Redundancy of Lossless Data Compression for Known Sources by Analytic Methods

Michael Drmota
TU Wien
michael.drmota@tuwien.ac.at

Wojciech Szpankowski
Purdue University
szpan@purdue.edu
Editorial Scope

Topics

Foundations and Trends® in Communications and Information Theory publishes survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2016, Volume 13, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328. Also available as a combined paper and online subscription.
Redundancy of Lossless Data Compression for Known Sources by Analytic Methods

Michael Drmota
TU Wien
michael.drmota@tuwien.ac.at

Wojciech Szpankowski
Purdue University
szpan@purdue.edu
Contents

1 Introduction 2

2 Preliminary Results 7
   2.1 Prefix Codes and Their Properties 8
   2.2 Redundancy 11

3 Redundancy of Shannon and Huffman VF Codes 13
   3.1 Average Redundancy for Memoryless Sources 13
   3.2 Shannon Code Redundancy for Markov Sources 26
   3.3 Maximal Redundancy for a Generalized Shannon Code 40

4 Redundancy of Tunstall and Khodak VF Codes 46
   4.1 Variable-to-Fixed Codes 46
   4.2 Redundancy of the Khodak VF Code 49
   4.3 Analysis of the Tunstall Code 60

5 Redundancy of Divide-and-Conquer VF Arithmetic Coding 63
   5.1 Redundancy of Boncelet’s Code 64
   5.2 Divide-and-Conquer Recurrence 68
   5.3 Central Limit Law for Boncelet’s Algorithms 82
# 6 Redundancy of VV Khodak Codes

6.1 Variable-to-Variable Codes and Redundancy Rate . . . . . 85
6.2 Redundancy of the Khodak VV Code . . . . . . . . . . . . 86
6.3 Explicit Construction of a Khodak VV Code . . . . . . . 95
6.4 Khodak's Redundancy for Almost All Sources . . . . . . 100

# 7 Redundancy of Non Prefix One-to-One Codes

7.1 Binary One-to-One Code . . . . . . . . . . . . . . . . . . . . . 110
7.2 Non-binary One-to-One Code . . . . . . . . . . . . . . . . . 117

# 8 Concluding Remarks

Acknowledgements 128
References 129
Abstract

Lossless data compression is a facet of source coding and a well studied problem of information theory. Its goal is to find a shortest possible code that can be unambiguously recovered. Here, we focus on rigorous analysis of code redundancy for known sources. The redundancy rate problem determines by how much the actual code length exceeds the optimal code length. We present precise analyses of three types of lossless data compression schemes, namely fixed-to-variable (FV) length codes, variable-to-fixed (VF) length codes, and variable-to-variable (VV) length codes. In particular, we investigate the average redundancy of Shannon, Huffman, Tunstall, Khodak and Boncelet codes. These codes have succinct representations as trees, either as coding or parsing trees, and we analyze here some of their parameters (e.g., the average path from the root to a leaf). Such trees are precisely analyzed by analytic methods, known also as analytic combinatorics, in which complex analysis plays decisive role. These tools include generating functions, Mellin transform, Fourier series, saddle point method, analytic poissonization and depoissonization, Tauberian theorems, and singularity analysis. The term analytic information theory has been coined to describe problems of information theory studied by analytic tools. This approach lies on the crossroad of information theory, analysis of algorithms, and combinatorics.
Introduction

The basic problem of source coding better known as (lossless) data compression is to find a binary code that can be unambiguously recovered with shortest possible description either on average or for individual sequences. Thanks to Shannon’s work we know that on average the number of bits per source symbol cannot be smaller than the source entropy rate. There are many codes asymptotically achieving the entropy rate, therefore one turns attention to redundancy. The average redundancy of a source code is the amount by which the expected number of binary digits per source symbol for that code exceeds entropy. One of the goals in designing source coding algorithms is to minimize the redundancy. In this survey, we discuss various classes of source coding and their corresponding redundancy. It turns out that such analyses often resort to studying certain intriguing trees such as Huffman, Tunstall, Khodak and Boncelet trees, as well as various algorithms such as divide-and-conquer approach. We study them using tools from the analysis of algorithms and analytic combinatorics\(^1\) to discover precise and minute behavior of lossless compression codes.

\(^1\)Andrew Odlyzko has argued that: “analytic methods are extremely powerful and when they apply, they often yield estimates of unparalleled precision.”
Lossless data compression comes in three flavors: fixed-to-variable (FV) length codes, variable-to-fixed (VF) length codes, and finally variable-to-variable (VV) length codes. The latter includes the previous two families of codes and is the least studied among all data compression schemes. Over years we have seen a resurgence of interest in redundancy rate for fixed-to-variable coding (cf. [25, 28, 29, 30, 66, 90, 91, 92, 101, 103, 124, 126, 130, 132, 131, 139, 140, 151, 152, 164, 173, 180, 176, 177]). Surprisingly there are only a handful of results for variable-to-fixed codes (cf. [77, 97, 112, 133, 131, 134, 156, 161, 185]) and an almost non-existing literature on variable-to-variable codes (cf. [42, 50, 80, 97]). While there is some work on universal VF codes [156, 161, 185], to the best of our knowledge redundancy for universal VF and VV codes were not studied with the exception of some work of the Russian school [97, 96] (cf. also [99]).

In the fixed-to-variable code, discussed in Chapter 3, the encoder maps fixed length blocks of source symbols into variable-length binary code strings. Two important fixed-to-variable length coding schemes are the Shannon code and the Huffman code. In this survey we follow [152, 114]. We first discuss precise analyses of Shannon code redundancy for memoryless and Markov sources. We show that the average redundancy either converges to an explicitly computable constant, as the block length increases, or it exhibits a very erratic behavior fluctuating between 0 and 1. We also observe a similar behavior for the worst case or maximal redundancy. Then we move to the Huffman code. Despite the fact that Huffman codes have been so well known for so long, it was only relatively recently that their redundancy was fully understood. In [1] Abrahams summarizes much of the vast literature on fixed-to-variable length codes. Here, we present a precise analysis from our work [152] of the Huffman average redundancy for memoryless sources. We show that the average redundancy either converges to an explicitly computable constant, as the block length increases, or it exhibits a very erratic behavior fluctuating between 0 and 1. Following [114] we also present similar results for Markov sources.

Next, in Chapter 4 we study variable-to-fixed codes. A VF encoder partitions the source string into variable-length phrases that belong to
a given dictionary $\mathcal{D}$. Often a dictionary is represented by a complete
tree (i.e., a tree in which every node has maximum degree), also known
as the parsing tree. The code assigns a fixed-length word to each dic-
tionary entry. An important example of a variable-to-fixed code is the
Tunstall code $\mathcal{T}$. Savari and Gallager $\mathcal{S}$ present an analysis of
the dominant term in the asymptotic expansion of the Tunstall code
redundancy. In this survey, following $\mathcal{F}$, we describe a precise analysis
of the phrase length (i.e., path from the root to a terminal node in the
 corresponding parsing tree) for such a code and its average redundancy.
We also discuss a variant of Tunstall code known as VF Khodak code.

In the next Chapter 5 we continue analyzing VF codes due to Bon-
celet $\mathcal{B}$ who used the divide-and-conquer principle to design a prac-
tical encoding. Boncelet’s algorithm is computationally fast and its
practicality stems from the divide and conquer strategy: It splits the
input (e.g., parsing tree) into several smaller subproblems, solving each
subproblem separately, and then knitting together to solve the original
 problem. We use this occasion to present a careful analysis of a
divide-and-conquer recurrence which is at foundation of several divide-
and-conquer algorithms such as heapsort, mergesort, discrete Fourier
transform, queues, sorting networks, compression algorithms, and so
forth $\mathcal{G}$.

In Chapter 6 we consider variable-to-variable codes. A variable-to-
variable (VV) code is a concatenation of variable-to-fixed and fixed-
to-variable codes. A variable-to-variable length encoder consists of a
parser and a string encoder. The parser, as in VF codes, segments the
source sequence into a concatenation of phrases from a predetermined
dictionary $\mathcal{D}$. Next, the string encoder in a variable-to-variable scheme
takes the sequence of dictionary strings and maps each one into its
corresponding binary codeword of variable length. Aside from the spe-
cial cases where either the dictionary strings or the codewords have
a fixed length, very little is known about variable-to-variable length
codes, even in the case of memoryless sources. In 1972 Khodak $\mathcal{K}$
described a VV scheme with small average redundancy that decreases
with the growth of phrase length. He did not offer, however, an explicit
VV code construction. We will remedy this situation and follow $\mathcal{K}$. 

Full text available at: http://dx.doi.org/10.1561/0100000090
Finally, in Chapter 7 we discuss redundancy of one-to-one codes that are not necessarily prefix or even uniquely decodable. Recall that non-prefix codes are such codes which are not prefix free and do not satisfy Kraft’s inequality. In particular, we analyze binary and non-binary one-to-one codes whose average lengths are smaller than the source entropy in defiance of the Shannon lower bound.

Throughout this survey, we study various intriguing trees describing Huffman, Tunstall, Khodak and Boneclet codes. These trees are studied by analytic techniques of analysis of algorithms \[47, 85, 86, 87, 153\]. The program of applying tools from analysis of algorithms to problems of source coding and in general to information theory lies at the crossroad of computer science and information theory. It is also known as analytic information theory. In fact, the interplay between information theory and computer science dates back to the founding father of information theory, Claude E. Shannon. His landmark paper “A Mathematical Theory of Communication” is hailed as the foundation for information theory. Shannon also worked on problems in computer science such as chess-playing machines and computability of different Turing machines. Ever since Shannon’s work on both information theory and computer science, the research at the interplay between these two fields has continued and expanded in many exciting ways. In the late 1960s and early 1970s, there were tremendous interdisciplinary research activities, exemplified by the work of Kolmogorov, Chaitin, and Solomonoff, with the aim of establishing algorithmic information theory. Motivated by approaching Kolmogorov complexity algorithmically, A. Lempel (a computer scientist), and J. Ziv (an information theorist) worked together in the late 1970s to develop compression algorithms that are now widely referred to as Lempel-Ziv algorithms. Analytic information theory is a continuation of these efforts.

Finally, we point out that this survey deals only with source coding for known sources. The more practical universal source coding (in which the source distribution is unknown) is left for our future book Analytic Information Theory. However, at the end of this survey we provide an extensive bibliography on the redundancy rate problem, including universal source coding.
This survey is organized as follows. In the next chapter, we present some preliminary results such as Kraft’s inequality, Shannon’s lower bound, and Barron’s lemma. In Section 3 we analyze Shannon and Huffman codes. Then we turn our attention in Section 4 to the Tunstall and VF Khodak codes. Finally, in Section 6 we discuss the VV code of Khodak and its interesting analysis. We conclude this survey with a chapter concerning the average redundancy for non-prefix codes such as one-to-one codes.
References


References


References


References


References


References


References


References


