is the least number $r \in \mathbb{R}\cup\{-\infty, \infty\}$ such that $a_m \leq r$ for $m \leq n$. The infimum $\inf_{m\geq n} a_m$ is the largest number $r \in \mathbb{R} \cup \{-\infty, \infty\}$ such that $a_m \geq r$ for $m \leq n$. The upper and the lower limits are defined as*

$$\limsup_{n\to\infty} a_n := \inf_{n\geq 1} \sup_{m\geq n} a_m, \tag{4.24}$$

$$\liminf_{n\to\infty} a_n := \sup_{n\geq 1} \inf_{m\geq n} a_m. \tag{4.25}$$

The values are the asymptotic upper and lower bounds for the oscillations of sequence $(a_n)_{n\in\mathbb{N}}$. *In general, we have* $\limsup_{n\to\infty} a_n \geq \liminf_{n\to\infty} a_n$. *If the oscillations asymptotically vanish, namely, if* $\limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = a$ *then we say that sequence* $(a_n)_{n\in\mathbb{N}}$ *has limit* a *and we write it as* $\lim_{n\to\infty} a_n = a$.

Having limits, we can define the sum of an infinite series as a limit

$$\sum_{n=1}^{\infty} a_n := \lim_{n\to\infty} \sum_{i=1}^{n} a_i. \tag{4.26}$$

This expression is defined in particular if $a_n \geq 0$. Now we can define a probability measure as a normalized, additive, and continuous function of events defined on a suitable domain.

Definition 4.14 (countably additive probability space) *A countably additive probability space* $(\Omega, \mathcal{J}, P)$ *is a triple where:*

- Ω, *called an event space, is a certain set.*
- $\mathcal{J} \subset 2^{\Omega}$, *called a* σ*-field, is a subset of subsets of* Ω *which satisfies*
  1. $\Omega \in \mathcal{J}$,
  2. $A \in \mathcal{J}$ *implies* $A^c \in \mathcal{J}$, *where* $A^c := \Omega \setminus A$,
  3. $A_1, A_2, A_3, ... \in \mathcal{J}$ *implies* $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{J}$.
- $P : \mathcal{J} \to [0, 1]$, *called a probability measure, is a function that satisfies*
  1. $P(\Omega) = 1$,
  2. $P(A) \geq 0$ *for* $A \in \mathcal{J}$,
  3. $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$ *for* $A_i \in \mathcal{J}$ *where* $A_i \cap A_j = \emptyset$.

In the above, we assume continuity of probability to guarantee a nice behavior. Moreover, it is a natural generalization of a prequential probability space since we have the following theorem.

Theorem 4.15 (Kolmogorov process theorem) *Let* $(\Omega, \mathcal{J}, P)$ *be a prequential probability space. Let* $\mathcal{J}'$ *be the intersection of all* σ*-fields that contain set* $\mathcal{J}$. *There is a unique function* $P' : \mathcal{J}' \to [0, 1]$ *such that* $(\Omega, \mathcal{J}', P')$ *is a countably additive probability space and* $P'(A) = P(A)$ *for* $A \in \mathcal{J}$.

Proof: (Omitted. See Theorems 36.1 and 36.2 of [8].) □

As a result, we have a wide range of countably additive probability spaces on which we can discuss general real random variables.

Definition 4.16 (real random variable) *Let $(\Omega, \mathcal{J}, P)$ be a countably additive probability space. Function $Y : \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$ is called an extended real random variable if for all $r \in \mathbb{R} \cup \{-\infty, \infty\}$, we have*

$$(Y \leq r) := \{\omega \in \Omega : Y(\omega) \leq r\} \in \mathcal{J}. \tag{4.27}$$

In particular, discrete extended real random variables are extended real random variables according to the above definition. It can be proved that if $Y_1, Y_2, ...$ are extended real random variables then so are $Y_1 + Y_2$, $Y_1 - Y_2$ and other continuous functions of Y_i. What is less trivial, if $Y_1, Y_2, ...$ are extended real random variables then also the supremum $\sup_{n\in\mathbb{N}} Y_n$ and the infimum $\inf_{n\in\mathbb{N}} Y_n$ are extended real random variables. Consequently, limits $\limsup_{n\to\infty} Y_n$ and $\liminf_{n\to\infty} Y_n$ are also extended real random variables.

For a discrete random variable $\Phi : \Omega \to \{\text{true}, \text{false}\}$ taking values in propositions, let us say the Φ holds almost surely, written Φ a.s., if

$$P(\Phi) = P(\{\omega \in \Omega : \Phi(\omega) \text{ is true}\}) = 1. \tag{4.28}$$

We have the following important result.

Theorem 4.17 (Borel-Cantelli lemma) *We have:*

- *If $\sum_{n=1}^{\infty} P(Y_n > Y) < \infty$ then $\limsup_{n\to\infty} Y_n \leq Y$ a.s.*
- *If $\sum_{n=1}^{\infty} P(Y_n < Y) < \infty$ then $\liminf_{n\to\infty} Y_n \geq Y$ a.s.*

Proof: We have

$$\begin{aligned} P(\limsup_{n\to\infty} Y_n > Y) &= P(\forall_{n\in\mathbb{N}} \exists_{k\geq n} Y_k > Y) \\ &= \lim_{n\to\infty} P(\exists_{k\geq n} Y_k > Y) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} P(Y_k > Y) = 0. \end{aligned} \tag{4.29}$$

Analogously, we demonstrate the other statement. □

For countably additive probability spaces, we can discuss this property.

Definition 4.18 (almost sure convergence) *For a real stochastic process $(Y_n)_{i\in\mathbb{N}}$, we say that Y_n converge almost surely to a real random variable Y if $\lim_{n\to\infty} Y_n = Y$ a.s.*

To prove the almost sure convergence, it is often convenient to consider the upper and the lower limit of random variables separately.

The almost sure convergence is stronger than convergence in probability. To show it, we need to generalize the notion of expectation to arbitrary real random variables. In the general case, the expectation is formally defined as the Lebesgue integral.

Definition 4.19 (expectation) *For a general extended real random variable $Y : \Omega \to [0, \infty]$ the expectation is*

$$\mathbf{E}\, Y := \sup_{X \le Y} \mathbf{E}\, X, \tag{4.30}$$

where the supremum is taken over all discrete real random variables X such that $X(\omega) \le Y(\omega)$. For real random variables $Y_1, Y_2 : \Omega \to [0, \infty]$, the expectation of random variable $Y_1 - Y_2$ is defined as

$$\mathbf{E}(Y_1 - Y_2) := \mathbf{E}\, Y_1 - \mathbf{E}\, Y_2 \tag{4.31}$$

if $\mathbf{E}\, Y_1 < \infty$ or $\mathbf{E}\, Y_2 < \infty$.

As in the discrete case, it can be shown that $\mathbf{E}(X + Y) = \mathbf{E}\, X + \mathbf{E}\, Y$ and $\mathbf{E}(aY + b) = a\, \mathbf{E}\, Y + b$ if all the expectations are defined.

Applying the upper and the lower limits, we can state three important theorems about sequences of expectations. These are: the monotone convergence, the Fatou lemma, and the dominated convergence.

Theorem 4.20 (monotone convergence) *Let $(Y_n)_{n \in \mathbb{N}}$ be a sequence of non-negative, $Y_n \ge 0$, and growing, $Y_{n+1} \ge Y_n$, real random variables. Then*

$$\sup_{n \in \mathbb{N}} \mathbf{E}\, Y_n = \mathbf{E} \sup_{n \in \mathbb{N}} Y_n. \tag{4.32}$$

Proof: (Omitted. See Theorem 16.2 of [8].) □

Theorem 4.21 (Fatou lemma) *Let $(Y_n)_{n \in \mathbb{N}}$ be a sequence of non-negative, $Y_n \ge 0$, real random variables. Then*

$$\liminf_{n \to \infty} \mathbf{E}\, Y_n \ge \mathbf{E} \liminf_{n \to \infty} Y_n. \tag{4.33}$$

Proof: Denote $X_n := \inf_{k \ge n} Y_k \le Y_n$. We have $X_{n+1} \ge X_n$ and $\liminf_{n \to \infty} Y_n = \sup_{n \in \mathbb{N}} X_n$. Hence by the monotone convergence, we have

$$\liminf_{n \to \infty} \mathbf{E}\, Y_n \ge \lim_{n \to \infty} \mathbf{E}\, X_n = \mathbf{E} \liminf_{n \to \infty} Y_n. \tag{4.34}$$

□

Theorem 4.22 (Lebesgue dominated convergence) *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of real random variables which satisfy $\mathbf{E}\sup_{n\in\mathbb{N}}|Y_n| < \infty$. If there exists limit $\lim_{n\to\infty} Y_n$ then*

$$\lim_{n\to\infty} \mathbf{E}\, Y_n = \mathbf{E} \lim_{n\to\infty} Y_n. \tag{4.35}$$

Proof: Let $X_m := |Y_m - \lim_{n\to\infty} Y_n|$ and $Z = \sup_{n\in\mathbb{N}}|Y_n|$. We have $0 \leq X_m \leq 2Z$. Hence by the Fatou lemma, we obtain

$$\begin{aligned}\mathbf{E}\, 2Z &= \mathbf{E} \liminf_{m\to\infty}(2Z - X_m) \\ &\leq \liminf_{m\to\infty} \mathbf{E}(2Z - X_m) = \mathbf{E}\, 2Z - \limsup_{m\to\infty} \mathbf{E}\, X_m\end{aligned} \tag{4.36}$$

Thus the claim follows by

$$0 = \limsup_{m\to\infty} \mathbf{E}\, X_m \geq \limsup_{m\to\infty} \left|\mathbf{E}\, Y_m - \lim_{n\to\infty} Y_n\right|. \tag{4.37}$$

□

The Lebesgue dominated convergence will be used quite often. In particular, using this theorem, we can prove that the almost sure convergence is stronger than the convergence in probability.

Theorem 4.23 (Riesz theorem) *If $\lim_{n\to\infty} Y_n = Y$ almost surely then it also holds in probability.*

Proof: Let us consider metric

$$d : \mathbb{R}\times\mathbb{R} \ni (x, y) \mapsto \min\{1, |x - y|\}. \tag{4.38}$$

By the Markov inequality, we have

$$\epsilon P(d(Y_n, Y) > \epsilon) \leq \mathbf{E}\, d(Y_n, Y) \leq P(d(Y_n, Y) > \epsilon) + \epsilon. \tag{4.39}$$

Hence $(Y_i)_{i\in\mathbb{N}}$ converges to Y in probability if and only if

$$\lim_{n\to\infty} \mathbf{E}\, d(Y_n, Y) = 0. \tag{4.40}$$

In contrast, the almost sure convergence is equivalent to

$$\lim_{n\to\infty} d(Y_n, Y) = 0 \text{ a.s.} \tag{4.41}$$

Condition (4.41) implies (4.40) by the Lebesgue dominated convergence. □

In general, convergence in probability does not imply the almost sure convergence as it will be demonstrated in the exercises. We have however quite many important cases when convergence in probability can be lifted to the almost sure convergence. First, we will present an application of the Borel-Cantelli lemma to demonstrate the strong law of large numbers.

Theorem 4.24 (strong law of large numbers) *Let $(Y_i)_{i\in\mathbb{N}}$ be a real IID process with expectation $\mathbf{E}\,Y_i = \mu$, variance $\mathbf{E}(Y_i-\mu)^2 = \sigma^2$, and fourth central moment $\mathbf{E}(Y_i-\mu)^4 = \kappa^4$, where $|\mu|, \sigma^2, \kappa^4 < \infty$. Then we have*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i = \mu \text{ a.s.} \tag{4.42}$$

Proof: Let an $\epsilon > 0$. By the Markov inequality and independence of Y_i, we obtain

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right| \geq \epsilon\right) &= P\left(\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right)^4 \geq \epsilon^4\right) \\ &\leq \frac{\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right)^4}{\epsilon^4} \qquad (4.43) \\ &= \frac{n\kappa^4}{n^4\epsilon^4} + \binom{4}{2}\frac{n(n-1)\sigma^4}{n^4\epsilon^4}. \qquad (4.44) \end{aligned}$$

Hence

$$\sum_{n=1}^{\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right| > \epsilon\right) < \infty. \tag{4.45}$$

Thus the Borel-Cantelli lemma yields

$$P\left(\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right| > \epsilon\right) = 0. \tag{4.46}$$

Hence, the claim follows since ϵ was chosen arbitrarily. □

One can wonder whether the strong law of large numbers can be generalized to IID processes that have no finite expectation, variance, or fourth central moment. In the following version of the strong law of large numbers, notice the lack of the assumption of finite expectation or variance. Instead, it is only assumed that the random variables are non-negative.

Theorem 4.25 (strong law of large numbers) *Let $(Y_i)_{i\in\mathbb{N}}$ be a real IID process with $Y_i \geq 0$. Then we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i = \mathbf{E}\, Y_i \ \ a.s. \tag{4.47}$$

Proof: The claim is a special case of the Birkhoff ergodic theorem, to be established via the Ivanov downcrossing inequality in Chapter 9. □

***

To recapitulate this chapter, we have learned about the law of large numbers. This law is central to the theory of learning since given a sample of data, we can estimate the unknown probabilities via empirical frequencies and use them for prediction or data compression. This second idea will give rise to universal coding in Chapter 5, where the uniquely decodable code adapts to the compressed data. By the way, information theory will pay back to probability calculus by motivating some constructive results.

Further reading

The weak law of large numbers for the Bernoulli process was proved by Jakob Bernoulli in book *Ars conjectandi* in 1713 [6]. The name "law of large numbers" was coined by Siméon Poisson [104]. The Markov inequality is due to Pafnuty Chebyshev [122], who was the teacher of Andrey Markov. The Borel-Cantelli lemma is due to Emile Borel [10] and Francesco Cantelli [16]. The further generalizations of the strong law of large numbers were due to Andrey Kolmogorov and Alexander Khinchin. Andrey Kolmogorov wrote also the first exposition of the measure-theoretic probability calculus [83]. The Fatou lemma was proved by Pierre Fatou, whereas the monotone and dominated convergence were shown by Henri Lebesgue. Classical textbooks in probability were written by William Feller [51], Patrick Billingsley [8], and Leo Breiman [14]. To gain a modern perspective, the book by Olav Kallenberg [78] can be also consulted.

Thinking exercises

1. *Monty Hall paradox:* A participant of the "Let's Make A Deal" quiz hosted by Monty Hall is exposed to three closed doors. Behind one of the doors there is an expensive car, behind two other doors there

are two goats. Monty Hall asks the participant to choose a door. It is known that there is a goat behind one of the not selected doors. This door is opened and the goat is shown. Now the participant is asked to choose one of the remaining two doors. He will get what is behind it. Should he choose the same door as before or the other one?

2. Show that $P(X \leq Z) \leq P(X \leq Y) + P(Y \leq Z)$.

3. Consider random variables $S_n = \sum_{i=1}^{n} Y_i$, where Y_i are not independent. Prove the Cauchy-Schwarz inequality

$$(\mathbf{E}\, XY)^2 \leq \mathbf{E}\, X^2\, \mathbf{E}\, Y^2 \tag{4.48}$$

and consequently, show that

$$\sqrt{\operatorname{Var} S_n} \leq \sum_{i=1}^{n} \sqrt{\operatorname{Var} Y_i}. \tag{4.49}$$

4. Show that $\lim_{n\to\infty} Y_n = Y$ i.p. and $\lim_{n\to\infty} Y_n = Y'$ i.p. imply $Y = Y'$ a.s.

5. For the event space $\Omega = [0, 1]$, the Lebesgue measure $P([a, b]) = b - a$, function $f(n) = \sqrt{n} - \lfloor \sqrt{n} \rfloor$ and random variables

$$Y_n(\omega) = \begin{cases} 1 & f(n) < \omega < f(n+1), \\ 1 & \omega < f(n+1) < f(n), \\ 1 & f(n+1) < f(n) < \omega, \\ 0 & \text{else.} \end{cases} \tag{4.50}$$

show that $\lim_{n\to\infty} Y_n = 0$ i.p. but $\lim_{n\to\infty} Y_n$ does not exist almost surely. Give a few other examples of functions $f(n)$ for which the same holds. What are the general conditions on such $f(n)$?

6. Let $(Y_i)_{i\in\mathbb{N}}$ be a real IID process with $Y_i \geq 0$. Using Theorem 4.24, the monotone convergence, and the Fatou lemma show that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i = \mathbf{E}\, Y_i \text{ a.s.} \tag{4.51}$$

Chapter 5

Universality

Empirical distribution and empirical entropy. Maximum likelihood. Superadditivity of empirical entropy. Shtarkov sum bound. Penalized maximum likelihood. Consistency of empirical entropy. Asymptotic equipartition for IID processes. Barron lemma. Universal codes for IID processes. Universality criterion. Laplace estimator. Multinomial coefficients and entropy. Stirling approximation.

In this chapter, we will study universal codes for IID processes. The problem of universal coding consists in constructing a single prefix-free code, which is sufficiently good for any process in a given class—unlike the Huffman code, which is optimal for a single fixed probability distribution. This comes of course at a certain cost. Namely, the universal code is sufficiently good for any process in a given class but is not exactly optimal for any of them. However, the difference of lengths for a universal code and the Huffman code usually grows much slower than any of these lengths.

Let us also note that the problem of universal compression falls under the scope of statistics. Indeed, the interest of statisticians lies in identifying parameters of a stochastic process based on the data generated by that process. The entropy of a IID process is an example of such a parameter. When we have a universal code then we may estimate the entropy as the encoding rate achieved by the code. The estimate, being the length of a universal code divided by the length of the coded string, converges to the true entropy as the string length grows unboundedly.

In the following, we will work with a finite alphabet consisting of digits, namely, $\mathbb{X} = \{1, 2, ..., m\}$. We also denote the Shannon entropy of a vector

of probabilities $p = (p_1, ...p_m)$,

$$H(p) := H(p_1, ..., p_m) := -\sum_{l=1}^{m} p_l \log p_l, \tag{5.1}$$

where $p_l \geq 0$, $\sum_{l=1}^{m} p_l = 1$, and $0 \log 0 := 0$.

An important example of a probability distribution is the empirical distribution of digits in a given data sequence, which is simply the relative frequency of a given digit.

Definition 5.1 (empirical distribution) *The empirical distribution of string $x_1^n \in \mathbb{X}^n$ is*

$$\hat{\pi}(l|x_1^n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i = l\}. \tag{5.2}$$

In the next step, we introduce the empirical entropy, which is the Shannon entropy of the empirical distribution.

Definition 5.2 (empirical entropy) *The empirical entropy of a string $x_1^n \in \mathbb{X}^n$ is*

$$\mathcal{H}(x_1^n) := H\left(\hat{\pi}(\cdot|x_1^n)\right). \tag{5.3}$$

There are several simple bounds for the empirical entropy. First, the empirical entropy of a sequence is upper bounded by the logarithm of the sequence length.

Theorem 5.3 *We have inequality $\mathcal{H}(x_1^n) \leq \log n$.*

Proof: Let $k_l = \sum_{i=1}^{n} \mathbf{1}\{x_i = l\}$. We derive

$$\mathcal{H}(x_1^n) = H\left(\frac{k_1}{n}, ..., \frac{k_m}{n}\right) = \log n - \sum_{l=1}^{m} k_l \log k_l \leq \log n \tag{5.4}$$

since $k_l \in \mathbb{N} \cup \{0\}$. □

The next bound is also important. It applies the maximum likelihood.

Definition 5.4 (maximum likelihood) *For a string $x_1^n \in \mathbb{X}^n$, we define the maximum likelihood (ML)*

$$\hat{\mathbb{P}}(x_1^n) := \max_{\pi} \prod_{i=1}^{n} \pi(x_i), \tag{5.5}$$

where the maximum is taken across all probability vectors $\pi : \{1, 2, ..., m\} \to [0, 1]$, where $\pi(l) \geq 0$ and $\sum_{l=1}^{m} \pi(l) = 1$.

In fact, the distribution π that maximizes the expression on the right hand side of (5.5) is exactly the empirical distribution.

Theorem 5.5 *For any probability vector $\pi : \{1, 2, ..., m\} \to [0, 1]$, we have*

$$\mathcal{H}(x_1^n) = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{\pi}(x_i|x_1^n) = -\frac{1}{n} \log \hat{\mathbb{P}}(x_1^n) \leq -\frac{1}{n}\sum_{i=1}^{n} \log \pi(x_i). \quad (5.6)$$

Proof: We may write

$$\begin{aligned} \mathcal{H}(x_1^n) = H(\hat{\pi}(\cdot|x_1^n)) &= -\sum_{l=1}^{m} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i = l\} \log \hat{\pi}(l|x_1^n) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \log \hat{\pi}(x_i|x_1^n). \end{aligned} \quad (5.7)$$

On the other hand,

$$\begin{aligned} 0 \leq D(\hat{\pi}(\cdot|x_1^n)\|\pi) &= \sum_{l=1}^{m} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i = l\} \log \frac{\hat{\pi}(l|x_1^n)}{\pi(l)} \\ &= -\frac{1}{n}\sum_{i=1}^{n} \log \pi(x_i) + \frac{1}{n}\sum_{i=1}^{n} \log \hat{\pi}(x_i|x_1^n), \end{aligned} \quad (5.8)$$

where the equality holds for $\hat{\pi}(\cdot|x_1^n) = \pi$. □

In particular, by the above representation, we obtain another two important theorems. The first one states superadditivity of empirical entropy.

Theorem 5.6 *We have*

$$k\mathcal{H}(x_1^k) + (n-k)\mathcal{H}(x_{k+1}^n) \leq n\mathcal{H}(x_1^n). \quad (5.9)$$

Proof: The claim follows by inequality

$$\left(\max_{\pi} \prod_{i=1}^{k} \pi(x_i)\right)\left(\max_{\pi} \prod_{i=k+1}^{n} \pi(x_i)\right) \geq \max_{\pi} \prod_{i=1}^{n} \pi(x_i). \quad (5.10)$$

□

The maximum likelihood is not a probability distribution since expression $\sum_{x_1^n} \hat{\mathbb{P}}(x_1^n)$, called the Shtarkov sum, is greater than 1. The second theorem provides an upper bound for this expression.

Theorem 5.7 (Shtarkov sum bound) *We have inequality*

$$\sum_{x_1^n} \hat{\mathbb{P}}(x_1^n) \leq (n+1)^m. \tag{5.11}$$

Proof: Let $\mathcal{P} := \{\hat{\pi}(\cdot|x_1^n) : x_1^n \in \mathbb{X}^n\}$ be the set of distinct empirical distributions. We notice that for any $\pi \in \mathcal{P}$ there holds inequality

$$\sum_{x_1^n : \hat{\pi}(\cdot|x_1^n)=\pi} \hat{\mathbb{P}}(x_1^n) \leq \sum_{x_1^n} \prod_{i=1}^{n} \pi(x_i) = 1. \tag{5.12}$$

Hence

$$\sum_{x_1^n} \hat{\mathbb{P}}(x_1^n) = \sum_{\pi \in \mathcal{P}} \sum_{x_1^n : \hat{\pi}(\cdot|x_1^n)=\pi} \hat{\mathbb{P}}(x_1^n) \leq \sum_{\pi \in \mathcal{P}} 1 \leq \# \mathcal{P}. \tag{5.13}$$

How many distinct empirical distributions are there? There are m coordinates of the probability vector. Each may assume values only from set $\{0/n, 1/n, ..., n/n\}$. Thus we may bound $\# \mathcal{P} \leq (n+1)^m$. □

Thus we may define an important incomplete distribution.

Definition 5.8 (penalized maximum likelihood) *For a string $x_1^n \in \mathbb{X}^n$, we define the penalized maximum likelihood (PML)*

$$\mathbb{P}(x_1^n) := \frac{\hat{\mathbb{P}}(x_1^n)}{(n+1)^m}. \tag{5.14}$$

Obviously, we have inequality $\sum_{x_1^n} \mathbb{P}(x_1^n) \leq 1$, so there exists a prefix-free Shannon-Fano code with respect to $\mathbb{P}$. As we will show further, this Shannon-Fano code is an example of a universal code. But first we need to develop some theory of what we are searching for exactly.

Let us consider an IID process $(X_i)_{i \in \mathbb{N}}$ with random variables $X_i : \Omega \to \mathbb{X} = \{1, 2, ..., m\}$. We can ask about the difference between the empirical entropy $\mathcal{H}(X_1^n)$ and the Shannon entropy $H(X_i)$. Theorem 5.3 asserts that empirical entropy $\mathcal{H}(X_1^n)$ is a poor estimate of the Shannon entropy $H(X_i)$ if there holds inequality

$$\log n < H(X_i) \leq \log m, \tag{5.15}$$

which is possible if the sample length is smaller than the alphabet, $n < m$. However, we may suppose that for a finite alphabet size m, the empirical entropy is a consistent estimator of the Shannon entropy. That is, we may suppose that random variable $\mathcal{H}(X_1^n)$ converges to parameter $H(X_i)$ when the sample size n tends to infinity. It is so indeed. The respective result follows by the continuity of the entropy function and the strong law of large numbers.

Theorem 5.9 (consistency of empirical entropy) *Let $(X_i)_{i\in\mathbb{N}}$ be an IID process with random variables $X_i : \Omega \to \mathbb{X}$. We have*

$$\lim_{n\to\infty} \mathcal{H}(X_1^n) = H(X_i) \text{ a.s.} \tag{5.16}$$

Proof: Let $K_l = \sum_{i=1}^n \mathbf{1}\{X_i = l\}$ be the frequency of digit l. We want to show that

$$\lim_{n\to\infty} H\left(\frac{K_1}{n}, ..., \frac{K_m}{n}\right) = H(X_i) \text{ a.s.} \tag{5.17}$$

But by the strong law of large numbers, we have

$$\lim_{n\to\infty} \frac{K_l}{n} = P(X_i = l) \text{ a.s.} \tag{5.18}$$

Moreover, Shannon entropy $H(p_1, ..., p_m)$ is a continuous function of probabilities $(p_1, ..., p_m)$. Hence (5.18) implies (5.17). □

Once we know that the Shannon entropy can be estimated by the empirical entropy, we may hope that there exist universal codes. As we have mentioned, a universal code is a single prefix-free code which is sufficiently good for any process in a given class. There are two results that suggest a reasonable definition: the Barron lemma, applying the Barron inequality from Chapter 2, and the asymptotic equipartition.

The Barron lemma states that any reasonable code length is greater than the pointwise entropy—for sufficiently long samples.

Theorem 5.10 (Barron lemma) *For any uniquely decodable code $B : \mathbb{X}^* \to \{0,1\}^*$ and any stochastic process $(X_i)_{i\in\mathbb{N}}$ with random variables $X_i : \Omega \to \mathbb{X}$, we have*

$$\lim_{n\to\infty} \left[|B(X_1^n)| + \log P(X_1^n)\right] = \infty \text{ a.s.} \tag{5.19}$$

Proof: By the Markov and Kraft inequalities, we obtain

$$\begin{aligned}
\sum_{n=1}^{\infty} P(|B(X_1^n)| \le -\log P(X_1^n) + M) &= \sum_{n=1}^{\infty} P\left(\frac{2^{-|B(X_1^n)|}}{P(X_1^n)} \ge 2^{-M}\right) \\
&\le \sum_{n=1}^{\infty} 2^M \,\mathbf{E}\left(\frac{2^{-|B(X_1^n)|}}{P(X_1^n)}\right) = 2^M \sum_{n=1}^{\infty}\sum_{x_1^n} P(X_1^n = x_1^n)\cdot\frac{2^{-|B(x_1^n)|}}{P(X_1^n = x_1^n)} \\
&\le 2^M \sum_{w\in\mathbb{X}^*} 2^{-|B(w)|} \le 2^M < \infty.
\end{aligned} \tag{5.20}$$

Hence by the strong Borel-Cantelli lemma, for any real number M, we have

$$\liminf_{n\to\infty} [|B(X_1^n)| + \log P(X_1^n)] \geq M \text{ a.s.} \tag{5.21}$$

This implies (5.19). □

The next result, called the asymptotic equipartition, states that asymptotically all samples with a positive probability are equally probable.

Theorem 5.11 (asymptotic equipartition) *For any IID process $(X_i)_{i\in\mathbb{N}}$ with random variables $X_i : \Omega \to \mathbb{X}$, we have*

$$\lim_{n\to\infty} \frac{[-\log P(X_1^n)]}{n} = H(X_i) \text{ a.s.} \tag{5.22}$$

Proof: First, we observe

$$\frac{[-\log P(X_1^n)]}{n} = \frac{1}{n}\left[-\log \prod_{i=1}^{n} P(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} [-\log P(X_i)]\,. \tag{5.23}$$

Thus the asymptotic equipartition (5.22) follows simply by the strong law of large numbers

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} [-\log P(X_i)] = \mathbf{E}\,[-\log P(X_i)] = H(X_i) \text{ a.s.}, \tag{5.24}$$

since alphabet $\mathbb{X}$ is finite. □

Thus, a reasonable definition of universal codes is as follows.

Definition 5.12 (universal code) *Let $\mathbb{X}$ be a finite alphabet. Let $B : \mathbb{X}^* \to \{0,1\}^*$ be a uniquely decodable code. Code B is called universal for IID processes over alphabet $\mathbb{X}$ if for any IID process $(X_i)_{i\in\mathbb{N}}$ with random variables $X_i : \Omega \to \mathbb{X}$, we have*

$$\lim_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)|}{n} = H(X_i), \tag{5.25}$$

$$\lim_{n\to\infty} \frac{|B(X_1^n)|}{n} = H(X_i) \text{ a.s.} \tag{5.26}$$

A sufficient criterion for a code to be universal is as follows.

Theorem 5.13 (universality criterion) *Let $B : \mathbb{X}^* \to \{0,1\}^*$ be a uniquely decodable code. Code B is universal for IID processes over a finite alphabet $\mathbb{X}$ if for any string $x_1^n \in \mathbb{X}^*$, we have*

$$|B(x_1^n)| \leq C(n) + n\mathcal{H}(x_1^n), \tag{5.27}$$

where $\lim_{n\to\infty} C(n)/n = 0$.

Proof: Let B be a uniquely decodable code and $(X_i)_{i\in\mathbb{N}}$ be an IID process. Hence by the Barron lemma and the asymptotic equipartition, we obtain the lower bound

$$\liminf_{n\to\infty} \frac{|B(X_1^n)|}{n} \geq \lim_{n\to\infty} \frac{[-\log P(X_1^n)]}{n} = H(X_i) \text{ a.s.} \tag{5.28}$$

Similarly, by the source coding inequality we obtain

$$\liminf_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)|}{n} \geq \lim_{n\to\infty} \frac{\mathbf{E}\,[-\log P(X_1^n)]}{n} = H(X_i). \tag{5.29}$$

It remains to prove the upper bounds. First, if (5.27) holds then we have

$$|B(X_1^n)| \leq C(n) - \log P(X_1^n). \tag{5.30}$$

Hence by the asymptotic equipartition

$$\limsup_{n\to\infty} \frac{|B(X_1^n)|}{n} \leq \lim_{n\to\infty} \frac{[-\log P(X_1^n)]}{n} = H(X_i) \text{ a.s.} \tag{5.31}$$

Similarly,

$$\limsup_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)|}{n} \leq \lim_{n\to\infty} \frac{\mathbf{E}\,[-\log P(X_1^n)]}{n} = H(X_i). \tag{5.32}$$

Thus we have established universality of B. □

Now we show that penalized maximum likelihood defines a universal code.

Theorem 5.14 *The prefix-free Shannon-Fano code with respect to penalized maximum likelihood $\mathbb{P}$ satisfies universality criterion (5.27).*

Proof: We have

$$\begin{aligned} -\log \mathbb{P}(x_1^n) &\leq m\log(n+1) - \log \hat{\mathbb{P}}(x_1^n) \\ &\leq m\log(n+1) + n\mathcal{H}(x_1^n), \end{aligned} \tag{5.33}$$

which is enough to assert criterion (5.27). □

There are many more examples of universal codes. Consider for instance this construction.

Definition 5.15 (Laplace estimator) *The Laplace estimator is the probability distribution for alphabet* $\mathbb{X} = \{1, 2, ..., m\}$ *iteratively defined as*

$$R_0(x_1) := \frac{1}{m}, \tag{5.34}$$

$$R_0(x_{n+1}|x_1^n) := \frac{\sum_{i=1}^n \mathbf{1}\{x_i = x_{n+1}\} + 1}{n+m}, \tag{5.35}$$

$$R_0(x_1^n) := R_0(x_1) \prod_{i=1}^{n-1} R_0(x_{i+1}|x_1^i). \tag{5.36}$$

To show universality of the Shannon-Fano code for the Laplace estimator, we need to exhibit a close relationship between multinomial coefficients and Shannon entropy. For $k_l \geq 0$ and $\sum_{l=1}^m k_l = n$, the multinomial coefficient is

$$\binom{n}{k_1, ..., k_m} := \frac{n!}{k_1!...k_m!}. \tag{5.37}$$

We will approximate the multinomial coefficient with Shannon entropy, as

$$\log \binom{n}{k_1, ..., k_m} \approx \log \frac{n^n}{k_1^{k_1}...k_m^{k_m}} = nH\left(\frac{k_1}{n}, ..., \frac{k_m}{n}\right). \tag{5.38}$$

To be sufficiently precise, we will derive a simple upper bound for the error of this approximation.

The first step are two inequalities.

Theorem 5.16 *For* $x > 0$*, we have*

$$\log x \leq (x-1)\log e. \tag{5.39}$$

Proof: Function $x \mapsto \log x$ is concave. Hence the tangent to its graph lies above it. The desired inequality arises when we take the tangent at $x = 1$. □

Theorem 5.17 (Stirling approximation) *For* $n = \mathbb{N} \cup \{0\}$*, we have*

$$\frac{n^n}{e^n} \leq n! \leq \frac{(n+1)^{n+1}}{e^n}, \tag{5.40}$$

where $0^0 := 1$.

Proof: Inequalities (5.40) are satisfied for $n = 0$. For $\ln x$ being the natural logarithm of x and $n \in \mathbb{N}$, we obtain

$$\ln n! = \sum_{j=1}^{n} \ln j! \in \left(\int_0^n \ln x dx, \int_1^{n+1} \ln x dx \right), \tag{5.41}$$

whereas

$$\int_a^b \ln x dx = \left[x \ln x - x \right]_a^b . \tag{5.42}$$

Hence we obtain inequalities (5.40). □

Now we can prove a bound for the multinomial coefficients.

Theorem 5.18 *We have the upper bound*

$$\log \binom{n}{k_1, ..., k_m} \leq (n+1) H \left(\frac{k_1}{n+1}, ..., \frac{k_m}{n+1}, \frac{1}{n+1} \right). \tag{5.43}$$

Proof: Using the Stirling approximation (5.40), we obtain

$$\begin{aligned} \log \binom{n}{k_1, ..., k_m} &\leq \log \frac{(n+1)^{n+1}}{k_1^{k_1} ... k_m^{k_m}} = \log \frac{(n+1)^{n+1}}{1^1 \cdot k_1^{k_1} ... k_m^{k_m}} \\ &= (n+1) H \left(\frac{k_1}{n+1}, ..., \frac{k_m}{n+1}, \frac{1}{n+1} \right). \end{aligned} \tag{5.44}$$

□

We need two more results about approximations of entropies.

Theorem 5.19 *For $n_i = \sum_{l=1}^{m_i} k_{il}$, we have the decomposition*

$$\begin{aligned} &(n_1 + n_2) H \left(\frac{k_{11}}{n_1 + n_2}, ..., \frac{k_{1m_1}}{n_1 + n_2}, \frac{k_{21}}{n_1 + n_2}, ..., \frac{k_{2m_2}}{n_1 + n_2} \right) \\ &= (n_1 + n_2) H \left(\frac{n_1}{n_1 + n_2}, \frac{n_2}{n_1 + n_2} \right) \\ &\quad + n_1 H \left(\frac{k_{11}}{n_1}, ..., \frac{k_{1m_1}}{n_1} \right) + n_1 H \left(\frac{k_{21}}{n_2}, ..., \frac{k_{2m_2}}{n_2} \right). \end{aligned} \tag{5.45}$$

Proof: (Left as an exercise.) □

Theorem 5.20 *We have the upper bound*

$$(n+r)H\left(\frac{n}{n+r},\frac{r}{n+r}\right)\leq r\left(\log\frac{n}{r}+3\right) \tag{5.46}$$

Proof: Using (5.39), we may write

$$\begin{aligned}(n+r)H\left(\frac{n}{n+r},\frac{r}{n+r}\right)&=r\log\frac{n+r}{r}+n\log\frac{n+r}{n}\\&\leq r\log\left(\frac{n}{r}+1\right)+r\log e\\&\leq r\left(\log\frac{n}{r}+3\right).\end{aligned} \tag{5.47}$$

□

Now we may prove that the Laplace estimator yields a universal code.

Theorem 5.21 *The prefix-free Shannon-Fano code with respect to Laplace estimator R_0 satisfies universality criterion (5.27).*

Proof: We can express

$$R_0(x_1^n)=\binom{n+m-1}{k_1,...,k_m,m-1}^{-1}, \tag{5.48}$$

where $k_l=\sum_{i=1}^n \mathbf{1}\{x_i=l\}$ is the frequency of digit l. By the previous three theorems, we obtain

$$\begin{aligned}-\log R_0(x_1^n)&\leq (n+m)H\left(\frac{k_1}{n+m},...,\frac{k_m}{n+m},\frac{m-1}{n+m},\frac{1}{n+m}\right)\\&\leq m\left(\log\frac{n}{m}+3\right)+\log(m-1)+3+nH\left(\frac{k_1}{n},...,\frac{k_m}{n}\right)\\&\leq m\left(\log n+6\right)+n\mathcal{H}(x_1^n),\end{aligned} \tag{5.49}$$

which is enough to assert criterion (5.27). □

***

To recapitulate this chapter, we have exhibited some universal codes for IID processes over a finite alphabet. It can be shown that there are no universal codes for a countably infinite alphabet. In the following chapters, we will construct other examples of universal codes and we will lift this concept to other processes that satisfy a generalized law of large numbers. This generalized law of large numbers is called the ergodic theorem. It holds for instance for irreducible Markov processes, to be discussed in Chapter 6.

Further reading

Maximum likelihood plays the central role in statistical inference. The complex history of its invention was described by Stephen Stigler [118]. The normalized rather than penalized maximum likelihood was invented by Yuri Shtarkov [116]. The Stirling approximation is named after James Stirling and it was discovered in his correspondence with Abraham de Moivre around 1729. The Laplace estimator is due to Pierre-Simon de Laplace [88]. The idea of universal coding dates back to Andrey Kolmogorov's seminal paper [84]. The importance of the asymptotic equipartition for information theory was noticed by Claude Shannon [114]. The non-existence of universal codes for a countably infinite alphabet was shown by John Kieffer in the more general setting of stationary ergodic processes [81]. Later this reasoning was simplified [62, 102]. A book on interactions between universal coding and mathematical statistics was written by Peter Grünwald [60]. The books by Thomas Cover and Joy Thomas [26] and by Imre Csiszár and János Körner [30] are also recommended. The idea of types, discussed in the exercises to this chapter, was developed by Imre Csiszár [29].

Thinking exercises

1. Let $(X_i)_{i\in\mathbb{N}}$ be an IID process over a finite alphabet. Show that
   (a) $\mathbf{E}\,\mathcal{H}(X_1^n) \le \mathbf{E}\,\mathcal{H}(X_1^{2n})$;
   (b) $\mathbf{E}\,\mathcal{H}(X_1^n) \le H(X_i)$;
   (c) $\lim_{n\to\infty} \mathbf{E}\,\mathcal{H}(X_1^n) = H(X_i)$.

2. We have two random variables X and Y with disjoint sets of values. Let Z take values $P(Z=0)=p$ and $P(Z=1)=1-p$ and be independent from X and Y. Compute the entropy of variable
$$U = \begin{cases} X, & \text{if } Z=0, \\ Y, & \text{if } Z=1. \end{cases} \tag{5.50}$$

3. Prove Theorem 5.19.

4. *Type code:* A sequence x_1^n such that $x_i \in \mathbb{X}$ is called a sequence of type $(k_1,...,k_m)$ for $k_l := \sum_{i=1}^n \mathbf{1}\{x_i = l\}$ being the frequencies of digits $l \in \mathbb{X}$. The type code $B : \mathbb{X}^* \to \{0,1\}^*$ is defined as
$$B(x_1^n) := \mathrm{una}''(n)\,\mathrm{bin}_{n+1}(k_1)...\,\mathrm{bin}_{n+1}(k_{m-1})\,\mathrm{bin}_T(t) \tag{5.51}$$

where T is the number of sequences of length n and type $(k_1, ..., k_m)$ and x_1^n is the t-th sequence of type $(k_1, ..., k_m)$ enumerated in some fixed order. Show that the type code is universal for IID processes.

5. Show that

$$\lim_{n\to\infty} \left[\ln n! - n \ln n + n - \ln \sqrt{2\pi n}\right] = 0. \tag{5.52}$$

Consequently, for an IID process $(X_i)_{i\in\mathbb{N}}$ over alphabet $\mathbb{X} = \{1, 2, ..., m\}$ and $K_l := \sum_{i=1}^n \mathbf{1}\{X_i = l\}$, show that

$$\lim_{n\to\infty} \left[\log \binom{n}{K_1, ..., K_m} - nH\left(\frac{K_1}{n}, ..., \frac{K_m}{n}\right) + \frac{m-1}{2}\log(2\pi n)\right] = -\frac{1}{2}\sum_{l=1}^m \log P(X_i = l) \text{ a.s.} \tag{5.53}$$

6. Consider three universal codes: the PML code, the Laplace estimator, and the type code. Which one of them is the shortest for large input string lengths? Is this result influenced by the size of the alphabet?

Chapter 6

Memory

Markov processes on a countable state space. Communicating classes. Finite and irreducible Markov processes. Invariant distributions. Uniqueness and existence of invariant distribution. Recurrence times. Markov and strong Markov property. Ergodic theorem for Markov processes. Higher order Markov processes. Asymptotic equipartition for Markov processes.

Whereas IID processes are central to theory of mathematical statistics, Markov processes are the simplest processes with some dependence on the past. Markov processes exhibit some rudimentary dependence—exactly only on the single directly preceding observation or symbol. The idea of limited dependence can be easily generalized to dependence on a fixed number of previous symbols. So generalized higher order Markov processes were proposed by as primitive statistical models of human language.

Although human language does not seem to a limited dependence in this sense, many intuitions that are developed for finite Markov processes can be generalized to more complex stochastic processes. In this chapter, we will study properties of Markov processes in more detail. The general theory of Markov processes is much richer than what we sketch in this chapter.

Let $\mathbb{X}$ be a countable set, whose elements will be called states. A vector $\pi : \mathbb{X} \to [0, 1]$ is called a distribution when it satisfies

$$\sum_{x \in \mathbb{X}} \pi(x) = 1, \tag{6.1}$$

whereas a matrix $\tau : \mathbb{X} \times \mathbb{X} \to [0, 1]$ is called stochastic when it satisfies

$$\sum_{x \in \mathbb{X}} \tau(y, x) = 1 \tag{6.2}$$

for all $y \in \mathbb{X}$. The following small modification of the IID process is called a Markov process. In this modification, each random variable depends only on the directly preceding random variable.

Definition 6.1 (Markov process) *A stochastic process $(X_i)_{i\in\mathbb{N}}$ over a countable alphabet $\mathbb{X}$ is called a Markov process if for all $n \in \mathbb{N}$, we have*

$$P(X_1^n = x_1^n) = \pi(x_1) \prod_{i=2}^{n} \tau(x_{i-1}, x_i) \tag{6.3}$$

for some vector $\pi : \mathbb{X} \to [0, 1]$, called the initial distribution, and some matrix $\tau : \mathbb{X} \times \mathbb{X} \to [0, 1]$, called the transition matrix.

It is obvious that initial distribution π above is a distribution, whereas transition matrix τ is a stochastic matrix.

The long-run behavior of a Markov process depends on the structure of the transition matrix. This behavior is relatively simple if the state space $\mathbb{X}$ is finite and coefficients of the transition matrix are all strictly positive, $\tau(y, x) > 0$ for all $y, x \in \mathbb{X}$. More complicated phenomena arise when this does not hold. To describe them, it pays off to study a certain equivalence relation on elements of the state space $\mathbb{X}$.

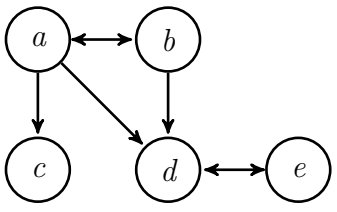
Definition 6.2 (communication) *For a Markov process with a given transition matrix τ, we say that x leads to y and write it as $x \to y$ if*

$$P(X_n = y \text{ for some } n \in \mathbb{N} | X_1 = x) > 0. \tag{6.4}$$

We also say that x communicates with y and write $x \leftrightarrow y$ if $x \to y$ and $y \to x$.

Example 6.3 *Consider a transition matrix and the corresponding graph of communicating states:*

τ	a	b	c	d	e
a	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0
b	$\frac{1}{6}$	0	0	$\frac{5}{6}$	0
c	0	0	1	0	0
d	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
e	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$



(6.5)

Theorem 6.4 *Relation $\leftrightarrow$ is an equivalence relation on $\mathbb{X}$, i.e., it is*

- *reflexive: $x \leftrightarrow x$,*

- *symmetric:* $x \leftrightarrow y$ *if and only if* $y \leftrightarrow x$,
- *transitive:* $x \leftrightarrow y$ *and* $y \leftrightarrow z$ *implies* $x \leftrightarrow z$.

Proof:

- The reflexivity follows by $P(X_1 = x|X_1 = x) = 1 > 0$.
- The symmetry follows by the definition of $\leftrightarrow$.
- The transitivity follows since $x \to y$ holds if and only if

$$\tau(x_1, x_2)\tau(x_2, x_3)...\tau(x_{n-1}, x_n) > 0 \tag{6.6}$$

for some $x_1, x_2, ..., x_n$ where $x_1 = x$ and $x_n = y$. To prove the latter, we observe that (6.6) holds if and only if

$$P(X_n = x_n|X_1 = x_1) = \tau^{n-1}(x_1, x_n) := \sum_{x_2, x_3, ..., x_{n-1}} \tau(x_1, x_2)\tau(x_2, x_3)...\tau(x_{n-1}, x_n) > 0. \tag{6.7}$$

In turn, (6.7) is equivalent to $x \to y$ since

$$\begin{aligned} P(X_n = x_n|X_1 = x_1) &\leq P(X_m = y \text{ for some } m \in \mathbb{N}|X_1 = x) \\ &\leq \sum_{n=1}^{\infty} P(X_n = x_n|X_1 = x_1). \end{aligned} \tag{6.8}$$

□

For an equivalence relation $\leftrightarrow$, an equivalence class is a set of arguments that are equivalent to a given argument:

$$[x] := \{y \in \mathbb{X} : x \leftrightarrow y\}. \tag{6.9}$$

The equivalence classes are disjoint and partition the state space $\mathbb{X}$. In case of the relation $\leftrightarrow$, the equivalence classes are called communicating classes.

Example 6.5 *For transition matrix (6.5), the communicating classes are:* $\{a, b\}$, $\{c\}$, $\{d, e\}$.

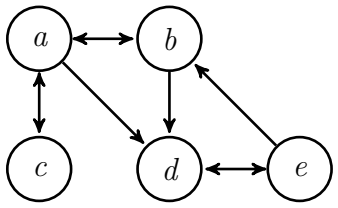
Definition 6.6 (closed class) *A communicating class* $C \subset \mathbb{X}$ *is called closed if there is no escape from it, i.e., if* $x \in C$ *and* $x \to y$ *implies* $y \in C$.

Example 6.7 *For transition matrix (6.5), the closed classes are:* $\{c\}$, $\{d, e\}$.

Definition 6.8 (irreducible Markov process) *A transition matrix* τ *or the respective Markov process are called irreducible if the state space* $\mathbb{X}$ *is the single communicating class.*

Example 6.9 *Transition matrix (6.5) is not irreducible. Here is an example of an irreducible transition matrix and the corresponding graph of communicating states:*

τ	a	b	c	d	e
a	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0
b	$\frac{1}{6}$	0	0	$\frac{5}{6}$	0
c	$\frac{1}{9}$	0	$\frac{8}{9}$	0	0
d	0	0	0	$\frac{2}{3}$	$\frac{1}{3}$
e	0	$\frac{1}{12}$	0	$\frac{1}{2}$	$\frac{5}{12}$



(6.10)

Definition 6.10 (finite Markov process) *A Markov process is called (in)finite if the state space* $\mathbb{X}$ *is (in)finite.*

Example 6.11 *Transition matrices (6.5) and (6.10) define finite Markov processes.*

Example 6.12 *Consider now* $\mathbb{X} = \mathbb{N}$, $\tau(n, 1) = 1/2$ *and* $\tau(n, n + 1) = 1/2$. *This transition matrix defines an infinite Markov process.*

The next important concept is an invariant distribution.

Definition 6.13 (invariant distribution) *A distribution* $\bar{\pi}$ *is called invariant for a given transition matrix* τ *if*

$$\sum_{y \in \mathbb{X}} \bar{\pi}(y)\tau(y, x) = \bar{\pi}(x) \tag{6.11}$$

for all $x \in \mathbb{X}$.

Obviously, if the initial distribution π is invariant then $P(X_i = x) = \pi(x)$ with $n \in \mathbb{N}$ holds for the respective Markov process.

Definition 6.14 (stationary Markov process) *Markov processes with an invariant initial distribution are called stationary.*

Another important interpretation of the invariant distribution is the equilibrium interpretation:

Theorem 6.15 *Let $\mathbb{X}$ be a finite state space and consider the n-th power τ^n of the transition matrix given by (6.7). Suppose that for some $z \in \mathbb{X}$ there exist limits $\lim_{n\to\infty} \tau^n(z,x)$ for all $x \in \mathbb{X}$. Then $\bar{\pi}(x) := \lim_{n\to\infty} \tau^n(z,x)$ is an invariant distribution.*

Proof: By the finiteness of $\mathbb{X}$, we can interchange the sums and the limits:

$$\sum_{x\in\mathbb{X}} \bar{\pi}(x) = \sum_{x\in\mathbb{X}} \lim_{n\to\infty} \tau^n(z,x) = \lim_{n\to\infty} \sum_{x\in\mathbb{X}} \tau^n(z,x) = 1, \tag{6.12}$$

$$\begin{aligned} \bar{\pi}(x) &= \lim_{n\to\infty} \tau^{n+1}(y,x) = \lim_{n\to\infty} \sum_{y\in\mathbb{X}} \tau^n(z,y)\tau(y,x) \\ &= \sum_{y\in\mathbb{X}} \lim_{n\to\infty} \tau^n(z,y)\tau(y,x) = \sum_{y\in\mathbb{X}} \bar{\pi}(y)\tau(y,x). \end{aligned} \tag{6.13}$$

So $\bar{\pi}$ is a distribution and is invariant. □

Existence of an invariant distribution is not always guaranteed.

Example 6.16 *Consider $\mathbb{X} = \mathbb{N}$ and $\tau(n, n+1) = 1$. Then $\bar{\pi}(n+1) = \bar{\pi}(n)$ and condition $\sum_{n\in\mathbb{N}} \bar{\pi}(n) = 1$ cannot be satisfied.*

Example 6.17 *Consider $\mathbb{X} = \mathbb{N}$ and $\tau(n,1) = 1/2$ and $\tau(n, n+1) = 1/2$. Then $\bar{\pi}(n+1) = \bar{\pi}(n)/2 = 2^{-n}\bar{\pi}(1)$, so $\bar{\pi}(1) = 1/(1 + 2^{-1} + 2^{-2} + ...) = 1/2$. Thus the invariant distribution is $\bar{\pi}(n) = 2^{-n}$.*

The above example suggests that the invariant distribution may fail to exist only if the state space is infinite. It is indeed so.

Theorem 6.18 *Let $(X_i)_{i\in\mathbb{N}}$ be a finite Markov process. Then invariant distribution exists but need not be unique.*

Proof: (Omitted. See Theorems 1.5.5, 1.5.6, and 1.7.6 of [101].) □

Example 6.19 *Consider transition matrix (6.5), which is not irreducible. One invariant distribution is $\bar{\pi}_1(c) = 1$, another is $\bar{\pi}_2(d) = \bar{\pi}_2(e) = 1/2$. Also linear combinations $\lambda_1\bar{\pi}_1 + \lambda_2\bar{\pi}_2$, where $\lambda_1, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$, are invariant distributions.*

The above example suggests that invariant distribution has to be unique if there is only one communicating class. It is indeed so.

Theorem 6.20 *Let $(X_i)_{i\in\mathbb{N}}$ be an irreducible Markov process. Then the invariant distribution is unique if it exists.*

Proof: (Omitted. See Theorem 1.7.7 of [101].) □

Resuming, for a finite and irreducible Markov process, the invariant distribution exists and is unique.

Now let us introduce some random variables which tell how long we have to wait for the subsequent occurrences of a given state $x \in \mathbb{X}$.

Definition 6.21 (passage and recurrence times) *Let $(X_i)_{i\in\mathbb{N}}$ be a Markov process. We inductively define random variables called passage times*

$$T_0^x := 0, \tag{6.14}$$

$$T_n^x := \inf\left\{n \in \mathbb{N} : n > T_{n-1}^x, X_n = x\right\}. \tag{6.15}$$

Having these, we define random variables called successive recurrence times

$$R_n^x := \begin{cases} T_{n+1}^x - T_n^x & \text{if } T_n^x < \infty, \\ \infty & \text{otherwise.} \end{cases} \tag{6.16}$$

Successive recurrence times are simply the time intervals between subsequent occurrences of a state $x \in \mathbb{X}$. It turns out that they are independent and identically distributed. The reason is that the process forgets its previous history when restarted in state x.

Theorem 6.22 *Random variables $R_1^x, R_2^x, R_3^x, ...$ form an IID process.*

Proof: The fact that $R_1^x, R_2^x, R_3^x, ...$ is an IID process is a direct consequence of the following strong Markov property: Let $T : \Omega \to \mathbb{N} \cup \{\infty\}$ be a random variable such that event $(T = n)$ depends only on X_1^n. Then process X_T^∞ conditioned on event $(T < \infty, X_T = x)$ is a Markov process with the initial distribution concentrated on x (i.e., $\pi(x) = 1$) and with the same transition matrix as process $(X_i)_{i\in\mathbb{N}}$. In the considered case, it suffices to take $T = T_n^x$. □

The following theorem, called the ergodic theorem for Markov processes, is an application of the strong law of large numbers for process $R_1^x, R_2^x, R_3^x, ...$ It states that the relative frequency of sampling a given state $x \in \mathbb{X}$ equals its invariant probability $\bar{\pi}(x)$. In particular, if the process is stationary then this invariant probability equals the marginal probability of the state, $\bar{\pi}(x) = \pi(x) = P(X_i = x)$.

Theorem 6.23 (ergodic theorem) *Let $(X_i)_{i\in\mathbb{N}}$ be an irreducible Markov process such that the invariant distribution $\bar{\pi}$ exists. Then*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i=x\}=\frac{1}{\mathbf{E}\,R_i^x}=\bar{\pi}(x)\ \ a.s. \tag{6.17}$$

Proof: By the strong law of large numbers for process $R_1^x, R_2^x, R_3^x, ...$ with $\mathbf{E}\,R_i^x=\mu(x)$, we have

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}R_i^x=\mu(x)\ \text{a.s.} \tag{6.18}$$

Denote $V_n:=\sum_{i=1}^{n}\mathbf{1}\{X_i=x\}$. We have

$$\sum_{i=0}^{V_n}R_i^x\geq n,\quad \sum_{i=0}^{V_n-1}R_i^x\leq n. \tag{6.19}$$

Hence

$$\frac{V_n}{\sum_{i=0}^{V_n}R_i^x}\leq\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i=x\}\leq\frac{V_n}{\sum_{i=0}^{V_n-1}R_i^x}. \tag{6.20}$$

If $(X_i)_{i\in\mathbb{N}}$ is an irreducible Markov process such that the invariant distribution $\bar{\pi}$ exists then it can be shown that $P(R_0^x<\infty)=1$ and $\lim_{n\to\infty}V_n=\infty$ a.s. Thus

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i=x\}=\frac{1}{\mu(x)}\ \text{a.s.} \tag{6.21}$$

It remains to show that $1/\mu(x)=\bar{\pi}(x)$. We will apply the Lebesgue dominated convergence theorem. Hence, we obtain

$$\begin{aligned}\frac{1}{\mu(x)}&=\mathbf{E}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i=x\}=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\,\mathbf{1}\{X_i=x\}\\&=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}P(X_i=x)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}P(X_{i+1}=x)\\&=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\mathbb{X}}P(X_i=y)P(X_{i+1}=x|X_i=y)\\&=\sum_{y\in\mathbb{X}}\frac{1}{\mu(y)}\tau(y,x).\end{aligned} \tag{6.22}$$

Thus, in view of the uniqueness of the invariant distribution, we derive $1/\mu(x)=\bar{\pi}(x)$. □

Let us tamper with the statement of the ergodic theorem a bit. We recall that Markov processes with an invariant initial distribution are called stationary. Let us generalize the concept of a stationary Markov process to other classes of processes.

Definition 6.24 (stationary process) *A stochastic process $(X_i)_{i\in\mathbb{N}}$ over a countable alphabet $\mathbb{X}$ is called stationary if for all $t, n \in \mathbb{N}$ and all strings $x_1^n \in \mathbb{X}^*$, we have*

$$P(X_{t+1}^{t+n} = x_1^n) = P(X_1^n = x_1^n). \tag{6.23}$$

It can be easily checked that a stationary Markov process is a stationary process in the above sense.

Now, let us define higher order Markov processes.

Definition 6.25 (higher order Markov process) *A stochastic process $(X_i)_{i\in\mathbb{N}}$ over a countable alphabet $\mathbb{X}$ is called a k-th order Markov process if for all $n > k$, we have*

$$P(X_n|X_1^{n-1}) = P(X_n|X_{n-k}^{n-1}). \tag{6.24}$$

Analogously, we call IID processes 0-th order Markov processes. We call a process a higher order Markov process if it is k-th order Markov for some $k \geq 0$.

In particular, a Markov process is a 1-st order Markov process and conversely. Any k-th order Markov process is a $(k+1)$-th order Markov process. Moreover, if $(X_i)_{i\in\mathbb{N}}$ is a k-th order Markov process then $(Y_i)_{i\in\mathbb{N}}$ with $Y_i = X_i^{i+k-1}$ is a Markov process. We will call this $(X_i)_{i\in\mathbb{N}}$ irreducible if $(Y_i)_{i\in\mathbb{N}}$ is irreducible and finite if so is $(Y_i)_{i\in\mathbb{N}}$. It can be checked that $(X_i)_{i\in\mathbb{N}}$ is stationary if and only if $(Y_i)_{i\in\mathbb{N}}$ is stationary.

In view of the above, we can restate the ergodic theorem as follows.

Theorem 6.26 (ergodic theorem) *Let $(X_i)_{i\in\mathbb{N}}$ be a stationary irreducible higher order Markov process over a finite alphabet $\mathbb{X}$ and let $f : \mathbb{X}^k \to [0, \infty]$ be a non-negative real function. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_{i+1}^{i+k}) = \mathbf{E}\, f(X_1^k) \ \ a.s. \tag{6.25}$$

Proof: Without loss of generality we may assume that k is the order of process $(X_i)_{i\in\mathbb{N}}$. From the ergodic theorem for the Markov process $(Y_i)_{i\in\mathbb{N}}$

with $Y_i = X_{i+1}^{i+k}$, we obtain

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} f(X_{i+1}^{i+k}) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} f(Y_i) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\sum_{y\in\mathbb{X}^k} f(y)\mathbf{1}\left\{Y_i = y\right\} \\ &= \sum_{y\in\mathbb{X}^k} f(y)\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{1}\left\{Y_i = y\right\} = \sum_{y\in\mathbb{X}^k} f(y)P(Y_1 = y) \\ &= \mathbf{E}\, f(X_1^k), \end{aligned} \tag{6.26}$$

where the limit and the summation can be exchanged since the summation consists of finitely many terms. □

A direct consequence of this restatement is the asymptotic equipartition.

Theorem 6.27 (asymptotic equipartition) *For any stationary irreducible k-th order Markov process $(X_i)_{i\in\mathbb{N}}$ over a finite alphabet $\mathbb{X}$, we have*

$$\lim_{n\to\infty}\frac{[-\log P(X_1^n)]}{n} = H(X_i|X_{i-k}^{i-1}) \ \textit{a.s.} \tag{6.27}$$

Proof: By definition $P(X_n|X_1^{n-1}) = P(X_n|X_{n-k}^{n-1})$, so

$$P(X_1^n) = P(X_1^k)\prod_{i=k+1}^{n} P(X_i|X_1^{i-1}) = P(X_1^k)\prod_{i=k+1}^{n} P(X_i|X_{i-k}^{i-1}). \tag{6.28}$$

Hence by the ergodic theorem we obtain

$$\begin{aligned}\lim_{n\to\infty}\frac{[-\log P(X_1^n)]}{n} &= \lim_{n\to\infty}\frac{1}{n}\left[-\log P(X_1^k) - \sum_{i=k+1}^{n}\log P(X_i|X_{i-k}^{i-1})\right] \\ &= \mathbf{E}\left[-\log P(X_i|X_{i-k}^{i-1})\right] = H(X_i|X_{i-k}^{i-1}) \text{ a.s.} \end{aligned} \tag{6.29}$$

□

***

Recapitulating this chapter, we have generalized the strong law of large numbers as the ergodic theorem and asymptotic equipartition for higher order Markov processes. These two theorems will play the fundamental role in theory of universal codes for processes with memory, to be discussed in Chapters 7 and 8. Besides higher order Markov processes, same universal codes are good also for arbitrary stationary ergodic processes as we will see in Chapter 10. Stationary and ergodic processes, to be detailed in Chapter 9, generalize Markov processes with stationary distributions and irreducible transition matrices, respectively.

Further reading

Andrey Markov published his first paper on Markov processes in 1906. More often cited in the popular literature is his attempt to model the frequencies of consonants and vowels in the poem *Eugene Onegin* by Alexander Pushkin [92, 93]—as a Markov process, of course. This idea was further pursued by Claude Shannon [114, 115], who considered approximating texts in natural language also by higher order Markov processes. In this chapter, we only presented discrete-time and discrete-space Markov processes. A concise modern introduction to theory of Markov processes, covering both discrete and non-discrete cases, was written by James Norris [101].

Thinking exercises

1. Tell which of the following transition matrices are irreducible:

$$\tau = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}; \qquad \tau = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/4 & 1/4 & 1/2 \\ 0 & 3/4 & 1/4 \end{pmatrix};$$

$$\tau = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 3/5 & 2/5 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/4 & 3/4 \end{pmatrix}.$$

2. For a Markov process $(X_i)_{i\in\mathbb{N}}$, show that variables X_i and X_k are conditionally independent given X_j if $i \leq j \leq k$.

3. For a Markov process $(X_i)_{i\in\mathbb{N}}$, prove that

$$I(X_i; X_k) \leq I(X_i; X_j) \text{ for } i \leq j \leq k. \tag{6.30}$$

4. For a stationary k-th order Markov process $(X_i)_{i\in\mathbb{N}}$, show that

$$\lim_{n\to\infty} \frac{H(X_1^n)}{n} = H(X_i|X_{i-k}^{i-1}) \text{ a.s.} \tag{6.31}$$

5. *Mixing processes:* A stationary process $(X_i)_{i\in\mathbb{N}}$ over countable alphabet $\mathbb{X}$ is called mixing if for all $k \in \mathbb{N}$ and all strings $x_1^k, y_1^k \in \mathbb{X}^*$ such that $P(X_1^k = y_1^k) > 0$, we have (6.32).

$$\lim_{n\to\infty} P(X_{n+1}^{n+k} = x_1^k | X_1^k = y_1^k) = P(X_1^k = x_1^k). \tag{6.32}$$

Show that a stationary Markov process $(X_i)_{i\in\mathbb{N}}$ over a countable alphabet $\mathbb{X}$ is mixing if $P(X_{i+1} = x|X_i = y) > 0$ for all $x, y \in \mathbb{X}$.

6. *Aperiodic Markov processes:* A stationary Markov process $(X_i)_{i\in\mathbb{N}}$ over a countable alphabet $\mathbb{X}$ is called aperiodic if for each $x \in \mathbb{X}$ there exists an $N \in \mathbb{N}$ such that for all $n \geq N$ we have $P(X_{i+n} = x|X_i = x) > 0$.

   (a) Show that for a stationary irreducible aperiodic Markov process $(X_i)_{i\in\mathbb{N}}$, we have

   $$\lim_{n\to\infty} P(X_{i+n} = x|X_i = y) = P(X_i = x) \tag{6.33}$$

   for all $x, y \in \mathbb{X}$.

   (b) Show that a stationary Markov process $(X_i)_{i\in\mathbb{N}}$ is mixing with $P(X_i = x) > 0$ for all $x \in \mathbb{X}$ if and only if it is irreducible aperiodic [36].

Chapter 7

Phrases

Universal codes. Universality criteria. Distinct parsing. Lempel-Ziv parsing. Lempel-Ziv code. Ziv inequality. Universality of the Lempel-Ziv code. Dictionary grammars. Grammar expansion. Minimal grammar-based code. Universality of the minimal grammar-based code.

In this chapter, we will discuss two simple examples of codes that are universal for higher order Markov processes. These codes are the Lempel-Ziv code and the minimal grammar-based code. The Lempel-Ziv code is highly practical, whereas the minimal grammar-based code is mostly an intellectual excursion due its computational intractability. Before we discuss these codes, we have to explain what universal codes are in general. We will state the definition and demonstrate two criteria that allow to check whether a given code is universal for higher order Markov processes.

Universality criteria

First, let us fix attention on the achievable lower bound of the code length in general. This will be done via the asymptotic equipartition. We will focus on processes that have a well-defined limit of the rate of pointwise entropy. Let $\mathbb{X}$ be a finite alphabet.

Definition 7.1 (equipartitioned process) *A process $(X_i)_{i\in\mathbb{N}}$ with random variables $X_i : \Omega \to \mathbb{X}$ is called an equipartitioned process if there exists a constant h, called the entropy rate, such that*

$$\lim_{n\to\infty} \frac{\mathbf{E}\left[-\log P(X_1^n)\right]}{n} = h, \tag{7.1}$$

$$\lim_{n\to\infty} \frac{[-\log P(X_1^n)]}{n} = h \ a.s. \tag{7.2}$$

Theorem 7.2 *Any stationary irreducible higher order Markov process is equipartitioned.*

Proof: Let $(X_i)_{i\in\mathbb{N}}$ be a stationary irreducible k-th order Markov process. Its entropy rate is $h = H(X_i|X_{i-k}^{i-1})$ by the asymptotic equipartition. □

More examples of equipartitioned processes—in the instance of stationary ergodic processes—will come in Chapter 10.

Now we define universal codes in general.

Definition 7.3 (universal code) *Let $\mathcal{C}$ be a subclass of equipartitioned processes $(X_i)_{i\in\mathbb{N}}$ with random variables $X_i : \Omega \to \mathbb{X}$. Let $B : \mathbb{X}^* \to \{0,1\}^*$ be a uniquely decodable code. Code B is called universal for class $\mathcal{C}$ if for any process $(X_i)_{i\in\mathbb{N}}$ from $\mathcal{C}$, we have*

$$\lim_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)|}{n} = h, \tag{7.3}$$

$$\lim_{n\to\infty} \frac{|B(X_1^n)|}{n} = h \ a.s. \tag{7.4}$$

Theorem 7.4 *For a uniquely decodable code $B : \mathbb{X}^* \to \{0,1\}^*$ and an equipartitioned process $(X_i)_{i\in\mathbb{N}}$, we have*

$$\liminf_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)|}{n} \geq h, \tag{7.5}$$

$$\liminf_{n\to\infty} \frac{|B(X_1^n)|}{n} \geq h \ a.s. \tag{7.6}$$

Proof: By the Barron lemma, we obtain the lower bound

$$\liminf_{n\to\infty} \frac{|B(X_1^n)|}{n} \geq \lim_{n\to\infty} \frac{[-\log P(X_1^n)]}{n} = h \text{ a.s.} \tag{7.7}$$

Similarly, by the source coding inequality, we obtain

$$\liminf_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)|}{n} \geq \lim_{n\to\infty} \frac{\mathbf{E}\,[-\log P(X_1^n)]}{n} = h \tag{7.8}$$

□

There are two sufficient criteria for a code to be universal with respect to higher order Markov processes. The first one is as follows.

Theorem 7.5 (universality criterion) *Let $B : \mathbb{X}^* \to \{0,1\}^*$ be a uniquely decodable code. Code B is universal for stationary irreducible higher order Markov processes over a finite alphabet $\mathbb{X}$ if for any conditional probability distribution $\tau : \mathbb{X}^{k+1} \to [0,1]$, where $\tau(x_{k+1}|x_1^k) \geq 0$ and $\sum_{x_{k+1}} \tau(x_{k+1}|x_1^k) = 1$, and for all strings $x_1^n \in \mathbb{X}^*$, we have*

$$|B(x_1^n)| \leq C(n,k) - \log \prod_{i=1}^{n} \tau(x_i|x_{i-k}^{i-1}), \tag{7.9}$$

where $\lim_{k\to\infty} \limsup_{n\to\infty} C(n,k)/n = 0$.

Proof: Let B be a uniquely decodable code and $(X_i)_{i\in\mathbb{N}}$ be a stationary irreducible k-th order Markov process. If (7.9) holds then by the ergodic theorem we obtain

$$\limsup_{n\to\infty} \frac{|B(X_1^n)| - C(n,k)}{n} \leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} \left[-\log P(X_i|X_{i-k}^{i-1}) \right] = H(X_i|X_{i-k}^{i-1}) \text{ a.s.} \tag{7.10}$$

This holds for any $k \geq 1$, so for stationary irreducible higher order Markov processes we have

$$\limsup_{n\to\infty} \frac{|B(X_1^n)|}{n} \leq \lim_{k\to\infty} \left[\limsup_{n\to\infty} \frac{C(n,k)}{n} + H(X_i|X_{i-k}^{i-1}) \right] = h \text{ a.s.} \tag{7.11}$$

Similarly, taking the expectations we obtain

$$\limsup_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)| - C(n,k)}{n} \leq \mathbf{E}\left[-\log P(X_i|X_{i-k}^{i-1}) \right] = H(X_i|X_{i-k}^{i-1}). \tag{7.12}$$

This also holds for any $k \geq 1$, so for stationary irreducible higher order Markov processes we have

$$\limsup_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)|}{n} \leq \lim_{k\to\infty} \left[\limsup_{n\to\infty} \frac{C(n,k)}{n} + H(X_i|X_{i-k}^{i-1}) \right] = h. \tag{7.13}$$

Thus in view of Theorem 7.4 we have established universality of B. □

The second universality criterion is based on a k-block empirical entropy rather than the k-order conditional empirical entropy.

Theorem 7.6 (universality criterion) *Let $B : \mathbb{X}^* \to \{0,1\}^*$ be a uniquely decodable code. Code B is universal for stationary irreducible higher order Markov processes over a finite alphabet $\mathbb{X}$ if for any block probability distribution $\pi : \mathbb{X}^k \to [0,1]$, where $\pi(x_1^k) \geq 0$ and $\sum_{x_1^k} \pi(x_1^k) = 1$, and for all strings $x_1^n \in \mathbb{X}^*$, we have*

$$|B(x_1^n)| \leq C(n,k) - \frac{1}{k} \log \prod_{i=0}^{n-k} \pi(x_{i+1}^{i+k}), \tag{7.14}$$

where $\lim_{k\to\infty} \limsup_{n\to\infty} C(n,k)/n = 0$.

Proof: Let B be a uniquely decodable code and $(X_i)_{i\in\mathbb{N}}$ be a stationary irreducible k-th order Markov process. If (7.14) holds then by the ergodic theorem we obtain

$$\limsup_{n\to\infty} \frac{|B(X_1^n)| - C(n,k)}{n} \leq \frac{1}{k} \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-k} \left[-\log P(X_{i+1}^{i+k})\right] = \frac{H(X_1^k)}{k} \text{ a.s.} \tag{7.15}$$

This holds for any $k \geq 1$, so for stationary irreducible higher order Markov processes we have

$$\limsup_{n\to\infty} \frac{|B(X_1^n)|}{n} \leq \lim_{k\to\infty} \left[\limsup_{n\to\infty} \frac{C(n,k)}{n} + \frac{H(X_1^k)}{k}\right] = h \text{ a.s.} \tag{7.16}$$

Similarly, taking the expectations we obtain

$$\limsup_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)| - C(n,k)}{n} \leq \frac{\mathbf{E}\left[-\log P(X_{i+1}^{i+k})\right]}{k} = \frac{H(X_1^k)}{k}. \tag{7.17}$$

This also holds for any $k \geq 1$, so for stationary irreducible higher order Markov processes we have

$$\limsup_{n\to\infty} \frac{\mathbf{E}\,|B(X_1^n)|}{n} \leq \lim_{k\to\infty} \left[\limsup_{n\to\infty} \frac{C(n,k)}{n} + \frac{H(X_1^k)}{k}\right] = h. \tag{7.18}$$

Thus in view of Theorem 7.4 we have established universality of B. □

Lempel-Ziv code

In this section, we will discuss an example of a simple and highly practical universal code called the Lempel-Ziv code. The Lempel-Ziv code operates by splitting the coded strings into smaller phrases. Such a splitting operation is called parsing. We will particularly interested in a distinct parsing, which is a parsing where all phrases are distinct.

Definition 7.7 (distinct parsing) *A sequence of phrases $(w_1, w_2, ..., w_c)$, where $w_i \in \mathbb{X}^*$ is called a distinct parsing of a string $u \in \mathbb{X}^*$ if $w_1 w_2 ... w_c = u$ and $w_i \neq w_j$ for $i \neq j$.*

An important instance of a distinct parsing is the Lempel-Ziv parsing. The Lempel-Ziv parsing is obtained by reading the string from the left and cutting off the shortest phrases that have not appeared before.

Definition 7.8 (Lempel-Ziv parsing) *The Lempel-Ziv parsing of a string $u \in \mathbb{X}^*$ is a distinct parsing $(w_1, w_2, ..., w_c)$ where $w_1 = \lambda$, $w_i = w_{p(i)} z_i \in \mathbb{X}^*$ for $i \geq 2$, $z_i \in \mathbb{X}$, and indices $p(i) < i$ are such that $w_{p(i)}$ is the longest phrase that can be selected.*

By construction, the Lempel-Ziv parsing, except for the last phrase, is a distinct parsing and is unique.

Now we can define the Lempel-Ziv code.

Definition 7.9 (Lempel-Ziv code) *For simplicity, we assume that the coded data are binary sequences, that is $\mathbb{X} = \{0, 1\}$. Consequently, the Lempel-Ziv code $\mathrm{LZ} : \{0,1\}^* \to \{0,1\}^*$ is defined as*

$$\mathrm{LZ}(x_1^n) := \mathrm{una}''(m)\, \mathrm{bin}_2(p(2)) z_2\, \mathrm{bin}_3(p(3)) z_3 ... \mathrm{bin}_{C_n}(p(C_n)) z_{C_n}, \qquad (7.19)$$

where $\mathrm{bin}_n(k)$ is the fixed-length code (1.6) and the Lempel-Ziv parsing of string x_1^n is $(w_1, w_2, ..., w_{C_n})$ with $w_i = w_{p(i)} z_i \in \{0,1\}^*$ for $i \geq 2$.

It can be easily seen that the Lempel-Ziv code is prefix-free and its length can be upper bounded as

$$|\mathrm{LZ}(x_1^n)| \leq (2 \log C_n + 1) + C_n(\log C_n + 2) \leq (C_n + 2)(\log C_n + 2), \quad (7.20)$$

where C_n is the number of Lempel-Ziv phrases for string x_1^n.

The first step to prove universality of the Lempel-Ziv code is the following theorem.

Theorem 7.10 *Let $(W_1, W_2, ..., W_{C_n})$ be a distinct parsing of a string $X_1^n \in \{0,1\}^*$. We have inequality*

$$\frac{C_n}{n} \leq \frac{1}{\log n - \log(\log n + 2) - 3}. \qquad (7.21)$$

Proof: Let $n_k = \sum_{j=1}^{k} j 2^j = (k-1)2^{k+1} + 2$ be the sum of lengths of distinct phrases that are not longer than k. The number of phrases in a distinct parsing will be maximal if the phrases are as short as possible. For $n_k \leq n <$

n_{k+1} this happens if we take all phrases of length $\leq k$ and $(n-n_k)/(k+1)$ phrases of length $k+1$. Then

$$C_n \leq \sum_{j=1}^{k} 2^j + \frac{n-n_k}{k+1} = 2^{k+1} - 2 + \frac{n-n_k}{k+1} \leq \frac{n_k}{k-1} + \frac{n-n_k}{k+1} \leq \frac{n}{k-1}. \tag{7.22}$$

In the following we will provide a bound for k given n. We have $n \geq n_k = (k-1)2^{k+1} + 2 \geq 2^k$, so

$$k \leq \log n. \tag{7.23}$$

Moreover $n < n_{k+1} = k2^{k+2} + 2 \leq (\log n + 2)2^{k+2}$. Hence

$$k + 2 > \log \frac{n}{\log n + 2}. \tag{7.24}$$

Further transformations yield $k - 1 > \log n - \log(\log n + 2) - 3$. Hence we obtain the claim. □

Thus to show universality of the Lempel-Ziv code, it suffices to prove that expression $C_n \log C_n$ falls below the pointwise entropy $-\log P(X_1^n)$ for any stationary irreducible k-th order Markov process. The subsequent important observation is as follows.

Theorem 7.11 (Ziv inequality) *Let $(X_i)_{i\in\mathbb{N}}$ be a k-th order Markov process. Assume that string X_1^n is parsed into distinct phrases $(W_1, W_2, ..., W_{C_n})$. Let U_j denote the k bits preceding W_j. Next, let C_n^{lu} denote the number of phrases W_i that have length l and context $U_i = k$. We have inequality*

$$\sum_{l,u} C_n^{lu} \log C_n^{lu} \leq -\log P(X_1^n | X_{-k+1}^0) \ \textit{a.s.} \tag{7.25}$$

Proof: Observe that

$$\begin{aligned} -\log P(X_1^n | X_{-k+1}^0) &= -\sum_{i=1}^{n} \log P(X_i | X_{i-k}^{i-1}) = -\sum_{j=1}^{C_n} \log P(W_j | U_j) \\ &= -\sum_{l,u} C_n^{lu} \cdot \frac{1}{C_n^{lu}} \sum_{j:|W_j|=l, U_j=u} \log P(W_j|U_j) \\ &\geq -\sum_{l,u} C_n^{lu} \log \left(\frac{1}{C_n^{lu}} \sum_{j:|W_j|=l, U_j=u} P(W_j|U_j) \right) \ \text{a.s.}, \end{aligned} \tag{7.26}$$

where the inequality follows from the Jensen inequality because the logarithm function is concave. Because the phrases W_j under the sum are distinct, we have $\sum_{j:|W_j|=l,U_j=u} P(W_j|U_j) \leq 1$. Hence the claim follows. □

Another useful auxiliary result is a bound for the Shannon entropy of a random variable taking values in natural numbers by its expectation.

Theorem 7.12 *For a random variable $N : \Omega \to \mathbb{N}$, we have*

$$H(N) \leq 2 \log \mathbf{E}(N+1), \tag{7.27}$$

Proof: Consider probability distribution

$$q(n) := \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)} \tag{7.28}$$

for $n \in \mathbb{N}$. By non-negativity of the Kullback-Leibler divergence and by the Jensen inequality, we have

$$\begin{aligned} 0 \leq \sum_{s=1}^{\infty} P(N=n) \log \frac{P(N=n)}{q(n)} &= \mathbf{E} \log N + \mathbf{E} \log(N+1) - H(N) \\ &\leq 2 \log \mathbf{E}(N+1) - H(N). \end{aligned} \tag{7.29}$$

Regrouping, we obtain the claim. □

The previous two observations can be resumed in the desired proposition. Namely, the Lempel-Ziv code is universal for higher order Markov processes over the binary alphabet.

Theorem 7.13 *The Lempel-Ziv code satisfies universality criterion (7.9).*

Proof: Let $(X_i)_{i \in \mathbb{N}}$ be a stationary irreducible k-th order Markov process over alphabet $\mathbb{X} = \{0, 1\}$. Assume that string X_1^n is parsed into distinct phrases $(W_1, W_2, ..., W_{C_n})$. In the following, we apply the notation from the Ziv inequality. Let L and U be random variables such that

$$P(L = l, U = u) = \frac{C_n^{lu}}{C_n}. \tag{7.30}$$

Using the Ziv inequality, we observe

$$\begin{aligned} C_n \log C_n &\leq C_n \sum_{l,u} \frac{C_n^{lu}}{C_n} \log \frac{C_n^{lu}}{C_n} + \sum_{l,u} C_n^{lu} \log C_n^{lu} \\ &\leq C_n H(L,U) - \log P(X_1^n | X_{-k+1}^0). \end{aligned} \tag{7.31}$$

Thus it suffices to show that

$$\lim_{n\to\infty} \frac{C_n H(L,U)}{n} = 0. \tag{7.32}$$

The expectation of L is

$$\mathbf{E}\, L = \sum_{l,u} \frac{l C_n^{lu}}{C_n} = \frac{n}{C_n}. \tag{7.33}$$

Hence by Theorem 7.12, we obtain

$$H(L) \le 2\log(\mathbf{E}\, L + 1) = 2\log\left(\frac{n}{C_n} + 1\right). \tag{7.34}$$

On the other hand, $H(U) \le k \log \# \mathbb{X} = k$, so

$$H(L,U) \le H(L) + H(U) \le 2\log\left(\frac{n}{C_n} + 1\right) + k. \tag{7.35}$$

Hence Theorem 7.10 yields (7.32). Thus the length of the Lempel-Ziv code $|\mathrm{LZ}(X_1^n)| \le (C_n + 2)(\log C_n + 2)$ satisfies universality criterion (7.9). □

Grammar-based codes

In this section we will develop some other universal codes which may seem quite natural. These codes, which we call the minimal grammar-based codes, can be constructed by first defining a recursive dictionary of substrings and then using binary pointers to these substrings to encode a given string. The shortest code of this form turns out to be universal—even if the binary pointers are far from being optimal.

In the following, we can start defining the minimal grammar-based code. Let the alphabet be $\mathbb{X} = \{1, 2, ..., m\}$.

Definition 7.14 (admissible grammar) *An admissible grammar is a function*

$$G : \{-V_G, -V_G + 1, ..., -1\} \to \{-V_G, -V_G + 1, ..., -1, 1, 2, ..., m\}^+ \tag{7.36}$$

such that for every $G(r) = (r_1, r_2, ..., r_p)$ we have $r_i > r$. Strings $G(r)$ for $r > -V_G$ are called secondary rules, whereas string $G(-V_G)$ is called the primary rule.

The positive numbers in the above are called terminal symbols, whereas the negative numbers are called non-terminal symbols. There is exactly one rule $G(r)$ per non-terminal symbol r and each non-terminal symbol r can be rewritten only onto greater symbols.

The production of a string by an admissible grammar can be also made precise in a simple way in the next definition.

Definition 7.15 (grammar expansion) *For an admissible grammar (7.36), we iteratively define its expansion function*

$$G^* : \{-V_G, -V_G + 1, ..., -1, 1, 2, ..., m\} \to \{1, 2, ..., m\}^+ \tag{7.37}$$

as $G^*(r) := r$ for $r > 0$ and concatenation $G^*(r) := G^*(r_1)G^*(r_2)...G^*(r_p)$ for $G(r) = (r_1, r_2, ..., r_p)$. We say that an admissible grammar G produces a string $u \in \{1, 2, ..., m\}^+$ if $G^*(-V_G) = u$.

The definition of the minimal grammar-based code is quite straightforward. We encode the grammars symbol by symbol and we consider all admissible grammars that encode a given string and we choose the shortest one according to some simple encoding of natural numbers.

Definition 7.16 (local grammar encoder) *Consider a prefix-free code for integers $\psi : \{..., -2, -1, 0, 1, 2, ..., m\} \to \{0, 1\}^*$. The local grammar encoder ψ^* for an admissible grammar G returns string*

$$\psi^*(G) := \psi^*(G(-V_G))\psi^*(G(-V_G + 1))...\psi^*(G(-1))\psi(0), \tag{7.38}$$

where $\psi^*(r_1, r_2, ..., r_p) := \psi(r_1)\psi(r_2)...\psi(r_p)\psi(0)$.

Definition 7.17 (minimal admissible code) *We define the ψ-minimal admissible grammar transform $\Gamma_\psi(u)$ as the admissible grammar G that produces string $u \in \{1, 2, ..., m\}^+$ and minimizes length $|\psi^*(G)|$. Subsequently, the ψ-minimal admissible code $B_\psi : \{1, 2, ..., m\}^+ \to \{0, 1\}^*$ is defined as*

$$B_\psi(u) = \psi^*(\Gamma_\psi(u)). \tag{7.39}$$

The exact theory of minimal grammars depends on the choice of code ψ.

Definition 7.18 (proper code) *A code $\psi : \{..., -2, -1, 0, 1, 2, ..., m\} \to \{0, 1\}^*$ is called m-proper if*

1. *ψ is prefix-free;*
2. *$|\psi(n)| = c_1$ for $0 \leq n \leq m$ and some $c_1 < \infty$;*

3. $|\psi(-n-1)| \geq |\psi(-n)|$ *for* $n \geq 0$;

4. $|\psi(-n)| \leq \log n + 2\log(\log n + 1) + c_2$ *for* $n \geq 1$ *and some* $c_2 < \infty$.

Succinctly, ψ-minimal grammars and codes with an m-proper code ψ will be called m-proper.

Proper codes exist by the Kraft inequality. In particular, there exists an m-proper code ψ such that

$$\psi(n) := \begin{cases} \mathrm{bin}_{m+2}(n), & 0 \leq n \leq m, \\ \mathrm{bin}_{m+2}(m+1)\,\mathrm{una}''(-n), & n < 0, \end{cases} \tag{7.40}$$

where $\mathrm{bin}_{m+2} : \{0, 1, 2, ..., m+1\} \to \{0,1\}^*$ is a fixed-length code with length

$$|\mathrm{bin}_{m+2}(n)| = 1 + \lfloor \log(m+2) \rfloor \tag{7.41}$$

and $\mathrm{una}'' : \{1, 2, ...\} \to \{0,1\}^*$ is the Elias delta code with length

$$|\mathrm{una}''(n)| = \lfloor \log n \rfloor + 2\lfloor \log(\lfloor \log n \rfloor + 1) \rfloor + 2. \tag{7.42}$$

Computing the minimal admissible grammar for a given string may be hard. To overcome the problem of tractability of minimal admissible codes, we may restrict the class of grammars over which we perform the minimization and hope to maintain the universality of the code. A sufficiently rich class is the class of block grammars.

Definition 7.19 (block grammar) *A k-block grammar is an admissible grammar G such that every secondary rule has form $G(r) = (R_1, R_2, ..., R_k)$ where $R_i > 0$ and the primary rule has form*

$$G(-V_G) = (R_1, R_2, ..., R_l, r_1, r_2, ..., r_p, R_{-l'}, R_{-l'+1}, ..., R_{-1}) \tag{7.43}$$

where $R_i > 0$, $r_i < 0$, and $l, l' < k$. An admissible grammar is called a block grammar if it is a k-block grammar for a certain k.

Definition 7.20 (minimal block code) *We define the ψ-minimal block grammar transform $\Gamma^{\#}_{\psi}(u)$ as the block grammar G that produces string $u \in \{1, 2, ..., m\}^+$ and minimizes length $|\psi^*(G)|$. Subsequently, the ψ-minimal block code $B^{\#}_{\psi} : \{1, 2, ..., m\}^+ \to \{0,1\}^*$ is defined as*

$$B^{\#}_{\psi}(u) = \psi^*(\Gamma^{\#}_{\psi}(u)). \tag{7.44}$$

The proper minimal block code can be computed in a time close to linear (with some logarithmic add-ons). For this goal, we have to consider all parsings of the input string into k-blocks and to minimize the code length over k. To determine the optimal code length for each of these parsings, we notice that by inequality $|\psi(-n-1)| \geq |\psi(-n)|$, the optimal secondary rules should be sorted according to the ranked empirical distribution of k-blocks. Once such sorting is performed, the resulted code is minimal within the class of block grammars since all rules have the same length after local encoding by equality $|\psi(n)| = c_1$ for $0 \leq n \leq m$.

Since the proper minimal block code is uniquely decodable and it achieves the constrained global minimum, we can demonstrate easily that this code is strongly universal. The key observation is inequality (7.45), which implies that ranked probabilities are upper bounded by the harmonic series.

Theorem 7.21 (harmonic bound) *Let $\pi : \mathbb{X} \to [0,1]$ be a probability distribution. Let $(x_1, x_2, ...)$ be a sequence of distinct $x_j \in \mathbb{X}$ such that $\pi(x_j) \geq \pi(x_{j+1})$. Then*

$$\pi(x_n) \leq \frac{1}{n}. \tag{7.45}$$

Proof: We have $n\pi(x_n) \leq \sum_{j=1}^{n} \pi(x_j) \leq 1$. □

Theorem 7.22 *The m-proper minimal block code satisfies universality criterion (7.14) for alphabet $\mathbb{X} = \{1, 2, ..., m\}$.*

Proof: It suffices to show that the ψ-minimal block code satisfies universality criterion (7.14). We will consider a sequence of k-block grammars G_l for string x_1^n indexed by index $l \in \{0, 1, ..., k-1\}$ such that:

- The secondary rules, regardless of index l, define all k-blocks in the order of ranking given by the distribution π:

$$G_l(r) \in \mathbb{X}^k \text{ for } -m^k \leq r < 0 \text{ and } \pi(G(r)) \geq \pi(G(r-1)). \tag{7.46}$$

- The primary rule of each grammar G_l defines string x_1^n using the identifiers for k-blocks shifted by l positions:

$$G_l(-m^k - 1) = (R_1, R_2, ..., R_l, r_1^l, r_2^l, ..., r_{p_l}^l, R_{-l'_l}, R_{-l'_l+1}, ..., R_{-1}) \tag{7.47}$$

where $0 < R_i < m$, $-m^k \leq r_i^l < 0$, and $0 \leq l, l'_l < k$.

We observe that none of these grammars can be better than the ψ-minimal block grammar for x_1^n. Hence, for any $l \in \{0, 1, ..., k-1\}$, we may bound

$$\left|B_\psi^\#(x_1^n)\right| \le |\psi^*(G_l)| \le C(k) + \sum_{i=1}^{p_l} \left|\psi(r_i^l)\right|, \tag{7.48}$$

where $C(k) := \left[m^k(k+1) + 2k + 2\right] |\psi(0)|$.

We have inequality $|\psi(-j)| \le \log j + 2\log(\log j + 1) + c_2$ for $j \ge 1$ by the hypothesis and inequality $\pi(G(-j)) \le 1/j$ by Lemma 7.21. Hence, we may further bound

$$\begin{aligned}\sum_{i=1}^{p_l} \left|\psi(r_i^l)\right| &\le \sum_{i=1}^{p_l} \left[\log(-r_i^l) + 2\log(\log m^k + 1) + c_2\right] \\ &\le \frac{n}{k}\left[2\log(\log m^k + 1) + c_2\right] - \sum_{i=1}^{p_l} \log \pi(G(r_i^l)).\end{aligned} \tag{7.49}$$

Denote $C(n,k) := C(k) + \frac{n}{k}\left[2\log(\log m^k + 1) + c_2\right]$. Then we may bound

$$\begin{aligned}\left|B_\psi^\#(x_1^n)\right| &\le C(n,k) - \min_{l\in\{0,1,...,k-1\}} \sum_{i=1}^{p_l} \log \pi(G(r_i^l)) \\ &\le C(n,k) - \frac{1}{k}\sum_{l=0}^{k-1}\sum_{i=1}^{p_l} \log \pi(G(r_i^l)) \\ &= C(n,k) - \frac{1}{k}\sum_{i=0}^{n-k} \log \pi(x_{i+1}^{i+k}).\end{aligned} \tag{7.50}$$

To conclude, we observe $\lim_{k\to\infty} \limsup_{n\to\infty} C(n,k)/n = 0$. □

Obviously the ψ-minimal admissible code is shorter than the ψ-minimal block code. In consequence, universality of the proper minimal admissible code follows by universality of the proper minimal block code.

***

To recapitulate this chapter, we have constructed the Lempel-Ziv code and the minimal grammar-based code. These codes are simple but their convergence to the entropy rate is not very fast. There are codes which compress particular sources better. Usually the better the compression, the harder the code to compute. The limit of efficient compression is set by the Kolmogorov complexity to be discussed in Chapter 11 but, as we will learn, it is not computable.

Further reading

The Lempel-Ziv code was invented by Abraham Lempel and Jacob Ziv in 1977 [130]. The Lempel-Ziv code is implemented in the ZIP program for file compression (or GZIP for the Linux operating system). It is worth stressing, however, that ZIP or GZIP do not make a fully universal code because of a limited buffer length. The minimal grammar-based code as presented in this chapter is based on earlier ideas of grammar-based codes and minimal block codes. Whereas the idea of minimal block codes comes from the work of David Neuhoff and Paul Shields [100], grammar-based coding was inspired by the doctoral thesis of Carl de Marcken [31] in computational linguistics. More theoretical insight in this domain was provided by John Kieffer and Enhui Yang [82], who proved universality of a wide class of grammar-based codes, and by Moses Charikar and others [20], who showed intractability of computing the minimal admissible grammar. The particular approach outlined in this chapter is inspired by my paper [33].

Thinking exercises

1. Find the Lempel-Ziv parsings for the sequences:
   (a) 010101010101010101...,
   (b) 1001000100001000001...,
   (c) 001001001001001001...,
   (d) 1011001100011000011....
2. Consider the constant sequence 00000000....
   (a) Produce the Lempel-Ziv parsing for this sequence.
   (b) Show that the number of bits per symbol for prefixes of that sequence tends to zero with the increasing length.
3. Produce a sequence for which the number of phrases in the Lempel-Ziv parsing grows as fast as possible.
4. Produce a sequence for which the number of phrases in the Lempel-Ziv parsing grows as slow as possible.
5. *Neuhoff-Shields code:* Let $\mathbb{X} = \{0, 1, 2, ..., m-1\}$. Let $B_k : (\mathbb{X}^k)^* \to \{0,1\}^*$ be the type codes (see exercises to Chapter 5) for alphabets

$\mathbb{X}^k$ where $k \geq 1$. For $x_i \in \mathbb{X}$, define prefix-free codes

$$E_{l,k}(x_1^n) := \mathrm{una}''(n)\,\mathrm{bin}_{n+1}(k)\,\mathrm{bin}_{n+1}(l)\,\mathrm{bin}_m(x_1)...\,\mathrm{bin}_m(x_l) \\ B_k(x_{l+1}^{l+k\lfloor n/k\rfloor})\,\mathrm{bin}_m(x_{l+k\lfloor n/k\rfloor+1})...\,\mathrm{bin}_m(x_n). \tag{7.51}$$

Let $L(x_1^n)$ and $K(x_1^n)$ be the minimal l and k such that the length of code word $E_{l,k}(x_1^n)$ is minimal. Show that the code defined as $E(x_1^n) := E_{L(x_1^n),K(x_1^n)}(x_1^n)$ satisfies universality criterion (7.14) [100].

6. Suppose that a process $(X_i)_{i\in\mathbb{N}}$ taking values in a finite set $\mathbb{X} = \{1, 2, ..., m\}$ satisfies

$$\lim_{n\to\infty} \frac{[-\log P(X_1^n)]}{n} = H \text{ a.s.} \tag{7.52}$$

for a certain random variable H. Show that

$$\lim_{n\to\infty} \frac{\mathbf{E}\,[-\log P(X_1^n)]}{n} = \mathbf{E}\,H. \tag{7.53}$$

Hint: Consider a code B with length

$$|B(x_1^n)| = 1 + \min\left\{n\left\lceil \log D\right\rceil, \left\lceil -\log P(X_1^n = x_1^n)\right\rceil\right\}. \tag{7.54}$$

Chapter 8

Mixtures

Universal distributions. Mixture and maximum distributions. Maximum likelihood and penalized maximum likelihood. Empirical entropy. Shtarkov sum bound. Universality of penalized maximum likelihood. Laplace estimator and prediction by partial matching distributions. Universality of prediction by partial matching.

In this chapter, we will exhibit two examples of probability distributions that are universal for higher order Markov processes. These two examples are called the penalized maximum likelihood (PML) and the prediction by partial matching (PPM). They generalize the ideas of two universal distributions for IID processes discussed in Chapter 5. In contrast, to the Lempel-Ziv code and the minimal grammar-based code from Chapter 7, the PML and the PPM are based on probabilistic considerations rather than string combinatorics. They also yield shorter code words for Markov processes.

Universal distributions

Universal distributions are defined analogously to universal codes.

Definition 8.1 (universal distribution) *An incomplete distribution $Q : \mathbb{X}^* \to [0, 1]$ is called universal for a given class of processes if the Shannon-Fano code with respect to Q is universal for the same class of processes.*

There are two general simple construction of distributions that are universal for higher order Markov processes given a sequence of distributions that are universal for k-th order Markov processes. These two constructions are the mixture and the maximum.

Theorem 8.2 (mixture and maximum distributions) *Suppose that incomplete distributions $Q_k : \mathbb{X}^* \to [0,1]$ are universal for stationary irreducible k-th order Markov processes each, where $k \in \{0,1,...\}$. Let coefficients $w_k > 0$ be such that $\sum_{k=0}^{\infty} w_k \leq 1$. Then incomplete distributions*

$$Q_{mix}(x_1^n) := \sum_{k=0}^{\infty} w_k Q_k(x_1^n), \tag{8.1}$$

$$Q_{max}(x_1^n) := \max_{k \geq 0} w_k Q_k(x_1^n) \tag{8.2}$$

are universal for stationary irreducible higher order Markov processes.

Proof: Consider an arbitrary $k \geq 0$. We have $Q_{\mathrm{mix}}(x_1^n), Q_{\max}(x_1^n) \geq w_k Q_k(x_1^n)$. Hence Q_{mix} and Q_{mix} are universal for k-th order Markov processes by the respective universality of Q_k. Since k was chosen arbitrarily then Q_{mix} and Q_{mix} are universal for higher order Markov processes. □

In the following, we will exhibit two different sequences of distributions which are universal for k-th order Markov processes over a finite alphabet $\mathbb{X} = \{1, 2, ..., m\}$. The first one is incomplete, whereas the second one is prequential. Both apply insights acquired in Chapter 5.

Maximum likelihood

Let us proceed to the first construction.

Definition 8.3 (maximum likelihood) *We define the maximum likelihood (ML) in the class of k-th order Markov processes over a finite alphabet $\mathbb{X}$ as*

$$\hat{\mathbb{P}}(x_1^n|k) := \max_{\tau} \prod_{i=k+1}^{n} \tau(x_i|x_{i-k}^{i-1}), \quad k < n, \tag{8.3}$$

where the maximum is taken across all k-th order transition matrices $\tau : \mathbb{X}^{k+1} \to [0,1]$. We assume that $\hat{\mathbb{P}}(x_1^n|k) := 1$ for $k \geq n$. The τ maximizing the expression on the right-hand side of (8.3) is called the maximum likelihood parameter for x_1^n and denoted $\hat{\tau}(\cdot|\cdot, x_1^n)$.

Let us evaluate $\hat{\mathbb{P}}(x_1^n|k)$ and $\hat{\tau}(\cdot|\cdot, x_1^n)$ explicitly. First, we adapt the definition of the empirical entropy from Chapter 5.

Definition 8.4 (empirical entropy) *Let us write the frequency of string a_1^l in string x_1^n as*

$$N(a_1^l|x_1^n) := \sum_{i=1}^{n-l+1} \mathbf{1}\left\{x_i^{i+l-1} = a_1^l\right\}. \tag{8.4}$$

Subsequently, let us denote the k-th order empirical entropy

$$\mathcal{H}(x_1^n|k) := \sum_{a_1^k} \frac{N(a_1^k|x_1^{n-1})}{n-k} \sum_{a_{k+1}} \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})} \left[-\log \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})}\right]. \tag{8.5}$$

Analogously, as in the case of IID processes, the empirical entropy is the minus logarithm of the maximum likelihood and the maximum likelihood parameter is the empirical distribution.

Theorem 8.5 *We have*

$$\hat{\tau}(a_{k+1}|a_1^k, x_1^n) = \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})}, \tag{8.6}$$

$$-\log \hat{\mathbb{P}}(x_1^n|k) = (n-k)\mathcal{H}(x_1^n|k). \tag{8.7}$$

Proof: Write succinctly $\hat{\pi}(a_1^k) := \frac{N(a_1^k|x_1^{n-1})}{n-k}$ and $\hat{\tau}(a_{k+1}|a_1^k) := \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})}$. By non-negativity of Kullback-Leibler divergence $D(\hat{\tau}\|\tau)$ we obtain

$$\begin{aligned}
\frac{1}{n-k}\prod_{i=k+1}^{n} \tau(x_{i-k}^{i-1}, x_i) &= \sum_{a_1^k} \hat{\pi}(a_1^k) \sum_{a_{k+1}} \hat{\tau}(a_{k+1}|a_1^k)\left[-\log \tau(a_{k+1}|a_1^k)\right] \\
&= \sum_{a_1^k} \hat{\pi}(a_1^k) \sum_{a_{k+1}} \hat{\tau}(a_{k+1}|a_1^k)\left[-\log \hat{\tau}(a_{k+1}|a_1^k)\right] \\
&\quad + \sum_{a_1^k} \hat{\pi}(a_1^k) \sum_{a_{k+1}} \hat{\tau}(a_{k+1}|a_1^k) \log \frac{\hat{\tau}(a_{k+1}|a_1^k)}{\tau(a_{k+1}|a_1^k)} \\
&\geq \sum_{a_1^k} \hat{\pi}(a_1^k) \sum_{a_{k+1}} \hat{\tau}(a_{k+1}|a_1^k)\left[-\log \hat{\tau}(a_{k+1}|a_1^k)\right] \\
&= \mathcal{H}(x_1^n|k),
\end{aligned} \tag{8.8}$$

where the inequality becomes equality for $\tau = \hat{\tau}$. □

Consider now the sum $\sum_{x_1^n \in \mathbb{X}^n} \hat{\mathbb{P}}(x_1^n|k)$. This sum is called the Shtarkov sum and is greater than 1. But we have the following upper bound.

Theorem 8.6 (Shtarkov sum bound) *For the class of k-th order Markov processes over alphabet $\{1, 2, ..., m\}$,*

$$\sum_{x_1^n \in \mathbb{X}^n} \hat{\mathbb{P}}(x_1^n|k) \leq Z(n|k) := \min\left\{m^n, m^k(n-k+1)^{m^{k+1}}\right\}. \tag{8.9}$$

Proof: Bound $\sum_{x_1^n \in \mathbb{X}^n} \hat{\mathbb{P}}(x_1^n|k) \leq m^n$ follows by $\hat{\mathbb{P}}(x_1^n|k) \leq 1$. Let us see how we can improve it. Let $\mathcal{P} := \{\hat{\tau}(\cdot|x_1^n) : x_1^n \in \mathbb{X}^n\}$ be the set of distinct maximum likelihood parameter values. We notice that for any $\tau \in \mathcal{P}$ there holds inequality

$$\sum_{x_1^n : \hat{\tau}(\cdot|x_1^n) = \tau} \hat{\mathbb{P}}(x_1^n|k) \leq m^k. \tag{8.10}$$

Hence

$$\sum_{x_1^n} \hat{\mathbb{P}}(x_1^n|k) = \sum_{\tau \in \mathcal{P}} \sum_{x_1^n : \hat{\tau}(\cdot|x_1^n) = \tau} \hat{\mathbb{P}}(x_1^n|k) \leq \sum_{\tau \in \mathcal{P}} m^k \leq m^k \, \# \, \mathcal{P}. \tag{8.11}$$

How many distinct maximum likelihood parameter values are there? There are m^{k+1} different strings a_1^{k+1}. Since $\hat{\tau}(a_{k+1}|a_1^k, x_1^n) = \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})}$ then for each a_1^{k+1} there are $\leq n-k+1$ distinct numerators. The denominators are simply the sums of numerators and are not independent parameters. Thus we may bound

$$\# \, \mathcal{P} \leq (n-k+1)^{m^{k+1}}. \tag{8.12}$$

□

Let us propose the following definition.

Definition 8.7 (penalized maximum likelihood) *The penalized maximum likelihood (PML) is*

$$\mathbb{P}(x_1^n|k) := \frac{\hat{\mathbb{P}}(x_1^n|k)}{Z(n|k)}. \tag{8.13}$$

In contrast to the maximum likelihood, the penalized maximum likelihood is an incomplete distribution.

Using the penalized maximum likelihood, we will exhibit a universal distribution which is good for stationary irreducible Markov processes of any order. The construction applies the already discussed idea of the maximum distribution.

Definition 8.8 (PML maximum) *The PML maximum is defined as*

$$\mathbb{P}(x_1^n) := \max_{k\geq 0} w_k \mathbb{P}(x_1^n|k), \quad w_k = \frac{1}{k+1} - \frac{1}{k+2}. \tag{8.14}$$

Maximum (8.14) can be effectively computed since $\mathbb{P}(x_1^n|k) = m^{-n}$ for $k \geq n$.

Theorem 8.9 *The PML maximum satisfies universality criterion (7.9).*

Proof: Consider a conditional probability distribution $\pi : \mathbb{X}^{k+1} \to [0,1]$, where $\pi(x_{k+1}|x_1^k) \geq 0$ and $\sum_{x_{k+1}} \pi(x_{k+1}|x_1^k) = 1$. Consider a string $x_1^n \in \{1,2,...,m\}^*$. We have

$$\begin{aligned} -\log \mathbb{P}(x_1^n|k) &= \log Z(n|k) - \log \hat{\mathbb{P}}(x_1^n|k) \\ &\leq \log Z(n|k) - \log \prod_{i=1}^{n} \pi(x_i|x_{i-k}^{i-1}). \end{aligned} \tag{8.15}$$

Criterion (7.9) is satisfied since

$$\log Z(n|k) = \min\left\{n \log m,\ k \log m + m^{k+1} \log(n-k+1)\right\}, \tag{8.16}$$

so $\lim\limits_{k\to\infty} \limsup\limits_{n\to\infty} \dfrac{1}{n} \left[-\log w_k + \log Z(n|k)\right] = 0.$ □

Laplace estimator

Now let us ask for a prequential universal distribution. Why should we care about prequential distributions? Intuitively, we may expect that when a distribution Q is both prequential and universal then conditional probabilities

$$Q(x_{n+1}|x_1^n) := \frac{Q(x_1^{n+1})}{Q(x_1^n)} \tag{8.17}$$

should converge in some sense to conditional probabilities $P(x_{n+1}|x_1^n)$. Consequently, distribution Q could be used not only for universal coding but also for some sort of universal prediction. However, universal prediction turns out a much more advanced topic than universal coding. In this chapter we will make just the first step. Namely, we will exhibit a prequential universal distribution. It will be demonstrated in Chapter 10 that this prequential universal distribution yields in fact a universal predictor.

Our point of departure is the Laplace estimator over alphabet $\{1, 2, ..., m\}$. The Laplace estimator reads

$$R_0(x_1) := \frac{1}{m}, \tag{8.18}$$

$$R_0(x_{n+1}|x_1^n) := \frac{N(x_{n+1}|x_1^n) + 1}{n + m}, \tag{8.19}$$

$$R_0(x_1^n) := R_0(x_1) \prod_{i=1}^{n-1} R_0(x_{i+1}|x_1^i). \tag{8.20}$$

Distribution R_0 is not only prequential but also universal for IID processes, as we know from Chapter 5. It estimates the probability of symbol x_{n+1} as the relative frequency of symbol x_{n+1} in the previously seen sample x_1^n. Another important observation is that the relative frequency is smoothed by adding 1 in the numerator and m in the denominator to make sure that R_0 is always defined and positive.

We may suppose that we may generalize this construction for k-th order Markov processes by estimating the probability of x_{n+1} as the relative frequency of x_{n+1} given context x_{n+1-k}^n in the previously seen sample x_1^n. These relative frequencies would be smoothed analogously, so we would obtain the following prequential distributions.

Definition 8.10 (PPM distributions) *For $k \geq 0$, we define the prediction by partial matching (PPM) distributions*

$$R_k(x_1^{k+1}) := \frac{1}{m^{k+1}}, \tag{8.21}$$

$$R_k(x_{n+1}|x_1^n) := \frac{N(x_{n+1-k}^{n+1}|x_1^n) + 1}{N(x_{n+1-k}^n|x_1^{n-1}) + m}, \quad n \geq k + 1, \tag{8.22}$$

$$R_k(x_1^n) := R_k(x_1^{k+1}) \prod_{i=k+1}^{n-1} R_k(x_{i+1}|x_1^i). \tag{8.23}$$

Using the PPM distributions, we will exhibit a universal distribution which is good for stationary irreducible Markov processes of any order. The construction applies the already discussed idea of the mixture distribution.

Definition 8.11 (PPM mixture) *We define the PPM mixture as*

$$R(x_1^n) := \sum_{k=0}^{\infty} w_k R_k(x_1^n), \quad w_k = \frac{1}{k+1} - \frac{1}{k+2}. \tag{8.24}$$

Sum (8.24) can be effectively computed since $R_k(x_1^n) = m^{-n}$ for $k \geq n$.

Theorem 8.12 *The PPM mixture satisfies universality criterion (7.9).*

Proof: Consider a conditional probability distribution $\pi : \mathbb{X}^{k+1} \to [0,1]$, where $\pi(x_{k+1}|x_1^k) \geq 0$ and $\sum_{x_{k+1}} \pi(x_{k+1}|x_1^k) = 1$. Consider a string $x_1^n \in \{1,2,...,m\}^*$. The first step is to observe that

$$\begin{aligned} R_k(x_1^n) &= \frac{1}{m^k} \prod_{a_1^k} \frac{\prod_{a_{k+1}} 1 \cdot 2 ... N(a_1^{k+1}|x_1^n)}{m \cdot (m+1) ... (N(a_1^k|x_1^{n-1}) + m - 1)} \\ &= \frac{1}{m^k} \prod_{a_1^k} \frac{(m-1)! \prod_{a_{k+1}} N(a_1^{k+1}|x_1^n)!}{(N(a_1^k|x_1^{n-1}) + m - 1)!}. \end{aligned} \tag{8.25}$$

Let us abbreviate $N_w := N(w|x_1^{n-1})$ and $N_{wj} := N(wj|x_1^n)$ for $j = 1,2,...,m$. Using techniques from the proof of Theorem 5.21, we can bound

$$\begin{aligned} -\log R_k(x_1^n) &= k \log m + \sum_{w \in \mathbb{X}^k} \log \binom{N_w + m - 1}{N_{w1}, ..., N_{wm}, m-1} \\ &\leq k \log m + \sum_{w \in \mathbb{X}^k} (N_w + m) H\left(\frac{N_{w1}}{N_w + m}, ..., \frac{N_{wm}}{N_w + m}, \frac{m-1}{N_w + m}, \frac{1}{N_w + m} \right) \\ &\leq k \log m + \sum_{w \in \mathbb{X}^k} \left[m(\log n + 6) + N_w H\left(\frac{N_{w1}}{N_w}, ..., \frac{N_{wm}}{N_w} \right) \right] \\ &= k \log m + \sum_{w \in \mathbb{X}^k} m(\log n + 6) + (n-k) \mathcal{H}(x_1^n|k) \\ &\leq k \log m + m^{k+1}(\log n + 6) - \log \prod_{i=1}^n \pi(x_i|x_{i-k}^{i-1}), \end{aligned} \tag{8.26}$$

where $\lim_{k \to \infty} \limsup_{n \to \infty} \frac{1}{n} \left[-\log w_k + k \log m + m^{k+1}(\log n + 6) \right] = 0$. Thus criterion (7.9) is satisfied. □

***

Recapitulating this chapter, we have constructed universal distributions for higher order Markov processes such as the penalized maximum likelihood and the prediction by partial matching. These two examples are based on the ideas of universal codes for IID processes, in contrast to the Lempel-Ziv code and the minimal grammar-based code from Chapter 7. These four codes together constitute the elementary theory of universal coding. In the following chapters, we will apply more advanced results to obtain stronger insights. We will present no new examples of codes but rather we will try to understand how powerful constructions we have developed so far.

Further reading

The PPM distributions and their mixture, under the name of R-measure, were proposed by Boris Ryabko [107, 112, 108], drawing on the Laplace estimator for IID processes by Pierre-Simon de Laplace [88] and the Krichevsky-Trofimov estimator by Raphail Krichevsky and Victor Trofimov [86]. Independently, a similar idea was proposed by John Cleary and Ian Witten in [23], from whom we borrow the more distinctive name PPM.

Thinking exercises

1. Let $(X_i)_{i\in\mathbb{N}}$ be a higher order stationary Markov process. Show that
$$\lim_{n\to\infty} \mathcal{H}(X_1^n|k) = H(X_i|X_{i-k}^{i-1}) \text{ a.s.} \tag{8.27}$$

2. Argue that each penalized maximum likelihood $\mathbb{P}(\cdot|k)$ and each PPM distribution $R_k(\cdot|k)$ is universal for stationary irreducible k-th order Markov processes over alphabet $\{1, 2, ..., m\}$.

3. *Maximal repetition length:* The maximal length of a repetition in string x_1^n is defined as
$$L(x_1^n) := \max\left\{l : x_{i+1}^{i+l} = x_{j+1}^{j+l} \text{ for some } 0 \le i < j \le n-l\right\}. \tag{8.28}$$
Show that the PPM mixture can be efficiently calculated as
$$R(x_1^n) = \sum_{k=0}^{L(x_1^n)} \left(\frac{1}{k+1} - \frac{1}{k+2}\right) R_k(x_1^n) + \frac{m^{-n}}{L(x_1^n)+2}. \tag{8.29}$$
Evaluate a similar truncation for the PML maximum.

4. Consider alphabet $\mathbb{X} = \mathbb{N}$ and let $\pi : \mathbb{N} \to [0, 1]$ be some probability distribution such that $\pi(l) > 0$ for all $l \in \mathbb{N}$. For $k \ge 0$, we define the prediction by partial matching distributions
$$R_k(x_1^{k+1}) := \prod_{i=1}^{k+1} \pi(x_i), \tag{8.30}$$
$$R_k(x_{n+1}|x_1^n) := \frac{N(x_{n+1-k}^{n+1}|x_1^n) + \pi(x_{n+1})}{N(x_{n+1-k}^{n}|x_1^{n-1}) + 1}, \quad n \ge k+1, \tag{8.31}$$
$$R_k(x_1^n) := R_k(x_1^{k+1}) \prod_{i=k+1}^{n-1} R_k(x_{i+1}|x_1^i). \tag{8.32}$$

We define the PPM mixture as

$$R(x_1^n) := \sum_{k=0}^{\infty} w_k R_k(x_1^n), \quad w_k = \frac{1}{k+1} - \frac{1}{k+2}. \tag{8.33}$$

Describe a class of processes with respect to which R is universal.

5. Let $\mathbb{X} = \{1, 2, ..., m\}$ and $c = m + 1$. Let $R_k(\cdot|\cdot)$ be the conditional PPM distributions for orders $k \geq 0$ and alphabet $\mathbb{X} \cup \{c\}$. Consider a sequence of strings $w_1^p := (w_1, w_2, ..., w_p)$ where $w_j \in \mathbb{X}^*$. Let $x_1^n = w_1 c w_2 c ... c w_p c$ and $w_i = x_{q_i+1}^{q_i+|w_i|}$. Consider function

$$Q(w_1^p) := \prod_{i=1}^{p} \prod_{j=0}^{|w_i|} R_j(x_{q_i+j+1}|x_1^{q_i+j}). \tag{8.34}$$

Let $Q(w_{p+1}|w_1^p) := Q(w_1^{p+1})/Q(w_1^p)$. Show that

$$-\log Q(w_{p+1}|w_1^p) \leq \log \frac{q_p + m}{N(w_{p+1}|w_1^p) + 1} + \frac{m\,|w_{p+1}|}{N(w_{p+1}|w_1^p) + 1}. \tag{8.35}$$

Is Q a prequential distribution? Is it universal for some processes over alphabet $\mathbb{X}^*$?

6. Let $R_k(\cdot|\cdot)$ be the conditional PPM distributions for orders $k \geq 0$. Define conditional entropies of these distributions as

$$h_k(x_1^n) := -\sum_{x_{n+1}} R_k(x_{n+1}|x_1^n) \log R_k(x_{n+1}|x_1^n). \tag{8.36}$$

Define random order $K(x_1^n)$ as the minimal k for which entropy $h_k(x_1^n)$ achieves the minimal value. Is $G(n) := n - K(x_1^n)$ a growing function of n? Consider function

$$Q(x_1^n) := \prod_{i=1}^{n} R_{K(x_1^{i-1})}(x_i|x_1^{i-1}). \tag{8.37}$$

Is Q a prequential distribution? Is it universal for some processes?

Chapter 9

Crossings

Crossings and convergence of sequences. Conditional expectation. Martingales. Prequential functions. Generalized Kraft equality. Stopping time. Doob optional stopping theorem. Doob upcrossing inequality. Doob convergence theorem. Lévy law. Azuma inequality and its corollary. Two-sided stationary processes. Ivanov downcrossing inequality. Birkhoff ergodic theorem. Ergodic processes. Ergodicity criterion. Ergodic decomposition. Breiman ergodic theorem.

This chapter is a preparation for Chapter 10, where we will study universal coding and universal prediction for general stationary ergodic processes. For this aim, we need to present basic properties of stationary processes and martingales, which are two important classes of well-behaved stochastic processes. Our study is focused on proving various laws of randomness, which state the almost sure convergence of some sequences of random variables.

Convergence criterion

Both for stationary processes and for martingales, we will apply an important proof technique which is based on upcrossings. The definition of upcrossings and downcrossings for finite and infinite sequences is as follows.

Definition 9.1 (upcrossing and downcrossing) *Consider a sequence of real numbers $(s_n)_{n\in\mathbb{A}}$ indexed by a subset of natural numbers $\mathbb{A} \subset \mathbb{N}$. We say that $(s_n)_{n\in\mathbb{A}}$ upcrosses an interval $[a, b]$ at least k times, written succinctly as $(s_n)_{n\in\mathbb{A}} \overset{k}{\rightsquigarrow} [a, b]$, if there are indices*

$$1 \leq i_1 < j_1 < i_2 < j_2 < ... < i_k < j_k \tag{9.1}$$

such that $i_l, j_l \in \mathbb{A}$, $s_{i_l} < a$, *and* $s_{j_l} > b$. *Analogously, we say that* $(s_n)_{n\in\mathbb{A}}$ *downcrosses an interval* $[a,b]$ *at least* k *times, written succinctly as* $(s_n)_{n\in\mathbb{A}} \overset{k}{\rightsquigarrow} [b,a]$ *if there are indices (9.1) such that* $i_l, j_l \in \mathbb{A}$, $s_{i_l} > b$, *and* $s_{j_l} < a$. *We also say that* $(s_n)_{n\in\mathbb{A}}$ *crosses an interval* $[a,b]$ *infinitely many times, written as* $(s_n)_{n\in\mathbb{A}} \overset{\infty}{\rightsquigarrow} [a,b]$, *if* $(s_n)_{n\in\mathbb{A}} \overset{k}{\rightsquigarrow} [a,b]$ *for all* $k \in \mathbb{N}$.

We recall that $\lim_{n\to\infty} s_n = s$ if and only if $\liminf_{n\to\infty} s_n = s$ and $\limsup_{n\to\infty} s_n = s$. Hence, it is quite easy to see that a given infinite sequence converges to a limit in extended real numbers, including $-\infty$ and $+\infty$, when it does not oscillate around any interval.

Theorem 9.2 (convergence criterion) *Let* $(s_n)_{n\in\mathbb{N}}$ *be a sequence of extended real numbers. Limit* $\lim_{n\to\infty} s_n$ *does not exist if and only if there exist* $a, b \in \mathbb{Q}$ *such that* $(s_n)_{n\in\mathbb{N}}$ *crosses interval* $[a,b]$ *infinitely many times.*

Proof: Limit $\lim_{n\to\infty} s_n$ does not exist if and only if there exists an interval $[a,b]$ where $a, b \in \mathbb{Q}$ and $a < b$ such that $\liminf_{n\to\infty} s_n < a$ and $\liminf_{n\to\infty} s_n > b$. Consequently, we notice that $\liminf_{n\to\infty} s_n < a$ and $\liminf_{n\to\infty} s_n > b$ hold if and only if $(s_n)_{n\in\mathbb{A}} \overset{\infty}{\rightsquigarrow} [a,b]$. □

Subsequently, we will present the probabilistic version of the above statement. Notice that we have used intervals with rational end-points $a, b \in \mathbb{Q}$ so as to be able to apply the countable additivity of the probability measure.

Theorem 9.3 (convergence criterion) *Let* $(S_n)_{n\in\mathbb{N}}$ *be a sequence of real random variables. Limit* $\lim_{n\to\infty} S_n$ *exists almost surely if and only if for any* $a, b \in \mathbb{Q}$ *where* $a < b$, *we have*

$$\inf_{k\in\mathbb{N}} P\left((S_n)_{n\in\mathbb{N}} \overset{k}{\rightsquigarrow} [a,b]\right) = 0. \tag{9.2}$$

Proof: By the previous result, we notice the equality of events

$$\left(\lim_{n\to\infty} S_n \text{ does not exist}\right) = \bigcup_{a,b\in\mathbb{Q}:\, a<b} \left((S_n)_{n\in\mathbb{N}} \overset{\infty}{\rightsquigarrow} [a,b]\right). \tag{9.3}$$

Hence by the countable additivity of the probability measure, limit $\lim_{n\to\infty} S_n$ exists almost surely if and only if for any $a, b \in \mathbb{Q}$ where $a < b$, we have

$$P\left((S_n)_{n\in\mathbb{N}} \overset{\infty}{\rightsquigarrow} [a,b]\right) = 0. \tag{9.4}$$

But, we also have the equality

$$\left((S_n)_{n\in\mathbb{N}} \overset{\infty}{\rightsquigarrow} [a,b]\right) = \bigcap_{k\in\mathbb{N}} \left((S_n)_{n\in\mathbb{N}} \overset{k}{\rightsquigarrow} [a,b]\right). \tag{9.5}$$

Thus, by the countable additivity of the probability measure, we obtain

$$P\left((S_n)_{n\in\mathbb{N}} \overset{\infty}{\rightsquigarrow} [a,b]\right) = \inf_{k\in\mathbb{N}} P\left((S_n)_{n\in\mathbb{N}} \overset{k}{\rightsquigarrow} [a,b]\right). \tag{9.6}$$

Hence the claim follows. □

What is surprising, the convergence criterion based on upcrossings or downcrossings can be effectively used as a proof technique in probability calculus. In the following, we will see two important applications thereof. The first one concerns martingales, whereas the second one concerns stationary processes.

Martingales

Martingales are an important concept in the analysis of universal coding and universal prediction. In order to define martingales, we need to define conditional expectations first. For a discrete random variable Z, we recall that random variable $P(X = x|Z)$ assumes value $P(X = x|Z = z)$ on event $(Z = z)$. Formally writing, we have

$$P(X = x|Z)(\omega) = P(X = x|Z = z) \iff Z(\omega) = z. \tag{9.7}$$

Definition 9.4 (conditional expectation) *Let Z be a discrete random variable. For a discrete real random variable $X : \Omega \to [0,\infty]$ the conditional expectation is defined as the random variable*

$$\mathbf{E}(X|Z) := \sum_{x:P(X=x|Z)>0} xP(X = x|Z). \tag{9.8}$$

For a general real random variable $Y : \Omega \to [0,\infty]$ the conditional expectation is defined as

$$\mathbf{E}(Y|Z) := \sup_{X\le Y} \mathbf{E}(X|Z), \tag{9.9}$$

where the supremum is taken over all discrete real random variables X such that $X(\omega) \le Y(\omega)$. For real random variables $Y_1, Y_2 : \Omega \to [0,\infty]$, the conditional expectation of random variable $Y_1 - Y_2$ is defined as

$$\mathbf{E}(Y_1 - Y_2|Z) := \mathbf{E}(Y_1|Z) - \mathbf{E}(Y_2|Z) \tag{9.10}$$

if $\mathbf{E}(Y_1|Z) < \infty$ or $\mathbf{E}(Y_2|Z) < \infty$.

It is illuminating to imagine the graph of random variable $\omega \mapsto \mathbf{E}(X|Z)(\omega)$ as interpolating between the graph of random variable $\omega \mapsto X(\omega)$ and the graph of constant value $\omega \mapsto \mathbf{E}\,X$. The more distinct values the random variable Z assumes, the closer is the graph of $\omega \mapsto \mathbf{E}(X|Z)(\omega)$ to the graph of $\omega \mapsto X(\omega)$. The fewer distinct values Z assumes, the closer is the graph of $\omega \mapsto \mathbf{E}(X|Z)(\omega)$ to the graph of $\omega \mapsto \mathbf{E}\,X$.

Example 9.5 *Suppose that the event space is $\Omega = [0,1]$ and the probability measure is the Lebesgue measure $P([a,b]) = b - a$. Then for*

$$Z(\omega) = \begin{cases} 0, & \omega < 1/2, \\ 1, & \omega \geq 1/2, \end{cases} \tag{9.11}$$

and $X(\omega) = \omega$, we have $\mathbf{E}\,X = 1/2$ and

$$\mathbf{E}(X|Z)(\omega) = \begin{cases} 1/4, & \omega < 1/2, \\ 3/4, & \omega \geq 1/2. \end{cases} \tag{9.12}$$

Example 9.6 *Let $Y = \mathbf{1}\{X = x\}$. Then $\mathbf{E}(Y|Z) = P(X = x|Z)$.*

Now we can define martingales.

Definition 9.7 (martingale) *A real number process $(Y_n)_{n\in\mathbb{N}}$ is called a martingale with respect to process $(X_n)_{n\in\mathbb{N}}$ over a finite alphabet if $Y_n = g(X_1^n)$ for a certain function g and $\mathbf{E}(Y_{n+1}|X_1^n) = Y_n$ for $n \geq 1$.*

To remember, each random variable $Y_n = \mathbf{E}(Y_{n+1}|X_1^n)$ is a a partially averaged version of the succeeding random variable Y_{n+1}. Hence martingale $(Y_n)_{n\in\mathbb{N}}$ is a sequence of random variables whose graphs become less and less averaged—like an approximation of a fractal.

An important result in theory of martingales is that this "fractal approximation" process can converge under some relatively general conditions. The approach discussed here is limited to martingales with respect to a process over a finite alphabet. First, we will show a characterization of such martingales in terms of prequential functions.

Definition 9.8 (prequential function) *For a finite alphabet $\mathbb{X}$, a function $Q : \mathbb{X}^* \to \mathbb{R}$ is called prequential if*

$$\sum_{x_{n+1}\in\mathbb{X}} Q(x_1^{n+1}) = Q(x_1^n). \tag{9.13}$$

Theorem 9.9 *Let $(Y_n)_{n\in\mathbb{N}}$ be a martingale with respect to process $(X_n)_{n\in\mathbb{N}}$ over a finite alphabet $\mathbb{X}$. There is a prequential function $Q : \mathbb{X}^* \to \mathbb{R}$ such that $P(X_1^n = x_1^n) = 0 \implies Q(x_1^n) = 0$ and*

$$Y_n = \frac{Q(X_1^n)}{P(X_1^n)}. \tag{9.14}$$

Proof: Let $Y_n = g(X_1^n)$. Define $Q(x_1^n) := P(X_1^n = x_1^n)g(x_1^n)$. We have

$$\begin{aligned}\sum_{x_{n+1}\in\mathbb{X}} Q(x_1^{n+1}) &= \sum_{x_{n+1}\in\mathbb{X}} g(x_1^{n+1})P(X_1^{n+1} = x_1^{n+1}) \\ &= P(X_1^n = x_1^n) \sum_{x_{n+1}\in\mathbb{X}} g(x_1^{n+1})P(X_{n+1} = x_{n+1}|X_1^n = x_1^n) \\ &= P(X_1^n = x_1^n)\,\mathbf{E}(Y_{n+1}|X_1^n) = P(X_1^n = x_1^n)Y_n \\ &= P(X_1^n = x_1^n)g(X_1^n) = Q(x_1^n) \end{aligned} \tag{9.15}$$

□

Now we will generalize the Kraft inequality to prequential functions. First, we have to define bounded, complete, and prefix-free sets.

Definition 9.10 (bounded complete prefix-free set) *A set of strings $A \subset \mathbb{X}^*$ over a finite alphabet $\mathbb{X}$ is called prefix-free if for any strings $x, y \in A$ and any string $u \in \mathbb{X}^*$ condition $y = xu$ implies $y = x$. A set of strings $A \subset \mathbb{X}^*$ is called complete if for any infinite sequence $y \in \mathbb{X}^{\mathbb{N}}$ there exists a string $x \in A$ and an infinite sequence $u \in \mathbb{X}^{\mathbb{N}}$ such that $y = xu$. A set of strings $A \subset \mathbb{X}^*$ is called bounded if there is natural number $n \in \mathbb{N}$ such that $|x| \leq n$ for any string $x \in A$.*

Theorem 9.11 (generalized Kraft equality) *Let A be a bounded complete prefix-free set of strings and Q be a prequential function over a finite alphabet $\mathbb{X}$. We have*

$$\sum_{x\in A} Q(x) = \sum_{x\in\mathbb{X}} Q(x). \tag{9.16}$$

Proof: The minimal number n such that $|x| \leq n$ for any string $x \in A$ is called the depth of A. We will proceed by induction on depth of sets A. Obviously (9.16) is satisfied for set A of depth 1, which is exactly $A = \mathbb{X}$. Suppose that (9.16) is satisfied for all sets A of depth n. Now let A' be a set of depth $n+1$. Let $A_1 := \{x : |x| \leq n, x \in A'\}$ and $A_2 := \{x : |x| = n, xu \in A', u \in \mathbb{X}\}$. We

have $A_1 \cap A_2 = \emptyset$ since A' is prefix-free and $A_1 \cup (A_2 \times \mathbb{X}) = A'$ since A' is complete. Moreover $A_1 \cup A_2$ is prefix-free and complete of depth n. Thus

$$\begin{aligned}\sum_{x\in A'} Q(x) &= \sum_{x\in A_1\cup(A_2\times\mathbb{X})} Q(x) = \sum_{x\in A_1} Q(x) + \sum_{x\in A_2}\sum_{u\in\mathbb{X}} Q(ux) \\ &= \sum_{x\in A_1} Q(x) + \sum_{x\in A_2} Q(x) = \sum_{x\in A_1\cup A_2} Q(x) = \sum_{x\in\mathbb{X}} Q(x). \qquad (9.17)\end{aligned}$$

Hence the claim is true in general. □

The subsequent developments apply a gambling metaphor. Imagine that a martingale is a series of prices for which a gambler buys or sells goods at a stock exchange. We will define stopping times which represent the gambler's decision to sell or buy goods at particular times that depend only on her knowledge of the past.

Definition 9.12 (stopping time) *A random variable $T : \Omega \to \mathbb{N}$ is called a stopping time with respect to a process $(X_n)_{n\in\mathbb{N}}$ over a finite alphabet $\mathbb{X}$ if there exists a complete prefix-free set $A \subset \mathbb{X}^*$ such that*

$$T = t \iff X_1^t \in A. \qquad (9.18)$$

Next, we will show that the gambler cannot gain or lose on average by selling initially bought goods at any stopping time if the martingale is bounded.

Theorem 9.13 (Doob optional stopping) *Let $(Y_n)_{n\in\mathbb{N}}$ be a martingale with $|Y_n| \leq M$ where $\mathbf{E}\, M < \infty$ and $T : \Omega \to \mathbb{N}$ be a stopping time with respect to a process $(X_n)_{n\in\mathbb{N}}$ over a finite alphabet. Then*

$$\mathbf{E}\, Y_T = \mathbf{E}\, Y_1. \qquad (9.19)$$

Proof: First, assume that T is bounded almost surely, namely, $T \leq t$ for some $t \in \mathbb{N}$. Let $Y_n = g(X_1^n)$ and $T = t \iff X_1^t \in A$. Define $Q(x_1^n) := P(X_1^n = x_1^n)g(x_1^n)$. Then

$$\mathbf{E}\, Y_T = \sum_{x\in A} Q(x) = \sum_{x\in\mathbb{X}} Q(x) = \mathbf{E}\, Y_1. \qquad (9.20)$$

Now let us consider an unbounded T. We have $Y_T = \lim_{t\to\infty} Y_{\min\{T,t\}}$ almost surely so $\mathbf{E}\, Y_T = \mathbf{E}\, Y_1$ follows from the analogous claim for bounded T by the Lebesgue dominated convergence. □

Applying the gambling metaphor further, the gambler will be buying and selling the same good at multiple stopping times in order to gain nothing. However, in this way, we will obtain the following bound for the number of upcrossings in bounded martingales.

Theorem 9.14 (Doob upcrossing inequality) *Let $(Y_n)_{n\in\mathbb{N}}$ be a martingale with respect to process $(X_n)_{n\in\mathbb{N}}$ over a finite alphabet. If $|Y_n| \leq M$ where $\mathbf{E}\, M < \infty$ then for any $a, b \in \mathbb{R}$ where $a < b$ we have*

$$P\left((Y_n)_{n\in\mathbb{N}} \overset{k}{\rightsquigarrow} [a,b]\right) \leq \frac{2\,\mathbf{E}\, M}{k(b-a)}. \tag{9.21}$$

Proof: Define random times

$$1 \leq I_1 < J_1 < I_2 < J_2 < ... \tag{9.22}$$

as the minimal numbers such that $I_l, J_l \in \mathbb{N}$, $Y_{I_l} < a$, and $Y_{J_l} > b$. It can be easily shown that I_l and J_l are stopping times and so are $\min\{I_l, n\}$ and $\min\{J_l, n\}$. Let K_m be the exact number of times that interval $[a,b]$ is upcrossed by sequence $(Y_n)_{1\leq n\leq m}$, namely,

$$K_m \geq k \iff (Y_n)_{n\leq m} \overset{k}{\rightsquigarrow} [a,b]. \tag{9.23}$$

Since $Y_{J_l} - Y_{I_l} > b - a$ and $Y_j - Y_i \geq -2M$ then we can bound

$$\sum_{l=1}^{m} \left(Y_{\min\{J_l,m\}} - Y_{\min\{I_l,m\}}\right) \geq \sum_{l=1}^{K_m} (Y_{J_l} - Y_{I_l}) - 2M > (b-a)K_m - 2M. \tag{9.24}$$

Hence, by the Doob optional stopping theorem, we infer

$$(b-a)\,\mathbf{E}\, K_m - 2\,\mathbf{E}\, M < \sum_{l=1}^{m} \left(\mathbf{E}\, Y_{\min\{J_l,m\}} - \mathbf{E}\, Y_{\min\{I_l,m\}}\right) = 0. \tag{9.25}$$

Let $K := \sup_{m\in\mathbb{N}} K_m$ be the total number of upcrossings. By the monotone convergence theorem, we derive $\mathbf{E}\, K = \sup_{m\in\mathbb{N}} \mathbf{E}\, K_m$. Consequently, by the Markov inequality, we obtain

$$P\left((Y_n)_{n\in\mathbb{N}} \overset{k}{\rightsquigarrow} [a,b]\right) = P(K \geq k) = \frac{\mathbf{E}\, K}{k} = \sup_{m\in\mathbb{N}} \frac{\mathbf{E}\, K_m}{k} \leq \frac{2\,\mathbf{E}\, M}{k(b-a)}. \tag{9.26}$$

□

Theorem 9.15 (Doob martingale convergence) *Let $(Y_n)_{n\in\mathbb{N}}$ be a martingale with respect to process $(X_n)_{n\in\mathbb{N}}$ over a finite alphabet. If $|Y_n| \leq M$ where $\mathbf{E}\, M < \infty$ then there exists limit $Y := \lim_{n\to\infty} Y_n$ a.s.*

Proof: By the Doob upcrossing inequality, for any $a, b \in \mathbb{Q}$ we have

$$\inf_{k\in\mathbb{N}} P\left((Y_n)_{n\in\mathbb{N}} \overset{k}{\rightsquigarrow} [a,b]\right) = 0. \tag{9.27}$$

Hence limit $Y := \lim_{n\to\infty} Y_n$ exists almost surely by Theorem 9.3. □

We note in passing that in Theorems 9.13–9.15, the assumption of a finite alphabet of process $(X_n)_{n\in\mathbb{N}}$ can be relaxed to an arbitrary set of values but this complicates the theory and we will omit this topic.

Now let us discuss some examples of martingales. Four important examples are as follows. First, the fractal approximation process converges to the fractal if the fractal exists.

Example 9.16 (complete martingale) *Let Y be a real random variable such that $\mathbf{E}\,|Y| < \infty$. Random variables $Y_n = \mathbf{E}(Y|X_1^n)$ exist and are a martingale with respect to process $(X_n)_{n\in\mathbb{N}}$. Moreover, the Doob martingale convergence theorem states that limit $\mathbf{E}(Y|X_1^\infty) := \lim_{n\to\infty} \mathbf{E}(Y|X_1^n)$ exists almost surely if $|Y_n| \le M$ where $\mathbf{E}\,M < \infty$.*

Second, an important sort of a fractal approximation are conditional probabilities given longer and longer blocks of random variables.

Example 9.17 *Random variables $Y_n := P(X_0 = x_0|X_{-n}^{-1})$ are a martingale with respect to process $(X_{-n})_{n\in\mathbb{N}}$. It is so since $Y_n = \mathbf{E}(\mathbf{1}\{X_0 = x_0\}\,|X_{-n}^{-1})$.*

As a result of the Doob martingale convergence, the conditional probabilities converge as well.

Theorem 9.18 (Lévy law) *Let $(X_i)_{i\in\mathbb{Z}}$ be a stochastic process over a finite alphabet $\mathbb{X}$. There exist limits*

$$P(X_i|X_{-\infty}^{i-1}) := \lim_{k\to\infty} P(X_i|X_{i-k}^{i-1}) \text{ a.s.} \tag{9.28}$$

Proof: Random variables $(P(X_i|X_{i-k}^{i-1}))_{k\in\mathbb{N}}$ form a bounded martingale. Thus the convergence holds by the Doob martingale convergence. □

Third, partial sums of an IID process $(X_n)_{n\in\mathbb{N}}$ itself with the removed expectation are obviously a martingale with respect to $(X_n)_{n\in\mathbb{N}}$.

Example 9.19 *Let $(X_n)_{n\in\mathbb{N}}$ be a discrete real IID process. Then random variables*

$$Y_n := \sum_{i=1}^{n} [X_i - \mathbf{E}\,X_i] \tag{9.29}$$

are a martingale with respect to process $(X_n)_{n\in\mathbb{N}}$.

Fourth, the previous idea of probabilistically independent increments $Y_n - Y_{n-1}$ can be generalized to arbitrary martingale increments.

Example 9.20 *Let $Z_n = g(X_1^n)$. Then random variables*

$$Y_n = \sum_{i=1}^{n} \left[Z_i - \mathbf{E}\left(Z_i \middle| X_1^{i-1} \right) \right] \tag{9.30}$$

are a martingale with respect to process $(X_n)_{n\in\mathbb{N}}$. Every martingale $(Y_n)_{n\in\mathbb{N}}$ has this representation since it suffices to put $Z_n = Y_n - Y_{n-1}$ but then $\mathbf{E}\left(Z_n \middle| X_1^{n-1} \right) = 0$. In particular, if series

$$Y = \lim_{n\to\infty} Y_n = \sum_{i=1}^{\infty} \left[Z_i - \mathbf{E}\left(Z_i \middle| X_1^{i-1} \right) \right] \tag{9.31}$$

converges almost surely then we have $Y_n = \mathbf{E}(Y|X_1^n)$.

An important result for considerations of Chapter 10 is the Azuma-Hoeffding inequality for martingales with bounded increments. To state it, we need first the Chernoff bound and the Hoeffding lemma.

Theorem 9.21 (Chernoff bound) *Let Z be a real random variable and $\lambda > 0$ be a constant. Then*

$$P(Z \geq a) \leq e^{-\lambda a}\, \mathbf{E}\, e^{\lambda Z} \tag{9.32}$$

Proof: The claim follows by $P(Z \geq a) = P\left(e^{\lambda Z} \geq e^{\lambda a} \right)$ and by the Markov inequality. □

Theorem 9.22 (Hoeffding lemma) *Let Z be a real random variable such that $a \leq Z \leq b$. Then*

$$\mathbf{E}\, e^{\lambda (Z - \mathbf{E}\, Z)} \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8} \right). \tag{9.33}$$

Proof: Without loss of generality, assume $\mathbf{E}\, Z = 0$. By convexity of function $x \mapsto e^{\lambda x}$ and the Jensen inequality, we obtain

$$\mathbf{E}\, e^{\lambda Z} \leq \frac{b - \mathbf{E}\, Z}{b-a} e^{\lambda a} + \frac{\mathbf{E}\, Z - a}{b-a} e^{\lambda b} = \frac{b e^{\lambda a} - a e^{\lambda b}}{b-a} \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8} \right), \tag{9.34}$$

where the last transition follows by the Taylor series approximation. □

Theorem 9.23 (Azuma-Hoeffding inequality) *Let $(X_n)_{n\in\mathbb{N}}$ be a process over a finite alphabet. Let $(Y_n)_{n\in\mathbb{N}}$ be a martingale process with respect to $(X_n)_{n\in\mathbb{N}}$ with increments bounded by $|Y_n - Y_{n-1}| \leq c_n$. Then*

$$P(|Y_n - Y_0| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^n c_i^2}\right). \tag{9.35}$$

Proof: By the Chernoff bound and the Hoeffding lemma, for $\lambda > 0$, we may write

$$\begin{aligned} P(|Y_n - Y_0| \geq \epsilon) &\leq e^{-\lambda\epsilon}\,\mathbf{E}\exp\left(\lambda\sum_{i=1}^n (Y_i - Y_{i-1})\right) \\ &= e^{-\lambda\epsilon}\,\mathbf{E}\exp\left(\lambda\sum_{i=1}^{n-1} (Y_i - Y_{i-1})\right)\mathbf{E}\left(e^{\lambda(Y_n - Y_{n-1})}\big|X_1^{n-1}\right) \\ &\leq e^{-\lambda\epsilon}\,\mathbf{E}\exp\left(\lambda\sum_{i=1}^{n-1} (Y_i - Y_{i-1})\right)\exp\left(\frac{\lambda^2 c_n^2}{2}\right) \\ &\leq e^{-\lambda\epsilon}\exp\left(\frac{\lambda^2}{2}\sum_{i=1}^n c_i^2\right). \end{aligned} \tag{9.36}$$

Minimizing the last expression with respect to λ yields the desired inequality. □

In Chapter 10 we will use the following corollary of the above statement: If the martingale increments are bounded by $c_n = Cn^{1/2-\epsilon}$ then we have convergence $\lim_{n\to\infty} Y_n/n = 0$ almost surely even if the martingale limit $\lim_{n\to\infty} Y_n$ does not exist. Notice also that in this case the average of the upper bounds diverges, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n c_i = \infty$, so the following corollary of the Azuma-Hoeffding inequality is a non-trivial probabilistic effect.

Theorem 9.24 *Let $(X_n)_{n\in\mathbb{N}}$ be a process over a finite alphabet. Let real random variables $(Z_n)_{n\in\mathbb{N}}$ satisfy $Z_n = g(X_1^n)$ and $|Z_n| \leq Cn^{1/2-\epsilon}$ for some $C < \infty$ and $\epsilon > 0$. Then*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \left[Z_i - \mathbf{E}\left(Z_i\big|X_1^{i-1}\right)\right] = 0 \text{ a.s.} \tag{9.37}$$

Proof: Define the martingale process

$$Y_n := \sum_{i=1}^n \left[Z_i - \mathbf{E}\left(Z_i\big|X_1^{i-1}\right)\right]. \tag{9.38}$$

By the Azuma-Hoeffding inequality for any $\delta > 0$ we obtain

$$P(|Y_n| \geq n\delta) \leq 2\exp\left(-\frac{\delta^2 n^2}{8C^2\sum_{i=1}^{n} i^{1-2\epsilon}}\right)$$
$$\leq 2\exp\left(-\frac{\delta^2 n^2}{8C^2 n^{2-2\epsilon}}\right) \leq 2\exp\left(-\frac{\delta^2 n^{2\epsilon}}{8C^2}\right). \quad (9.39)$$

Hence $\sum_{n=1}^{\infty} P(|Y_n| \geq n\delta) < \infty$ so the claim follows by the Borel-Cantelli lemma. □

Ergodic processes

Properties such as the ergodic theorem and the asymptotic equipartition can be generalized to a subclass of stationary processes that resemble irreducible Markov processes. This well-behaved subclass of stationary processes is called ergodic processes. The standard theory of ergodic processes is quite technically advanced. In this section, we will try to convey essential intuitions and present some less standard results that give a general insight.

First of all, we need to work with processes that extend in both directions such as process $(X_i)_{i\in\mathbb{Z}}$ indexed by integers rather than process $(X_i)_{i\in\mathbb{N}}$ indexed by natural numbers. We adapt the definition from Chapter 6:

Definition 9.25 (stationary process) *A stochastic process $(X_i)_{i\in\mathbb{Z}}$ over a countable alphabet $\mathbb{X}$ is called stationary if for all $t \in \mathbb{Z}$, all $k \in \mathbb{N}$ and all strings $x_1^k \in \mathbb{X}^*$, we have*

$$P(X_{t+1}^{t+k} = x_1^k) = P(X_1^k = x_1^k). \quad (9.40)$$

Let us seek for a generalization of the ergodic theorem for arbitrary stationary processes. For this goal, we will investigate first the number of downcrossings for shifted rates of a given finite non-decreasing real function $f(x)$ on an interval $(0, L)$. Formally, we require $f(y) \geq f(x)$ for $y \geq x$. We will count oscillations of rates $\left((f(x) - f(s)) / (x - s)\right)_{s<x<L}$ around interval $[a, b]$ where $0 < a < b$ with an arbitrary starting point $0 < s < L$. Extending our previous notation, we say that there are at least k downcrossings to the right of s, written as

$$\left(\frac{f(x) - f(s)}{x - s}\right)_{s<x<L} \overset{k}{\rightsquigarrow} [b, a], \quad (9.41)$$

if and only if there exist numbers

$$s < x_1 < y_1 < x_2 < y_2 < \ldots < x_k < y_k < L \quad (9.42)$$

such that $f(x_l) - f(s) > b(x_l - s)$ and $f(y_l) - f(s) < a(y_l - s)$.

There is a clever geometric bound for the relative measure of points that are followed by at least k downcrossings. We will investigate this relative measure for a restricted class of non-decreasing piecewise constant functions that we call Ivanov functions.

Definition 9.26 (Ivanov function) *A function $f : (0, L) \to \mathbb{R}$ is called an Ivanov function if there exists a sequence $(L_k)_{k\in\mathbb{N}}$ such that $L_1 := L$, $L_k > L_{k+1}$, $\lim_{k\to\infty} L_k = 0$, and $f(x) = f(L_{k+1}) \geq f(L_{k+2})$ for $x \in [L_{k+1}, L_k)$.*

Theorem 9.27 (continuous Ivanov downcrossing inequality) *Consider an Ivanov function $f : (0, L) \to \mathbb{R}$. For $a, b \in \mathbb{R}$ where $0 < a < b$, we have*

$$\frac{1}{L}\int_0^L \mathbf{1}\left\{\left(\frac{f(x) - f(s)}{x - s}\right)_{s<x<L} \overset{k}{\rightsquigarrow} [b, a]\right\} ds \leq \left(\frac{a}{b}\right)^k. \tag{9.43}$$

Proof: (Omitted. See Theorems 2.1 and 3.1 of [24].) □

Instead of presenting the known proof of Theorem 9.27, which is quite long and complex, we will exhibit a simple function that achieves this bound.

Example 9.28 (maximal Ivanov function) *Let $0 < a < b$ and $L_k := \left(\frac{a}{b}\right)^{k-1} L$. We consider function $f(x) := aL_k = bL_{k+1}$ for $x \in [L_{k+1}, L_k)$ and $k \geq 1$. Simply speaking, the graph of function $y = f(x)$ is self-similar and sandwiched between two straight lines $y = bx$ and $y = ax$. For this function, there are at least k downcrossings to the right of point s if and only if $s < L_k$. Thus the left hand side of (9.43) equals $L_{k+1}/L = \left(\frac{a}{b}\right)^k$.*

We hope that the above example will inspire the reader to invent a shorter proof of Theorem 9.27 that could fit the style of this textbook.

Ivanov's bound can be specialized to non-decreasing sequences. Namely, the following statement is also true.

Theorem 9.29 (discrete Ivanov downcrossing inequality) *Consider a sequence $c_0 \leq c_1 \leq c_2 \leq ... \leq c_L$. For $a, b \in \mathbb{R}$ where $0 < a < b$, we have*

$$\frac{1}{L}\sum_{s=0}^{L-1} \mathbf{1}\left\{\left(\frac{c_n - c_s}{n - s}\right)_{s<n\leq L} \overset{k}{\rightsquigarrow} [b, a]\right\} \leq 2k\left(\frac{2k+1}{2k-1}\right)^k \left(\frac{a}{b}\right)^k. \tag{9.44}$$

Proof: It suffices to consider an Ivanov function $f(x) := c_{\lfloor x \rfloor}$ and perform a few simple bounds. First, we choose an $m > 1$. For $x > s + m$, we obtain

$$a(\lfloor x \rfloor - \lfloor s \rfloor) \leq a(x + 1 - s) \leq \frac{a(m+1)}{m}(x - s), \tag{9.45}$$

$$b(\lfloor x \rfloor - \lfloor s \rfloor) \geq a(x - 1 - s) \geq \frac{b(m-1)}{m}(x - s), \tag{9.46}$$

Further, we consider integers

$$s < i_1 < j_1 < i_2 < j_2 < ... < i_k < j_k \le s + m \tag{9.47}$$

where $b(i_l - s) < c_{i_l} - c_s$ and $c_{j_l} - c_s < a(j_l - s)$. Since terms c_n are non-decreasing, we obtain

$$b(i_l - s) < c_{i_l} - c_s \le c_{j_l} - c_s < a(j_l - s) \le a(i_{l+1} - s). \tag{9.48}$$

Hence $m \ge j_k - s > (b/a)(i_k - s) > (b/a)^k(i_1 - s) \ge (b/a)^k$ so

$$k \le k(m) := (\log m)/(\log b - \log a). \tag{9.49}$$

We denote the left hand side of (9.44) as p. In view of the above bounds, we derive

$$\begin{aligned}
p &\le \frac{1}{L}\sum_{s=0}^{L-1} \mathbf{1}\left\{ \left(\frac{c_n - c_s}{n-s}\right)_{s+m<n\le L} \overset{k-k(m)}{\rightsquigarrow} [b,a] \right\} \\
&= \frac{1}{L}\int_0^L \mathbf{1}\left\{ \left(\frac{c_{\lfloor x\rfloor} - c_{\lfloor s\rfloor}}{\lfloor x\rfloor - \lfloor s\rfloor}\right)_{s+m<x<L} \overset{k-k(m)}{\rightsquigarrow} [b,a] \right\} ds \\
&\le \frac{1}{L}\int_0^L \mathbf{1}\left\{ \left(\frac{c_{\lfloor x\rfloor} - c_{\lfloor s\rfloor}}{x-s}\right)_{s+m<x<L} \overset{k-k(m)}{\rightsquigarrow} \left[\frac{b(m-1)}{m}, \frac{a(m+1)}{m}\right] \right\} ds \\
&\le \frac{1}{L}\int_0^L \mathbf{1}\left\{ \left(\frac{c_{\lfloor x\rfloor} - c_{\lfloor s\rfloor}}{x-s}\right)_{s<x<L} \overset{k-k(m)}{\rightsquigarrow} \left[\frac{b(m-1)}{m}, \frac{a(m+1)}{m}\right] \right\} ds \\
&\le \left(\frac{a(m+1)}{b(m-1)}\right)^{k-k(m)}.
\end{aligned} \tag{9.50}$$

Finally, observing that

$$\left(\frac{a(m+1)}{b(m-1)}\right)^{-k(m)} \le \left(\frac{a}{b}\right)^{-k(m)} = m \tag{9.51}$$

yields

$$p \le \min_{m\ge 1} m \left(\frac{m+1}{m-1}\right)^k \left(\frac{a}{b}\right)^k \le 2k \left(\frac{2k+1}{2k-1}\right)^k \left(\frac{a}{b}\right)^k, \tag{9.52}$$

where $m = k + \sqrt{k^2 - 1}$ is the exact minimizer. □

In consequence, a probabilistic bound for stationary processes follows.

Theorem 9.30 (probabilistic Ivanov downcrossing inequality) *Let $(Y_i)_{i\in\mathbb{Z}}$ be a real stationary process with $Y_i \geq 0$. For $a, b \in \mathbb{R}$ where $0 < a < b$, we have*

$$P\left(\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)_{n\in\mathbb{N}} \overset{k}{\rightsquigarrow} [b,a]\right) \leq 2k\left(\frac{2k+1}{2k-1}\right)^k \left(\frac{a}{b}\right)^k. \tag{9.53}$$

Proof: We denote the left hand side of (9.53) as p. Since $(Y_i)_{i\in\mathbb{Z}}$ is a non-negative process, we may introduce a non-decreasing random sequence $C_n := \sum_{i=1}^{n} Y_i$. Then applying stationarity and Theorem 9.29, we may write

$$\begin{aligned}
p_M &:= P\left(\left(\frac{C_n}{n}\right)_{0<n\leq M} \overset{k}{\rightsquigarrow} [b,a]\right) \\
&= \frac{1}{L}\sum_{s=0}^{L-1} \mathbf{E}\,\mathbf{1}\left\{\left(\frac{C_n - C_s}{n-s}\right)_{s<n\leq M+s} \overset{k}{\rightsquigarrow} [b,a]\right\} \\
&= \frac{1}{L}\,\mathbf{E}\sum_{s=0}^{L-1} \mathbf{1}\left\{\left(\frac{C_n - C_s}{n-s}\right)_{s<n\leq M+s} \overset{k}{\rightsquigarrow} [b,a]\right\} \\
&\leq \frac{1}{L}\,\mathbf{E}\sum_{s=0}^{L+M-1} \mathbf{1}\left\{\left(\frac{C_n - C_s}{n-s}\right)_{s<n\leq L+M} \overset{k}{\rightsquigarrow} [b,a]\right\} \\
&\leq \left(\frac{L+M}{L}\right) 2k\left(\frac{2k+1}{2k-1}\right)^k \left(\frac{a}{b}\right)^k \xrightarrow[L\to\infty]{} 2k\left(\frac{2k+1}{2k-1}\right)^k \left(\frac{a}{b}\right)^k.
\end{aligned} \tag{9.54}$$

To conclude, we notice that $p = \lim_{M\to\infty} p_M$. □

As a corollary we obtain this famous fact, which states that the time average of a real stationary process exists and is equal to its expectation if it is constant.

Theorem 9.31 (Birkhoff ergodic theorem) *Let $(Y_i)_{i\in\mathbb{Z}}$ be a real stationary process with $Y_i \geq 0$. There exists limit*

$$\bar{Y} := \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i \ \text{a.s.} \tag{9.55}$$

Moreover, we have $\mathbf{E}\,\bar{Y} = \mathbf{E}\,Y_i$.

Proof: By the Ivanov downcrossing inequality, for $a, b \in \mathbb{Q}$ where $0 < a < b$ we have

$$\lim_{k\to\infty} P\left((\bar{Y}_n)_{n\in\mathbb{N}} \overset{k}{\rightsquigarrow} [b,a]\right) \leq \lim_{k\to\infty} 2k\left(\frac{2k+1}{2k-1}\right)^k \left(\frac{a}{b}\right)^k = 0, \tag{9.56}$$

where $\bar{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i$. Hence limit $\bar{Y} = \lim_{n\to\infty} \bar{Y}_n$ exists almost surely by the crossing convergence criterion.

Let $0 < M < \infty$. If $0 \le Y_i \le M$ then by the Lebesgue dominated convergence,

$$\mathbf{E}\,\bar{Y} = \mathbf{E} \lim_{n\to\infty} \bar{Y}_n = \lim_{n\to\infty} \mathbf{E}\,\bar{Y}_n = \mathbf{E}\,Y_i. \tag{9.57}$$

Suppose next that $Y_i \ge 0$ cannot be upper bounded. By the Fatou lemma,

$$\mathbf{E}\,\bar{Y} = \mathbf{E} \liminf_{n\to\infty} \bar{Y}_n \le \liminf_{n\to\infty} \mathbf{E}\,\bar{Y}_n = \mathbf{E}\,Y_i. \tag{9.58}$$

Now let $Y_i^M := \min\{Y_i, M\}$. Write $\bar{Y}_n^M := \frac{1}{n}\sum_{i=1}^n Y_i^M$. We know that limit $\bar{Y}^M := \lim_{n\to\infty} \bar{Y}_n^M$ exists almost surely but it need not be constant. But we can easily see that $\bar{Y} \ge \bar{Y}^M$ almost surely. Hence

$$\mathbf{E}\,\bar{Y} \ge \mathbf{E}\,\bar{Y}^M = \mathbf{E} \lim_{n\to\infty} \bar{Y}_n^M = \lim_{n\to\infty} \mathbf{E}\,\bar{Y}_n^M = \mathbf{E}\,Y_i^M. \tag{9.59}$$

It suffices to note that the sandwich bound $\mathbf{E}\,Y_i^M \le \mathbf{E}\,\bar{Y} \le \mathbf{E}\,Y_i$ gets arbitrarily tight since $\lim_{M\to\infty} \mathbf{E}\,Y_i^M = \mathbf{E}\,Y_i$ by the monotone convergence. □

As we have stated above, the time average of a real stationary process is equal to its expectation if it is constant. Subsequently, we will adopt an unusual definition of an ergodic process which directly appeals to this property.

Definition 9.32 (ergodic process) *A stationary process $(X_i)_{i\in\mathbb{Z}}$ over a countable alphabet $\mathbb{X}$ is called ergodic if for all $k \in \mathbb{N}$ and all strings $x_1^k \in \mathbb{X}^*$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}\left\{X_{i+1}^{i+k} = x_1^k\right\} = P(X_1^k = x_1^k) \ \text{a.s.} \tag{9.60}$$

The standard definition of an ergodic process involves the shift-invariant σ-field, which we do not want to explain here. Our definition is equivalent to that one. Let us state some theorem which allows to effectively check which stationary processes are ergodic.

Theorem 9.33 (ergodicity criterion) *A stationary process $(X_i)_{i\in\mathbb{Z}}$ over a countable alphabet $\mathbb{X}$ is ergodic if and only if for all $k \in \mathbb{N}$ and all strings $x_1^k \in \mathbb{X}^*$ such that $P(X_1^k = x_1^k) > 0$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P(X_{i+1}^{i+k} = x_1^k | X_1^k = x_1^k) = P(X_1^k = x_1^k). \tag{9.61}$$

Proof: Let us write $A_i := (X_{i+1}^{i+k} = x_1^k)$ and $\mathbf{1}\{A_i\} := \mathbf{1}\left\{X_{i+1}^{i+k} = x_1^k\right\}$. We have $\mathbf{E}\,\mathbf{1}\{A_i\} = P(A_i)$. Assume first that condition (9.60) holds. Let us rewrite condition (9.60) multiplying both sides by $\mathbf{1}\{A_0\}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}\{A_i\}\,\mathbf{1}\{A_0\} = P(A_0)\mathbf{1}\{A_0\} \text{ a.s.} \tag{9.62}$$

Applying the Lebesgue dominated convergence yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P(A_i \cap A_0) = P(A_0)P(A_0). \tag{9.63}$$

Dividing both sides by $P(A_0)$, we obtain (9.61).

Suppose now that condition (9.60) does not hold. Denote the random variable corresponding to the varying limit

$$\bar{Y} := \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}\{A_i\} - P(A_0). \tag{9.64}$$

We have $\mathbf{E}\,\bar{Y} = 0$ and $\mathbf{E}\,\bar{Y}^2 > 0$. We may derive by stationarity of process $(X_i)_{i\in\mathbb{Z}}$ and shift-invariance of $\bar{Y}$ that $P(A_i|\bar{Y} \geq y) = P(A_0|\bar{Y} \geq y)$. Analogously to the first part of the proof, for $y > 0$, we obtain

$$P(A_0|\bar{Y} \geq y) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P(A_i|\bar{Y} \geq y) \geq P(A_0) + y, \tag{9.65}$$

$$P(A_0|\bar{Y} \leq -y) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P(A_i|\bar{Y} \leq -y) \leq P(A_0) - y. \tag{9.66}$$

Since $\mathbf{E}\,\bar{Y} = \int_0^1 P(\bar{Y} \geq y)dy - \int_0^1 P(\bar{Y} \leq -y)dy$, we have

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P(A_i|A_0) - P(A_0) = \mathbf{E}(\bar{Y}|A_0) \\
&= \int_0^1 P(\bar{Y} \geq y|A_0)dy - \int_0^1 P(\bar{Y} \leq -y|A_0)dy \\
&= \frac{\int_0^1 P(\bar{Y} \geq y, A_0)dy - \int_0^1 P(\bar{Y} \leq -y, A_0)dy}{P(A_0)} \\
&= \frac{\int_0^1 P(\bar{Y} \geq y)P(A_0|\bar{Y} \geq y)dy - \int_0^1 P(\bar{Y} \leq -y)P(A_0|\bar{Y} \leq -y)dy}{P(A_0)}
\end{aligned}$$

$$\geq \frac{\int_0^1 (y + P(A_0))P(\bar{Y} \geq y)dy + \int_0^1 (y - P(A_0))P(\bar{Y} \leq -y)dy}{P(A_0)}$$
$$= \frac{\int_0^1 yP(\bar{Y} \geq y)dy + \int_0^1 yP(\bar{Y} \leq -y)dy}{P(A_0)} = \frac{\mathbf{E}\,\bar{Y}^2}{2P(A_0)} > 0. \tag{9.67}$$

Hence the ergodicity criterion (9.61) does not hold either. □

In view of condition (9.61), an ergodic process forgets "on average" its initial state. Let us discuss a few examples of stationary processes, specifying some conditions for their ergodicity:

- *IID processes:* We can verify easily that they are all ergodic.

- *Periodic processes:* For a periodic sequence $(y_i)_{i\in\mathbb{N}}$ where $y_{i+p} = y_i$, a periodic process $(X_i)_{i\in\mathbb{Z}}$ is the stationary process such that

$$P(X_1^k = x_1^k) = \frac{1}{p}\sum_{i=0}^{p-1} \mathbf{1}\left\{y_{i+1}^{i+k} = x_1^k\right\}. \tag{9.68}$$

  Obviously, a periodic process is ergodic.

- *Markov processes of order $k \in \mathbb{N}$:* As we discussed in Chapter 6, these processes are ergodic if they are irreducible. It can be shown that this is both a necessary and a sufficient condition.

Just like any stationary Markov process can be decomposed into irreducible Markov processes, any stationary process can be decomposed into ergodic processes. However, the important difference is that a stationary Markov process can decompose into countably many irreducible components, at most, whereas a general stationary process can decompose into uncountably many ergodic components. This makes the general theory of stationary processes technically complicated so we will not pursue this topic further.

The Birkhoff ergodic theorem can be stated also as follows.

Theorem 9.34 (Birkhoff ergodic theorem) *Let $(X_i)_{i\in\mathbb{Z}}$ be a stationary ergodic process over a finite alphabet $\mathbb{X}$ and let $f_k : \mathbb{X}^k \to [0, \infty]$ be non-negative real functions. Then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f_k(X_{i+1}^{i+k}) = \mathbf{E}\, f_k(X_1^k) \ \ a.s. \tag{9.69}$$

Proof: By the definition of process $(X_i)_{i\in\mathbb{Z}}$, we obtain

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} f(X_{i+1}^{i+k}) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\sum_{y\in\mathbb{X}^k} f(y)\mathbf{1}\left\{X_{i+1}^{i+k}=y\right\}\\ &= \sum_{y\in\mathbb{X}^k} f(y)\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{1}\left\{X_{i+1}^{i+k}=y\right\}\\ &= \sum_{y\in\mathbb{X}^k} f(y)P(X_1^k=y) = \mathbf{E}\, f(X_1^k), \qquad (9.70)\end{aligned}$$

where we can exchange the limit and the summation since the summation consists of finitely many terms. □

Moreover, several times in Chapter 10, we will need a generalization of the ergodic theorem that allows to take the time averages over an appropriately converging sequence of functions. The following strengthening of the dominated convergence is a prerequisite.

Theorem 9.35 (Lebesgue dominated convergence) *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of real random variables which satisfy $\mathbf{E}\sup_{n\in\mathbb{N}}|Y_n|<\infty$. If there exists limit $\lim_{n\to\infty} Y_n$ then*

$$\inf_{t\in\mathbb{N}}\mathbf{E}\sup_{m>t}\left|Y_m-\lim_{n\to\infty}Y_n\right| = 0. \qquad (9.71)$$

Proof: Let $X_t := \sup_{m>t}|Y_m-\lim_{n\to\infty}Y_n|$ and $Z=\sup_{n\in\mathbb{N}}|Y_n|$. We have $0\le X_t\le 2Z$ and $\lim_{t\to\infty}X_t=0$. Hence by the Fatou lemma, we obtain

$$\begin{aligned}\mathbf{E}\,2Z &= \mathbf{E}\liminf_{t\to\infty}(2Z-X_t)\\ &\le \liminf_{t\to\infty}\mathbf{E}(2Z-X_t) = \mathbf{E}\,2Z-\limsup_{t\to\infty}\mathbf{E}\,X_t \qquad (9.72)\end{aligned}$$

Thus the claim follows by $0\ge\limsup_{t\to\infty}\mathbf{E}\,X_t\ge\inf_{t\in\mathbb{N}}\mathbf{E}\,X_t$. □

The generalization of the ergodic theorem is as follows.

Theorem 9.36 (Breiman ergodic theorem) *Let $(X_i)_{i\in\mathbb{Z}}$ be a stationary ergodic process over a finite alphabet $\mathbb{X}$. Let $f_k:\mathbb{X}^k\to[0,\infty]$ be non-negative real functions. Suppose that*

$$\mathbf{E}\sup_{k\in\mathbb{N}} f_k(X_{i-k}^{i-1})<\infty \qquad (9.73)$$

and there exist limits

$$f(X_{-\infty}^{i-1}) := \lim_{k\to\infty} f_k(X_{i-k}^{i-1}) \ \ a.s. \tag{9.74}$$

Then we have the generalized ergodic theorem

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f_k(X_1^k) = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(X_{-\infty}^k) = \mathbf{E}\, f(X_{-\infty}^i) \ \ a.s. \tag{9.75}$$

Proof: By the monotone convergence, we have

$$\begin{aligned}
&\mathbf{E} \sup_{s>t} \inf_{m\in\mathbb{N}} \sup_{n>m} \left| \frac{1}{n}\sum_{k=1}^{n} f_s(X_{k-s+1}^k) - \frac{1}{n}\sum_{k=1}^{n} f(X_{-\infty}^k) \right| \\
&\le \mathbf{E} \sup_{s>t} \sup_{n\in\mathbb{N}} \left| \frac{1}{n}\sum_{k=1}^{n} f_s(X_{k-s+1}^k) - \frac{1}{n}\sum_{k=1}^{n} f(X_{-\infty}^k) \right| \\
&= \sup_{s>t} \sup_{n\in\mathbb{N}} \mathbf{E} \left| \frac{1}{n}\sum_{k=1}^{n} f_s(X_{k-s+1}^k) - \frac{1}{n}\sum_{k=1}^{n} f(X_{-\infty}^k) \right| \\
&\le \sup_{s>t} \sup_{n\in\mathbb{N}} \mathbf{E} \frac{1}{n}\sum_{k=1}^{n} \left| f_s(X_{k-s+1}^k) - f(X_{-\infty}^k) \right| \\
&= \sup_{s>t} \mathbf{E} \left| f_s(X_{i-s+1}^i) - f(X_{-\infty}^i) \right|.
\end{aligned} \tag{9.76}$$

Hence, by Theorem 9.35, we obtain

$$\begin{aligned}
&\mathbf{E} \limsup_{s\to\infty} \limsup_{n\to\infty} \left| \frac{1}{n}\sum_{k=1}^{n} f_s(X_{k-s+1}^k) - \frac{1}{n}\sum_{k=1}^{n} f(X_{-\infty}^k) \right| \\
&\le \inf_{t\in\mathbb{N}} \mathbf{E} \sup_{s>t} \sup_{n\in\mathbb{N}} \left| \frac{1}{n}\sum_{k=1}^{n} f_s(X_{k-s+1}^k) - \frac{1}{n}\sum_{k=1}^{n} f(X_{-\infty}^k) \right| \\
&\le \inf_{t\in\mathbb{N}} \mathbf{E} \sup_{s>t} \left| f_s(X_{i-s+1}^i) - f(X_{-\infty}^i) \right| = 0.
\end{aligned} \tag{9.77}$$

As a result, by Theorem 9.34, we derive

$$\begin{aligned}
&\mathbf{E} \limsup_{n\to\infty} \left| \frac{1}{n}\sum_{k=1}^{n} f(X_{-\infty}^k) - \mathbf{E}\, f(X_{-\infty}^i) \right| \\
&\le \mathbf{E} \limsup_{s\to\infty} \limsup_{n\to\infty} \left| \frac{1}{n}\sum_{k=1}^{n} f(X_{-\infty}^k) - \frac{1}{n}\sum_{k=1}^{n} f_s(X_{k-s+1}^k) \right|
\end{aligned}$$

$$+ \mathbf{E} \limsup_{s\to\infty} \limsup_{n\to\infty} \left| \frac{1}{n} \sum_{k=1}^{n} f_s(X_{k-s+1}^k) - \mathbf{E}\, f_s(X_{i-s+1}^i) \right|$$
$$+ \inf_{t\in\mathbb{N}} \mathbf{E} \sup_{s>t} \left| f_s(X_{i-s+1}^i) - f(X_{-\infty}^i) \right| = 0. \tag{9.78}$$

Consequently, we obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_{-\infty}^k) = \mathbf{E}\, f(X_{-\infty}^i) \text{ a.s.} \tag{9.79}$$

since $\mathbf{E}\, Y = 0$ for $Y \geq 0$ implies $Y = 0$ almost surely.

To derive the second equality, we consider function

$$g_{p,t}(X_{k-p}^{k-1}) := \max_{t\leq r,s\leq p} \left| f_r(X_{k-r}^{k-1}) - f_s(X_{k-s}^{k-1}) \right|. \tag{9.80}$$

This function satisfies

$$\mathbf{E} \sup_{p\in\mathbb{N}} g_{p,t}(X_{k-p}^{k-1}) < \infty \tag{9.81}$$

and there exist limits

$$g_t(X_{-\infty}^{k-1}) := \sup_{r,s\geq t} \left| f_r(X_{k-r}^{k-1}) - f_s(X_{k-s}^{k-1}) \right| = \lim_{p\to\infty} g_{p,t}(X_{k-p}^{k-1}) \text{ a.s.} \tag{9.82}$$

In view of this, we derive

$$\begin{aligned}
&\mathbf{E} \limsup_{n\to\infty} \left| \frac{1}{n} \sum_{k=1}^{n} f_k(X_1^k) - \frac{1}{n} \sum_{k=1}^{n} f(X_{-\infty}^k) \right| \\
&\leq \mathbf{E} \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \left| f_k(X_1^k) - f(X_{-\infty}^k) \right| \\
&\leq \inf_{t\in\mathbb{N}} \mathbf{E} \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g_t(X_{-\infty}^k) = \inf_{t\in\mathbb{N}} \mathbf{E}\, g_t(X_{-\infty}^i) \\
&\leq 2 \inf_{t\in\mathbb{N}} \mathbf{E} \sup_{r\geq t} \left| f_r(X_{-r}^{-1}) - f(X_{-\infty}^{-1}) \right| = 0
\end{aligned} \tag{9.83}$$

by (9.79) applied to $f = g_t$ and by Theorem 9.35. Consequently, we obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f_k(X_1^k) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_{-\infty}^k) \text{ a.s.} \tag{9.84}$$

since $\mathbf{E}\, Y = 0$ for $Y \geq 0$ implies $Y = 0$ almost surely. □

***

Recapitulating this chapter, we have studied processes that generalize IID processes, namely, martingales and ergodic processes. We have developed powerful results for their probabilistic convergence. In Chapter 10 we will apply these results to study universal coding and prediction for stationary ergodic processes, which generalize irreducible higher order Markov processes.

Further reading

The theory of martingales and the technique of upcrossings were developed by Joseph Doob. His book [42] was influential. The Lévy law is due to Paul Lévy. The Azuma-Hoeffding inequality was originally stated by Wassily Hoeffding for bounded IID processes and later generalized to martingales by Kazuoki Azuma [3]. The first proof of the Birkhoff ergodic theorem was due to George Birkhoff [9]. It was considerably shortened by Adriano Garsia [54]. Leo Breiman proved the Breiman ergodic theorem [13]. Its proof can be also found in the paper by Paul Algoet [1]. Vladimir Ivanov found out a proof of the Birkhoff ergodic theorem based on downcrossings [72, 73]. Its proof was shortened by Pierre Collet and Jean-Pierre Eckmann, but their version is still quite complex [24]. This result is useful in the study of algorithmic randomness to be mentioned in Chapter 12. An assortment of various theorems pertaining to stationary and ergodic processes can be found in the books by Robert Gray [58] and by Łukasz Dębowski [36].

Thinking exercises

1. Show that

$$\mathbf{E}(\mathbf{E}(Y|Z)|g(Z)) = \mathbf{E}(Y|g(Z)), \tag{9.85}$$

$$\mathbf{E}\,\mathbf{E}(Y|Z) = \mathbf{E}\,Y, \tag{9.86}$$

$$\mathbf{E}(g(Z)|Z) = g(Z), \tag{9.87}$$

$$\mathbf{E}(\mathbf{E}(Y|g(Z))|Z) = \mathbf{E}(Y|g(Z)). \tag{9.88}$$

2. *Incomplete martingale:* Let $(X_n)_{n\in\mathbb{N}}$ be an IID process such that $P(X_n = 0) = P(X_n = 2) = 1/2$. Show that process $(Y_n)_{n\in\mathbb{N}}$ such that $Y_n := \prod_{i=1}^n X_i$ is a martingale with respect to $(X_n)_{n\in\mathbb{N}}$. Show that there is no random variable Y such that $Y_n = \mathbf{E}(Y|X_1^n)$.

3. Let $(X_i)_{i\in\mathbb{Z}}$ be a process over a finite alphabet. Let $Z_n = g(X_1^n)$ be random variables that satisfy

$$\mathbf{E}(Z_{n+1}|X_1^n) = Y_n := \frac{1}{n}\sum_{i=1}^{n} Z_i. \tag{9.89}$$

Show that process $(Y_i)_{i\in\mathbb{Z}}$ is a martingale with respect to $(X_i)_{i\in\mathbb{Z}}$.

4. Show that a stationary Markov process $(X_i)_{i\in\mathbb{Z}}$ is ergodic with $P(X_i = x) > 0$ for all $x \in \mathbb{X}$ if and only if it is irreducible [36].

5. *Mixing processes:* To recall, a stationary process $(X_i)_{i\in\mathbb{Z}}$ over a countable alphabet $\mathbb{X}$ is called mixing if for all $k \in \mathbb{N}$ and all strings $x_1^k, y_1^k \in \mathbb{X}^*$ such that $P(X_1^k = y_1^k) > 0$, we have (6.32). Show that every mixing process is ergodic but not every ergodic process is mixing.

6. *Recurrence times:* By an analogy to Markov processes, for a process $(X_i)_{i\in\mathbb{Z}}$ and a set B such that $P(X_0 \in B) > 0$, let us define passage times

$$T_0 := 0, \tag{9.90}$$
$$T_n := \inf\{n \in \mathbb{N} : n > T_{n-1}, X_n \in B\}, \tag{9.91}$$

and successive recurrence times

$$R_n := \begin{cases} T_{n+1} - T_n & \text{if } T_n < \infty, \\ \infty & \text{otherwise.} \end{cases} \tag{9.92}$$

Show the following properties of successive recurrence times for a stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ [77, 21]:

(a) the Poincaré recurrence theorem:

$$P(R_k < \infty \text{ for all } k \in \mathbb{N}|X_0 \in B) = 1; \tag{9.93}$$

(b) the Kac lemma:

$$\mathbf{E}(R_k|X_0 \in B) = \frac{1}{P(X_0 \in B)} \text{ for } k \in \mathbb{N}; \tag{9.94}$$

(c) the ergodic theorem:

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} R_i = \frac{1}{P(X_0 \in B)} \text{ a.s.}; \tag{9.95}$$

(d) the conditional stationarity:

$$P(R_{t+1}^{t+k} = r_1^k|X_0 \in B) = P(R_1^k = r_1^k|X_0 \in B) \text{ for } t \in \mathbb{N}; \tag{9.96}$$

Chapter 10

Limits

Risk functions. Analogies between coding and prediction. Induced predictor. Entropy rate. Shannon-McMillan-Breiman theorem. Universal codes and distributions. Unpredictability rate. Universal predictors. Pinsker inequality. Universal predictor induced by a universal prequential distribution.

In this chapter, we will generalize the constructions of universal codes from Chapters 7 and 8 to arbitrary stationary ergodic processes. In particular, we will study universal coding as contrasted and combined with the problem of universal prediction. The problems of universal coding and universal prediction are quite similar. In both, we seek for a single procedure that would be optimal within a class of probabilistic sources but in each problem we apply a different risk function.

Risk functions

Let $\mathbb{X}$ be a finite alphabet. In particular, in the problem of universal coding, we try to find an incomplete distribution $Q : \mathbb{X}^* \to [0, 1]$ such that the risk function

$$\ell(Q, x_1^n) := -\log Q(x_1^n) = -\sum_{i=0}^{n-1} \log Q(x_{i+1}|x_1^i) \tag{10.1}$$

is in some sense minimal across a wide class of infinite sequences $(x_i)_{i\in\mathbb{N}}$, such as typical outcomes of an IID or higher order Markov process. By contrast, in the problem of universal prediction, we seek for a function $f : \mathbb{X}^* \to \mathbb{X}$,

called a predictor, such that the risk function

$$\ell(f, x_1^n) := \sum_{i=0}^{n-1} \mathbf{1}\left\{x_{i+1} \neq f(x_1^i)\right\} \tag{10.2}$$

is minimal in the same sense across some typical sequences $(x_i)_{i\in\mathbb{N}}$.

It is natural to consider the following predictor which returns the most probable guess for a given prequential distribution. It is called the induced predictor.

Definition 10.1 (induced predictor) *Consider a finite alphabet $\mathbb{X} \subset \mathbb{N}$. Let $Q : \mathbb{X}^* \to [0, 1]$ be a prequential distribution. Then we define conditional probabilities*

$$Q(x_{n+1}|x_1^n) := \frac{Q(x_1^{n+1})}{Q(x_1^n)} \tag{10.3}$$

and the induced predictor

$$f_Q(x_1^n) := \arg\max_{x_{n+1}\in\mathbb{X}} Q(x_{n+1}|x_1^n), \tag{10.4}$$

where $\arg\max_{x\in\mathbb{X}} g(x) := \min\{a \in \mathbb{X} : g(a) \geq g(x)$ *for all* $x \in \mathbb{X}\}$.

Consequently, we may ask whether the induced predictor $f = f_Q$ minimizes risk $\ell(f, x_1^n)$ if prequential distribution Q minimizes risk $\ell(Q, x_1^n)$. The ultimate goal of this chapter is to show that the PPM mixture R introduced in Chapter 8 solves both problems indeed in the class of stationary ergodic processes over a finite alphabet $\{1, 2, ..., m\}$.

Coding

Let us begin with a formal statement of the problem of universal coding. The lower bound for risk $\ell(Q, x_1^n)$ in the class of stationary ergodic processes is given by the entropy rate. Let us recall that limits $P(X_i|X_{-\infty}^{i-1}) := \lim_{k\to\infty} P(X_i|X_{i-k}^{i-1})$ exist almost surely by the Lévy law. Hence some way of defining the entropy rate is as follows.

Definition 10.2 (entropy rate) *For a stationary process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet $\mathbb{X}$, the entropy rate is*

$$h := \mathbf{E}\left[-\log P(X_i|X_{-\infty}^{i-1})\right]. \tag{10.5}$$

We can ask a natural question whether the entropy rate is the limit of conditional entropies $h = \lim_{k\to\infty} H(X_i|X_{i-k}^{i-1})$. The first step to establish this is a uniform bound for conditional pointwise entropies.

Theorem 10.3 *Let $(X_i)_{i\in\mathbb{Z}}$ be a stochastic process over a finite alphabet. We have*

$$\mathbf{E}\sup_{k\in\mathbb{N}}\left[-\log P(X_i|X_{i-k}^{i-1})\right] \leq H(X_i) + \frac{1}{\ln 2}. \tag{10.6}$$

Proof: Let us take

$$Y := \sup_{k\in\mathbb{N}}\left[-\log P(X_i|X_{i-k}^{i-1})\right], \tag{10.7}$$

$$K_y := \inf_{k\in\mathbb{N}}\left\{k : P(X_i = x|X_{i-k}^{i-1}) < 2^{-y}\right\}. \tag{10.8}$$

We have $(Y > y) = (K_y < \infty)$. Variable K_y is a stopping time for bounded martingale $(P(X_i = x, K_y < \infty|X_{i-k}^{i-1}))_{k\in\mathbb{N}}$ with respect to process $(X_{i-k})_{k\in\mathbb{N}}$. Thus, by the Doob optional stopping theorem (Theorem 9.13), we obtain

$$\begin{aligned} P(X_i = x, Y > y) &= P(X_i = x, K_y < \infty) \\ &= \mathbf{E}\, P(X_i = x, K_y < \infty|X_{i-K_y}^{i-1}) \leq 2^{-y}. \end{aligned} \tag{10.9}$$

Hence, since $Y \geq 0$, we may write

$$\begin{aligned} \mathbf{E}\, Y &= \int_0^\infty P(Y > y)dy = \int_0^\infty \sum_{x\in\mathbb{X}} P(X_i = x, Y > y)dy \\ &\leq \sum_{x\in\mathbb{X}} \int_0^\infty \min\left\{P(X_i = x), 2^{-y}\right\} dy \\ &= \sum_{x\in\mathbb{X}} P(X_i = x)\left[-\log P(X_i = x) + \frac{1}{\ln 2}\right] = H(X_i) + \frac{1}{\ln 2}. \end{aligned} \tag{10.10}$$

□

Applying the above, entropy rate is a limit of conditional entropies.

Theorem 10.4 *For a stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet with $H(X_i) < \infty$, we have*

$$h = \lim_{k\to\infty} H(X_i|X_{i-k}^{i-1}) = \lim_{n\to\infty} \frac{H(X_1^n)}{n}. \tag{10.11}$$

Proof: We have $\mathbf{E} \sup_{k\in\mathbb{N}} \left[-\log P(X_i|X_{i-k}^{i-1}) \right] \leq H(X_i) + \frac{1}{\ln 2} < \infty$. Hence by the Lebesgue dominated convergence we derive

$$\mathbf{E} \lim_{k\to\infty} \left[-\log P(X_i|X_{i-k}^{i-1}) \right] = \lim_{k\to\infty} \mathbf{E} \left[-\log P(X_i|X_{i-k}^{i-1}) \right], \tag{10.12}$$

where the left-hand side is h and the right-hand side is $\lim_{k\to\infty} H(X_i|X_{i-k}^{i-1})$.

The remaining identity follows by observing

$$H(X_1^n) = \sum_{k=1}^{n} H(X_i|X_1^{i-1}) = \sum_{k=0}^{n-1} H(X_0|X_{-k}^{-1}) \tag{10.13}$$

and by the general fact

$$\lim_{k\to\infty} a_k = a \implies \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} a_k = a, \tag{10.14}$$

which is called the Toeplitz theorem (left as an exercise). □

As a result we obtain this generalization of asymptotic equipartition.

Theorem 10.5 (Shannon-McMillan-Breiman theorem) *For a stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet with $H(X_i) < \infty$, we have*

$$\lim_{n\to\infty} \frac{[-\log P(X_1^n)]}{n} = h \text{ a.s.} \tag{10.15}$$

Proof: We have $P(X_1^n) = \prod_{i=1}^{n} P(X_i|X_1^{i-1})$. Limits $P(X_i|X_{-\infty}^{i-1}) := \lim_{k\to\infty} P(X_i|X_{i-k}^{i-1})$ exist by the Lévy law and $\mathbf{E} \sup_{k\in\mathbb{N}} \left[-\log P(X_i|X_{i-k}^{i-1}) \right] \leq H(X_i) + \frac{1}{\ln 2} < \infty$. Hence the Breiman ergodic theorem (Theorem 9.36) yields

$$\begin{aligned} \lim_{n\to\infty} \frac{[-\log P(X_1^n)]}{n} &= \lim_{n\to\infty} \frac{1}{n} \left[-\sum_{i=1}^{n} \log P(X_i|X_1^{i-1}) \right] \\ &= \mathbf{E} \left[-\log P(X_i|X_{-\infty}^{i-1}) \right] = H(X_i|X_{-\infty}^{i-1}) = h \text{ a.s.} \end{aligned} \tag{10.16}$$

□

In view of the Shannon-McMillan-Breiman theorem, stationary ergodic processes are equipartitioned and it makes sense to consider universal codes with respect to this class of processes. Fortunately, in order to exhibit such universal codes and distributions, we do not need to invent anything new. The following criterion is the same as for Markov processes!

Theorem 10.6 (universality criterion) *Let $B : \mathbb{X}^* \to \{0,1\}^*$ be a uniquely decodable code. Code B is universal for stationary ergodic processes over a finite alphabet $\mathbb{X}$ if it satisfies criterion (7.9) or criterion (7.14).*

Proof: The proof is the same as of Theorems 7.5 and 7.6, except for invoking the Birkhoff ergodic theorem instead of the ergodic theorem for Markov processes. □

Hence the Lempel-Ziv code and the minimal grammar-based code defined in Chapter 7 as well as the PML maximum and the PPM mixture defined in Chapter 8 are all universal also for stationary ergodic processes over a finite alphabet.

Prediction

Let us proceed to a formal statement of the problem of universal prediction. The lower bound for risk $\ell(f, x_1^n)$ in the class of stationary ergodic processes is given by the unpredictability rate. This quantity may be defined as follows.

Definition 10.7 (unpredictability rate) *For a stationary process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet $\mathbb{X}$, the unpredictability rate is*

$$u := \mathbf{E}\left[1 - \max_{x_i \in \mathbb{X}} P(x_i | X_{-\infty}^{i-1})\right]. \tag{10.17}$$

By the Lévy law $P(X_i|X_{-\infty}^{i-1}) = \lim_{k\to\infty} P(X_i|X_{i-k}^{i-1})$ and the Lebesgue dominated convergence, we also have $u = \lim_{k\to\infty} \mathbf{E}\left[1 - \max_{x_i\in\mathbb{X}} P(x_i|X_{i-k}^{i-1})\right]$.

Comparing the formulas for the unpredictability rate u and the entropy rate h, it is natural to ask whether these values are related. We have this bound, which implies $h \to 0$ if and only if $u \to 0$ for a finite alphabet.

Theorem 10.8 (Fano inequality) *For a stationary process $(X_i)_{i\in\mathbb{Z}}$ over alphabet $\{1, 2, ..., m\}$, we have*

$$\frac{u\, H(1/m)}{1 - 1/m} \leq h \leq H(u) + u \log(m-1), \tag{10.18}$$

where we abbreviate the Shannon entropy $H(r) := H(r, 1-r)$ for an $r \in [0,1]$.

Proof: Let $\hat{X}_i = \arg\max_{x_i\in\{1,2,...,m\}} P(x_i|X_{-\infty}^{i-1})$. We have $u = 1 - P(X_i = \hat{X}_i)$ and $h = H(X_i|X_{-\infty}^{i-1})$. Let $Y = \mathbf{1}\left\{X_i = \hat{X}_i\right\}$. As for the right inequality,

we obtain

$$\begin{aligned} H(X_i|X_{-\infty}^{i-1}) &\le H(X_i|\hat{X}_i) = H(X_i, Y|\hat{X}_i) = H(Y|\hat{X}_i) + H(X_i|Y, \hat{X}_i) \\ &\le H(Y) + H(X_i|Y, \hat{X}_i) \\ &\le H(P(X_i = \hat{X}_i)) + [1 - P(X_i = \hat{X}_i)] \log(m-1). \end{aligned} \tag{10.19}$$

As for the left inequality, by concavity of function $p \mapsto H(p)$ we have

$$H(p) \ge H(q)\frac{1-p}{1-q} + H(1)\frac{p-q}{1-q} = H(q)\frac{1-p}{1-q} \tag{10.20}$$

for $p \in [q, 1]$. In particular $P(X_i = \hat{X}_i|X_{-\infty}^{i-1}) \ge 1/m$, so

$$\begin{aligned} H(X_i|X_{-\infty}^{i-1}) &\ge H(Y|X_{-\infty}^{i-1}) \ge \mathbf{E}\left[H(P(X_i = \hat{X}_i|X_{-\infty}^{i-1}))\right] \\ &\ge \frac{H(1/m)}{1-1/m}\,\mathbf{E}\left[1 - P(X_i = \hat{X}_i|X_{-\infty}^{i-1})\right] \\ &= \frac{H(1/m)}{1-1/m}[1 - P(X_i = \hat{X}_i)]. \end{aligned} \tag{10.21}$$

□

The above result suggests that we may pursue analogies between coding and prediction further. Now, as an analogue of the source coding theorem, we will show that no predictor can beat the induced predictor f_P and the error rate committed by the latter equals the unpredictability rate.

Theorem 10.9 (source prediction) *Let $f : \mathbb{X}^* \to \mathbb{X}$ be a predictor. For any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ with random variables $X_i : \Omega \to \mathbb{X}$, we have*

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbf{1}\left\{X_{i+1} \ne f(X_1^i)\right\} \ge u \text{ a.s.} \tag{10.22}$$

where the equality holds with $\liminf = \lim$ *for $f = f_P$.*

Proof: The rate of the prediction risk is bounded by $0 \le \mathbf{1}\{X_{i+1} \ne f(X_1^i)\} \le 1$. Hence, from Theorem 9.24 for any predictor f, we derive

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \left[\mathbf{1}\left\{X_{i+1} \ne f(X_1^i)\right\} - P(X_{i+1} \ne f(X_1^i)|X_1^i)\right] = 0 \text{ a.s.} \tag{10.23}$$

Moreover, we have

$$P(X_{i+1} \neq f(X_1^i)|X_1^i) \geq 1 - \max_{x_{i+1} \in \mathbb{X}} P(x_{i+1}|X_1^i) \tag{10.24}$$

where the inequality becomes equality for $f = f_P$. Subsequently, we observe that limits $\lim_{n\to\infty} P(x_0|X_{-n}^{-1})$ exist almost surely by the Lévy law. Thus by the Breiman ergodic theorem and the dominated convergence, almost surely we obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[1 - \max_{x_{i+1} \in \mathbb{X}} P(x_{i+1}|X_1^i)\right] = \mathbf{E}\left[1 - \max_{x_0 \in \mathbb{X}} P(x_0|X_{-\infty}^{-1})\right] = u \text{ a.s.} \tag{10.25}$$

Hence the claimed inequality follows by (10.23), (10.24) and (10.25). □

Thus let us postulate the following concept of a universal predictor.

Definition 10.10 (universal predictor) *Let $f : \mathbb{X}^* \to \mathbb{X}$ be a predictor. Predictor f is called universal for stationary ergodic processes over a finite alphabet $\mathbb{X}$ if for any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ with random variables $X_i : \Omega \to \mathbb{X}$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}\left\{X_{i+1} \neq f(X_1^i)\right\} = u \text{ a.s.} \tag{10.26}$$

We do not need to specially care about universality in expectation since this follows from the almost sure universality by the boundedness of risk $0 \leq \mathbf{1}\{X_{i+1} \neq f(X_1^i)\} \leq 1$.

Do universal predictors exist? Can they be induced by universal prequential distributions? Subsequently, we will show that each universal prequential distribution induces a universal predictor under a mild condition. This condition is satisfied in particular by the PPM mixture R defined in Definition 8.11. On our way, we will apply the Breiman ergodic theorem, the Azuma corollary, the Pinsker inequality and a few other results.

The first stage of preparations includes assertions which can be called the smoothed equipartition and the smoothed universality.

Theorem 10.11 (smoothed equipartition) *Let $\mathbb{X}$ be a finite alphabet. For any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ with random variables $X_i : \Omega \to \mathbb{X}$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[-\sum_{x_{i+1} \in \mathbb{X}} P(x_{i+1}|X_1^i) \log P(x_{i+1}|X_1^i)\right] = h \text{ a.s.} \tag{10.27}$$

Proof: Let us write the conditional entropy

$$h(x_1^n) := -\sum_{x_{n+1}\in\mathbb{X}} P(x_{n+1}|x_1^n)\log P(x_{n+1}|x_1^n). \tag{10.28}$$

We have $0 \leq h(x_1^n) \leq \log m$ with m being the cardinality of the alphabet. Moreover by the Lévy law, there exists limit

$$h(X_{-\infty}^{-1}) := \lim_{n\to\infty} h(X_{-n}^{-1}) = -\sum_{x_0\in\mathbb{X}} P(x_0|X_{-\infty}^{-1})\log P(x_0|X_{-\infty}^{-1}) \text{ a.s.} \tag{10.29}$$

Hence by the Breiman ergodic theorem, we obtain

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} h(X_1^i) = \mathbf{E}\, h(X_{-\infty}^{-1}) = \mathbf{E}\left[-\log P(X_0|X_{-\infty}^{-1})\right] = h \text{ a.s.} \tag{10.30}$$

□

Theorem 10.12 (smoothed universality) *Let $\mathbb{X}$ be a finite alphabet. Let $Q : \mathbb{X}^* \to [0,1]$ be a prequential distribution which is universal for stationary ergodic processes and satisfies*

$$-\log Q(x_{n+1}|x_1^n) \leq Cn^{1/2-\epsilon} \tag{10.31}$$

for some $C < \infty$ and $\epsilon > 0$. Then for any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ with random variables $X_i : \Omega \to \mathbb{X}$, we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\left[-\sum_{x_{i+1}\in\mathbb{X}} P(x_{i+1}|X_1^i)\log Q(x_{i+1}|X_1^i)\right] = h \text{ a.s.} \tag{10.32}$$

Proof: Let us write the conditional pointwise entropy $Z_i := -\log Q(X_{i+1}|X_1^i)$. We have

$$\mathbf{E}\left(Z_i\middle|X_1^i\right) = -\sum_{x_{i+1}\in\mathbb{X}} P(x_{i+1}|X_1^i)\log Q(x_{i+1}|X_1^i). \tag{10.33}$$

Now suppose that distribution Q is universal and satisfies (10.31). Then by the Azuma corollary,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{E}\left(Z_i\middle|X_1^i\right) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} Z_i = \lim_{n\to\infty}\frac{1}{n}\left[-\log Q(X_1^n)\right] = h \text{ a.s.} \tag{10.34}$$

□

In the second stage of our preparations, we will prove the famous Pinsker inequality and yet another inequality, called the prediction inequality. The Pinsker inequality compares the Kullback-Leibler divergence with the total variation distance.

Theorem 10.13 (Pinsker inequality) *Let p and q be probability distributions over a countable alphabet $\mathbb{X}$. We have*

$$\left[\sum_{x\in\mathbb{X}} |p(x)-q(x)|\right]^2 \leq (2\ln 2)\sum_{x\in\mathbb{X}} p(x)\log\frac{p(x)}{q(x)}. \tag{10.35}$$

Proof: First, we will prove inequality (10.35) for $\mathbb{X}=\{0,1\}$. Let us denote $a := p(1)$ and $b := q(1)$ and assume without loss of generality that $a \geq b$. The difference between the right hand side and the left hand side of (10.35) is

$$g(a,b) = (2\ln 2)\left[a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}\right] - 4(a-b)^2. \tag{10.36}$$

Since $g(a,b)=0$ for $a=b$ and

$$\frac{\partial g(a,b)}{\partial b} = -2\left[\frac{a}{b}+\frac{1-a}{1-b}\right] + 8(a-b) = -2(a-b)\left[\frac{1}{b(1-b)} - 4\right] \leq 0 \tag{10.37}$$

then $g(a,b)\geq 0$.

Assume now that $\mathbb{X}$ is general. Let $U=\{x\in\mathbb{X} : p(x)\geq q(x)\}$ and

$$\tilde{p}(1) := \sum_{x\in U} p(x), \qquad \tilde{p}(0) := \sum_{x\notin U} p(x), \tag{10.38}$$

$$\tilde{q}(1) := \sum_{x\in U} q(x), \qquad \tilde{q}(0) := \sum_{x\notin U} q(x). \tag{10.39}$$

By a property of the Kullback-Leibler divergence which generalizes the data-processing inequality, we obtain

$$\begin{aligned}(2\ln 2)\sum_{x\in\mathbb{X}} p(x)\log\frac{p(x)}{q(x)} &\geq (2\ln 2)\sum_{x\in\{0,1\}} \tilde{p}(x)\log\frac{\tilde{p}(x)}{\tilde{q}(x)} \\ &\geq \left[\sum_{x\in\{0,1\}} |\tilde{p}(x)-\tilde{q}(x)|\right]^2 = \left[\sum_{x\in\mathbb{X}} |p(x)-q(x)|\right]^2.\end{aligned} \tag{10.40}$$

□

By contrast, the prediction inequality connects the total variation distance with error probabilities of the induced predictors.

Theorem 10.14 (prediction inequality) *Let p and q be two probability distributions over a countable alphabet $\mathbb{X}$. For $x_p = \arg\max_{x\in\mathbb{X}} p(x)$ and $x_q = \arg\max_{x\in\mathbb{X}} q(x)$, we have inequality*

$$0 \leq p(x_p) - p(x_q) \leq \sum_{x\in\mathbb{X}} |p(x) - q(x)| \,. \tag{10.41}$$

Proof: Without loss of generality, assume $x_p \neq x_q$. By the definition of x_p and x_q, we have $p(x_p) - p(x_q) \geq 0$ and $q(x_q) - q(x_p) \geq 0$. Hence we obtain

$$\begin{aligned} 0 &\leq p(x_p) - p(x_q) \leq p(x_p) - p(x_q) - q(x_p) + q(x_q) \\ &\leq |p(x_p) - q(x_p)| + |p(x_q) - q(x_q)| \leq \sum_{x} |p(x) - q(x)| \,. \end{aligned} \tag{10.42}$$

□

Now we can show that every universal distribution which satisfies condition (10.31) induces a universal predictor.

Theorem 10.15 (induced prediction) *Let $\mathbb{X}$ be a finite alphabet. Let $Q : \mathbb{X}^* \to [0,1]$ be a prequential distribution which is universal for stationary ergodic processes and satisfies condition (10.31). Then the induced predictor $f_Q : \mathbb{X}^* \to \mathbb{X}$ is universal for stationary ergodic processes.*

Proof: By the smoothed equipartition and the smoothed universality, we derive

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[\sum_{x_{i+1}} P(x_{i+1}|X_1^i) \log \frac{P(x_{i+1}|X_1^i)}{Q(x_{i+1}|X_1^i)} \right] = 0 \text{ a.s.} \tag{10.43}$$

Hence applying the Pinsker inequality yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[\sum_{x_{i+1}} \left| P(x_{i+1}|X_1^i) - Q(x_{i+1}|X_1^i) \right| \right]^2 = 0 \text{ a.s.} \tag{10.44}$$

Subsequently, the Cauchy-Schwarz inequality $\mathbf{E}\, Y^2 \geq (\mathbf{E}\, Y)^2$ implies

$$\begin{aligned} 0 &\geq \lim_{n\to\infty} \left[\frac{1}{n} \sum_{i=0}^{n-1} \sum_{x_{i+1}} \left| P(x_{i+1}|X_1^i) - Q(x_{i+1}|X_1^i) \right| \right]^2 \\ &= \left[\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \sum_{x_{i+1}} \left| P(x_{i+1}|X_1^i) - Q(x_{i+1}|X_1^i) \right| \right]^2 \geq 0 \text{ a.s.} \end{aligned} \tag{10.45}$$

As a result we infer

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\sum_{x_{i+1}}\left|P(x_{i+1}|X_1^i)-Q(x_{i+1}|X_1^i)\right|=0 \text{ a.s.} \tag{10.46}$$

Consequently, combining this with the prediction inequality yields

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\left[P(X_{i+1}\neq f_Q(X_1^i)|X_1^i)-P(X_{i+1}\neq f_P(X_1^i)|X_1^i)\right]=0 \text{ a.s.} \tag{10.47}$$

Now, we notice that by (10.23), we have almost surely

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\left[\mathbf{1}\left\{X_{i+1}\neq f_Q(X_1^i)\right\}-P(X_{i+1}\neq f_Q(X_1^i)|X_1^i)\right]=0 \text{ a.s.}, \tag{10.48}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\left[\mathbf{1}\left\{X_{i+1}\neq f_P(X_1^i)\right\}-P(X_{i+1}\neq f_P(X_1^i)|X_1^i)\right]=0 \text{ a.s.} \tag{10.49}$$

Combining the three above observations completes the proof. □

Thus, in order to exhibit a universal predictor, we do not need to invent anything new, either.

Theorem 10.16 *Let the cardinality of the alphabet be $m \geq 2$. The PPM mixture R defined in Definition 8.11 satisfies*

$$-\log R(x_1^n) \leq 2\log(n+2)+n\log m, \tag{10.50}$$

$$-\log R(x_{n+1}|x_1^n) \leq 3\log(n+m). \tag{10.51}$$

Hence the induced predictor f_R is universal for stationary ergodic processes over alphabet $\{1,2,...,m\}$.

Proof: Observe that $R_k(x_1^n)=m^{-n}$ for $k\geq n$. Hence by $R(x_1^n)\geq w_nR_n(x_1^n)$ and $w_n=(n+1)^{-1}(n+2)^{-1}$, we obtain claim (10.50). The derivation of claim (10.51) is slightly longer. First, by the definition of R_k, we have

$$-\log R_k(x_{n+1}|x_1^n)\leq\log\left[N(x_{n+1-k}^n|x_1^{n-1})+m\right]\leq\log(n+m). \tag{10.52}$$

for any $k=0,1,....$ Now let

$$g:=\arg\max_{k\in\mathbb{N}}R_k(x_1^n). \tag{10.53}$$

We have $g \leq n$, since $R_k(x_1^n) = m^{-n}$ for $k \geq n$. Moreover, we have $R(x_1^n) \leq R_g(x_1^n)$. Combining this with $R(x_1^{n+1}) \geq w_g R_g(x_1^{n+1})$ yields

$$\begin{aligned} -\log R(x_{n+1}|x_1^n) &= -\log R(x_1^{n+1}) + \log R(x_1^n) \\ &\leq -\log w_g - \log R_g(x_1^{n+1}) + \log R_g(x_1^n) \\ &\leq 2\log(g+2) + \log R_g(x_{n+1}|x_1^n) \leq 3\log(n+m). \quad (10.54) \end{aligned}$$

Thus universality of predictor f_R follows by universality of the PPM mixture R and Theorem 10.15. □

***

To recapitulate this chapter, we have seen that universal codes constructed in Chapters 7 and 8, which are good for higher order Markov processes, are also suitable for general stationary ergodic processes. Moreover, the PPM mixture from Chapter 8 solves not only the problem of universal data compression but also the problem of universal prediction. This provides yet another link between information theory and learning. In the following Chapters 11 and 12, we will make some further excursion to connect information theory with theory of computation.

Further reading

The concept of the entropy rate was introduced by Claude Shannon in paper [114]. In paper [115], he also tried to estimated the entropy rate of a text in English, which amounted to 1 bit per letter. Leo Breiman proved the Shannon-McMillan-Breiman theorem using the Breiman ergodic theorem [13]. The more popular proof of the Shannon-McMillan-Breiman theorem uses a sandwich bound and was discovered by Paul Algoet and Thomas Cover [2]. The most general case of equipartitioned processes are not stationary ergodic processes but asymptotically mean stationary (AMS) ergodic processes, which generalize the concept of non-stationary Markov processes. The theory of AMS processes was developed by Robert Gray and John Kieffer [59]. The Fano inequality was discovered by Robert Fano [47], whereas its converse can be found in my book [36]. The Pinsker inequality in a weaker form was discovered by Mark Pinsker, whereas the form presented in this chapter is due to Solomon Kullback, Imre Csiszár, and Johannes Kemperman. See also the book by Imre Csiszár and János Körner [30]. The theory of universal prediction bears to the works of Ray Solomonoff [117], David Bailey [4], Donald Ornstein [103], and Paul Algoet [1]. First practical universal predictors were constructed by László Györfi, Gábor Lugosi, and Gusztav

Morvai [63, 61]. Interactions between universal coding and universal prediction were studied by Boris Ryabko [112, 108, 109] and Joe Suzuki [119]. Daniil Ryabko connected the existence of universal codes with separability of the space of probability measures [110]. The theory of induced predictors in the shape discussed in this chapter was introduced by Łukasz Dębowski and Tomasz Steifer [40]. There is also a recent survey by Gusztav Morvai and Benjamin Weiss [99]. The PPM distribution has multiple applications in statistical inference from stationary processes. These applications cover not only source prediction but also density estimation as it can be found in [111] and [39].

Thinking exercises

1. *Toeplitz theorem:* Prove that $\lim_{k\to\infty} a_k = a$ implies $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} a_k = a$. Show that the converse is not true.

2. *Fekete lemma:* Consider four conditions:

   (a) Sequence $(a_n)_{n\in\mathbb{N}}$ has decreasing increments if $(a_n - a_{n-1})_{n\in\mathbb{N}}$ is decreasing.

   (b) Sequence $(a_n)_{n\in\mathbb{N}}$ has decreasing n-ths if $(a_n/n)_{n\in\mathbb{N}}$ is decreasing.

   (c) Sequence $(a_n)_{n\in\mathbb{N}}$ is subadditive if $a_{n+m} \leq a_n + a_m$.

   (d) Sequence $(a_n)_{n\in\mathbb{N}}$ has descending n-ths if $\lim_{n\to\infty} a_n/n = \inf_{n\in\mathbb{N}} a_n/n$.

   Show that (a) $\implies$ (b) $\implies$ (c) $\implies$ (d) but the converse is not true. Moreover, show that if a sequence $(a_n)_{n\in\mathbb{N}}$ has decreasing increments then this sequence is concave, i.e., $\lambda a_n + (1-\lambda)a_m \leq a_{\lambda n+(1-\lambda)m}$ for $0 \leq \lambda \leq 1$. Which of conditions (a)–(d) are satisfied by the entropy $H(X_1^n)$ of a stationary process $(X_i)_{i\in\mathbb{Z}}$? See also [50].

3. Let $(X_i)_{i\in\mathbb{N}}$ be a stationary process with entropy rate h.

   (a) Let N be a random variable assuming values in natural numbers, where events $(N = n)$ and $(X_1^n = x_1^n)$ are independent. Show that

   $$H(X_1^N) \leq H(N) + H\left(X_1^{\lceil \mathbf{E}\, N\rceil}\right). \tag{10.55}$$

(b) Let N be a random variable assuming values in natural numbers, where events $(N = n)$ and $(X_{n+1}^m = x_{n+1}^m)$ are independent. Show that

$$H(X_1^N) \geq H(N|(X_i)_{i\in\mathbb{N}}) + h\,\mathbf{E}\,N, \tag{10.56}$$

where $H(N|(X_i)_{i\in\mathbb{N}}) := \lim_{n\to\infty} H(N|X_1^n)$,

4. *Cesàro mean distribution:* Let $Q : \mathbb{X}^* \to [0,1]$ be a prequential distribution. The Cesàro mean distribution [62, 39] is defined as

$$\bar{Q}(x_n|x_1^{n-1}) := \frac{1}{n}\sum_{i=1}^{n-1} Q(x_n|x_{n-i}^{n-1}), \tag{10.57}$$

$$\bar{Q}(x_1^n) := \prod_{i=1}^{n} \bar{Q}(x_i|x_1^{i-1}). \tag{10.58}$$

Show that:

(a) For a stationary process $(X_i)_{i\in\mathbb{Z}}$ over a countable alphabet $\mathbb{X}$, we have

$$\frac{1}{n}\,\mathbf{E}\left[-\log Q(X_1^n)\right] \geq \mathbf{E}\left[-\log \bar{Q}(X_n|X_1^{n-1})\right]. \tag{10.59}$$

(b) If distribution Q is universal for stationary ergodic processes then the induced predictor $f_{\bar{Q}} : \mathbb{X}^* \to \mathbb{X}$ with respect to the Cesàro mean distribution $\bar{Q}$ is universal for stationary ergodic processes.

5. *Jensen-Shannon divergence:* The Jensen-Shannon divergence between probability distributions p and q is defined as

$$\mathrm{JSD}(p,q) := \frac{1}{2}D\left(p\middle\|\frac{p+q}{2}\right) + \frac{1}{2}D\left(q\middle\|\frac{p+q}{2}\right). \tag{10.60}$$

Show that

(a) $\mathrm{JSD}(p,q) = H\left(\frac{p+q}{2}\right) - \frac{1}{2}H(p) - \frac{1}{2}H(q)$;
(b) $0 \leq \mathrm{JSD}(p,q) \leq 1$;
(c) $\mathrm{JSD}(p,q) \leq \frac{1}{2}\sum_x |p(x) - q(x)|$.

6. *Jensen-Shannon distance:* The Jensen-Shannon divergence between probability distributions p and q is defined as $\sqrt{\mathrm{JSD}(p,q)}$. Show that this is a metric on probability distributions, i.e., we have

(a) $\sqrt{\mathrm{JSD}(p,q)} = 0 \implies p = q$;
(b) $\sqrt{\mathrm{JSD}(p,q)} = \sqrt{\mathrm{JSD}(q,p)}$;
(c) $\sqrt{\mathrm{JSD}(p,q)} \leq \sqrt{\mathrm{JSD}(p,r)} + \sqrt{\mathrm{JSD}(r,q)}$.

Chapter 11

Computation

Register machine. Total, partial, and computable functions. Functions of natural numbers, rational numbers, and strings. Computable functions. Universal function. Halting problem. Computable and computably enumerable sets. Semi-computable functions. Kolmogorov complexity. Uncomputability of Kolmogorov complexity.

In this chapter, we will study the theory of computation and will define the main concept of the algorithmic information theory which is called the Kolmogorov complexity. Kolmogorov complexity is a beautiful theoretical approach to the definition of the amount of information, which turns out infeasible in practice. However, despite being infeasible, this approach sets out a hard lower bound for the length of any practical universal code. At least for that reason, it should be remembered.

Before we define the Kolmogorov complexity itself, we need to provide an introduction to the theory of computable functions. The more frequently traveled path to define computable functions goes via Turing machines. Here we will use an alternative approach, called register machines, which assumes that the reader has a certain familiarity with imperative programming languages, i.e., languages with an operation of substitution such as "$x := x+1$". We suppose that this approach is more natural nowadays.

The definition of a register machine is as follows.

Definition 11.1 (register machine) *The register machine is the interpreter of the following programming language. First, there are variables taking values in natural numbers, called registers:* $R_1, R_2, R_3, ... \in \mathbb{N}$. *Second, an admissible program is a finite list* $\pi = (\pi_1, \pi_2, ..., \pi_{|\pi|})$ *of commands* π_l *of form:*

- $R_j := 1$ *(the value of register* R_j *becomes* 1*);*
- $R_j := R_j + 1$ *(the value of register* R_j *gets incremented by* 1*);*
- $R_j := R_k$ *(the value of register* R_j *becomes the value of* R_k*);*
- $R_j = R_k \Rightarrow m$ *(if* R_j *equals* R_k *then go to the* m*-th command).*

Definition 11.2 (state of register machine) *The state of the register machine during the interpretation of program* $\pi = (\pi_1, \pi_2, ..., \pi_{|\pi|})$ *is a list* $x = (x_0, x_1, x_2, ...)$*, where* $x_i \in \mathbb{N}$*, value* x_0 *is the present command number and* x_j *for* $j \in \mathbb{N}$ *are the present values of registers* R_j*. Formally, we define relation* $x \overset{\pi}{\to} x'$ *linking the states directly succeeding in computation of program* π*, namely, we have* $(x_0, x_1, x_2, ...) \overset{\pi}{\to} (x'_0, x'_1, x'_2, ...)$ *if and only if*

- *if* $\pi_{x_0} =$ "$R_j := 1$" *then* $x'_0 = x_0 + 1$*,* $x'_j = 1$*, and* $x'_l = x_l$ *else;*
- *if* $\pi_{x_0} =$ "$R_j := R_j + 1$" *then* $x'_0 = x_0 + 1$*,* $x'_j = x_j + 1$*, and* $x'_l = x_l$ *else;*
- *if* $\pi_{x_0} =$ "$R_j := R_k$" *then* $x'_0 = x_0 + 1$*,* $x'_j = x_k$*, and* $x'_l = x_l$ *else;*
- *if* $x_j = x_k$ *and* $\pi_{x_0} =$ "$R_j = R_k \Rightarrow m$" *then* $x'_0 = m$ *and* $x'_l = x_l$ *else;*
- *if* $x_j \neq x_k$ *and* $\pi_{x_0} =$ "$R_j = R_k \Rightarrow m$" *then* $x'_0 = x_0 + 1$ *and* $x'_l = x_l$ *else.*

We denote $\overset{\pi}{\to}^*$ *for the transitive closure of relation* $\overset{\pi}{\to}$*, namely, proposition* $x \overset{\pi}{\to}^* x'$ *holds if* $x = x'$ *or* $x \overset{\pi}{\to}^* x'' \overset{\pi}{\to} x'$ *for some* x''*.*

The register machine can be used to define computable functions of natural numbers but a certain care is needed. Let us observe that there need not hold $(x_0, x_1, x_2, ...) \overset{\pi}{\to}^* (|\pi| + 1, x'_1, x'_2, ...)$ since the register machine may not halt on a program π—it can get stuck into an infinite loop of computations. If we want to establish a strict correspondence between programs and functions, we need to introduce a special convention dealing explicitly with computations that do not halt.

Definition 11.3 (total and partial functions) *Let* $\perp$ *be a special symbol corresponding to an undefined value. Let* A *and* B *be some sets that do not contain* $\perp$*. Any function* $F : A \to B$ *is called a total function. Any function* $F : A \to B \cup \{\perp\}$ *is called a partial function. Property* $F : A \to B \cup \{\perp\}$ *will be written briefly as* $F : A \overset{o}{\to} B$*. For* $x \in A$*, we say* $F(x)$ *halts and write* $F(x)\downarrow$ *if* $F(x) \neq \perp$ *and we say* $F(x)$ *does not halt and write* $F(x)\uparrow$ *if* $F(x) = \perp$*.*

Definition 11.4 (partial computable function) *The natural function computed by a program π is the function $F_\pi : \mathbb{N} \stackrel{o}{\to} \mathbb{N}$ defined as*

$$F_\pi(x) := \begin{cases} y & \text{if } (1, x, 1, 1, ...) \stackrel{\pi}{\to}^* (|\pi| + 1, y, ...), \\ \perp & \text{else.} \end{cases} \tag{11.1}$$

A partial function $F : \mathbb{N} \stackrel{o}{\to} \mathbb{N}$ is called computable if $F = F_\pi$ for a certain program π.

The theory of computation is intimately linked with coding discrete objects as other discrete objects. Important examples of discrete objects are not only natural numbers but also binary strings and programs. To speak of computable functions of binary strings, we will apply some codes for natural numbers introduced in Chapter 1. In particular, let us consider the non-singular military code $\text{mil} : \mathbb{N} \to \{0,1\}^*$ and the prefix-free unary code $\text{una} : \mathbb{N} \to \{0,1\}^*$.

- Using the military code $\text{mil} : \mathbb{N} \to \{0,1\}^*$, we will speak of computable functions of binary strings—understanding them as the coded versions of computable functions of natural numbers. Since $\text{mil} : \mathbb{N} \to \{0,1\}^*$ is a one-to-one mapping, we will apply it to convert strings into numbers and numbers into strings whenever necessary—overloading notations $F(\text{mil}^{-1}(x)) := F(x)$ and $F(x) := \text{mil}(F(x))$ wherever appropriate.

- As for the prefix-free unary code $\text{una} : \mathbb{N} \to \{0,1\}^*$, we will use a useful notation for the following non-singular code for pairs of binary strings or natural numbers:

$$\langle u_1, u_2 \rangle := \text{una}(|w_1| + 1) w_1 w_2, \tag{11.2}$$

  where $w_i = u_i$ for $u_i \in \{0,1\}^*$ and $w_i = \text{mil}(u_i)$ for $u_i \in \mathbb{N}$. This notation will be iterated:

$$\langle u_1, ..., u_n \rangle := \langle \langle u_1, ..., u_{n-1} \rangle, u_n \rangle . \tag{11.3}$$

- Last but not least, the programs for a register machine can be also encoded as binary strings (or as natural numbers, if needed). To determine some fixed correspondence between binary strings and programs, we will define the binary code word $B^*(\pi) \in \{0,1\}^*$ for a program $\pi = (\pi_1, \pi_2, ..., \pi_{|\pi|})$ as

$$B^*(\pi) := B(\pi_1) B(\pi_2) ... B(\pi_{|\pi|}), \tag{11.4}$$

  where

- $B(\pi_l) := 00\,\mathrm{una}(j)$ if $\pi_l =$ "$R_j := 1$";
- $B(\pi_l) := 01\,\mathrm{una}(j)$ if $\pi_l =$ "$R_j := R_j + 1$";
- $B(\pi_l) := 10\,\mathrm{una}(j)\,\mathrm{una}(k)$ if $\pi_l =$ "$R_j := R_k$";
- $B(\pi_l) := 11\,\mathrm{una}(j)\,\mathrm{una}(k)\,\mathrm{una}(m)$ if $\pi_l =$ "$R_j = R_k \Rightarrow m$".

It can be checked that code $\pi \mapsto B^*(\pi)$ is non-singular since code $\pi_l \mapsto B(\pi_l)$ is prefix-free. Again, overloading notation, we will write π instead of $B^*(\pi)$ and F_π.

Now we can make a practical use of all these coding correspondences and we can use them to construct some important theoretical entities.

Definition 11.5 (universal function) *The universal function* $U : \{0,1\}^* \stackrel{o}{\to} \{0,1\}^*$ *is the partial function such that for all programs* π *we have*

$$U(w) := \begin{cases} \pi(n) & \text{if } w = \langle \pi, n \rangle\,, \\ \bot & \text{else.} \end{cases} \tag{11.5}$$

We can meaningfully ask whether this function is computable.

Theorem 11.6 *The universal function is computable.*

Proof: The proof is tedious. It boils down to writing an interpreter of programs for the register machine, which given a code word B_π decodes it to program π and executes its commands π_l in a virtual memory space. □

Let us exhibit an example of function which is not computable. In fact, this function answers which computable functions halt on particular inputs. The respective result is by Alan Turing.

Theorem 11.7 (halting problem) *The total function* $H : \{0,1\}^* \to \{0,1\}$ *such that*

$$H(w) := \begin{cases} 1 & \text{if } U(w)\downarrow, \\ 0 & \text{if } U(w)\uparrow, \end{cases} \tag{11.6}$$

is not computable.

Proof: Suppose by contradiction that function H is total computable. Then we may define a partial computable function

$$\pi(n) := \begin{cases} 1 & \text{if } H(\langle n, n \rangle) = 0, \\ \bot & \text{if } H(\langle n, n \rangle) = 1. \end{cases} \tag{11.7}$$

(To remember: π is the negation of H taken on the diagonal, like in Georg Cantor's famous proof that the set of real numbers is uncountable.)

Let us write explicitly

$$H(\langle \pi, n \rangle) = \begin{cases} 1 & \text{if } \pi(n) \downarrow, \\ 0 & \text{if } \pi(n) \uparrow . \end{cases} \tag{11.8}$$

Let us deduce the value of $H(\langle \pi, \pi \rangle)$:

- First, if $H(\langle \pi, \pi \rangle) = 0$ then $\pi(\pi) = 1$ and so $H(\langle \pi, \pi \rangle) = 1$.
- Second, if $H(\langle \pi, \pi \rangle) = 1$ then $\pi(\pi) = \perp$ and so $H(\langle \pi, \pi \rangle) = 0$.

Since we have obtained a contradiction in both cases then function H cannot be total computable. □

Undecidability of the halting problem is an important fact. If function H were computable, we could use it to solve many mysteries of number theory and other branches of pure mathematics.

Example 11.8 (Goldbach conjecture) *The Goldbach conjecture states that every even number greater than* 2 *is the sum of two prime numbers. It is still unknown whether it is true. The counterexamples for the Goldbach conjecture can be sought by a simple search procedure: we iterate through consecutive numbers* $n = 2, 3, 4, ..,$ *we check whether* $2n$ *is the sum of two prime numbers, and we abandon this search when the answer is negative. This procedure corresponds to a certain partial computable function* F_π*. Thus, if we could compute* $H(B_\pi)$ *then we would know whether the Goldbach conjecture is true.*

Although the halting problem is not decidable, its solution can be effectively approximated in an uncontrolled way. To approach this topic, we will introduce computable and computably enumerable sets.

Definition 11.9 (computable and computably enumerable sets) *A set* $A \subset \mathbb{N}$ *is called computable if there exists a total computable function* $F : \mathbb{N} \to \{0, 1\}$ *such that*

$$n \in A \iff F(n) = 1. \tag{11.9}$$

A set $A \subset \mathbb{N}$ *is called computably enumerable if there exists a partial computable function* $G : \mathbb{N} \to \{1, \perp\}$ *such that*

$$n \in A \iff G(n) = 1. \tag{11.10}$$

Every computable set is computably enumerable since it suffices to take $G(n) = 1 \iff F(n) = 1$. We also notice that the halting set

$$\mathcal{H} := \{w : U(w)\downarrow\} \tag{11.11}$$

is not computable, as we have shown, but it is computably enumerable. It is so since we can put function $G(w) = 1 \iff U(w)\downarrow$ to satisfy the definition of a computably enumerable set.

In contrast, set $\{0,1\}^* \setminus \mathcal{H}$, i.e., the complement of the halting set is neither computable nor computably enumerable. It is so since we have this general fact:

Theorem 11.10 (Post theorem) *Sets A and $\mathbb{N} \setminus A$ are both computably enumerable if and only if they are both computable.*

Proof: Let sets A and $\mathbb{N} \setminus A$ be both computably enumerable Let $G, G' : \mathbb{N} \to \{1, \perp\}$ satisfy $n \in A \iff G(n) = 1$ and $n \in \mathbb{N} \setminus A \iff G'(n) = 1$. Then we can construct a computable total function $F : \mathbb{N} \to \{0,1\}$ such that $F(n) = G(n)$ if $G(n)\downarrow$ and $F(n) = 1 - G'(n)$ if $G'(n)\downarrow$. This function satisfies $n \in A \iff F(n) = 1$ so set A is computable. Analogously we prove computability of set $\mathbb{N} \setminus A$.

In contrast, if sets A and $\mathbb{N} \setminus A$ are computable then they are computably enumerable since every computable set is computably enumerable. □

Now let us look into a problem of approximating the graph of a function of natural numbers by computably enumerable sets. For a function $F : \mathbb{N} \to \mathbb{N}$ let us introduce sets

$$L_F := \{\langle n, q\rangle : q \leq F(n),\ q, n \in \mathbb{N}\}, \tag{11.12}$$

$$U_F := \{\langle n, q\rangle : q \geq F(n),\ q, n \in \mathbb{N}\}. \tag{11.13}$$

There is an easy application of the previous result.

Theorem 11.11 *A total function $F : \mathbb{N} \to \mathbb{N}$ is computable if and only if sets L_F and U_F are both computably enumerable.*

Proof: In view of Theorem 11.10, it suffices to show that $F : \mathbb{N} \to \mathbb{N}$ is computable if and only if sets L_F and U_F are both computable. To show it, we introduce functions $G, G' : \mathbb{N} \to \{0,1\}$ such that

$$G(\langle q, n\rangle) = 1 \iff q \leq F(n), \tag{11.14}$$

$$G'(\langle q, n\rangle) = 1 \iff q \geq F(n). \tag{11.15}$$

If F is total computable then G and G' are total computable which proves computability of sets L_F and U_F. If sets L_F and U_F are both computable then functions G and G' are total computable and hence function

$$F(n) = \min \{q \in \mathbb{N} : G(\langle q, n \rangle) = G'(\langle q, n \rangle)\} \tag{11.16}$$

is also total computable. □

Otherwise, we have these definitions.

Definition 11.12 (lower and upper semi-computable functions) *A total function $F : \mathbb{N} \to \mathbb{N}$ is called:*

- *lower semi-computable if set L_F is computably enumerable,*
- *upper semi-computable if set U_F is computably enumerable.*

In plain words, a function is lower semi-computable if it can be computably approximated from below, whereas a function is upper semi-computable if it can be computably approximated from above.

An important example of a function which is not computable but is upper semi-computable is the Kolmogorov complexity. As we have mentioned, the Kolmogorov complexity is some measure of information of a great theoretical importance. In the following, the arguments of the universal function will be also called programs. The Kolmogorov complexity is simply defined as the length of the shortest program for a universal function which outputs a given string or another given discrete object such as a natural number.

Definition 11.13 (Kolmogorov complexity) *The Kolmogorov complexity is the function $C : \{0,1\}^* \to \mathbb{N} \cup \{0\}$ such that*

$$C(w) := \min \{|p| : U(p) = w\}, \tag{11.17}$$

where $U : \{0,1\}^* \xrightarrow{o} \{0,1\}^*$ is the universal function.

The reason for using the universal function is that it can invoke any other computable function, and if that function gives a short description of the string then so does the universal function—up to an additive constant.

Although Kolmogorov complexity can be approximated in an uncontrolled way, finding a non-trivial lower bound for the Kolmogorov complexity is impossible for a concrete string. Intuitively, it is so since it requires finding the minimum over a finite set of programs which do not necessarily halt and we know already that the halting problem is undecidable. The rigorous proof is as follows.

Theorem 11.14 *The Kolmogorov complexity* $C : \{0,1\}^* \to \mathbb{N} \cup \{0\}$ *is upper semi-computable but it is not computable.*

Proof: We can perform a computable exhaustive search over all programs p such that $U(p) = w$ and $|p| \leq q$. This search halts if and only if $q \geq C(w)$. Hence there exists a partial computable function $G : \mathbb{N} \cup \{0\} \times \{0,1\}^* \to \{1, \perp\}$ such that $G(\langle w, q \rangle) = 1$ if and only if $q \geq C(w)$. Thus set

$$U_C = \{\langle w, q \rangle : q \geq C(w),\, q \in \mathbb{N} \cup \{0\}\,, w \in \{0,1\}^*\} \tag{11.18}$$

is computably enumerable, which proves that the Kolmogorov complexity is upper semi-computable.

Uncomputability of the Kolmogorov complexity will be shown by contradiction, considering it as a function of natural numbers. Assume that function $C : \mathbb{N} \to \mathbb{N} \cup \{0\}$ is computable. Then we may construct a total computable function

$$G(l) := \min \{n \in \mathbb{N} : C(n) \geq l\}\,. \tag{11.19}$$

Since function G is computable, there is a string $u \in \{0,1\}^*$ such that $U(u\,\mathrm{mil}(l)) = G(l)$ for all $l \in \mathbb{N}$. Hence for $n = G(l)$ we obtain

$$l \leq C(n) \leq |u\,\mathrm{mil}(l)| = |u| + \lfloor \log l \rfloor\,. \tag{11.20}$$

This inequality is supposed to hold for all $l \in \mathbb{N}$ but this is impossible since $l > |u| + \lfloor \log l \rfloor$ for sufficiently large l. Hence the Kolmogorov complexity is not computable. □

The above proof can be adapted to state an even more negative result by Gregory Chaitin, which is a version of the famous incompleteness theorem by Kurt Gödel. Namely, suppose that we have an inference system I consisting of a finite number axioms and inference rules. Suppose, moreover, that inference system I is sound, i.e., it allows only to prove only true theorems. Then it can be shown that there is a number M such that property $C(w) \geq M$ cannot be proved for any concrete string w in the inference system I.

***

Recapitulating this chapter, we have presented basic results from the theory of computation. The introduced concept of Kolmogorov complexity is the base of an algebra of algorithmic information measures to be developed in Chapter 12 that resembles the Shannon information measures developed in Chapter 3. In contrast to the Shannon measures, which pertain to random variables, these algorithmic measures pertain to individual binary strings.

Further reading

The theory of computation historically branched off mathematical logic. Kurt Gödel proved the first incompleteness theorems, which showed that every sufficiently rich and sound inference system contains true statements which cannot be proved in this system [57]. This result inspired Alan Turing to investigate universal computers and the halting problem and to lay foundations of theoretical computer science [123]. Both results by Kurt Gödel and Alan Turing applied the diagonal argument by Georg Cantor, originally used to prove that real numbers cannot be enumerated [17]. After the seminal works of Claude Shannon [114, 115], Ray Solomonoff [117] and Andrey Kolmogorov [84] independently proposed to measure the information content of a string by the length of the shortest program to generate it. Algorithmic information theory was developed since then and, in particular, Gregory Chaitin showed that incompressibility of strings cannot be proved in general [19]. His proof formalizes Berry's paradox from mathematical logic, whereas Gödel's proof of the incompleteness theorem formalizes the liar's paradox. The most popular textbook in algorithmic information theory is by Ming Li and Paul Vitányi [90]. Basic topics thereof are also present in the book by Thomas Cover and Joy Thomas [26].

Thinking exercises

1. *Post correspondence problem:* Let $\mathbb{X}$ be a finite alphabet that contains at least two symbols. Consider two n-element lists of strings $\alpha_1, ..., \alpha_n \in \mathbb{X}^*$ and $\beta_1, ..., \beta_n \in \mathbb{X}^*$, where n is arbitrary. This pair of lists is called a domino. A solution for the domino is a finite sequence of indices $(i_l)_{1 \leq l \leq k}$, where $k \geq 1$ and $1 \leq i_l \leq n$, such that

$$\alpha_{i_1} ... \alpha_{i_k} = \beta_{i_1} ... \beta_{i_k}. \tag{11.21}$$

   The decision function is the total function that for a given domino returns 1 if a solution exists and returns 0 if there is no solution. Show that this decision function is not computable. (Similar dominoes or jigsaw puzzles are often used to prove uncomputability of various decision functions.)

2. *Computably enumerable sets:* Show that a set $A \subset \mathbb{N}$ is computably enumerable if and only if there exists such a computable total function $G : \mathbb{N} \to A$ that for each $a \in A$ there exists such an $n \in \mathbb{N}$ that $G(n) = a$.

3. *Invariance theorem:* We can generalize the concept of the universal function in the following way: We will call an arbitrary partial computable function $G : \{0,1\}^* \stackrel{o}{\to} \{0,1\}^*$ also universal if for any program π there exists a string p_π such that

$$G(p_\pi w) = F_\pi(w). \tag{11.22}$$

Subsequently, we can generalize Kolmogorov complexity as

$$C_G(w) := \min\{|p| : G(p) = w\}. \tag{11.23}$$

Show that the generalized Kolmogorov complexity C_G is not computable, either. Show also that for each universal G there exists a constant $c < \infty$ such that for all strings $w \in \{0,1\}^*$ we have

$$|C_G(w) - C(w)| \leq c. \tag{11.24}$$

Thus it does not matter much which universal function G we use.

4. *Incompleteness theorem:* Without reading Chaitin's paper [19], try to figure out the proof of Chaitin's incompleteness theorem. Namely, for any formal inference system I that allows only to prove only true theorems, there is a number M such that property $C(w) \geq M$ cannot be proved for any concrete string w in the inference system I.

   Hint: Given an inference system, the set of its proofs is computably enumerable.

5. *Continuity of Kolmogorov complexity:* Show that for natural numbers n, m we have $|C(n+m) - C(n)| \leq 2\log m + c$.

6. Let a binary string w satisfy $C(w) \geq n - c$, where $n = |w|$. Show that $C(u), C(v) \geq n/2 - c$ for $w = uv$ and $|u| = |v|$.

Chapter 12

Complexity

Information as a password. Kolmogorov complexity as the length of a non-singular code. Coding bound. Incompressible strings. Counting bound. Oscillations of Kolmogorov complexity. Probabilistic bound. High complexity infinite sequences. Conditional Kolmogorov complexity. Bounds for conditional complexity. Chain rule for conditional complexity. Algorithmic mutual information. Bounds for algorithmic information. Chain rule for algorithmic information. Data-processing inequality.

As we already know from Chapter 11, Kolmogorov complexity is the length of the shortest program which outputs a given string or another discrete object. This function is an appealing absolute measure of information contained in a discrete object but it is uncomputable. However, Kolmogorov complexity can be computably approximated in an uncontrolled way and, as we will see in this chapter, it sets out a lower bound for all practical definitions of the amount of information.

Having adopted such a definition, we may suppose that the amount of information is equal to the amount of irreducible novelty that a given object conveys. In fact, if some part of the shortest program can be inferred from other parts of the shortest program then we may omit this redundant part. Thus the definition of information as the length of the shortest program equates information with unpredictability. Unpredictability is closely related to randomness. Should information and randomness be the same?

Equating information with randomness may seem counterintuitive when we ask about the information content of a seminal work of a human mind. Here we should distinguish between useless and useful information. Both

are random, but to avoid pessimistic connotations, we propose that useful information should be rather imagined as a kind of a *password*, i.e., the unique string of random bits that unlocks the sesame. Not every random string is the right password but a good password should be a random string. In fact, there are awfully many random strings which are not the right password. Moreover, there should be no simple way of checking that a given random string is the right password other than feeding it to the sesame. Such conceptualization seems to agree with our intuitions.

In this chapter, we will make a more systematic exposition of the algorithmic information theory, which revolves around the concept of Kolmogorov complexity. First, we will link the Kolmogorov complexity with non-singular codes discussed in Chapter 1. We observe that Kolmogorov complexity is the length of a non-singular binary code that cannot be computed but can be effectively decoded.

Example 12.1 (Kolmogorov code) *Let notation w^* denote the first shortest program such that $U(w^*) = w$. The length of string w^* is the Kolmogorov complexity of w, $|w^*| = C(w)$. Code $B : \{0,1\}^* \ni w \mapsto w^* \in \{0,1\}^*$ is non-singular but not computable since its length is not computable. By contrast, the inverse function $B^{-1} : \{0,1\}^* \ni w^* \mapsto w \in \{0,1\}^*$ is obviously computable since to compute w it suffices to execute program w^*.*

In general, the length of any binary code that can be effectively decoded provides an upper bound for the Kolmogorov complexity, up to a constant corresponding to the complexity of the inverse code.

Theorem 12.2 (coding bound) *Let $B : \{0,1\}^* \to A \subset \{0,1\}^*$ be a non-singular code. Suppose that the inverse function $B^{-1} : A \ni B(w) \mapsto w \in \{0,1\}^*$ is computable. Then there exist a constant $c < \infty$ such that for all strings $w \in \{0,1\}^*$ the Kolmogorov complexity $C(w)$ is bounded as*

$$C(w) \leq |B(w)| + c. \tag{12.1}$$

Proof: By computability of B^{-1}, there is a string u such that $U(uB(w)) = w$ for all strings w, where U is the universal function introduced in Chapter 11. Plugging this into the definition of the Kolmogorov complexity we obtain $C(w) \leq |uB(w)| \leq |B(w)| + |u|$. □

In particular, the length of a string itself sets an upper bound for its Kolmogorov complexity. For this aim, it suffices to consider the trivial code $w \mapsto w$.

Theorem 12.3 (length bound) *There is a constant $c < \infty$ such that for all strings $w \in \{0,1\}^*$ we have*

$$C(w) \leq |w| + c. \tag{12.2}$$

Proof: The identity function $w \mapsto w$ is computable. Hence there is a string u such that $U(uw) = w$ for all strings w. Plugging this into the definition of Kolmogorov complexity, we obtain $C(w) \leq |uw| \leq |w| + |u|$. □

Now let us proceed to the phenomenon of incompressibility, which is converse to the length bound.

Definition 12.4 (incompressible string) *A binary string w is called c-incompressible if*

$$C(w) \geq |w| - c. \tag{12.3}$$

Thus, a string w is incompressible when the shortest program to compute is close to command "print w".

As we remarked in Chapter 11, we cannot prove incompressibility of almost any concrete string. However, we can easily show that there must be quite many incompressible strings since there are not so many short programs compared to the number of all strings.

Theorem 12.5 (counting bound) *There exist at least $2^n - 2^{n-c} + 1$ distinct c-incompressible binary strings of length n.*

Proof: The number of c-incompressible strings of length n is greater than the difference between the number of all strings of length n and the number of distinct programs of length strictly smaller than $n - c$. There exists 2^n distinct strings of length n and there exist $1 + 2 + ... + 2^{n-c-1} = 2^{n-c} - 1$ distinct programs of length strictly smaller than $n - c$. Hence the claim follows. □

In particular, the majority of strings of a given length are 1-incompressible. What may be surprising, though, there are no infinite sequences whose all prefixes are c-incompressible for any fixed c. This phenomenon is known as oscillations of Kolmogorov complexity. Let $x_j^k := x_j x_{j+1} ... x_k$ denote a substring of an infinite sequence $(x_i)_{i \in \mathbb{N}}$.

Theorem 12.6 (oscillations of complexity) *Consider an arbitrary infinite binary sequence $(x_i)_{i \in \mathbb{N}}$. There exists a constant $c < \infty$ such that for infinitely many $n \in \mathbb{N}$ we have*

$$C(x_1^n) \leq n - \log n + c. \tag{12.4}$$

Proof: Consider an arbitrary number $m \in \mathbb{N}$. Let x_1^m be the military code for number k. Then we have $C(x_1^{m+k}) \leq C(x_{m+1}^{m+k}) + c \leq k + 2c$ because we may compute prefix x_1^m given the length of x_{m+1}^{m+k}. Put $n = m + k$. We have

$$\begin{aligned} m = \lfloor \log k \rfloor &= \lfloor \log(n-m) \rfloor = \lfloor \log(n - \lfloor \log(n-m) \rfloor) \rfloor \\ &\geq \log(n - \log n) - 1 = \log n - \log\left(1 - \frac{\log n}{n}\right) - 1 \\ &\geq \log n - \log\left(1 - \frac{\log 3}{3}\right) - 1. \end{aligned} \tag{12.5}$$

Consequently, $C(x_1^n) \leq k + 2c \leq n - \log n + 2c + \log\left(1 - \frac{\log 3}{3}\right)$. □

For further considerations, it is convenient to introduce this notation for functions of strings, which ignores ugly error terms.

Definition 12.7 (rough inequality and equality) *For a real function g of strings $u_i \in \{0,1\}^*$, we write $g(u_1, ..., u_n) \lesssim 0$ if there exists a $c < \infty$ such that for all $u_i \in \{0,1\}^*$ we have*

$$g(u_1, ..., u_n) \leq c \max\left\{1, \log|u_1|, ..., \log|u_n|\right\}. \tag{12.6}$$

Further, we write $g(u_1, ..., u_n) \lesssim f(u_1, ..., u_n)$ if and only if $g(u_1, ..., u_n) - f(u_1, ..., u_n) \lesssim 0$ and we write $g(u_1, ..., u_n) \approx f(u_1, ..., u_n)$ if and only if $|g(u_1, ..., u_n) - f(u_1, ..., u_n)| \lesssim 0$.

Having this notation, we can easily state that the Kolmogorov is lower bounded by the pointwise entropy for any stochastic process almost surely.

Theorem 12.8 (probabilistic bound) *Consider an arbitrary stochastic process $(X_i)_{i \in \mathbb{N}}$ over a finite alphabet $\{0,1\}^*$. We have*

$$\mathbf{E}\, C(X_1^n) \gtrsim \mathbf{E}\left[-\log P(X_1^n)\right] = H(X_1^n), \tag{12.7}$$

$$C(X_1^n) \gtrsim -\log P(X_1^n) \text{ a.s.} \tag{12.8}$$

Proof: The Kolmogorov complexity is the length of a non-singular code $w \mapsto w^*$ for strings of an arbitrary length, where $|w^*| = C(w)$. Hence there exists a prefix-free code $B : \{0,1\}^* \to \{0,1\}^*$ such that $|B(w)| = \mathrm{una}''(C(w)+1)w^*$. Since $C(x_1^n) \leq n + c$ then

$$|B(x_1^n)| \leq C(x_1^n) + 2\log n + c \tag{12.9}$$

for a $c < \infty$. Now, the source coding inequality and the Barron lemma state that

$$\mathbf{E}\left[|B(X_1^n)| + \log P(X_1^n)\right] \geq 0, \tag{12.10}$$

$$\lim_{n \to \infty} \left[|B(X_1^n)| + \log P(X_1^n)\right] = \infty \text{ a.s.} \tag{12.11}$$

Hence, considering the displayed formulas, we derive the claim. □

The above theorem inspires the following concept of a high complexity infinite binary sequence.

Theorem 12.9 (high complexity sequences) *There exist infinite binary sequences $(x_i)_{i\in\mathbb{N}}$ such that $C(x_1^n) \approx n$. In fact, $C(X_1^n) \approx n$ holds almost surely if $(X_i)_{i\in\mathbb{N}}$ is the binary process with a uniform measure.*

Proof: We have the uniform upper bound $C(x_1^n) \leq n + c$. By the previous theorem for the uniform probability measure $P(X_1^n = x_1^n) = 2^{-n}$, where $x_i \in \{0, 1\}$, we obtain

$$C(X_1^n) \gtrsim -\log P(X_1^n) = n \text{ a.s.} \tag{12.12}$$

Hence $C(X_1^n) \approx n$ holds almost surely. □

High complexity sequences inspire theory of algorithmic randomness, which is being intensely developed in recent years but dates back to the work of Per Martin-Löf. The guiding idea is that an infinite sequence is *not* algorithmically random when there are sufficiently many regular patterns, like 010101010101, which can be used to compress its description significantly. Using Kolmogorov complexity is an ingredient contributing to the success of this project. High complexity sequences are some simple-minded approximations of rigorously defined algorithmically random sequences.

Subsequently, we will introduce the conditional Kolmogorov complexity and will we demonstrate an approximate chain rule for that quantity which resembles the chain rule for conditional entropy. The conditional Kolmogorov complexity is the length of the shortest program that produces a given string when another string is given for free.

Definition 12.10 (conditional complexity) *The conditional Kolmogorov complexity is the function $C : \{0, 1\}^* \times \{0, 1\}^* \to \mathbb{N} \cup \{0\}$ such that*

$$C(w|u) := \min \{|p| : U(\langle u, p\rangle) = w\}, \tag{12.13}$$

where $U : \{0, 1\}^* \xrightarrow{o} \{0, 1\}^*$ is the universal function.

The conditional Kolmogorov complexity of w given u is essentially less than the unconditional complexity of w. The conditional complexity of w is also negligible if we condition on w.

Theorem 12.11 *For strings $u, w \in \{0, 1\}^*$, we have*

$$C(w|u) \lesssim C(w), \tag{12.14}$$

$$C(w|w) \approx 0. \tag{12.15}$$

Proof: There is a program for computing w given u which ignores u and applies the shortest program for w as a subroutine. Hence follows inequality $C(w|u) \leq C(w) + c$. There is also a program which computes w given w by copying the input to the output. Hence follows inequality $C(w|w) \leq c$. $\square$

We can also prove the following chain rule: The complexity of a pair (u, w) equals approximately the complexity of u plus the complexity of w given u. This is relationship is analogous to the chain rule $H(X)+H(Y|X) = H(X, Y)$ but only holds as the rough equality.

Theorem 12.12 (chain rule) *For strings $u, w \in \{0,1\}^*$, we have*

$$C(u) + C(w|u) \approx C(\langle u, w \rangle). \tag{12.16}$$

Proof: In the following c denotes an arbitrary constant whose value may change between equations but it does not depend on the considered strings. Let p be the shortest program such that $U(p) = u$ and let q be the shortest program such that $U(\langle u, q \rangle) = w$. Function $\langle p, q \rangle \mapsto \langle u, w \rangle$ is computable. Hence

$$\begin{aligned} C(\langle u, w \rangle) &\leq |\langle p, q \rangle| + c \\ &\leq 2 \log |p| + C(u) + C(w|u) + c \\ &\leq C(u) + C(w|u) + c \log \max \{|u|, |w|\}. \end{aligned} \tag{12.17}$$

The proof of the converse bound is more difficult. In the following, we apply this lemma: If set A is computably enumerable given an object w then the Kolmogorov complexity of the element $x \in A$ which is enumerated as the r-th one is bounded by $C(x|w) \leq \log r + c$.

To begin the proper proof, let us construct a $2^{|u|} \times 2^{|w|}$ matrix with rows indexed by strings $x \in \{0,1\}^{|u|}$ and columns indexed by strings $y \in \{0,1\}^{|w|}$. Let $s = C(\langle u, w \rangle)$. We define that cell (x, y) of the matrix contains 1 if $C(\langle x, y \rangle) \leq s$ and it contains 0 otherwise. The number of cells in the matrix that contain 1's is not greater than $2^{s+1} \geq 1 + 2 + ... + 2^s$, being the number of programs of length $\leq s$.

Let t be such that the number of 1's in the row indexed by u lies in interval $(2^{t-1}, 2^t]$. Given u, $|w|$, and s, the set of columns y such that cell (u, y) contains 1 is computably enumerable. Since one of these columns is column w, say the q-th one where $1 \leq q \leq 2^t$, then we obtain the bound

$$\begin{aligned} C(w|u) &\leq |\langle |w|, s, q \rangle| + c \\ &\leq 2(\log |w| + \log s) + t + c \\ &\leq t + c \log \max \{|u|, |w|\}, \end{aligned} \tag{12.18}$$

since $s \leq 2\,|u| + |w| + c$.

Now consider the set of rows that contain at least 2^{t-1} 1's. The number of such rows is upper bounded by $2^{s+1}/2^{t-1} = 2^{s-t+2}$. Given $|u|$, $|w|$, t and s, the set of these rows is computably enumerable. Since one of these rows is row u, say the p-th one where $1 \leq p \leq 2^{s-t+2}$, then we obtain the bound

$$\begin{aligned} C(u) &\leq |\langle |u|\,, |w|\,, t, s, p\rangle| + c \\ &\leq 2(\log |u| + \log |w| + \log t + \log s) + s - t + c \\ &\leq s - t + c \log \max\{|u|\,, |w|\}\,, \end{aligned} \tag{12.19}$$

since $t \leq |w|$.

Adding the previous two bounds, we obtain

$$\begin{aligned} C(u) + C(w|u) &\leq s + c \log \max\{|u|\,, |w|\} \\ &= C(\langle u, w\rangle) + c \log \max\{|u|\,, |w|\}\,. \end{aligned} \tag{12.20}$$

Combining this with (12.17), we obtain the claim. □

Let us play more with the differences of Kolmogorov complexity. An important concept in the algorithmic information theory is the algorithmic mutual information. It is defined as follows.

Definition 12.13 (algorithmic information) *The algorithmic mutual information is the function $J : \{0,1\}^* \times \{0,1\}^* \to \mathbb{Z}$ defined as*

$$J(u; w) := C(u) + C(w) - C(\langle u, w\rangle). \tag{12.21}$$

Analogously to the bounds for conditional complexity, we can state some bounds for algorithmic mutual information. First, the mutual information is essentially non-negative. Second, it is essentially symmetric. Third, the mutual information between a string and its copy essentially equals its Kolmogorov complexity.

Theorem 12.14 *For strings $u, w \in \{0,1\}^*$, we have*

$$J(u; w) \gtrsim 0, \tag{12.22}$$

$$J(u; w) \approx J(w; u), \tag{12.23}$$

$$J(w; w) \approx C(w). \tag{12.24}$$

Proof: We have

$$C(\langle u, w\rangle) \approx C(u) + C(w|u) \lesssim C(u) + C(w). \tag{12.25}$$

Hence the first claim follows. Next, function $\langle u, w\rangle \mapsto \langle w, u\rangle$ is computable, so

$$|C(\langle u, w\rangle) - C(\langle u, w\rangle)| \leq c. \tag{12.26}$$

Hence the second claim follows. Next, function $w \mapsto |w, w|$ and its inverse are both computable, so

$$|C(\langle w, w\rangle) - C(w)| \leq c. \tag{12.27}$$

Hence the third claim follows. □

We also have a chain rule for algorithmic mutual information, which is a simple consequence of the chain rule for conditional complexity.

Theorem 12.15 (chain rule) *For strings $u, w \in \{0,1\}^*$, we have*

$$C(u) - C(u|w) \approx J(u; w). \tag{12.28}$$

Proof: We have

$$C(u) - C(u|w) - J(u; w) = C(w) + C(u|w) - C(\langle u, w\rangle) \approx 0 \tag{12.29}$$

by the chain rule for conditional complexity. □

We can also ask how conditional complexity and algorithmic mutual information behave when we plug in a computable function instead of one of their original arguments. It turns out that they are essentially monotone.

Theorem 12.16 (data-processing inequality) *For a total computable function $G : \{0,1\}^* \to \{0,1\}^*$ and for strings $u, w \in \{0,1\}^*$, we have*

$$C(G(w)|u) \lesssim C(w|u), \tag{12.30}$$
$$C(w|G(u)) \gtrsim C(w|u), \tag{12.31}$$
$$J(u; G(w)) \lesssim J(u; w). \tag{12.32}$$

Proof: There is a program for computing $G(w)$ given u which applies the shortest program for w given u as a subroutine and then applies function G. Hence follows inequality $C(G(w)|u) \leq C(w|u) + c$, where c depends on G.

There is also a program for computing w given u which computes $G(u)$ and then applies the shortest program for w given $G(u)$ as a subroutine. Hence follows inequality $C(w|u) \leq C(w|G(u)) + c$.

The third claim follows by calculation

$$J(u; G(w)) \lesssim C(u) - C(u|G(w)) \lesssim C(u) - C(u|w) \approx J(u; w). \tag{12.33}$$

□

In general, the problems of encoding ordered tuples are the source of some error terms in the algebraic properties of the Kolmogorov complexity discussed in this chapter. These error terms can be significantly reduced if we define the Kolmogorov complexity as the length of the shortest program for a universal function which halts only on a prefix-free set of programs. That version of Kolmogorov complexity is known as the prefix-free complexity, whereas the Kolmogorov complexity discussed in Chapter 11 and in this one is called the plain complexity. In general, the prefix-free Kolmogorov complexity satisfies somewhat neater algebraic identities than the plain complexity but its theory is more difficult.

***

To recapitulate this chapter, we have developed the algebra of algorithmic information measures that is analogous to the algebra of Shannon information measures. In Chapters 13 and 14, we will see an application of these calculi to study sublinear effects in universal coding and yield some insights into statistical modeling of natural language.

Further reading

Algebraic properties of the plain Kolmogorov complexity were already studied by Andrey Kolmogorov and described by Alexander Zvonkin and Leonid Levin in the survey paper [132]. The simple proof of the chain rule for the plain Kolmogorov complexity exhibited in this chapter is taken from the work of Marius Zimand [128]. The constructions developed in this chapter have much more elegant counterparts in the theory of prefix-free Kolmogorov complexity, exposed neatly by Gregory Chaitin [18]. However, the chain rules for the prefix-free Kolmogorov complexity were independently derived earlier by Peter Gács [53] and later by Gregory Chaitin [18]. The idea of high complexity infinite sequences was developed by Per Martin-Löf as algorithmically random sequences [94]. Martin-Löf's algorithmically random sequences have a simple characterization in terms of the prefix-free complexity as shown by Claus-Peter Schnorr (unpublished). The relevant book sources are the textbook by Ming Li and Paul Vitányi [90] and the monograph by Rodney Downey and Denis Hirschfeldt [43].

Thinking exercises

1. *Minimal programs are random:* A binary string p is called a minimal program if $U(p) = w$ and $|p| = C(w)$ for some binary string w.

Show that there is a constant c such that all minimal programs are c-incompressible.

2. *Normalized information distance:* The normalized information distance between strings u and w is defined as

$$d(u,w) := \frac{\max\{C(u|w), C(w|u)\}}{\max\{C(u), C(w)\}}. \tag{12.34}$$

Show that $d(u,w)$ is almost a metric, i.e., it satisfies

- $d(u,w) \approx d(w,u)$;
- $d(u,w) \lesssim d(u,v) + d(v,w)$.

3. *Coding bound:* Let $A \subset \{0,1\}^*$ be a computable subset of binary strings. Let $A_n := \{w \in A : |w| = n\}$ be its subset of strings of length n. Let $\# A$ be the number of elements in A. Show that for any string $w \in A_n$, we have

$$C(w) \lesssim C(A_n) + \log \# A_n. \tag{12.35}$$

4. *Effective strong law of large numbers:* Consider the set of strings

$$A_n(\epsilon) = \left\{ x_1^n \in \{0,1\}^n : \left| \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{2} \right| \geq \epsilon \right\}. \tag{12.36}$$

Argue by the Azuma-Hoeffding inequality (Theorem 9.23) that

$$\log \# A_n(\epsilon) \leq n + 1 - \frac{2n\epsilon}{\ln 2}. \tag{12.37}$$

Consequently, prove that any high complexity binary sequence $(x_i)_{i\in\mathbb{N}}$ such that $C(x_1^n) \approx n$ satisfies the effective strong law of large numbers

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{2}. \tag{12.38}$$

5. *Effective Birkhoff ergodic theorem:* A binary sequence $(x_i)_{i\in\mathbb{N}}$ is called algorithmically random with respect to a prequential distribution $Q : \{0,1\}^* \to [0,1]$ if

$$C(x_1^n) \approx -\log Q(x_1^n). \tag{12.39}$$

Show that the set of sequences that are algorithmically random with respect to Q has probability 1 with respect to Q. Applying the ideas from the previous two exercises and the Ivanov downcrossing inequality (Theorem 9.30), demonstrate that if distribution Q is both computable as a function and stationary ergodic then a sequence $(x_i)_{i\in\mathbb{N}}$ that is algorithmically random with respect to Q satisfies the effective Birkhoff ergodic theorem

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i = Q(x_i). \tag{12.40}$$

See also [7, 52].

6. *Halting probability:* For a program $p \in \{0,1\}^*$, denote the discounted length $l(p) := |p| + 2\lfloor \log(|p|+1) \rfloor + 2$. Define the halting probability

$$\Omega := \sum_{p\in\{0,1\}^* : U(p)\downarrow} 2^{-l(p)}. \tag{12.41}$$

Obviously, $\Omega \in [0,1]$ by the Kraft inequality. The bits of the halting probability are digits $\Omega_k \in \{0,1\}$ such that

$$\Omega = \sum_{k=1}^{\infty} \Omega_k 2^{-k}. \tag{12.42}$$

Show that:

(a) There is a partial computable function that, given the first n bits $\Omega_1^n := (\Omega_1, ..., \Omega_n)$ and a program p such that $l(p) \leq n$ as an input, computes whether universal function U halts on p. (Ω is an oracle.)

(b) $C(\Omega_1^n) \approx n$. (Ω is incompressible.)

(c) Ω is an irrational number in range $(0,1)$. (Ω is a probability.)

These results mean in particular that mathematical knowledge is highly compressible: The list of all mathematical theorems—of a certain form—that can be stated using roughly less than n bits can be compressed to at most n independent binary facts (kind of axioms?).

Chapter 13

Excess

Hilberg exponent. Excess bound. Hilberg exponents for mutual information. Markov order. Markov order estimator and its consistency. Bounds for mutual information applying penalized maximum likelihood. Bounds for mutual information applying grammar-based codes.

In Chapter 10, we discussed asymptotic equipartition and universal distributions for stationary ergodic processes. In this chapter, we will sharpen these results in another direction. We will be interested in sublinear effects in universal coding, namely, the rates of convergence of the encoding rate to the entropy rate. For this goal, we will introduce some measures of this convergence such as excess entropy and Hilberg exponents. It will turn out that the lengths of universal codes can converge to the entropy limit at a somewhat different speed than the pointwise entropy. In particular, we will apply the notion of the Kolmogorov complexity discussed in Chapters 11 and 12. This quantity sets a lower bound for achievable redundancy rates of any efficiently computable code. We will also discuss the problem of consistent estimation of the Markov order and we will revisit the minimal grammar-based code.

Excess bounds

To begin, let us consider a stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet $\{1, 2, ..., m\}$. Denote the entropy rate $h := \mathbf{E}\left[-\log P(X_i|X_{-\infty}^{i-1})\right]$ and the coding risk $\ell(Q, x_1^n) := -\log Q(x_1^n)$. We showed in Chapter 10 that

$$\lim_{n\to\infty} \frac{\mathbf{E}\,\ell(Q, X_1^n)}{n} = h, \tag{13.1}$$

$$\lim_{n\to\infty} \frac{\ell(Q, X_1^n)}{n} = h \text{ a.s.}, \tag{13.2}$$

where we may take $Q = P$ or any universal distribution such as $Q = \mathbb{P}$ or $Q = R$ introduced in Chapter 8 in particular. Since these universal distributions have computable lengths of their Shannon-Fano codes, then by the coding bound and the probabilistic bound for the Kolmogorov complexity from Chapter 12, we may also take $Q = C$ in equations (13.1)–(13.2) when we formally define the Kolmogorov coding risk $\ell(C, x_1^n) := C(x_1^n)$.

Denote the expected pointwise entropy or Kolmogorov complexity as

$$\ell(Q, n) := \mathbf{E}\, \ell(Q, X_1^n) = \mathbf{E}\left[-\log Q(X_1^n)\right], \tag{13.3}$$

$$\ell(C, n) := \mathbf{E}\, \ell(C, X_1^n) = \mathbf{E}\, C(X_1^n). \tag{13.4}$$

We notice that $\ell(\mathbb{P}, n) \gtrsim \ell(C, n) \gtrsim \ell(P, n) \geq nh$. We may ask how fast ratio $\ell(Q, X_1^n)/n$ converges to the limit h depending whether $Q = \mathbb{P}$, $Q = C$, or $Q = P$. To quantify the rate of convergence $\ell(Q, n)/n \to h$, we may consider the growth of deviation $\ell(Q, n) - nh$. We observe the following fact.

Theorem 13.1 (excess bound) *Denote* $\nabla S(n) := 2S(n) - S(2n)$ *for a real function* $S : \mathbb{N} \to \mathbb{R}$. *If* $\lim_{n\to\infty} S(n)/n = s$ *for an* $s \in \mathbb{R}$ *then*

$$\limsup_{n\to\infty} (S(n) - ns) \leq \limsup_{n\to\infty} \nabla S(n), \tag{13.5}$$

$$\liminf_{n\to\infty} (S(n) - ns) \geq \liminf_{n\to\infty} \nabla S(n). \tag{13.6}$$

Proof: We have the finite telescope sum

$$\sum_{k=0}^{K-1} \frac{\nabla S(2^k n)}{2^{k+1}} = S(n) - n \cdot \frac{S(2^K n)}{2^K n}. \tag{13.7}$$

For K tending to infinity, the above identity implies the infinite telescope sum

$$\sum_{k=0}^{\infty} \frac{\nabla S(2^k n)}{2^{k+1}} = S(n) - ns. \tag{13.8}$$

Write $J^+ = \limsup_{n\to\infty} \nabla S(n)$ and $J^- = \liminf_{n\to\infty} \nabla S(n)$. Without loss of generality, we may assume $-\infty < J^+, J^- < \infty$. Observe that $J^- - \epsilon \leq \nabla S(n) \leq J^+ + \epsilon$ for all but finitely many n for any $\epsilon > 0$. Hence by the telescope sum (13.8) we obtain for sufficiently large n that

$$J^- - \epsilon = \sum_{k=0}^{\infty} \frac{J^+ + \epsilon}{2^{k+1}} \leq S(n) - ns \leq \sum_{k=0}^{\infty} \frac{J^+ + \epsilon}{2^{k+1}} = J^+ + \epsilon. \tag{13.9}$$

Hence the claims follow. □

By virtue of the demonstrated excess bounds, the rate of convergence $\ell(Q,n)/n \to h$ is linked to the rate of pointwise mutual information

$$J_Q(u;v) := \log \frac{Q(uv)}{Q(u)Q(v)} \tag{13.10}$$

and the algorithmic mutual information

$$J_C(u;v) := C(u) + C(v) - C(uv). \tag{13.11}$$

Namely by stationarity,

$$\nabla\ell(Q,n) = \mathbf{E}\, J_Q(X_{-n+1}^0; X_1^n), \tag{13.12}$$
$$\nabla\ell(C,n) = \mathbf{E}\, J_C(X_{-n+1}^0; X_1^n). \tag{13.13}$$

Notice that for $Q = P$ we have the Shannon entropy

$$H(X_1^n) = \ell(P,n) = \mathbf{E}\left[-\log P(X_1^n)\right] \tag{13.14}$$

and the Shannon mutual information

$$I(X_{-n+1}^0; X_1^n) = \nabla\ell(P,n) = \mathbf{E}\, J_P(X_{-n+1}^0; X_1^n). \tag{13.15}$$

Shannon mutual information $I(X_{-n+1}^0; X_1^n)$ is a non-decreasing function of n by the data-processing inequality. Hence, limit $\lim_{n\to\infty} I(X_{-n+1}^0; X_1^n)$ exists and by the excess bound, we may define excess entropy

$$E := \lim_{n\to\infty} (H(X_1^n) - nh) = \lim_{n\to\infty} I(X_{-n+1}^0; X_1^n). \tag{13.16}$$

However limit $\lim_{n\to\infty} \mathbf{E}\, J_Q(X_{-n+1}^0; X_1^n)$ need not exist for $Q \neq P$.

To tackle with this problem, let us introduce the following concept.

Definition 13.2 (Hilberg exponent) *The Hilberg exponent of a real function $S : \mathbb{N} \to \mathbb{R}$ is defined as*

$$\operatorname*{hilb}_{n\to\infty} S(n) := \limsup_{n\to\infty} \max\left\{\frac{\log S(n)}{\log n}, 0\right\}. \tag{13.17}$$

The Hilberg exponent captures the asymptotic power-law growth of the respective function, for example

$$\operatorname*{hilb}_{n\to\infty} n^\beta = \beta \text{ for } \beta \geq 0. \tag{13.18}$$

The decay of the respective function is ignored.

We have an analogical excess bound for Hilberg exponents.

Theorem 13.3 (excess bound) *Denote* $\nabla S(n) := 2S(n) - S(2n)$ *for a real function* $S : \mathbb{N} \to \mathbb{R}$. *If* $\lim_{n\to\infty} S(n)/n = s$ *for an* $s \in \mathbb{R}$ *then*

$$\operatorname*{hilb}_{n\to\infty} (S(n) - ns) \le \operatorname*{hilb}_{n\to\infty} \nabla S(n) \tag{13.19}$$

with the equality if $S(n) \ge ns$.

Proof: Write $\beta = \operatorname{hilb}_{n\to\infty} \nabla S(n)$. Since the left hand side of (13.19) side is not greater than 1, it is enough to prove this inequality for $\beta < 1$. Observe that $\nabla S(n) \le n^{\beta+\epsilon}$ for all but finitely many n for any $\epsilon > 0$. Then for $\epsilon < 1 - \beta$, by the telescope sum (13.8) we obtain for sufficiently large n that

$$S(n) - ns \le \sum_{k=0}^{\infty} \frac{n^{\beta+\epsilon} 2^{k(\beta+\epsilon)}}{2^{k+1}} = n^{\beta+\epsilon} \sum_{k=0}^{\infty} 2^{(\beta+\epsilon-1)k-1} = \frac{n^{\beta+\epsilon}}{2(1 - 2^{\beta+\epsilon-1})}. \tag{13.20}$$

Since ϵ can be taken arbitrarily small, we obtain (13.19).

Now assume that $S(n) \ge ns$. Then we have

$$S(n) - ns = \frac{\nabla S(n)}{2} + \frac{S(2n) - 2ns}{2} \ge \frac{\nabla S(n)}{2}. \tag{13.21}$$

Hence $\beta \le \operatorname{hilb}_{n\to\infty} (S(n) - ns)$. Thus we obtain the equality in (13.19). □

Consequently, we may define these Hilberg exponents

$$\beta_P := \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\,[-\log P(X_1^n)] - nh) = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J_P(X_{-n+1}^0; X_1^n), \tag{13.22}$$

$$\beta_C := \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\, C(X_1^n) - nh) = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J_C(X_{-n+1}^0; X_1^n), \tag{13.23}$$

$$\beta_{\mathbb{P}} := \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\,[-\log \mathbb{P}(X_1^n)] - nh) = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J_{\mathbb{P}}(X_{-n+1}^0; X_1^n), \tag{13.24}$$

where $0 \le \beta_P \le \beta_C \le \beta_{\mathbb{P}} \le 1$ by the probabilistic and coding bounds on Kolmogorov complexity (Theorems 12.2 and 12.8).

Markov order estimation

Excess entropy E can be infinite, whereas exponents β_P, β_C, and $\beta_{\mathbb{P}}$ may differ from zero. However, in the following, we will show that for higher order Markov processes over a finite alphabet and for the universal PML maximum distribution $\mathbb{P}$ introduced in Definition 8.8, we have

$$E < \infty \text{ and } \beta_P = \beta_C = \beta_{\mathbb{P}} = 0. \tag{13.25}$$

As we will see, this topic is intimately connected with consistent estimation of the Markov order of a process.

Let us define the Markov order of a stationary process formally.

Definition 13.4 (Markov order) *Let $(X_i)_{i\in\mathbb{N}}$ be a stationary ergodic process over a finite alphabet $\mathbb{X}$. The Markov order of the process is defined as*

$$M := \inf\{\infty\} \cup \left\{k \geq 0 : H(X_i|X_{i-k}^{i-1}) = h\right\}. \tag{13.26}$$

According to the above, IID processes are 0-th order Markov processes.

An obvious application of the data-processing inequality is as follows.

Theorem 13.5 *For any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet $\mathbb{X}$, we have inequality*

$$E = \sup_{n\in\mathbb{N}} I(X_{-n+1}^0; X_1^n) \leq M \log m. \tag{13.27}$$

Hence $\beta_P = 0$ if $M < \infty$.

Proof: Shannon mutual information $I(X_{-n+1}^0; X_1^n)$ is a non-decreasing function of n by the data-processing inequality. In particular, if $M < n < \infty$ then by the Markov property, we can write

$$\begin{aligned} I(X_{-n+1}^0; X_1^n) &= I(X_{-M+1}^0; X_1^n) + I(X_{-n+1}^{-M}; X_1^n | X_{-M+1}^0) \\ &= I(X_{-M+1}^0; X_1^n) \leq H(X_{-M+1}^0) \leq M \log m. \end{aligned} \tag{13.28}$$

To conclude, observe that $I(X_{-n+1}^0; X_1^n) = \mathbf{E}\, J_P(X_{-n+1}^0; X_1^n)$, so $\beta_P = 0$. $\square$

The next question is whether we can estimate the Markov order of a stationary ergodic process. In order to propose an estimator, we will apply the universal PML maximum distribution introduced in Chapter 8. Let us fix a finite alphabet $\{1, 2, ..., m\}$. To recall, the maximum likelihood (ML) is denoted $\hat{\mathbb{P}}(x_1^n|k)$, the Shtarkov sum bound is

$$\log Z(n|k) := \min\left\{n \log m,\ k \log m + m^{k+1} \log(n-k+1)\right\}, \tag{13.29}$$

the penalized maximum likelihood (PML) is

$$\mathbb{P}(x_1^n|k) := \frac{\hat{\mathbb{P}}(x_1^n|k)}{Z(n|k)}, \tag{13.30}$$

and the PML maximum distribution is

$$\mathbb{P}(x_1^n) := \max_{k\geq 0} w_k \mathbb{P}(x_1^n|k), \quad w_k := \frac{1}{k+1} - \frac{1}{k+2}. \tag{13.31}$$

The candidate for an estimator of the Markov order is as follows.

Definition 13.6 (Markov order estimator) *The Markov order estimator for a finite alphabet is defined as*

$$\mathbb{M}(x_1^n) := \inf\left\{k \geq 0 : \hat{\mathbb{P}}(x_1^n|k) \geq w_n \mathbb{P}(x_1^n)\right\}. \tag{13.32}$$

We have $\mathbb{M}(x_1^n) \leq n$ since for $k \geq n$ we obtain $\hat{\mathbb{P}}(x_1^n|k) = 1 \geq w_n \mathbb{P}(x_1^n)$.

The above Markov order estimator is asymptotically unbiased and consistent indeed.

Theorem 13.7 (consistency of Markov order estimator) *For any stationary ergodic process $(X_i)_{i\in\mathbb{N}}$ over a finite alphabet, we have*

$$\lim_{n\to\infty} \mathbf{E}\, \mathbb{M}(X_1^n) = M, \tag{13.33}$$

$$\lim_{n\to\infty} \mathbb{M}(X_1^n) = M \ \ a.s. \tag{13.34}$$

Proof: Our proof is split into impossibility of overestimation, impossibility of underestimation, and asymptotic unbiasedness.

To show the impossibility of overestimation, we assume $M < \infty$ without loss of generality. The bound for the overestimation probability is received by inequality

$$\hat{\mathbb{P}}(X_1^n|M) \geq P(X_{M+1}^n|X_1^M) \geq P(X_1^n). \tag{13.35}$$

Hence applying the Barron inequality, we receive

$$P\left(\mathbb{M}(X_1^n) > M\right) = P\left(\frac{w_n \mathbb{P}(X_1^n)}{\hat{\mathbb{P}}(X_1^n|M)} > 1\right) \leq P\left(\frac{w_n \mathbb{P}(X_1^n)}{P(X_1^n)} > 1\right) \leq w_n. \tag{13.36}$$

Since $\sum_{n=1}^{\infty} w_n = 1$ then by the Borel-Cantelli lemma, we obtain

$$\limsup_{n\to\infty} \mathbb{M}(X_1^n) \leq M \text{ a.s.} \tag{13.37}$$

Now, we demonstrate the impossibility of underestimation. Observe that for any $k < M$, we have $H(X_i|X_{i-k}^{i-1}) > h$ and

$$\lim_{n\to\infty} \frac{1}{n}\left[-\log \hat{\mathbb{P}}(X_1^n|k)\right] = \lim_{n\to\infty} \mathcal{H}(X_1^n|k) = H(X_i|X_{i-k}^{i-1}) \text{ a.s.} \tag{13.38}$$

whereas by universality of the PML maximum $\mathbb{P}$, we derive

$$\lim_{n\to\infty} \frac{1}{n}\left[-\log w_n \mathbb{P}(X_1^n)\right] = h \text{ a.s.} \tag{13.39}$$

Hence $\mathbb{M}(X_1^n)$ exceeds any $k < M$ for sufficiently large n almost surely. In other words,

$$\liminf_{n\to\infty} \mathbb{M}(X_1^n) \geq M \text{ a.s.} \tag{13.40}$$

Subsequently, let us prove the asymptotic unbiasedness. By $\mathbb{M}(x_1^n) \leq n$ and by the overestimation bound (13.36) we have

$$\mathbf{E}\,\mathbb{M}(X_1^n) \leq M + nP(\mathbb{M}(x_1^n) > M) \leq M + \frac{1}{n}. \tag{13.41}$$

On the other hand, by the Fatou lemma,

$$M = \mathbf{E} \liminf_{n\to\infty} \mathbb{M}(X_1^n) \leq \liminf_{n\to\infty} \mathbf{E}\,\mathbb{M}(X_1^n). \tag{13.42}$$

Hence claim (13.33) follows. □

Using the Markov order estimator, we obtain a simple upper bound for the pointwise mutual information with respect to the PML maximum distribution.

Theorem 13.8 *We have the bound*

$$J_{\mathbb{P}}(x_1^n; x_{n+1}^{2n}) \leq 2 \log Z(n|\mathbb{M}(x_1^{2n})) - 3 \log w_{2n}. \tag{13.43}$$

Proof: By definition, the maximum log-likelihood is subadditive,

$$\frac{\hat{\mathbb{P}}(x_1^{2n}|k)}{\hat{\mathbb{P}}(x_1^n|k)\hat{\mathbb{P}}(x_{n+1}^{2n}|k)} \leq \frac{\hat{\mathbb{P}}(x_1^{2n}|k)}{\hat{\mathbb{P}}(x_1^n|k)\hat{\mathbb{P}}(x_{n-k}^{n+k}|k)\hat{\mathbb{P}}(x_{n+1}^{2n}|k)} \leq 1. \tag{13.44}$$

Let us put $k = \mathbb{M}(x_1^{2n}) \leq 2n$. We observe by the definition of the PML maximum, by the definition of order estimator, by the definition of penalized maximum likelihood, and by subadditivity (13.44) that

$$\begin{aligned} \log \frac{\mathbb{P}(x_1^{2n})}{\mathbb{P}(x_1^n)\mathbb{P}(x_{n+1}^{2n})} &\leq \log \frac{\hat{\mathbb{P}}(x_1^{2n}|k)/w_{2n}}{w_k \mathbb{P}(x_1^n|k) w_k \mathbb{P}(x_{n+1}^{2n}|k)} \\ &= \log \left(\frac{[Z(n|k)]^2}{w_{2n} w_k^2} \cdot \frac{\hat{\mathbb{P}}(x_1^{2n}|k)}{\hat{\mathbb{P}}(x_1^n|k)\hat{\mathbb{P}}(x_{n+1}^{2n}|k)} \right) \\ &\leq 2 \log Z(n|k) - 3 \log w_{2n}. \end{aligned} \tag{13.45}$$

□

In consequence, the pointwise mutual information with respect to the PML maximum distribution grows only logarithmically for Markov processes.

Theorem 13.9 *For any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ over alphabet $\{1,2,...,m\}$, we have inequalities*

$$\lim_{n\to\infty} \frac{\mathbf{E}\, J_{\mathbb{P}}(X_{-n+1}^0; X_1^n)}{\log n} \leq 2m^{M+1} + 6 \tag{13.46}$$

$$\lim_{n\to\infty} \frac{J_{\mathbb{P}}(X_{-n+1}^0; X_1^n)}{\log n} \leq 2m^{M+1} + 6 \ \textit{a.s.} \tag{13.47}$$

Hence $\beta_C = \beta_{\mathbb{P}} = 0$ if $M < \infty$.

Proof: Thus, we obtain

$$\begin{aligned}\lim_{n\to\infty} \frac{J_{\mathbb{P}}(X_{-n+1}^0; X_1^n)}{\log n} &\leq \lim_{n\to\infty} \frac{2\log Z(n|\mathbb{M}(x_1^{2n})) - 3\log w_{2n}}{\log n} \\ &\leq \lim_{n\to\infty} \frac{2\log Z(n|M) - 3\log w_{2n}}{\log n} \\ &\leq 2m^{M+1} + 6 \text{ a.s.}\end{aligned} \tag{13.48}$$

The bound in expectation follows by the overestimation bound (13.36), which implies

$$\begin{aligned}\mathbf{E}\log Z(n|\mathbb{M}(x_1^{2n})) &\leq \log Z(n|M) + (n\log m)P(\mathbb{M}(x_1^{2n}) > M) \\ &\leq \log Z(n|M) + nw_n \log m \\ &\leq \log Z(n|M) + \frac{\log m}{2n}.\end{aligned} \tag{13.49}$$

Consequently equality $\beta_C = \beta_{\mathbb{P}} = 0$ follows since $\mathrm{hilb}_{n\to\infty} \log n = 0$. □

Grammar-based codes

Now let us consider two universal grammar-based codes introduced in Definitions 7.17 and 7.20. Let G and S be admissible grammars. We will write the adjoining operation

$$G = S \oplus w \iff V_G = V_S + 1 \text{ and } G(r) = \begin{cases} S(r), & -V_S \leq r \leq -1, \\ w, & r = -V_G, \end{cases} \tag{13.50}$$

where V_G is the number of non-terminals in grammar G. Grammar S is called the secondary part or, succinctly, the vocabulary of G. In the following, we allow for infinite vocabularies. We assume that code ψ is m-proper so $|\psi(n)| = c_1$ for $0 \leq n \leq m$ and $|\psi(-n)| \leq \log n + 2\log(\log n + 1) + c_2$ for $n \geq 1$. We denote $\|w\| := |\psi^*(w)|$ for an argument of the local grammar encoder ψ^*. We define also the diameter of a vocabulary S as

$$d(S) := \sup_{-V_S \leq r \leq -1} \|S^*(r)\| . \tag{13.51}$$

Mind that $d(S)$ may be infinite if S is infinite.

We recall that $\Gamma_\psi(u)$ is the minimal admissible grammar that produces string u and $\Gamma^\#_\psi(u)$ is the minimal block grammar that produces string u.

Definition 13.10 (admissible grammar decomposition) *For the minimal admissible grammar $\Gamma_\psi(u)$, we denote its vocabulary as $\Sigma_\psi(u)$. We also define the minimal admissible primary rule $\Pi_\psi(u|S)$ as the string $w \in \mathbb{Z}^+$ such that $S \oplus w$ is an admissible grammar that produces string u and minimizes length $\|w\|$.*

Definition 13.11 (block grammar decomposition) *For the minimal block grammar $\Gamma^\#_\psi(u)$, we denote its vocabulary as $\Sigma^\#_\psi(u)$. We also define the minimal block primary rule $\Pi^\#_\psi(u|S)$ as the string $w \in \mathbb{Z}^+$ such that $S \oplus w$ is a block grammar that produces string u and minimizes length $\|w\|$.*

Obviously, we have

$$\|\Gamma_\psi(u)\| = \|\Pi_\psi(u|\Sigma_\psi(u))\| + \|\Sigma_\psi(u)\| , \tag{13.52}$$

$$\left\|\Gamma^\#_\psi(u)\right\| = \left\|\Pi^\#_\psi(u|\Sigma_\psi(u))\right\| + \left\|\Sigma^\#_\psi(u)\right\| . \tag{13.53}$$

Having these notations, we observe an important fact.

Theorem 13.12 *For m-proper minimal codes, we have*

$$c_1 \leq \|\Pi_\psi(u|S)\| + \|\Pi_\psi(v|S)\| - \|\Pi_\psi(uv|S)\| \leq d(S), \tag{13.54}$$

$$\left\|\Pi^\#_\psi(u|S)\right\| + \left\|\Pi^\#_\psi(v|S)\right\| - \left\|\Pi^\#_\psi(uv|S)\right\| \leq d(S). \tag{13.55}$$

Proof: To prove the lower bound in (13.54), we put $w := \Pi_\psi(u|S)\Pi_\psi(v|S)$. Obviously $S \oplus w$ is an admissible grammar that produces string uv. Hence

$$\begin{aligned} \|\Pi_\psi(u|S)\| + \|\Pi_\psi(v|S)\| - c_1 &= \|w\| \\ &\geq \|\Pi_\psi(uv|S)\| . \end{aligned} \tag{13.56}$$

Regrouping yields the claim.

To prove the upper bound in (13.54), we may write $\Pi_\psi(uv|S) = xZy$, where $Z < 0$, $S^*(Z) = pq$, $S \oplus xp$ is an admissible grammar that produces string u, and $S \oplus qy$ is an admissible grammar that produces string v. Hence

$$\begin{aligned}\left\|\Pi^{\#}_\psi(uv|S)\right\| + d(S) &\geq \left\|\Pi^{\#}_\psi(uv|S)\right\| - \|Z\| + \|pq\| + c_1 \\ &= \|xp\| + \|qy\| \\ &\geq \|\Pi_\psi(u|S)\| + \|\Pi_\psi(v|S)\|\,. \end{aligned} \tag{13.57}$$

Regrouping yields the claim.

The proof of inequality (13.55) is analogous. The lower bound is not guaranteed for the reason that the concatenation of two block primary rules need not be a block primary rule. □

Let us denote the pointwise mutual information

$$J_\psi(u;v) := \|\Gamma_\psi(u)\| + \|\Gamma_\psi(v)\| - \|\Gamma_\psi(uv)\|\,. \tag{13.58}$$

Theorem 13.12 has a corollary that bounds these quantities.

Theorem 13.13 *We have*

$$J_\psi(u;v) \leq \|\Sigma_\psi(uv)\| + d(\Sigma_\psi(uv)). \tag{13.59}$$

Proof: We have

$$\|\Gamma_\psi(u)\| \leq \|\Pi_\psi(u|\Sigma_\psi(uv))\| + \|\Sigma_\psi(uv)\|\,, \tag{13.60}$$

$$\|\Gamma_\psi(v)\| \leq \|\Pi_\psi(v|\Sigma_\psi(uv))\| + \|\Sigma_\psi(uv)\|\,, \tag{13.61}$$

$$\|\Gamma_\psi(uv)\| = \|\Pi_\psi(uv|\Sigma_\psi(uv))\| + \|\Sigma_\psi(uv)\|\,. \tag{13.62}$$

Hence (13.59) follows from (13.54). □

On the other hand, we can bound the redundancies in this way.

Theorem 13.14 *Consider a stationary process $(X_i)_{i\in\mathbb{Z}}$ over alphabet $\{1, 2, ..., m\}$. We have*

$$\mathbf{E}\,\|\Gamma_\psi(X_1^n)\| - hn \geq \mathbf{E}\,\|\Sigma_\psi(X_1^n)\| - H(\Sigma_\psi(X_1^n)). \tag{13.63}$$

Proof: By the source coding inequality, we have

$$\begin{aligned}\mathbf{E}\,\|\Pi_\psi(X_1^n|\Sigma_\psi(X_1^n))\| &\geq H(X_1^n|\Sigma_\psi(X_1^n)) = H(X_1^n) - I(X_1^n;\Sigma_\psi(X_1^n)) \\ &\geq hn - H(\Sigma_\psi(X_1^n)). \end{aligned} \tag{13.64}$$

Hence we obtain (13.63). □

In this way, we obtain this characterization of mutual information for a stationary process $(X_i)_{i\in\mathbb{Z}}$ over alphabet $\{1, 2, ..., m\}$. Suppose that

$$\limsup_{n\to\infty} \frac{H(\Sigma_\psi(X_1^n))}{\mathbf{E}\,\|\Sigma_\psi(X_1^n)\|} < 1, \qquad \limsup_{n\to\infty} \frac{d(\Sigma_\psi(X_1^n))}{\mathbf{E}\,\|\Sigma_\psi(X_1^n)\|} < \infty. \tag{13.65}$$

Then we may define the Hilberg exponent

$$\beta_\psi := \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\,\|\Gamma_\psi(X_1^n)\| - hn) = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J_\psi(X_{-n+1}^0; X_1^n) = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,\|\Sigma_\psi(X_1^n)\|\,. \tag{13.66}$$

Analogous results can be derived for the minimal block grammar-based code.

Let $V_\psi(u)$ be the number of rules in the minimal admissible vocabulary $\Sigma_\psi(u)$. Similarly, let $V_\psi^\#(u)$ denote the number of rules in the minimal block vocabulary $\Sigma_\psi^\#(u)$. Moreover, let $L(u)$ be the maximal length of a repetition in u, namely,

$$L(u) := \max\{|v| : u = xvz = x'vz' \text{ for some } x \neq x', z \neq z'\}\,, \tag{13.67}$$

defined also in (8.28). If the maximal length of a repetition grows much slower than the string length then the rate of mutual information is dictated by the rate of the number of rules.

Theorem 13.15 *We have bounds*

$$\frac{d(\Sigma_\psi(u))}{c_1} \leq L(u) + 1, \tag{13.68}$$

$$V_\psi(u) \leq \frac{\|\Sigma_\psi(u)\|}{c_1} \leq V_\psi(u)[L(u) + 1], |u|\,, \tag{13.69}$$

$$\frac{d(\Sigma_\psi^\#(u))}{c_1} \leq L(u) + 1, \tag{13.70}$$

$$V_\psi^\#(u) \leq \frac{\left\|\Sigma_\psi^\#(u)\right\|}{c_1} \leq V_\psi^\#(u)[L(u) + 1], |u|\,. \tag{13.71}$$

Proof: We have $d(\Sigma_\psi(u)) \leq c_1[L(u) + 1]$ since each secondary rule in the minimal admissible grammar must be used at least twice in the primary rule. For a bit different reason, we have $d(\Sigma_\psi^\#(u)) \leq c_1[L(u) + 1]$ because at least one secondary rule in the minimal block grammar must be used at least twice in the primary rule. Obviously, $c_1 V_\psi(u) \leq \|\Sigma_\psi(u)\| \leq V_\psi(u) d(\Sigma_\psi(u))$ since each rule is written in the minimal admissible vocabulary in the most succinct form. Moreover, we have $\|\Sigma_\psi(u)\| \leq \|\Gamma_\psi(u)\| \leq c_1\,|u|$ by minimality.

For a bit different reason, $c_1 V_\psi^{\#}(u) \le \left\|\Sigma_\psi^{\#}(u)\right\| \le V_\psi^{\#}(u) d(\Sigma_\psi^{\#}(u))$ since each rule has the same length and is equal to its expansion in the minimal block grammar. Besides, $\left\|\Sigma_\psi^{\#}(u)\right\| \le \left\|\Gamma_\psi^{\#}(u)\right\| \le c_1 |u|$ holds by the constrained minimality. □

***

Recapitulating this chapter, we have investigated some sublinear effects in universal coding. We also showed that the Markov order of a stationary process can be consistently estimated. In Chapter 14, we will exhibit a simple example of stationary processes for which Hilberg exponents for the Shannon entropy and the Kolmogorov complexity can be different. This topic is connected with statistical modeling of natural language.

Further reading

The concept of excess entropy as related to stationary processes can be mostly attributed to the works of James Crutchfield and David Feldman. A good starting point is paper [28]. Some of their inspiration came from the paper by Wolfgang Hilberg [70], who supposed that the mutual information for natural language grows roughly like a power law. Some research concerning mathematical surroundings and linguistic implications of Hilberg's hypothesis was done by Łukasz Dębowski [36], see also papers [33, 37]. The idea of the Markov order estimator based on comparing the maximum likelihood with universal codes was proposed by Neri Merhav, Michael Gutman, and Jacob Ziv [96] and generalized later to hidden Markov processes [131].

Thinking exercises

1. For a stationary process $(X_i)_{i\in\mathbb{Z}}$, show that

$$H(X_0) - h \le E, \tag{13.72}$$

$$H(X_0|X_{-k}^{-1}) - h \le \frac{E}{k+1}, \quad k \ge 1. \tag{13.73}$$

$$\operatorname*{hilb}_{k\to\infty} k \left[H(X_0|X_{-k}^{-1}) - h\right] \le \beta_P. \tag{13.74}$$

2. For a stationary process $(X_i)_{i\in\mathbb{Z}}$, prove that

$$\frac{1}{n}\, \mathbf{E}\left[-\log \hat{\mathbb{P}}(x_1^n|k)\right] \le H(X_0|X_{-k}^{-1}). \tag{13.75}$$

3. For a stationary process $(X_i)_{i\in\mathbb{Z}}$ over alphabet $\{1,2,...,m\}$, define

$$f(n) := H(X_0|X_{-n}^{-1}) - h, \tag{13.76}$$

$$f_m(n) := \min_{k\in\mathbb{N}} \left(f(k) + \frac{m^k \log(n+1)}{n} \right), \tag{13.77}$$

$$g(n) := \frac{\mathbf{E}\, C(X_1^n)}{n} - h. \tag{13.78}$$

Demonstrate that $g(n) \le f_m(n) + \frac{3\log n}{n} + c$. Estimate $f_m(n)$ for

(a) $f(n) = 0$ for $n \ge q \in \mathbb{N}$.
(b) $f(n) = r^n$, $r < 1$,
(c) $f(n) = n^\gamma$, $\gamma < 0$.

4. Show that for any sequence $(x_i)_{i\in\mathbb{N}}$,

$$\sum_{k=0}^{\infty} \left[-\log \hat{\mathbb{P}}(x_1^{n+k}|k) \right] \le n \log n. \tag{13.79}$$

Consequently, for a stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$, demonstrate that the Markov order estimator satisfies

$$\limsup_{n\to\infty} \frac{\mathbb{M}(X_1^n)}{\log n} \le \frac{1}{h} \text{ a.s.} \tag{13.80}$$

5. Let $L(x_1^n)$ be the maximal length of a repetition in string x_1^n, defined in (13.67). For a finite alphabet, show that:

(a) $\lim_{n\to\infty} L(x_1^n) = \infty$ for any sequence $(x_i)_{i\in\mathbb{N}}$.
(b) $\lim_{n\to\infty} L(X_1^n)/n = 0$ almost surely for any stationary ergodic process $(X_i)_{i\in\mathbb{N}}$.
(c) $\liminf_{n\to\infty} L(X_1^n)/\log n \ge 1/h$ almost surely for any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$.

6. Consider a string $u \in \{1,2,...,m\}^*$. Let $L(u)$ be maximal length of a repetition in string u, defined in (13.67). The number of repetition types in string u is

$$R(u) := \#\left\{ v : u = xvz = x'vz' \text{ for some } x \ne x', z \ne z' \right\}. \tag{13.81}$$

Show that

$$R(u) \ge \sqrt{\frac{|u|}{L(u)} - 1} - m. \tag{13.82}$$

Chapter 14

Words

Hilberg's law. Zipf's law. Zipf processes. Monkey-typing explanation of Zipf's law. Miller processes. Herdan-Heaps' law. Large number of rare events. Hapax rate. Santa Fe processes. Hilberg exponents for Santa Fe processes.

The ultimate goal of statistical language modeling is to construct a stochastic process that with a high probability produces sequences of words that could be taken by us, humans, as texts written by us, humans. Engineers in machine learning made a huge leap in making computers produce such artificial texts. Highly publicized in the media was the family of GPT-n neural language models, which surprised the general public with a sheer amount of made-up surreal stories concerning, for instance, unicorns in the Andes that spoke perfect English. It would be advisable to understand the dynamics of such statistical language models from a theoretical point.

Approaching this topic from a certain distance, in this chapter we will make an excursion to quantitative linguistics. The goal of this chapter is to exhibit a simple example of stochastic processes, called Santa Fe processes, which enjoy Hilberg exponent $\beta_P = 0$ and arbitrarily large Hilberg exponents $\beta_C \leq \beta_{\mathbb{P}} \in (0,1)$. Interestingly, this topic is connected with statistical modeling of natural language. The hypothesis that inequality $\beta_C > 0$ holds for natural language is called Hilberg's hypothesis or Hilberg's law and it can be linked with the famous empirical Zipf law for the distribution of words.

Suppose that we sort the words of a natural language $(w_1, w_2, ...)$ according to their probabilities $\pi(w_1) \geq \pi(w_2) \geq ...$. We may also define $\pi(w_k)$ as the relative frequencies in a particular text and speak of an empirical probability distribution. The harmonic bound stated in Theorem 7.21 says that these probabilities satisfy $\pi(w_k) \leq 1/k$. Number k is called the rank

of word w_k. Zipf's law is an empirical fact about texts in natural language that each probability $\pi(w_k)$ is in a sense close to the upper bound $1/k$. Since the harmonic series is insummable, $\sum_{k=1}^{\infty} 1/k = \infty$, Zipf's law can be only an approximation. In the realm of pure mathematics, we may consider the following highly simplified model of a text.

Definition 14.1 (Zipf process) *An IID process $(K_i)_{i\in\mathbb{Z}}$ where $K_i : \Omega \to \mathbb{N}$ is called the Zipf process with a parameter $\alpha > 1$ when if*

$$P(K_i = k) = \frac{k^{-\alpha}}{\zeta(\alpha)}, \quad \zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}. \tag{14.1}$$

The marginal distribution (14.1) is called the Zipf distribution. Function $\alpha \mapsto \zeta(\alpha)$ is called the Riemann zeta function.

For texts in natural language of a moderate length and $P(K = k)$ estimated as the relative frequency of the word of rank k, the estimates of parameter α based on Zipf's law (14.1) are usually close to 1. However, the reader should be aware that the estimates of parameter α depend significantly on the text size, which is a sign that formula (14.1) is only an idealization. There is a way of correcting Zipf's distribution (14.1) so that the further discussed number of types and rate of hapaxes agree with the empirical data for natural language. This topic is a bit too technical for this chapter since there is no simple formula for probabilities $P(K = k)$.

What may be surprising, Zipf's law is also approximately obeyed by the following model called the monkey-typing model. Suppose that we have an IID process over alphabet $\{0, 1, 2\}$ and we define words in this stream of symbols as binary strings delimited by symbols 2. It turns out that the distribution of such words obeys Zipf's law approximately. This observation was made independently by Benoît Mandelbrot and George Miller. Since the Zipf distribution is sometimes called the Zipf-Mandelbrot distribution, let us call the monkey-typing models after George Miller solely.

Definition 14.2 (Miller processes) *An IID process $(X_i)_{i\in\mathbb{Z}}$ over alphabet $\{0, 1, 2\}$ is called a Miller letter process with a parameter $\theta \in (0, 1)$ when $P(X_i = 0) = P(X_i = 1) = \theta/2$ and $P(X_i = 2) = 1 - \theta$. For that process $(X_i)_{i\in\mathbb{Z}}$, let process $(Y_i)_{i\in\mathbb{Z}}$ be the sequence of binary strings separated by symbols 2:*

$$(X_i)_{i\in\mathbb{Z}} = ...2Y_{-2}2Y_{-1}2Y_0 2Y_1 2Y_2 2... \tag{14.2}$$

We call process $(Y_i)_{i\in\mathbb{Z}}$ the Miller word process. Let $\phi = \mathrm{mil}^{-1}$ be the inverse of the military code: $\phi(\lambda) = 1$, $\phi(0) = 2$, $\phi(1) = 3$, $\phi(00) = 4$, etc. We call process $(K_i)_{i\in\mathbb{Z}}$ where $K_i = \phi(Y_i)$ the Miller rank process.

We can easily see that the Miller word process $(Y_i)_{i\in\mathbb{Z}}$ is an IID process with the distribution

$$P(Y_i = y) = (1-\theta)\left(\frac{\theta}{2}\right)^{|y|}. \tag{14.3}$$

Consequently, the Miller rank process $(K_i)_{i\in\mathbb{Z}}$ is an IID process with the distribution

$$P(K_i = k) = (1-\theta)\left(\frac{\theta}{2}\right)^{\lfloor \log k \rfloor} \in \left[\frac{(1-\theta)}{k^\alpha}, \frac{2^\alpha(1-\theta)}{k^\alpha}\right], \quad \alpha = 1 - \log\theta. \tag{14.4}$$

Hence the Miller process is approximately equivalent to the Zipf process.

In the following, we will study some properties of processes that take an countably infinite number of values. Let $(K_i)_{i\in\mathbb{Z}}$ be a process where $K_i : \Omega \to \mathbb{K}$ and set $\mathbb{K}$ is countable. The elements of set $\mathbb{K}$ will be called types. The frequencies of types in sample K_1^n are random variables

$$F_k(n) := \sum_{i=1}^{n} \mathbf{1}\left\{K_i = k\right\}. \tag{14.5}$$

For an IID process $(K_i)_{i\in\mathbb{Z}}$, random variables $F_k(n)$ follow the Bernoulli distribution, which for large n can be approximated by the Poisson distribution. Namely, we have

$$\begin{aligned} P(F_k(n) = l) &= \binom{n}{l} P(K_i = k)^l (1 - P(K_i = k))^{n-l} \\ &\approx \frac{[nP(K_i = k)]^l}{l!} \exp\left(-nP(K_i = k)\right). \end{aligned} \tag{14.6}$$

Let us investigate how the number of observed types grows with the observed sample. The total number of types in sample K_1^n is

$$V(n) := \sum_{k\in\mathbb{K}} \mathbf{1}\left\{F_k(n) > 0\right\}. \tag{14.7}$$

In the following,

$$\Gamma(z) := \int_0^\infty t^{z-1} e^{-t} dt, \quad z > 0, \tag{14.8}$$

is the gamma function. We have $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$ in particular.

Theorem 14.3 (Herdan-Heaps law) *For the Zipf process $(K_i)_{i\in\mathbb{Z}}$, we have*

$$\lim_{n\to\infty} \frac{\mathbf{E}\,V(n)}{n^{1/\alpha}} = \frac{\Gamma(1-1/\alpha)}{[\zeta(\alpha)]^{1/\alpha}}. \tag{14.9}$$

Proof: Let us write $\beta = 1/\alpha$. We have

$$\begin{aligned}
\mathbf{E}\,V(n) &= \sum_{k=1}^{\infty} P(F_k(n) > 0) = \sum_{k=1}^{\infty}(1 - P(K_i \neq k)^n) \\
&= \sum_{k=1}^{\infty} \left(1 - \left(1 - \frac{k^{-\alpha}}{\zeta(\alpha)}\right)^n\right) \approx \int_1^{\infty} \left(1 - \left(1 - \frac{k^{-\alpha}}{\zeta(\alpha)}\right)^n\right) dk \\
&= \beta \left(\frac{n}{\zeta(\alpha)}\right)^{\beta} \int_{(1-1/\zeta(\alpha))^n}^{1} \frac{(1-u)du}{u^{1-1/n}[n(1-u^{1/n})]^{\beta+1}} \quad \left\{u := \left(1 - \frac{k^{-\alpha}}{\zeta(\alpha)}\right)^n\right\} \\
&= \beta \left(\frac{n}{\zeta(\alpha)}\right)^{\beta} \int_0^1 f_n(u)du,
\end{aligned} \tag{14.10}$$

where we denote functions

$$f_n(u) := \frac{(1-u)\mathbf{1}\left\{u \geq (1-1/\zeta(\alpha))^n\right\}}{u^{1-1/n}[n(1-u^{1/n})]^{\beta+1}}. \tag{14.11}$$

These functions tend to limit

$$\lim_{n\to\infty} f_n(u) = f(u) := \frac{(1-u)}{u(-\ln u)^{\beta+1}}. \tag{14.12}$$

We notice the upper bound $f_n(u) \leq f(u)$ for $u \in (0,1)$. Moreover, function $f(u)$ is integrable on $u \in (0,1)$. Indeed, putting $t := -\ln u$ and integrating by parts yields

$$\begin{aligned}
\int_0^1 f(u)du &= \int_0^{\infty} (1 - e^{-t})\, t^{-\beta-1} dt \\
&= (1 - e^{-t})(-\beta^{-1})t^{-\beta}|_0^{\infty} + \int_0^{\infty} e^{-t}\beta^{-1}t^{-\beta}dt = \beta^{-1}\Gamma(1-\beta).
\end{aligned} \tag{14.13}$$

Hence, by the Lebesgue dominated convergence, we derive

$$\lim_{n\to\infty} \frac{\mathbf{E}\,V(n)}{n^{\beta}} = \frac{\beta}{[\zeta(\alpha)]^{\beta}} \int_0^1 f(u)du = \frac{\Gamma(1-\beta)}{[\zeta(\alpha)]^{\beta}}. \tag{14.14}$$

□

For texts in natural language of a moderate length, the estimates of parameter α based on Herdan-Heaps' law (14.9) are usually close to 1.3. The reader should be aware that the estimates of parameter α depend strongly on the text size so Zipf's distribution (14.1) is only an idealization. Once we modify the marginal distribution (14.1) appropriately, the Poisson approximation (14.6) works surprisingly well.

Zipf processes exhibit the phenomenon of a large number of rare events (LNRE). Namely, the number of types that appear a few times is asymptotically a non-negative fraction of the total number of types. Let us denote the number of types that appear l times in sample K_1^n as

$$V_l(n) := \sum_{k\in\mathbb{K}} \mathbf{1}\left\{F_k(n) = l\right\}. \tag{14.15}$$

Types of frequency 1 are called hapaxes so $V_1(n)$ is called the number of hapaxes. Let us derive the hapax rate for Zipf processes.

Theorem 14.4 (hapax rate) *For the Zipf process $(K_i)_{i\in\mathbb{Z}}$, we have*

$$\lim_{n\to\infty} \frac{\mathbf{E}\, V_1(n)}{\mathbf{E}\, V(n)} = \frac{1}{\alpha}. \tag{14.16}$$

Proof: Let us write $\beta = 1/\alpha$. We have

$$\begin{aligned}
\mathbf{E}\, V_1(n) &= \mathbf{E}\sum_{k=1}^{\infty} \mathbf{1}\left\{F_k(n) = 1\right\} = \sum_{k=1}^{\infty} \binom{n}{1} P(K_i = k) P(K_i \neq k)^{n-1} \\
&= \sum_{k=1}^{\infty} n \left(\frac{k^{-\alpha}}{\zeta(\alpha)}\right) \left(1 - \frac{k^{-\alpha}}{\zeta(\alpha)}\right)^{n-1} \approx \int_1^{\infty} n \left(\frac{k^{-\alpha}}{\zeta(\alpha)}\right) \left(1 - \frac{k^{-\alpha}}{\zeta(\alpha)}\right)^{n-1} dk \\
&= \beta \left(\frac{n}{\zeta(\alpha)}\right)^{\beta} \int_{(1-1/\zeta(\alpha))^n}^{1} \frac{du}{[n(1-u^{1/n})]^{\beta}} \quad \left\{u := \left(1 - \frac{k^{-\alpha}}{\zeta(\alpha)}\right)^n\right\} \\
&= \beta \left(\frac{n}{\zeta(\alpha)}\right)^{\beta} \int_0^1 f_n(u) du,
\end{aligned} \tag{14.17}$$

where we denote functions

$$f_n(u) := \frac{\mathbf{1}\left\{u \geq (1 - 1/\zeta(\alpha))^n\right\}}{[n(1-u^{1/n})]^{\beta}}. \tag{14.18}$$

These functions tend to limit

$$\lim_{n\to\infty} f_n(u) = f(u) := \frac{1}{(-\ln u)^{\beta}}. \tag{14.19}$$

Putting $t := -\ln u$ and integrating yields

$$\int_0^1 f(u)du = \int_0^\infty e^{-t}\, t^\beta dt = \Gamma(1-\beta). \tag{14.20}$$

Hence, we derive

$$\lim_{n\to\infty} \frac{\mathbf{E}\, V_1(n)}{n^\beta} = \frac{\beta}{[\zeta(\alpha)]^\beta} \int_0^1 f(u)du = \frac{\beta\Gamma(1-\beta)}{[\zeta(\alpha)]^\beta}. \tag{14.21}$$

Plugging this into the previous theorem, we obtain the claim. □

For texts in natural language of a moderate length, naive estimates of parameter α based on the hapax rate law (14.16) are usually close to 2. Since these estimates depend on the text size then Zipf's distribution (14.1) is only an idealization. In fact, for natural language, the hapax rate is close to the log-linear function,

$$V_1(n)/V(n) \approx a + b \log n, \tag{14.22}$$

where $b < 0$. There is a method of inverting an arbitrary function $\mathbf{E}\, V_1(n)/\mathbf{E}\, V(n)$, to derive the expected number of types $\mathbf{E}\, V(n)$ and the expected rank-frequency plot which corrects Zipf's distribution (14.1). This is all feasible under the Poisson approximation (14.6), which works surprisingly well—also for natural language, which breaks the IID assumption.

In the following, we would like provide some sandwich bounds for the number of types. Let us consider a process $(K_i)_{i\in\mathbb{Z}}$ where the types are natural numbers, $\mathbb{K} = \mathbb{N}$. We will introduce two symmetric quantities

$$U(n) := \min\{k \in \mathbb{N} : F_k(n) = 0\}, \tag{14.23}$$

$$M(n) := \max\{k \in \mathbb{N} : F_k(n) > 0\}. \tag{14.24}$$

These quantities bound the number of types as

$$U(n) - 1 \le V(n) \le \min\{M(n), n\}. \tag{14.25}$$

They can be bounded themselves as follows.

Theorem 14.5 *For the Zipf process $(K_i)_{i\in\mathbb{Z}}$, we have*

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U(n) \ge \frac{1}{\alpha}, \tag{14.26}$$

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E} \min\{M(n), n\} \le \min\left\{\frac{2}{\alpha}, 1\right\}. \tag{14.27}$$

Proof: We have

$$\mathbf{E}\, U(n) \geq r - rP(U(n) \leq r), \tag{14.28}$$
$$\mathbf{E} \min\{M(n), n\} \leq r + nP(M(n) \geq r). \tag{14.29}$$

Putting $r = n^{1/\alpha-\epsilon}$ where $\epsilon > 0$ and using inequality $\ln(1+x) \leq x$, we can bound

$$\begin{aligned} P(U(n) \leq r) &\leq \sum_{k=1}^{r} P(F_k(n) = 0) = \sum_{k=1}^{r} \left(1 - \frac{k^{-\alpha}}{\zeta(\alpha)}\right)^n \\ &\leq r\left(1 - \frac{r^{-\alpha}}{\zeta(\alpha)}\right)^n = r\exp\left(n\ln\left(1 - \frac{r^{-\alpha}}{\zeta(\alpha)}\right)\right) \\ &\leq r\exp\left(-\frac{nr^{-\alpha}}{\zeta(\alpha)}\right) = n^{1/\alpha-\epsilon}\exp\left(-\frac{r^{\alpha\epsilon}}{\zeta(\alpha)}\right). \end{aligned} \tag{14.30}$$

Hence $\mathrm{hilb}_{n\to\infty} \mathbf{E}\, U(n) \geq 1/\alpha - \epsilon$. Taking $\epsilon \to 0$ yields (14.26).

By contrast, using inequality $(1+x)^n \geq 1 + nx$, we can bound

$$\begin{aligned} P(M(n) \geq r) &\leq \sum_{k=r}^{\infty} P(F_k(n) > 0) = \sum_{k=r}^{\infty} \left(1 - \left(1 - \frac{k^{-\alpha}}{\zeta(\alpha)}\right)^n\right) \\ &\leq \sum_{k=r}^{\infty} \frac{nk^{-\alpha}}{\zeta(\alpha)} \approx \int_r^{\infty} \frac{nk^{-\alpha}}{\zeta(\alpha)} dk = \frac{nr^{1-\alpha}}{(\alpha-1)\zeta(\alpha)}. \end{aligned} \tag{14.31}$$

Plugging $r = n^{2/\alpha}$, we obtain

$$\begin{aligned} \mathbf{E} \min\{M(n), n\} &\leq r + nP(M(n) \geq r) \\ &\leq n^{2/\alpha} + \frac{n^2 n^{2(1-\alpha)/\alpha}}{(\alpha-1)\zeta(\alpha)} \\ &= n^{2/\alpha}\left(1 + \frac{1}{(\alpha-1)\zeta(\alpha)}\right). \end{aligned} \tag{14.32}$$

Hence we obtain (14.27). □

In the following, we will present a simple example of a stationary process that has exponent $\beta_P = 0$ and arbitrarily large exponents $\beta_C \leq \beta_{\mathbb{P}} \in (0, 1)$. This kind of a process is called a Santa Fe process. The Santa Fe process is obtained by combining the Zipf process with a fixed binary sequence of a high Kolmogorov complexity.

Definition 14.6 (Santa Fe processes) *Consider a fixed sequence $(z_k)_{k\in\mathbb{N}}$ where $z_k \in \{0,1\}$ and the Kolmogorov complexity of the prefixes is high, i.e., $C(z_1^n) \approx n$. The Santa Fe rank process $(X_i)_{i\in\mathbb{Z}}$ is a sequence of pairs*

$$X_i := (K_i, z_{K_i}), \tag{14.33}$$

where $(K_i)_{i\in\mathbb{Z}}$ is the Zipf process. Consider the prefix-free code $\mathrm{una}'' : \mathbb{N} \to \{0,1\}^*$. *The Santa Fe word process $(Y_i)_{i\in\mathbb{Z}}$ is a sequence of binary strings*

$$Y_i := \mathrm{una}''(K_i) z_{K_i}. \tag{14.34}$$

We can interpret Santa Fe processes as a toy model of a text in natural language that conveys an infinite number of elementary meanings in a repetitive way. Namely, these processes can be interpreted as sequences of random statements (k, z) that assert for a randomly chosen index k that the k-th fact z_k equals z. This description, although indices k are scattered at random, is never contradictory: If statements (k, z) and (k', z') describe the same fact, i.e., $k = k'$, then both statements assign the same value to it, i.e., $z = z'$. Moreover, by the strong law of large numbers, for any k, the description of k-th fact will appear in a sufficiently long text almost surely.

The Santa Fe word process $(Y_i)_{i\in\mathbb{Z}}$ is IID so

$$\beta_P = \operatorname*{hilb}_{n\to\infty} I(Y_{-n+1}^0; Y_1^n) = 0. \tag{14.35}$$

By contrast, the Hilberg exponent for the Kolmogorov complexity is positive. In this way, we obtain a simple model of Hilberg's law.

Theorem 14.7 (Hilberg law) *For the Santa Fe word process $(Y_i)_{i\in\mathbb{Z}}$, we have*

$$\beta_C = \frac{1}{\alpha}. \tag{14.36}$$

Proof: By the previous results, it suffices to show that

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U(n) \le \beta_C \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, V(n) \tag{14.37}$$

Let us demonstrate the left inequality in (14.37). We have

$$\beta_C = \operatorname*{hilb}_{n\to\infty} \left(\mathbf{E}\, C(Y_1^n) - nh\right), \tag{14.38}$$

where $C(u)$ is the Kolmogorov complexity of string u and h is the entropy rate of process $(Y_i)_{i\in\mathbb{Z}}$. We also observe that there is a computable function G such

that $G(Y_1^n) = \left\langle K_1^n, z_1^{U(n)} \right\rangle$, where $(K_i)_{i\in\mathbb{Z}}$ is the Zipf process. Hence by the data processing inequality and the chain rule for the conditional complexity (Theorems 12.16 and 12.12), we obtain

$$\begin{aligned} C(Y_1^n) &\gtrsim C(G(Y_1^n)) = C(\left\langle K_1^n, z_1^{U(n)} \right\rangle) \gtrsim C(z_1^{U(n)}) + C(K_1^n | z_1^{U(n)}) \\ &\gtrsim C(z_1^{U(n)}) + C(K_1^n | z_1^n, U(n)) \gtrsim C(z_1^{U(n)}) + C(K_1^n | z_1^n) - C(U(n) | z_1^n) \\ &\gtrsim C(z_1^{U(n)}) + C(K_1^n | z_1^n) \approx U(n) + C(K_1^n | z_1^n). \end{aligned} \tag{14.39}$$

In fact, h is also the entropy rate of process $(K_i)_{i\in\mathbb{Z}}$, which is IID. Hence

$$\begin{aligned} \beta_C &= \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\, C(Y_1^n) - nh) \\ &\geq \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\, U(n) + \mathbf{E}\, C(K_1^n | z_1^n) - nh) \\ &\geq \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\, U(n) + H(K_1^n) - nh) = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U(n). \end{aligned} \tag{14.40}$$

Now let us demonstrate the right inequality in (14.37). Let

$$\mathcal{V}(n) := \{(k, z_k) : k \in K_1^n\} \tag{14.41}$$

be the set of types appearing in Y_1^n. Given K_1^n, to describe each type (k, z_k) in $\mathcal{V}(n)$, we need $\lfloor \log n \rfloor$ bits to describe k and 1 bit to describe z_k. Hence by the coding bound (Theorem 12.2), we have

$$C(\mathcal{V}(n) | K_1^n) \lesssim V(n)(1 + \log n). \tag{14.42}$$

On the other hand, process $(K_i)_{i\in\mathbb{Z}}$ is the Zipf process which has a computable distribution. Applying the coding bound (Theorem 12.2), where the code is the Shannon-Fano code for the Zipf process, we obtain

$$C(K_1^n) \lesssim -\log P(K_1^n). \tag{14.43}$$

Since together set $\mathcal{V}(n)$ and sequence K_1^n carry the same information as Y_1^n, we may write

$$\begin{aligned} \beta_C &= \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\, C(Y_1^n) - nh) \\ &= \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\, C(\mathcal{V}(n) | K_1^n) + \mathbf{E}\, C(K_1^n) - nh) \\ &\leq \operatorname*{hilb}_{n\to\infty} (\mathbf{E}\, V(n)(1 + \log n) + H(K_1^n) - nh) = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, V(n). \end{aligned} \tag{14.44}$$

□

As for the Hilberg exponent for the PML maximum from Definition 8.8, we obtain a sandwich bound.

Theorem 14.8 *For the Santa Fe word process $(Y_i)_{i\in\mathbb{Z}}$, we have*

$$\frac{1}{\alpha} \leq \beta_{\mathbb{P}} \leq \min\left\{\frac{2}{\alpha}, 1\right\}. \tag{14.45}$$

Proof: Since $\beta_C \leq \beta_{\mathbb{P}}$, by the previous results, it suffices to show that

$$\beta_{\mathbb{P}} \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E} \min\{M(n), n\}. \tag{14.46}$$

We can express probability $P(Y_1^n)$ as the probability of a Markov process over alphabet $\{0,1\}$ of order $R(n) := |\mathrm{una}''(M(n))| + 1$. In consequence,

$$\hat{\mathbb{P}}(Y_1^n | R(n)) \geq P(Y_1^n). \tag{14.47}$$

Hence we obtain

$$\begin{aligned} P\left(\mathbb{M}(Y_1^n) > R(n)\right) &\leq P\left(\hat{\mathbb{P}}(Y_1^n|R(n)) < w_{|Y_1^n|}\mathbb{P}(Y_1^n)\right) \\ &\leq P\left(\frac{w_{|Y_1^n|}\mathbb{P}(Y_1^n)}{P(Y_1^n)} > 1\right) \leq \sum_{y_1^n} w_{|y_1^n|}\mathbb{P}(y_1^n) \\ &\leq \sum_{i=n}^{\infty} w_i = \frac{1}{n}. \end{aligned} \tag{14.48}$$

Since $R(n) \leq \log M(n) + 2\log\log M(n) + c$ for some $c < \infty$ then

$$\begin{aligned} \mathbf{E}\min\left\{2^{\mathbb{M}(Y_1^n)}, n\right\} &\leq \mathbf{E}\min\left\{2^{R(n)}, n\right\} + nP\left(\mathbb{M}(Y_1^n) > R(n)\right) \\ &\leq \mathbf{E}\min\left\{2^{\log M(n) + 2\log\log M(n) + c}, n\right\} + n\cdot\frac{1}{n} \\ &= \mathbf{E}\min\left\{2^c M(n)(\log M(n))^2, n\right\} + 1. \end{aligned} \tag{14.49}$$

In particular, we may derive

$$\begin{aligned} \beta_{\mathbb{P}} &= \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J_{\mathbb{P}}(Y_1^n; Y_{n+1}^{2n}) \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\log Z(n|\mathbb{M}(Y_1^n)) \\ &\leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\min\left\{2^{\mathbb{M}(Y_1^n)}, n\right\} \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\min\left\{2^c M(n)(\log M(n))^2, n\right\} \\ &\leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\min\{M(n), n\}. \end{aligned} \tag{14.50}$$

□

***

To recapitulate this chapter, we have exhibited some stationary processes that satisfy Zipf's law and obey different Hilberg exponents for different measures of information. We hope that these simple examples can shed some light onto fundamental problems of statistical language modeling. We hope that we have raised some interest in the intersection of information theory and quantitative linguistics. Since this is the last chapter of this book, we also hope to have proved that universal coding is a live paradigm for studying not only limits of data compression but also problems of statistical inference.

Further reading

It was Frederick Jelinek [74, 75] who introduced stochastic processes to natural language engineering. Present statistical language models apply vector representations of words, called embeddings [97], and artificial neural networks of a special structure, called transformers [124]. They can produce texts that constitute plausible short stories and highly relevant replies to questions but usually they hallucinate facts so they cannot be trusted as a source of knowledge [105, 15]. There is an open problem of understanding what these models are capable of in general and how to improve them further. From this perspective, it may be fruitful to investigate statistical laws of language. The most famous one, Zipf's law, was discovered by Jean-Baptiste Estoup [45] and Edward Condon [25] and later popularized by George Zipf in book [129]. In particular, Zipf's law implies Herdan-Heaps' law investigated by Gustav Herdan [67] and Harold Heaps [65], as well as the phenomenon of large number of rare events researched by Estate Khmaladze [80]. That Zipf's law can be generated by pressing keyboard keys at random was discovered by Benoît Mandelbrot [91] and George Miller [98]. Researching excess entropy and stochastic processes, see [33, 36, 37] for an overview, Łukasz Dębowski showed that Zipf's law can be linked with the hypothesis by Wolfgang Hilberg about the power-law growth of mutual information [70] via Santa Fe processes and similar processes. The Hilberg exponent $\beta_{\mathbb{P}}$ for natural language equals approximately 0.8 as shown by Ryosuke Takahira, Kumiko Tanaka-Ishii, and Łukasz Dębowski [120, 121], see also later experiments involving large statistical langugae models [69, 64, 12, 79, 66, 68].

Thinking exercises

1. For a sequence of random variables $(Y_n)_{n\in\mathbb{N}}$ such that $Y_{n+1} \geq Y_n$, show

$$\operatorname*{hilb}_{n\to\infty} Y_n \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, Y_n \text{ a.s.} \tag{14.51}$$

Consequently, argue that

$$\beta_P \geq \operatorname*{hilb}_{n\to\infty} J_P(X^0_{-n+1}; X^n_1) \text{ a.s.}, \tag{14.52}$$

$$\beta_C \geq \operatorname*{hilb}_{n\to\infty} J_C(X^0_{-n+1}; X^n_1) \text{ a.s.} \tag{14.53}$$

2. Consider the Santa Fe word process $(Y_i)_{i\in\mathbb{Z}}$. Show that

$$\operatorname*{hilb}_{n\to\infty} U(n) = \operatorname*{hilb}_{n\to\infty} \left(C(Y^n_1) + \log P(Y^n_1)\right) = \operatorname*{hilb}_{n\to\infty} V(n) = \frac{1}{\alpha} \text{ a.s.} \tag{14.54}$$

3. Consider an IID process $(K_i)_{i\in\mathbb{Z}}$ taking values in natural numbers. Assume that n is large. Let $V(n)$ be the number of types in sample K^n_1 and let $V_l(n)$ be the number of types that appear l times in K^n_1.

   (a) Show that we may approximate:

$$\mathbf{E}\, V(n) \approx g(n) := \sum_{k=1}^{\infty} \left(1 - e^{-nP(K_i=k)}\right), \tag{14.55}$$

$$\mathbf{E}\, V_l(n) \approx g_l(n) := -\frac{(-n)^l}{l!} \frac{d^l g(n)}{dn^l}. \tag{14.56}$$

   Argue that $g(1) \leq 1$.

   (b) Show that we may approximate:

$$\operatorname{Var} V(n) \approx g(2n) - g(n) \leq g(n), \tag{14.57}$$

$$\operatorname{Var} V_l(n) \approx g_l(2n) - g_l(n) \leq g_l(n). \tag{14.58}$$

   (c) Show that we may compute the number of types from the hapax rate:

$$g(n) = g(1) \exp\left(\int_0^{\log n} h(u)du\right), \quad h(u) := \frac{g_1(\exp u)}{g(\exp u)}. \tag{14.59}$$

   In particular, we have Herdan-Heaps' law for a constant hapax rate $h(u) = 1/\alpha$.

4. Consider a distribution $p : \mathbb{N} \to [0, 1]$. Show that the Shannon entropy satisfies $H(p) < \infty$ if $\sum_{n=1}^{\infty} p(n)\, n < \infty$. Suppose additionally that the probabilities of consecutive numbers decrease, $p(n+1) \leq p(n)$. Show respectively that $H(p) < \infty$ if and only if $\sum_{n=1}^{\infty} p(n) \log n < \infty$.

5. Consider a function

$$p(n) = \frac{C}{n(\log n)^\beta}, \quad n \in \{2, 3, 4, ...\}, \tag{14.60}$$

where $C > 0$ and $\beta \geq 0$. Show that

$$\sum_{n=2}^{\infty} p(n) = \infty \text{ for } \beta \in [0, 1], \tag{14.61}$$

$$\sum_{n=2}^{\infty} p(n) < \infty \text{ for } \beta > 1. \tag{14.62}$$

Assume that parameter C is chosen so that $\sum_{n=2}^{\infty} p(n) = 1$ for $\beta > 1$. Show that the Shannon entropy satisfies

$$H(p) = \infty \text{ for } \beta \in (1, 2], \tag{14.63}$$

$$H(p) < \infty \text{ for } \beta > 2. \tag{14.64}$$

6. *Multiperiodic sequence:* Zipf's law is exhibited also by deterministic sequences called multiperiodic. Their construction is as follows [38]:

Let $\pi_r \in \mathbb{N}$ be certain natural number parameters, which we call periods of natural numbers $r \in \mathbb{N}$. Let $\sigma_r \in \{1, 2, ..., \pi_r\}$ be another sequence of parameters, which we call seeds of natural numbers $r \in \mathbb{N}$. To define sequence $(k_t)_{t\in\mathbb{N}}$, let $(k_t^{(r)})_{t\in\mathbb{N}}$ be the subsequence of $(k_t)_{t\in\mathbb{N}}$ from which we have removed all tokens $k_t < r$. Then we require a partial periodicity of this decimated subsequence, namely, that exactly every π_r-th token equals r. In formula, we require

$$k_t^{(r)} = r \iff t \equiv \sigma_r \mod \pi_r. \tag{14.65}$$

If the above constraint leaves a certain k_t undefined then we put $k_t = \infty$. Sequence $(k_t)_{t\in\mathbb{N}}$ given by these two conditions is called the multiperiodic sequence with periods $(\pi_r)_{r\in\mathbb{N}}$ and seeds $(\sigma_r)_{r\in\mathbb{N}}$.

For example, let $\pi_r = 1 + r$ and $\sigma_r = 1$ for all $r \in \mathbb{N}$. Then the multiperiodic sequence is

$$1, 2, 1, 3, 1, 4, 1, 2, 1, 5, 1, 6, 1, 2, 1, 3, 1, 7, 1, 2, 1, 8, 1, ... \tag{14.66}$$

Show that for periods $\pi_r \approx cr$, the relative frequency of a natural number r is asymptotically proportional to $r^{-\alpha}$ with $\alpha = (c + 1)/c$.

Programming exercises

1. *Establishing a text repository:* Make a repository of experimental text files for further programming exercises. Collect some texts in natural language (can be different human languages), DNA sequences, music in the MIDI format, sequences generated by pseudo-random number generators, and other discrete symbolic sequences. The suggested volume of the repository should be between 10 an 100 megabytes and the amount of text of each kind should be roughly the same.

2. *Codes for natural numbers:* Write a program that for a given natural number computes its military code and all Elias codes mentioned in Chapter 1, namely, the alpha, beta, gamma, delta, and omega codes. Generate the table of respective code words for numbers in range $[1, 20]$.

3. *Shannon entropy vs. Shannon-Fano and Huffman codes:* Write a program that for a given probability distribution computes the Huffman code, the Shannon-Fano code, and the Shannon entropy. For each file in your repository of text files, apply this program to the empirical probability distribution of characters, i.e., the probabilities are defined as the relative frequencies of characters.

4. *Empirical decay of mutual information:* For each kind of a text in your repository of text files, find the mutual information between two characters separated by k characters according to the empirical probability distribution. Plot the mutual information in a doubly logarithmic scale and try to fit some function to the data points.

5. *The law of large numbers and the central limit theorem:* Consider nine Bernoulli(θ) processes with $\theta \in \{0.1, 0.2, ..., 0.9\}$. For each θ, generate one sample y_1^n drawn IID from the respective distribution (using a pseudo-random number generator). Check how large n do we need so that $\left|\frac{1}{n}\sum_{i=1}^{n} y_i - \theta\right| \leq 0.01$. Fix some θ and generate 1000 samples y_1^n,

where $n = 100$. Plot a histogram of $z_n = (\frac{1}{n}\sum_{i=1}^{n} y_i - \theta)/\sqrt{n\theta(1-\theta)}$. Compare it with the normal distribution function

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{14.67}$$

6. *Convergence of empirical entropy:* Generate randomly 10 distributions $p(x)$, where $x \in \{a, b, ..., z\}$. For each of these distributions p, compute the Shannon entropy $H(p)$. For each of these distributions p, generate a sample of x_1^n drawn IID from the respective distribution. For each sample x_1^n, compute the empirical entropy $\mathcal{H}(x_1^n)$. Check how large n suffices so that $\mathcal{H}(x_1^n) - H(p) \leq H(p)/100$.

7. *Ergodic properties for Markov processes:* Generate randomly one hundred 5×5 transition matrices τ with randomly scattered zeros. Check automatically which of these transition matrices are irreducible. Check whether $\lim_{n\to\infty} \tau^n$ converges. Check whether $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \tau^i$ converges. Estimate numerically the invariant distribution.

8. *Implementation of universal codes:* Write a program that, for a given text, computes the Lempel-Ziv code, the minimal block code, the PML maximum, and the PPM mixture. For each kind of a text in your repository of text files, evaluate each of these four universal codes and report in a table the respective entropy rate estimate achieved by these codes with the best one bolded out.

9. *Statistical language models:* From your repository, select some texts in your native language or music files, compute the maximum likelihood transition matrices $\hat{\tau}$ for Markov orders $k = 0, 1, 2, ..., 20$. Generate some artificial texts according to these transition matrices. Do they resemble texts in your native language? Is the artificial music pleasant to listen? Can you describe qualitatively what is good and what is wrong depending on the Markov order k?

10. *Casinos and martingales:* Simulate numerically a casino roulette, i.e., a wheel with 37 pockets numbered from 0 to 36. Allow to place bets on odd or even numbers (without 0!). If you bet n dollars on either odd or even numbers and the ball falls on the respective number then you receive $2n$ dollars back, otherwise you receive 0 dollars. If there were no 0 pocket, the capital earned in this game would be a martingale process. If the 0 pocket is taken into account, the casino has a positive chance of income in each round.

The original meaning of word "martingale" described a gambling strategy in which the bet is doubled after each loss and halved after each winning. Starting with a capital of n dollars and an initial bet of 1 dollar, try to estimate numerically the number of rounds needed for the gambler's ruin in the martingale strategy. How does it depend on n?

11. *Normalized information distance:* Define complexity $\tilde{C}(u)$ as the length of a certain universal code for string u. Define the conditional complexity $\tilde{C}(w|u) := \tilde{C}(uw) - \tilde{C}(u)$. Consider the normalized information distance between strings u and w, defined as

$$d(u,w) := \frac{\max\left\{\tilde{C}(u|w), \tilde{C}(w|u)\right\}}{\max\left\{\tilde{C}(u), \tilde{C}(w)\right\}}. \tag{14.68}$$

Can we use it as a metric on strings and produce some sensible classification of symbolic sequences regardless of their length? Apply the normalized information distance to texts from your text repository and check whether it can be used for clustering similar texts (for example by the k-means algorithm). See also [22].

12. *Markov order estimates and maximal repetition length:* Write a program that, for a given text, compute the the Markov order estimator and the maximal length of a repetition. Check empirically for texts in your repository that the Markov order estimator is smaller than the the maximal length of a repetition. Check also how fast these two quantities grow with respect to the text size.

13. *Hilberg's law:* Consider the four codes implemented in Exercise 8 and the texts from your text repository. Try to estimate the Hilberg exponents of these codes for these texts.

14. *Zipf's and Herdan-Heaps' laws:* Generate samples of the Zipf process, the Miller rank process, and a prefix of the multiperiodic sequence. Compare how the number of types grows for these three sources with the length of the sample. Compare also the empirical rank-frequency plots for these sources with the theoretical distributions.

Bibliography

[1] P. H. Algoet. The strong law of large numbers for sequential decisions under uncertainty. *IEEE Transactions on Information Theory*, 40(3): 609–633, 1994.

[2] P. H. Algoet and T. M. Cover. A sandwich proof of the Shannon-McMillan-Breiman theorem. *The Annals of Probability*, 16:899–909, 1988.

[3] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[4] D. H. Bailey. *Sequential schemes for classifying and predicting ergodic processes*. PhD thesis, Stanford University, 1976.

[5] A. R. Barron. *Logically Smooth Density Estimation*. PhD thesis, Stanford University, 1985.

[6] J. Bernoulli. *The Art of Conjecturing, together with Letter to a Friend on Sets in Court Tennis*. Baltimore: Johns Hopkins University Press, 2005.

[7] L. Bienvenu, A. R. Day, M. Hoyrup, I. Mezhirov, and A. Shen. A constructive version of Birkhoff's ergodic theorem for Martin-Löf random points. *Information and Computation*, 210:21–30, 2012.

[8] P. Billingsley. *Probability and Measure*. New York: Wiley & Sons, 1979.

[9] G. D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences of the United States of America*, 17:656–660, 1932.

[10] E. Borel. Les probabilités dénombrables et leurs applications arithmetiques. *Rendiconti del Circolo Matematico di Palermo*, 2(27):247–271, 1909.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge: Cambridge University Press, 2004.

[12] M. Braverman, X. Chen, S. M. Kakade, K. Narasimhan, C. Zhang, and Y. Zhang. Calibration, entropy rates, and memory in language models. In *2020 International Conference on Machine Learning (ICML)*, 2020.

[13] L. Breiman. The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, 28:809–811, 1957.

[14] L. Breiman. *Probability*. Philadephia: SIAM, 1992.

[15] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, G. K. Ariel Herbert-Voss, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, M. L. Eric Sigler, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *2020 Conference on Neural Information Processing Systems (NIPS)*, 2020.

[16] F. P. Cantelli. Sulla probabilità come limite della frequenza. *Atti dell'Accademia Nazionale dei Lincei*, 26(1):39–45, 1917.

[17] G. Cantor. Über eine elementare Frage der Mannigfaltigkeitslehre. In W. B. Ewald, editor, *From Immanuel Kant to David Hilbert: A Source Book in the Foundations of Mathematics*, volume 2, pages 920–922. Oxford: Oxford University Press, 1996.

[18] G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22:329–340, 1975.

[19] G. J. Chaitin. Randomness and mathematical proof. *Scientific American*, 232(5):47–52, 1975.

[20] M. Charikar, E. Lehman, A. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, and A. Shelat. The smallest grammar problem. *IEEE Transactions on Information Theory*, 51:2554–2576, 2005.

[21] S.-T. Chen Moy. Successive recurrence times in a stationary process. *The Annals of Mathematical Statistics*, 30(4):1254–1257, 1959.

[22] R. Cilibrasi and P. M. B. Vitanyi. Automatic meaning discovery using Google. `http://www.arxiv.org/abs/cs.CL/0412098`, 2004.

[23] J. G. Cleary and I. H. Witten. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, 32:396–402, 1984.

[24] P. Collet and J.-P. Eckmann. Oscillations of observables in 1-dimensional lattice systems. `https://arxiv.org/abs/cond-mat/9705175`, 1997.

[25] E. U. Condon. Statistics of vocabulary. *Science*, 67(1733):300–300, 1928.

[26] T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd ed.* New York: Wiley & Sons, 2006.

[27] T. M. Cover, P. Gacs, and R. M. Gray. Kolmogorov's contributions to information theory and algorithmic complexity. *The Annals of Probability*, 17:840–865, 1989.

[28] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos*, 15:25–54, 2003.

[29] I. Csiszar. The method of types. *IEEE Transactions on Information Theory*, 44:2505–2523, 1998.

[30] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge: Cambridge University Press, 2011.

[31] C. G. de Marcken. *Unsupervised Language Acquisition*. PhD thesis, Massachussetts Institute of Technology, 1996.

[32] Ł. Dębowski. A general definition of conditional information and its application to ergodic decomposition. *Statistics and Probability Letters*, 79:1260–1268, 2009.

[33] Ł. Dębowski. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Transactions on Information Theory*, 57:4589–4599, 2011.

[34] Ł. Dębowski. *Information Theory and Statistics*. Institute of Computer Science, Polish Academy of Sciences, 2013.

[35] Ł. Dębowski. Approximating information measures for fields. *Entropy*, 22(1):79, 2020.

[36] Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. New York: Wiley & Sons, 2021.

[37] Ł. Dębowski. A refutation of finite-state language models through Zipf's law for factual knowledge. *Entropy*, 23:1148, 2021.

[38] Ł. Dębowski. A simplistic model of neural scaling laws: Multiperiodic Santa Fe processes. `https://arxiv.org/abs/2302.09049`, 2023.

[39] Ł. Dębowski. Universal densities exist for every finite reference measure. *IEEE Transactions on Information Theory*, 69(8):5277–5288, 2023.

[40] Ł. Dębowski and T. Steifer. Universal coding and prediction on ergodic random points. *The Bulletin of Symbolic Logic*, 28(2):387–412, 2022.

[41] R. L. Dobrushin. A general formulation of the fundamental Shannon theorems in information theory. *Uspekhi Matematicheskikh Nauk*, 14(6):3–104, 1959. In Russian.

[42] J. L. Doob. *Stochastic processes*. New York: Wiley & Sons, 1953.

[43] R. G. Downey and D. R. Hirschfeldt. *Algorithmic Randomness and Complexity*. New York: Springer, 2010.

[44] P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21:194–203, 1975.

[45] J. B. Estoup. *Gammes sténographiques*. Paris: Institut Stenographique de France, 1916.

[46] R. M. Fano. The transmission of information. Technical Report 65, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1949.

[47] R. M. Fano. *Transmission of Information*. Cambridge, MA: The MIT Press, 1961.

[48] S. Fehr and S. Berens. On the conditional Rényi entropy. *IEEE Transactions on Information Theory*, 60:6801–6810, 2014.

[49] R. Feistel and W. Ebeling. *Physics of Self-Organization and Evolution*, chapter Self-Organization of Information and Symbols. New York: Wiley & Sons, 2011.

[50] M. Fekete. Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. *Mathematische Zeitschriften*, 17:228–249, 1923.

[51] W. Feller. *An introduction to probability theory and its application.* New York: Wiley & Sons, 1967.

[52] J. N. Y. Franklin, N. Greenberg, J. S. Miller, and K. M. Ng. Martin-Löf random points satisfy Birkhoff's ergodic theorem for effectively closed sets. *Proceedings of the American Mathematical Society*, 140: 3623–3628, 2012.

[53] P. Gács. On the symmetry of algorithmic information. *Doklady Akademii Nauk*, 15:1477–1480, 1974.

[54] A. M. Garsia. A simple proof of E. Hopf's maximal ergodic theorem. *Journal of Mathematics and Mechanics*, 14:381–382, 1965.

[55] I. M. Gelfand, A. N. Kolmogorov, and A. M. Yaglom. Towards the general definition of the amount of information. *Doklady Akademii Nauk*, 111:745–748, 1956. In Russian.

[56] D. Gillman and R. L. Rivest. Complete variable-length "fix-free" codes. *Designs, Codes and Cryptography*, 5:109–114, 1995.

[57] K. Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I. *Monatshefte für Mathematik und Physik*, 38(1):173–198, 1931.

[58] R. M. Gray. *Probability, Random Processes, and Ergodic Properties.* New York: Springer, 2009.

[59] R. M. Gray and J. C. Kieffer. Asymptotically mean stationary measures. *The Annals of Probability*, 8:962–973, 1980.

[60] P. D. Grünwald. *The Minimum Description Length Principle.* Cambridge, MA: The MIT Press, 2007.

[61] L. Györfi and G. Lugosi. Strategies for sequential prediction of stationary time series. In M. Dror, P. L'Ecuyer, and F. Szidarovszky, editors, *Modeling Uncertainty: An examination of its theory, methods, and applications.* Dordrecht: Kluwer Academic Publishers, 2001.

[62] L. Györfi, I. Páli, and E. C. van der Meulen. There is no universal source code for infinite alphabet. *IEEE Transactions on Information Theory*, 40:267–271, 1994.

[63] L. Györfi, G. Lugosi, and G. Morvai. A simple randomized algorithm for consistent sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, 45(7):2642–2650, 1999.

[64] M. Hahn and R. Futrell. Estimating predictive rate-distortion curves via neural variational inference. *Entropy*, 21:640, 2019.

[65] H. S. Heaps. *Information Retrieval—Computational and Theoretical Aspects*. New York: Academic Press, 1978.

[66] T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. `https://arxiv.org/abs/2010.14701`, 2020.

[67] G. Herdan. *Quantitative Linguistics*. London: Butterworths, 1964.

[68] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish. Scaling laws for transfer. `https://arxiv.org/abs/2102.01293`, 2021.

[69] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. `https://arxiv.org/abs/1712.00409`, 2017.

[70] W. Hilberg. Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente? *Frequenz*, 44:243–248, 1990.

[71] D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the Institute of Radio Engineers*, 40:1098–1101, 1952.

[72] V. V. Ivanov. Geometric properties of monotone functions and probabilities of random fluctuations. *Sibirskii Matematicheskii Zhurnal*, 37: 117–150, 1996. In Russian.

[73] V. V. Ivanov. Oscillations of means in the ergodic theorem. *Doklady Akademii Nauk*, 347:736–738, 1996. In Russian.

[74] F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, 1976.

[75] F. Jelinek. *Statistical Methods for Speech Recognition*. Cambridge, MA: The MIT Press, 1997.

[76] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906.

[77] M. Kac. On the notion of recurrence in discrete stochastic processes. *The Bulletin of American Mathematical Society*, 53:1002–1010, 1947.

[78] O. Kallenberg. *Foundations of Modern Probability*. New York: Springer, 1997.

[79] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. `https://arxiv.org/abs/2001.08361`, 2020.

[80] E. Khmaladze. The statistical analysis of large number of rare events. Technical Report MS-R8804. Centrum voor Wiskunde en Informatica, Amsterdam, 1988.

[81] J. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24:674–682, 1978.

[82] J. C. Kieffer and E. Yang. Grammar-based codes: A new class of universal lossless source codes. *IEEE Transactions on Information Theory*, 46:737–754, 2000.

[83] A. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer, 1933.

[84] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

[85] L. Kraft. A device for quantizing, grouping, and coding amplitude modulated pulses. Master's thesis, Massachusetts Institute of Technology, 1949.

[86] R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, IT-27:199–207, 1981.

[87] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[88] P.-S. Laplace. *Essai philosophique sur les probabilités*. Paris: Courcier, 1814.

[89] G. Leibniz. Explication de l'Arithmétique Binaire. In C. Gerhardt, editor, *Die Mathematische Schriften*, volume 7, page 223. Berlin, 1879.

[90] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications, 3rd ed.* New York: Springer, 2008.

[91] B. Mandelbrot. Structure formelle des textes et communication. *Word*, 10:1–27, 1954.

[92] A. A. Markov. Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain. *Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg*, 7:153–162, 1913.

[93] A. A. Markov. An example of statistical investigation of the text 'Eugene Onegin' concerning the connection of samples in chains. *Science in Context*, 19:591–600, 2006.

[94] P. Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[95] B. McMillan. Two inequalities implied by unique decipherability. *IEEE Transactions on Information Theory*, IT-2:115–116, 1956.

[96] N. Merhav, M. Gutman, and J. Ziv. On the estimation of the order of a Markov chain and universal data compression. *IEEE Transactions on Information Theory*, 35(5):1014–1019, 1989.

[97] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119, 2013.

[98] G. A. Miller. Some effects of intermittent silence. *American Journal of Psychology*, 70:311–314, 1957.

[99] G. Morvai and B. Weiss. On universal algorithms for classifying and predicting stationary processes. *Probability Surveys*, 18:77–131, 2021.

[100] D. Neuhoff and P. C. Shields. Simplistic universal coding. *IEEE Transactions on Information Theory*, IT-44:778–781, 1998.

[101] J. R. Norris. *Markov Chains*. Cambridge: Cambridge University Press, 1997.

[102] A. Orlitsky, N. P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50:1469–1481, 2004.

[103] D. S. Ornstein. Guessing the next output of a stationary process. *Israel Journal of Mathematics*, 30(3):292–296, 1978.

[104] S. D. Poisson. *Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Paris: Bachelier, 1837.

[105] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. `https://openai.com/blog/better-language-models/`, 2019.

[106] A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. 1961.

[107] B. Ryabko. Twice-universal coding. *Problems of Information Transmission*, 20(3):173–177, 1984.

[108] B. Ryabko. Compression-based methods for nonparametric density estimation, on-line prediction, regression and classification for time series. In *2008 IEEE Information Theory Workshop, Porto*, pages 271–275. Institute of Electrical and Electronics Engineers, 2008.

[109] B. Ryabko. Compression-based methods for nonparametric prediction and estimation of some characteristics of time series. *IEEE Transactions on Information Theory*, 55(9):4309–4315, 2009.

[110] B. Ryabko. Applications of universal source coding to statistical analysis of time series. In I. Woungang, S. Misra, and S. C. Misra, editors, *Selected Topics in Information and Coding Theory*, Series on Coding and Cryptology. World Scientific Publishing, 2010.

[111] B. Ryabko, J. Astola, and M. Malyutov. *Compression-Based Methods of Statistical Analysis and Prediction of Time Series*. New York: Springer, 2016.

[112] B. Y. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24(2):87–96, 1988.

[113] A. A. Sardinas and G. W. Patterson. A necessary and sufficient condition for the unique decomposition of coded messages. In *Convention Record of the IRE, 1953 National Convention, Part 8: Information Theory*, pages 104–108. 1953.

[114] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 30:379–423,623–656, 1948.

[115] C. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64, 1951.

[116] Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(2):3–17, 1987.

[117] R. J. Solomonoff. A formal theory of inductive inference, part 1 and part 2. *Information and Control*, 7:1–22, 224–254, 1964.

[118] S. M. Stigler. The epic story of maximum likelihood. *Statistical Science*, 22(4):508–620, 2007.

[119] J. Suzuki. Universal prediction and universal coding. *Systems and Computers in Japan*, 34(6):1–11, 2003.

[120] R. Takahira, K. Tanaka-Ishii, and Ł. Dębowski. Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10):364, 2016.

[121] K. Tanaka-Ishii. *Statistical Universals of Language: Mathematical Chance vs. Human Choice*. New York: Springer, 2021.

[122] P. Tchebichef. Démonstration élémentaire d'une proposition générale de la théorie des probabilités. *Journal für die Reine und Angewandte Mathematik*, 33:259–267, 1846.

[123] A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230–265, 1936.

[124] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[125] A. D. Wyner. A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38:51–59, 1978.

[126] R. W. Yeung. *First Course in Information Theory*. Dordrecht: Kluwer Academic Publishers, 2002.

[127] Z. Zhang and R. W. Yeung. A non-Shannon-type conditional inequality of information quantities. *IEEE Transactions on Information Theory*, 43:1982, 1997.

[128] M. Zimand. Symmetry of information: A closer look. In M. J. Dinneen, B. Khoussainov, and A. Nies, editors, *Computation, Physics and Beyond. WTCS 2012*, Lecture Notes in Computer Science 7160. New York: Springer, 2012.

[129] G. K. Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston: Houghton Mifflin, 1935.

[130] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343, 1977.

[131] J. Ziv and N. Merhav. Estimating the number of states of a finite-state source. *IEEE Transactions on Information Theory*, 38(1):61–65, 1992.

[132] A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25: 83–124, 1970.

Index