On the Construction of (Explicit) Khodak's Code and Its Analysis[*]

May 10, 2008

Yann Bugeaud[†]
Dépt. de Mathématiques
Université Louis Pasteur
F-67084 Strasbourg
France
bugeaud@math.u-strasbg.fr

Michael Drmota[‡]
Inst. Diskr. Math. u. Geometrie
TU Wien
A-1040 Wien,
Austria
michael.drmota@tuwien.ac.at

Wojciech Szpankowski[§]
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

Variable-to-variable codes are very attractive yet not well understood data compression schemes. In 1972 Khodak claimed to provide upper and lower bounds for the achievable redundancy rate, however, he did not offer explicit construction of such codes. In this paper, we first present a constructive and transparent proof of Khodak's result showing that for memoryless sources there exists a code with the average redundancy bounded by $D^{-5/3}$, where D is the average delay (e.g., the average length of a dictionary entry). We also describe an algorithm that constructs a variable-to-variable length code with a small redundancy rate for large D. Then, we discuss several generalizations. We prove that the worst case redundancy does not exceed $D^{-4/3}$. Furthermore we provide similar upper bounds for Markov sources (of order 1). Finally, we consider bounds that are valid for *almost all* memoryless and Markov sources for which the set of exceptional source parameters has zero measure. In particular, for all memoryless sources outside this exceptional class, we prove there exists a variable-to-variable code with the average redundancy rate bounded by $D^{-4/3-m/3+\varepsilon}$ and the worst case redundancy rate bounded by $D^{-1-m/3+\varepsilon}$, where m is the cardinality of the alphabet. We complete our analysis with a lower bound showing that for all variable-to-variable codes the average and the worst case redundancy rates are at least $D^{-2m-1-\varepsilon}$ for almost all memoryless sources in the sense that the set of exceptional source parameters has zero measure. We prove these results using techniques of Diophantine approximations.

Index Terms: Variable-to-variable length codes, average and maximal redundancy rates, metric Diophantine approximations.

[*] A preliminary version of some parts of this paper was presented at *2004 International Symposium on Information Theory*, Chicago 2004, and *42-nd Allerton Conference*, Illinois, 2004.

[†] The work of this author was supported the Austrian Science Foundation FWF, Grant M882-N12.

[‡] The work of this author was supported the Austrian Science Foundation FWF Grant No. S9604.

[§] The work of this author was supported by NSF Grants CCF-0513636, and DMS-0503742, NIH Grant R01 GM068959-01, the AFOSR Grant FA8655-08-1-3018, and NSA Grant 07G-044.

1 Introduction

A variable-to-variable (VV) length code partitions a source sequence into variable length phrases that are encoded into strings of variable lengths. While it is well known that every VV (prefix) code is a concatenation of a variable-to-fixed length code (e.g., Tunstall code) and a fixed-to-variable length encoding (e.g., Huffman code), an optimal VV code has not yet been found. Fabris [9] proved that greedy, step by step, optimization (that is, a concatenation of Tunstall and Huffman codes) does not lead to an optimal VV code. In order to assess performance of VV codes, one needs to evaluate (at least asymptotically) the redundancy rate of (optimal) VV codes, which is still unknown. By *redundancy rate* we mean the excess of the code length over the optimal code length per source symbol. Our goal is to shed some light on the (average and maximal) redundancy rates of VV codes by re-examining and expanding a thirty year old paper by Khodak [14], who in 1972 claimed to provide upper and lower bounds for the achievable redundancy rate of VV length codes. However, Khodak did not offer explicit VV length codes that satisfy these bounds. Here, we present a transparent (and simplified) proof, generalize Khodak's results (i.e., we analyze maximal redundancy, Markov sources of order 1, *typical sources* in the sense that the exceptional set in the parameter space has zero measure), and describe an explicit algorithm that constructs a VV code with redundancy rates decaying to zero as the average delay increases.

Let us first briefly describe a VV encoder. A VV encoder has two components, a *parser* and a *string encoder*. The parser partitions the source sequence x into phrases $x^1, x^2, \ldots$ from a predetermined dictionary $\mathcal{C}$. We shall write d or d_i for a dictionary entry, and by D we denote the average dictionary (phrase) length also known as the average delay. A convenient way of representing the dictionary $\mathcal{C}$ is by a complete tree that we shall call the *parsing tree*. Next, the string encoder in a VV scheme maps each dictionary phrase into its corresponding binary codeword $C(d)$ of length $|C(d)| = \ell(d)$. Throughout this paper, we assume that the string encoder is a slightly modified Shannon code[1] and we concentrate on building a parsing tree for which $\log P(d)$ $(d \in \mathcal{C})$ is close to an integer. This allows us to construct a VV code with redundancy rates (per symbol) approaching zero as the average delay increases.

More precisely, for large delay D we shall show in Theorem 1 that there exist VV codes such that for memoryless sources the average redundancy rates *decay* as $D^{-5/3}$. This result basically belongs to Khodak [14], except we present here a transparent proof and an easily constructible VV code. Next, we extend this result in several directions. First, we show that for such codes the worst case redundancy rates decay as $D^{-4/3}$. Similar bounds hold also for Markov sources. More importantly, we study new bounds for *almost all* memoryless and Markov sources, that is, we prove bounds that hold for all possible source parameters with an exception of a set that has zero measure in the parameter space.[2] In particular, we

[1] A variant of Shannon code that is used here assigns to $d \in \mathcal{C}$ a binary word of length $\ell(d)$ close to $-\log P(d)$ when $\log P(d)$ is slightly larger or smaller than an integer. Naturally, Kraft's inequality will not be automatically satisfied but this is handled in Lemma 6 when proving Theorem 1.

[2] For example, if we consider memoryless sources with parameters $p_1, p_1, \ldots, p_m$, then the term "almost

show that for almost all memoryless sources there exists a VV code such that its average redundancy rate is bounded by $D^{-4/3-m/3+\varepsilon}$ and the worst case redundancy by $D^{-1-m/3+\varepsilon}$, where m is the alphabet size. We conclude our analysis with a lower bound showing that for all VV codes and for almost all memoryless sources the average and the worst case redundancy rates are at least $D^{-2m-1-\varepsilon}$. The latter result seems to contradict one of the lower bounds proposed by Khodak.

The results of this paper should be compared to redundancy rates of fixed-to-variable (FV) code lengths (e.g., Shannon and Huffman codes) and variable-to-fixed code lengths (e.g., Tunstall codes). Abrahams [1] discusses literature on fixed-to-variable length codes. For a memoryless source, [21] provides an asymptotic analysis of the Huffman and other codes for fixed length blocks of source symbols. While it has been known since Shannon that the redundancy rate (per symbol) for such codes is $O(1/D)$ (in this case D is fixed and equal to the block length), in [21] it is shown that the average redundancy rate either converges to a c/D for some constant c (e.g., $0.5/D$ in the case of the Shannon code) or it exhibits very erratic behavior fluctuating between 0 and $1/D$. For variable-to-fixed codes Savari and Gallager [17] present precise analysis of the dominant term in the asymptotic expansion of the Tunstall code redundancy. Basically, it was shown that the average redundancy rate decays as $O(1/D)$ (cf. [8] for some recent results). From this brief discussion, we conclude that while FV and VF codes waste a fraction of a bit per source symbol, we construct a VV code that loses a negligible information per symbol.

There is scarcely any literature on VV codes with a few exceptions such as [9, 10, 14, 18]. The most interesting, as already mentioned, is a thirty year old work by Khodak [14]. To the best of our knowledge not much was done since then, except that Fabris [9] (cf. also [10, 18]) analyzed Tunstall–Huffman VV code and provided a simple bound on their redundancy rate.

Finally, we say a word about our proof techniques. The main tool is the Diophantine approximation [5, 19]. This theory shows how to find a good approximation of linear forms like $k_1\gamma_1 + \cdots + k_m\gamma_m$ by rationals where k_i are integers and γ_i are irrational numbers. In the present context we have to construct a parsing tree for which $\log P(d)$ is close to an integer. Here $\log P(d)$ is of the form $k_1 \log p_1 + \cdots + k_m \log p_m$. Therefore it is natural to apply techniques from Diophantine approximation. Since $p_1 + ... + p_m = 1$, the coefficients $\log p_1, ..., \log p_m$ in the linear form are not independent and our almost sure results require non-trivial results on metric Diophantine approximation on manifolds.

The paper is organized as follows. In the next section, we first briefly discuss precise definitions of the average and the worst case redundancy rates for VV codes, followed by the presentation of our main results. We first consider redundancy rates for all (memoryless or Markov) sources (cf. Theorem 1) and then for *almost all* (memoryless or Markov) sources (cf. Theorem 2). To underline our constructive approach, we also briefly describe an algorithm that builds a VV code with vanishing redundancy rates as the average phrase

all sources" means that the set of $(p_1, p_1, \ldots, p_m) \in \mathbb{R}^m$ with $p_j > 0$ and $p_1 + \cdots + p_m = 1$ for which our statement does not hold has zero Lebesgue measure on the $(m-1)$-dimensional hyperplane $x_1 + \cdots + x_m = 1$. The statement "almost all Markov sources" has to be interpreted in a similar way. Here we use the Lebesgue measure on the corresponding parameter space of the transition probabilities p_{ij}.

length increases. We finish this section with a lower bound on redundancy rates valid for all VV codes and almost all sources (cf. Theorem 3) and an extension of our results to Markov sources (cf. Theorem 4). The next two sections, Sections 3 and 4, are devoted to the proofs of Lemma 3 and Lemma 4 which are the main ingredients for the proofs of Theorem 1 and Theorem 2. Finally, in the last Section 5 we prove Theorem 4 for Markov sources.

2 Main Results and Their Consequences

In this section we first define the average and the maximal redundancy rates for VV length codes. Then we present our main results valid for *all sources* (cf. Theorem 1) on the average and the maximal redundancy rates. We also propose an explicit algorithm that constructs a VV code with small redundancy rates. *Almost* all sources are discussed next (cf. Theorem 2). Finally, we present some lower bounds for the redundancy (cf. Theorem 3) and extend our results to Markov sources (cf. Theorem 4).

2.1 Redundancy Rates for VV Codes

Let us first formally introduce redundancy rates for VV codes by defining (asymptotic) *average* redundancy rate and *maximal* or *worst case* (i.e., for individual sequences) redundancy rate. To the best of our knowledge the worst case redundancy was not discussed before for VV codes.

Let $\mathcal{A} = \{a_1, \dots, a_m\}$ be the input alphabet of $m \geq 2$ symbols with *known* probabilities $p_1, \dots, p_m$. A memoryless source $\mathcal{S}$ generates a sequence X with the underlying probability $P_{\mathcal{S}}$. We denote by $P(d) := P_{\mathcal{C}}(d)$ the probability induced by the dictionary $\mathcal{C}$ and define the *average delay* or the *average phrase* length D as

$$D = \sum_{d \in \mathcal{C}} P_{\mathcal{C}}(d)|d|, \tag{1}$$

where $|d|$ is the length of $d \in \mathcal{C}$. The (asymptotic) average redundancy rate $\overline{r}$ is usually defined as

$$\overline{r} = \lim_{n \to \infty} \frac{\sum_{|x|=n} P_{\mathcal{S}}(x)(L(x) + \log P_{\mathcal{S}}(x))}{n}, \tag{2}$$

where $L(x)$ is the code length assigned to the source sequence x of length $|x| = n$. We shall call $\overline{r}$ the *average redundancy rate*. Using renewal reward theory as in [18] we arrive at

$$\lim_{n \to \infty} \frac{\sum_{x:|x|=n} P_{\mathcal{S}}(x) L(x)}{n} = \frac{\sum_{d \in \mathcal{C}} P_{\mathcal{C}}(d)\ell(d)}{D}. \tag{3}$$

An application of the *Conservation of Entropy Theorem* [15, 16, 20], as in [18], leads to

$$\overline{r} \;=\; \frac{\sum_{d \in \mathcal{C}} P_{\mathcal{C}}(d)\ell(d) - H_{\mathcal{C}}}{D} = \frac{\sum_{d \in \mathcal{C}} P_{\mathcal{C}}(d)(\ell(d) + \log P(d))}{D}, \tag{4}$$

which we adopt as our definition of the average redundancy rate.[3] Above, $H_{\mathcal{C}}$ denotes the entropy of $P_{\mathcal{C}}$. Furthermore, since we mostly deal with the probability induced by the dictionary, so we shall write $P = P_{\mathcal{C}}$.

Observe that (4) decomposes the redundancy rate of the VV length code into two terms. The denominator represents the expected length of a dictionary phrase and the numerator is the redundancy of a fixed-to-variable length code over an auxiliary source with "symbol" probabilities P. Therefore, by analogy we define the *maximal* redundancy rate r^* as follows

$$r^* = \frac{\max_{d\in\mathcal{C}}[\ell(d) + \log P(d)]}{D}. \tag{6}$$

The main purpose of this work is to construct a (complete) prefix free set (dictionary) $\mathcal{C}$ (i.e., a complete tree) on the input alphabet $\mathcal{A}$ and a bijective mapping C (a VV code) to another prefix free set on the binary alphabet $\{0,1\}$ with small average and maximal redundancy rates that decay to zero as the average delay increases.

2.2 Redundancy Rates for All Sources

We now start constructing a VV code with small redundancy rates. We recall that a VV coder consists of a parser and a string encoder. We fix throughout the string encoder to be a slightly modified Shannon code that assigns to a dictionary word $d \in \mathcal{C}$ the code length that is close to $-\log d$. Our goal is to build a dictionary (i.e., a complete parsing tree) that achieves this objective.

For every $d \in \mathcal{C}$ we can represent $P(d)$ as $P(d) = p_1^{k_1}\cdots p_m^{k_m}$, where $k_i = k_i(d)$ is the number of times symbol a_i appears in d. In what follows we will also use the notation $\text{type}(d) = (k_1, k_2, \ldots, k_m)$ for all strings with this probability. The numerator of the average redundancy rate for the Shannon code is

$$\begin{aligned} R &= \sum_{d\in\mathcal{C}} P(d)[\lceil -\log P(d)\rceil + \log P(d)] \\ &= \sum_{d\in\mathcal{C}} P(d)\cdot\langle k_1(d)\gamma_1 + k_2(d)\gamma_2 + \cdots + k_m(d)\gamma_m\rangle \end{aligned}$$

where $\gamma_i = \log p_i$ and $\langle x\rangle = x - \lfloor x\rfloor$ is the fractional part of x. We are to find integers $k_1 = k_1(d), \ldots k_m = k_m(d)$ such that the linear form $k_1\gamma_1 + k_2\gamma_2 + \cdots + k_m\gamma_m$ is close to an integer. Actually, we will do a little better by not using exactly the Shannon code with $\ell(d) = \lceil -\log P(d)\rceil$ but a variant of it in which $\ell(d)$ is the closest integer to $-\log P(d)$. Nevertheless, we will need some properties, discussed below, of the distribution of $\langle k_1\gamma_1 + k_2\gamma_2 + \cdots + k_m\gamma_m\rangle$ when at least one of γ_i is irrational. We first need to introduce the notion of *dispersion* and recall some properties of *continued fractions.*

[3] Observe that in (4) we ignore the rate of convergence in (3) since the redundancy rate (2) is explicitly defined as a limit.

Continued Fraction. A finite continued fraction expansion is a rational number of the form (cf. [2])

$$c_0 + \cfrac{1}{c_1 + \frac{1}{c_2 + \frac{1}{c_3 + \ddots + \frac{1}{c_n}}}},$$

where c_0 is an integer and c_j are *positive* integers for $j \geq 1$. We denote this rational number as $[c_0, c_1, \ldots, c_n]$. With help of the Euclidean algorithm, it is easy to see that every rational number has a finite continued fraction expansion.[4] Furthermore, if c_j is a given sequence of integers (that are positive for $j > 0$), then the limit $\theta = \lim_{n\to\infty}[c_0, c_1, \ldots, c_n]$ exists and is denoted by the infinite continued fraction expansion $\theta = [c_0, c_1, c_2 \ldots]$. Conversely, if θ is a real irrational number and if we recursively set

$$\theta_0 = \theta, \quad c_j = \lfloor \theta_j \rfloor, \quad \theta_{j+1} = 1/(\theta_j - c_j),$$

then $\theta = [c_0, c_1, c_2 \ldots]$. In particular, every irrational number has a unique infinite continued fraction expansion.

The *convergents* of an irrational number θ with infinite continued fraction expansion $\theta = [c_0, c_1, c_2 \ldots]$ are defined as

$$\frac{p_n}{q_n} = [c_0, c_1, \ldots, c_n],$$

where integers p_n, q_n are coprime. These integers can be recursively determined by

$$p_n = c_n p_{n-1} + p_{n-2}, \qquad q_n = c_n q_{n-1} + q_{n-2}.$$

In particular, p_n and q_n are growing exponentially quickly. Furthermore, the convergents $\frac{p_n}{q_n}$ are the best rational approximations of θ in the sense that

$$|q_n\theta - p_n| < \min_{0<q<q_n,\ p\in\mathbb{Z}} |q\theta - p|.$$

In particular one has [5]

$$\left|\theta - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2}. \tag{7}$$

The denominators q_n are called *best approximation denominators.*

Dispersion. Let $\|x\| = \min(\langle x\rangle, \langle -x\rangle) = \min(\langle x\rangle, 1 - \langle x\rangle)$ be the distance to the nearest integer. The *dispersion* $\delta(X)$ of the set $X \subseteq [0, 1)$ is defined as

$$\delta(X) = \sup_{0\leq y<1} \inf_{x\in X} \|y - x\|,$$

that is, for every $y \in [0, 1)$ there exists $x \in X$ with $\|y-x\| \leq \delta(X)$. Since $\|y+1\| = \|y\|$, the same assertion holds for all real y. Dispersion tells us that points of X are at most $2\delta(X)$ apart in $[0, 1]$. Therefore, there exist distinct points $x_1, x_2 \in X$ with $\langle y - x_1\rangle \leq 2\delta(X)$ and $\langle y - x_2\rangle \leq 2\delta(X)$.

The following property will be used throughout this paper.

[4] This finite continued fraction expansion is unique if we assume that $c_n > 1$. There is only one alternative representation given by $[c_0, c_1, \ldots, c_n - 1, 1]$.

Lemma 1. *Suppose that θ is an irrational number and let $N = q_n$ be a best approximation denominator (i.e., $p_n/q_n = [c_0, c_1, \ldots, c_n]$ is a convergent of the continued fraction expansion of $\theta = [c_0, c_1, c_2, \ldots]$). Then*

$$\delta\left(\{\langle k\theta\rangle : 0 \le k < N\}\right) \le \frac{2}{N}.$$

Proof. For $N = q_n$ we find from (7)

$$\left|\theta - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2}$$

or

$$\theta = \frac{p_n}{q_n} + \frac{\eta}{q_n^2}$$

for some $|\eta| < 1$. Consequently the numbers $\langle k\theta\rangle$ $(0 \le k < N = q_n)$ are quite close to the numbers $0, 1/N, 2/N, \ldots, (N-1)/N$, in particular, for every $k < N$ there exists $l = kp_n \bmod N < N$ with $\|k\theta - l/N\| < (N-1)/N^2 < 1/N$. Since p_n and $N = q_n$ are coprime it also follows that for every $l < N$ there exists $k < N$ with $\|k\theta - l/N\| < 1/N$.

Consequently, if y in an arbitrary number in $[0, 1)$ then there exists $l < N$ with $< 1/(2N)$ and another number $k\theta$ with $k < N$ and $< 1/N$. Thus

$$\|y - k\theta\| \le \|y - l/N\| + \|k\theta - l/N\| < \frac{2}{N}.$$

In conclusion, the dispersion of the set $\{\langle k\theta\rangle : 0 \le k < N\}$ is bounded by $2/N$. ■

Remark. The proof of Lemma 1 shows that we can work with every N that satisfies

$$\left|\theta - \frac{M}{N}\right| < \frac{1}{N^2} \tag{8}$$

for some integer M. It is well known that Dirichlet's approximation theorem (cf. [2, 5]) ensures the existence of infinitely many N for which (8) is satisfied. (A simple but non-constructive proof uses the pigeonhole principle.) The advantage of continued fraction theory is that the convergent p_n/q_n, satisfies (8) and it can be effectively computed.

The first consequence of Lemma 1 is the following property.

Lemma 2. *Let $(\gamma_1, \ldots, \gamma_m)$ be an m-vector of real numbers such that at least one of its coordinates is irrational. Let N be the best approximation denominator of the irrational number. Then the dispersion of the set*

$$X = \{\langle k_1\gamma + \cdots + k_m\gamma\rangle : 0 \le k_j < N \ (1 \le j \le m)\}$$

is bounded by

$$\delta(X) \le \frac{2}{N}.$$

Existence of a VV Code. The central step of all existence results of this paper is the observation that a bound on the dispersion of linear forms of $\log_2 p_j$ implies the existence of a VV code with small redundancy.

Our main result of this section follows directly from the below lemma whose proof is presented in Section 3.

Lemma 3. *Let $p_j > 0$ $(1 \le j \le m)$ with $p_1 + \cdots + p_m = 1$ be given and suppose that for some $N \ge 1$ and $\eta \ge 1$ the set*

$$X = \{\langle k_1' \log_2 p_1 + \cdots + k_m' \log_2 p_m \rangle : 0 \le k_j' < N \ (1 \le j \le m)\},$$

has dispersion

$$\delta(X) \le \frac{2}{N^\eta}. \tag{9}$$

Then there exists a VV code with the average code length $D = \Theta(N^3)$, the maximal length of order $\Theta(N^3 \log N)$, and the average redundancy rate

$$\overline{r} \le c_m' \cdot D^{-\frac{4+\eta}{3}}.$$

Furthermore, there exists another VV code with the average code length $D = \Theta(N^3)$ (and possible infinite maximal length) and the maximal redundancy rate

$$r^* \le c_m'' \cdot D^{-1-\frac{\eta}{3}},$$

where the constants $c_m', c_m'' > 0$ depend on m.

Clearly, Lemma 2 and Lemma 3 directly imply our main result presented below by setting $\eta = 1$ if one of the $\log_2 p_j$ is irrational. (If all $\log_2 p_j$ are rational, then the construction presented in section 3 is much simpler; see the Remark at the end of section 3).

Theorem 1. *Let $m \ge 2$ and $\mathcal{S}$ be a memoryless source on an alphabet of size m. Then for every $D_0 \ge 1$, there exists a VV code with average delay $D \ge D_0$ such that its average redundancy rate satisfies*

$$\overline{r} = O(D^{-5/3}), \tag{10}$$

and the average code length is $O(D \log D)$. Furthermore, there also exists a VV code with average delay $D \ge D_0$ such that worst case redundancy rate satisfies

$$r^* = O(D^{-4/3}), \tag{11}$$

however, maximal code length might be infinite.

The estimate (10) for $\overline{r}$ is the same as in Khodak [14]. However, the proof presented in [14] is rather sketchy and complicated. Our method uses similar ideas as that of [14] but is more transparent and leads to an explicit construction of a VV code with small redundancy rates that we discuss next.

2.3 Algorithm

In what follows we present an algorithm for designing a VV-code with arbitrarily large average dictionary length D; given a memoryless source with probability distribution $p_1, \ldots, p_m > 0$ on alphabet $a_1, \ldots, a_m$, the algorithm achieves redundancy rate smaller than $cD^{-4/3}$. In fact, we construct a code with redundancy $\overline{r} \le \varepsilon/D$, where $\varepsilon > 0$ is given and $D \ge c/\varepsilon^3$ (for some constant c). Note that we will not use the full strength of Theorem 1 that guarantees the existence of a code with the average redundancy smaller than $cD^{-5/3}$. This allows, however, some simplification of the algorithm, in particular we just use the (standard) Shannon code.

We will also make the assumption that all p_j are given rational numbers. (Otherwise we would have to assume that p_j is known to an arbitrary precision.) We then know that $\log_2 p_j$ is either irrational or an integer (which means that $p_j = 2^{-k}$). Thus, we can immediately decide whether all $\log_2 p_j$ are rational or not. If all p_j are negative powers of 2, then we can use a perfect code with zero redundancy. Thus, we only have to treat the case where p_m is not a negative powers of 2. We also assume that continued fraction expansion of $\log_2 p_m = [c_0, c_1, c_2, \ldots]$ is given and one determines a convergent $[c_0, c_1, c_2, \ldots, c_n] = M/N$ for which the denominator N satisfies $N > 4/\varepsilon$.

The main goal of the algorithms is to construct a prefix free set of words d with the property that for *most words* $\langle \log_2 P(d) \rangle$ is small. The reason for this *philosophy* is that if one uses the Shannon code as the string encoder, that is $\ell(d) = \lceil -\log_2 P(d) \rceil$, then the difference $\ell(d) - \log(1/P(d)) = \langle \log_2 P(d) \rangle$ is small and gives only a small contribution to the redundancy.

The main step of the algorithm is a loop of the same subroutine, The input is a pair $\mathcal{C}$, $\mathcal{B}$ of sets of words with the property that $\mathcal{C} \cup \mathcal{B}$ is a prefix free set. Words d in $\mathcal{C}$ are already *good* in the sense that $\langle \log_2 P(d) \rangle \le \frac{3}{4}\varepsilon$, whereas words r in $\mathcal{B}$ are *bad* because they do not satisfy this condition. In the first step of the subroutine, one chooses a word $r \in \mathcal{B}$ of minimal length and computes an integer k with $0 \le k < N$ that satisfies

$$\frac{1}{N} \le \langle kM/N + x + \log_2 P(r) \rangle \le \frac{2}{N}.$$

Here x is an abbreviation of $x = \sum_{j=1}^m k_j^0 \log_2 p_j$, where $k_j^0 = \lfloor p_j N^2 \rfloor$, $1 \le j \le m$. The computation of k can be done by solving the congruence $kM \equiv 1 - \lfloor (x + \log_2 P(r))N \rfloor \bmod N$ (e.g., with help of the Euclidean algorithm). This choice of k ensures that

$$0 \le \langle k \log_2 p_m + x + \log_2 P(r) \rangle \le 3/N \le \frac{3}{4}\varepsilon.$$

For this k we determine the set $\mathcal{C}'$ of all words d of type$(d) = (k_1^0, \ldots, k_{m-1}^0, k_m^0 + k)$. By construction all $d' \in \mathcal{C}'$ satisfy

$$\langle \log_2 P(r \cdot d') \rangle = \langle k \log_2 p_m + x + \log_2 P(r) \rangle \le \frac{3}{4}\varepsilon.$$

We now replace $\mathcal{C}$ by $\mathcal{C} \cup r \cdot \mathcal{C}'$ and $\mathcal{B}$ by $(\mathcal{B} \setminus \{r\}) \cup r \cdot (A^n \setminus \mathcal{C}')$. This construction ensures that (again) all word in $d \in \mathcal{C}$ satisfy

$$\langle \log_2 P(d) \rangle \le \frac{3}{4}\varepsilon.$$

The algorithm terminates when $P(\mathcal{C}) > 1 - \varepsilon/4$; that is, *most words* in $\mathcal{C} \cup \mathcal{B}$ are *good*. (The proof of Theorem 1 shows that this actually occurs when the average dictionary length D is of order $O(N^3)$. In particular, the special choice of integers $k_j^0 = \lfloor p_j N^2 \rfloor$ ensures that the probability $P(\mathcal{C})$ increases step by step as quickly as possible, compare with (23).)

As already mentioned, we finally use the Shannon code $C : \mathcal{C} \cup \mathcal{B} \to \{0,1\}^*$, that is $\ell(d) = \lceil -\log_2 P(d) \rceil$ for all $d \in \mathcal{C} \cup \mathcal{B}$. The redundancy can be estimated by

$$\begin{aligned}
\overline{r} &= \frac{1}{D} \sum_{d \in \mathcal{C} \cup \mathcal{B}} P(d) \left(\ell(d) - \log_2 \frac{1}{P(d)} \right) \\
&= \frac{1}{D} \sum_{d \in \mathcal{C} \cup \mathcal{B}} P(d) \, \langle \log_2 P(d) \rangle \\
&= \frac{1}{D} \left(\sum_{d \in \mathcal{C}} P(d) \langle \log_2 P(d) \rangle + \sum_{d \in \mathcal{B}} P(d) \langle \log_2 P(d) \rangle \right) \\
&\le \frac{1}{D} \left(P(\mathcal{C}) \frac{3}{4} \varepsilon + P(\mathcal{B}) \right) \\
&\le \frac{1}{D} \left(\frac{3}{4} \varepsilon + \frac{1}{4} \varepsilon \right) = \frac{\varepsilon}{D}.
\end{aligned}$$

Thus we have constructed a parsing tree and a VV code with a small redundancy rate. A more formal description of the algorithm follows.

Algorithm KhodCode:

Input: (i) m, an integer ≥ 2; (ii) positive rational numbers $p_1, \ldots, p_m$ with $p_1 + \cdots + p_m = 1$, p_m is not a power of 2; (iii) ε, a positive real number < 1.

Output: A VV-code, that is, a complete prefix free set on an m-ary alphabet and a prefix code $C : \mathcal{C} \to \{0,1\}^*$, with redundancy $\overline{r} \le \varepsilon/D$, where the average dictionary code length D satisfies $D \ge c(m, p_1, \ldots, p_m)/\varepsilon^3$ (for some constant $c(m, p_1, \ldots, p_m)$).

Notation: For a word $w \in \mathcal{A}^*$ that consists of k_j copies of a_j ($1 \le j \le m$) we set $P(w) = p_1^{k_1} \cdots p_m^{k_m}$ for the probability of w and $\mathrm{type}(w) = (k_1, \ldots, k_m)$. By ω we denote the empty word and set $P(\omega) = 1$.

```
1. Calculate the convergent $\frac{M}{N} = [c_0, c_1, \ldots, c_n]$ of the irrational number $\log_2 p_m$ for
   which $N > 4/\varepsilon$ (cf. the continued fraction expansions discussed in the previous
   subsection).

2. Set $k_j^0 = \lfloor p_j N^2 \rfloor$ $(1 \le j \le m)$, $x = \sum_{j=1}^m k_j^0 \log_2 p_j$, and $n_0 = \sum_{j=1}^m k_j^0$.

3. Set $\mathcal{C} = \emptyset$, $\mathcal{B} = \{\omega\}$, and $p = 0$
   while $p < 1 - \varepsilon/4$ do
           Choose $r \in \mathcal{B}$ of minimal length
           $b \leftarrow \log_2 P(r)$
           Find $0 \le k < N$ that solves the congruence $kM \equiv 1 - \lfloor (x + b)N \rfloor \bmod N$
```

```
        $n \leftarrow n_0 + k$
        $\mathcal{C}' \leftarrow \{d \in A^n : \text{type}(d) = (k_1^0, \ldots, k_{m-1}^0, k_m^0 + k)\}$
        $\mathcal{C} \leftarrow \mathcal{C} \cup r \cdot \mathcal{C}'$
        $\mathcal{B} \leftarrow (\mathcal{B} \setminus \{r\}) \cup r \cdot (A^n \setminus \mathcal{C}')$
        $p \leftarrow p + P(r)P(\mathcal{C}')$, where

            $P(\mathcal{C}') = \frac{n!}{k_1^0! \cdots k_{m-1}^0!(k_m^0 + k)!} p_1^{k_1^0} \cdots p_{m-1}^{k_{m-1}^0} p_m^{k_m^0 + k}.$

    end while.

4.  $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{B}$.

5.  Construct a Shannon code $C : \mathcal{C} \to \{0,1\}^*$ with $\ell(d) = \lceil -\log_2 P(d) \rceil$ for all $d \in \mathcal{C}$.
```

Let us consider an example.

Example. Assume $m = 2$ with $p_1 = 2/3$ and $p_2 = 1/3$. In the first iteration of the algorithm we assume that both $\mathcal{B}$ and $\mathcal{C}$ are empty. Easy computations show that

$$\log(1/3) = [-2, 2, 2, 2, 3, \ldots], \quad \text{and} \quad [-2, 2, 2, 2] = -\frac{19}{12},$$

hence $M = -19$ and $N = 12$. Let us set $\varepsilon = 0.4$ so $4/\varepsilon = 10 < 12 = N$. Therefore, $k_1^0 = 96$, $k_2^0 = 48$ so that $n_0 = 144 = N^2$. Solving the congruence

$$-19k = 1 + 1587 \mod 12$$

gives $k = 8$ and therefore

$$\mathcal{C}' = \{d \in A^{152} : \text{ type}(d) = (96, 56)\}$$

with $P(\mathcal{C}') = 0.04425103411$. Observe that $\mathcal{B} = A^{152} \setminus \mathcal{C}$.

In the second iteration we can pick up any string from $\mathcal{B}$, say the string $r = 00\ldots0$ with 152 zeros. We find, solving the congruence with $b = 152 \log_2(2/3) = -88.91430011$, that $k = 5$. Hence $\mathcal{C}' = \{d \in A^{149} : \text{ type}(d) = (96, 53)\}$ and $\mathcal{C} = \{d \in A^{152} : \text{ type}(d) = (96, 56)\} \cup r \cdot \mathcal{C}'$. We continue along the same path until the total probability of all “good” strings in $\mathcal{C}$ reaches the value $3/4 \cdot \varepsilon = 0.3$, which may take some time.

2.4 Redundancy Rates for Almost All Memoryless Sources

In this section we present better estimates for the redundancy rates but valid only for *almost all* memoryless sources. This means that the set of exceptional p_j (i.e., those p_j with $\sum_{j=1}^m p_j = 1$ and $p_j > 0$ for all $1 \le j \le m$ that do not satisfy the proposed property) has zero Lebesgue measure on the $(m-1)$-dimensional hyperplane $x_1 + \cdots + x_m = 1$. From a mathematical point of view, these results are more challenging.

While Lemma 1 and 2 laid foundation for Theorem 1, the next lemma, which we prove in Section 4, is crucial for our main result of this section.

Lemma 4. *Suppose that $\varepsilon > 0$. Then for almost all p_j $(1 \le j \le m)$ with $p_j > 0$ and $p_1 + p_2 + \cdots + p_m = 1$ the set*

$$X = \{\langle k_1 \log_2 p_1 + \cdots + k_m \log_2 p_m \rangle : 0 \le k_j < N \ (1 \le j \le m)\}$$

has dispersion

$$\delta(X) \le \frac{1}{N^{m-\varepsilon}} \tag{12}$$

for sufficiently large N. In addition, for almost all $p_j > 0$ there exists a constant $C > 0$ such that

$$\|k_1 \log_2 p_1 + \cdots + k_m \log_2 p_m\| \ge C \left(\max_{1 \le j \le m} |k_j| \right)^{-m-\varepsilon} \tag{13}$$

for all non-zero integer vectors $(k_1, \ldots, k_m)$.

We should point out that for $m = 2$ we shall slightly improve the estimate of the lemma. Indeed, we shall show that for almost all $p_1 > 0, p_2 > 0$ with $p_1 + p_2 = 1$ there exists a constant κ and infinitely many N such that the set $X = \{\langle k_1 \log_2 p_1 + k_2 \log_2 p_2 \rangle : 0 \le k_1, k_2 < N\}$ has dispersion

$$\delta(X) \le \frac{\kappa}{N^2}. \tag{14}$$

The estimate (14) is a little bit sharper than (12). However, it is only valid for infinitely many N and not for all but finitely many.[5]

By combining Lemma 3 and Lemma 4 we directly obtain our second main result valid for almost all sources.

Theorem 2. *Let $m \ge 2$ and $\mathcal{S}$ be a memoryless source on an alphabet of size m. Then for almost all source parameters, and for every sufficiently large D_0, there exists a VV code with the average delay D satisfying $D_0 \le D \le 2D_0$ such that its average redundancy rate is bounded by*

$$\overline{r} \le D^{-\frac{4}{3} - \frac{m}{3} + \varepsilon}, \tag{15}$$

where $\varepsilon > 0$ and maximal length is $O(D \log D)$.
Also, there exists a VV code with the average delay D satisfying $D_0 \le D \le 2D_0$ such that maximal redundancy is bounded by

$$r^* \le D^{-1 - \frac{m}{3} + \varepsilon}. \tag{16}$$

for any $\varepsilon > 0$.

This theorem shows that the *typical* best possible average redundancy $\overline{r}$ can be measured in terms of negative powers of D that are linearly decreasing in the alphabet size m. However, it seems to be a very difficult problem to obtain the optimal exponent (almost surely). Nevertheless, these bounds are best possible through the methods we applied.

[5] We point out that (12) and (14) are optimal. Since the set X consists of N^m points the dispersion must satisfy $\delta(X) \ge \frac{1}{2} N^{-m}$.

2.5 Lower Bound for Almost All Sources

We now present a lower bound for redundancy rates which is valid for almost all sources. It will follow from (13) of Lemma 4 and the following simple lower bound (cf. Corollary 1 in [14]).

Lemma 5. *Let $\mathcal{C}$ be a finite set with probability distribution P. Then*

$$\overline{r} \geq \frac{1}{2}\frac{1}{D}\sum_{d\in D} P(d)\|\log_2 P(d)\|^2.$$

Proof. Suppose that $|x| \leq 1$. Then we have $2^{-x} = 1 - x\log 2 + \eta(x)$ with $((\log 4)/4)x^2 \leq \eta(x) \leq (\log 4)x^2$. Thus, by using the representation

$$x = (1 - 2^{-x} + \eta(x))/(\log 2)$$

we obtain

$$\begin{aligned}\overline{r} &= \frac{1}{D}\sum_{d\in D} P(d)(\ell(d) + \log_2 P(d)) \\ &= \frac{1}{D\log 2}\sum_{d\in D} P(d)\left(1 - 2^{-\ell(d)-\log_2 P(d)} + \eta(\ell(d) + \log_2 P(d))\right) \\ &= \frac{1}{D\log 2}\left(1 - \sum_{d\in D} 2^{-\ell(d)}\right) + \frac{1}{D\log 2}\sum_{d\in D} P(d)\eta(\ell(d) + \log_2 P(d)).\end{aligned}$$

Hence, by Kraft's inequality and by the observation

$$\eta(x) \geq \min\{\eta(\langle x\rangle), \eta(\langle 1-x\rangle)\} \geq \frac{\log 4}{4}\|x\|^2$$

the result follows immediately. ∎

We are now in a position to present our finding regarding a lower bound on the redundancy rates for almost all sources.

Theorem 3. *Let S be a memoryless source on an alphabet of size $m \geq 2$. Then for almost all source parameters, and for every VV code with average delay $D \geq D_0$ (where D_0 is sufficiently large) we have*

$$r^* \geq \overline{r} \geq D^{-2m-1-\varepsilon}, \tag{17}$$

where $\varepsilon > 0$.

Proof. By Lemma 5 we have

$$\overline{r} \geq \frac{1}{2D}\sum_{d\in D} P(d)\|\log_2 P(d)\|^2.$$

Suppose that $P(d) = p_1^{k_1}\cdots p_m^{k_m}$ holds, that is

$$\log_2 P(d) = k_1\log_2 p_1 + \cdots + k_m\log_2 p_m.$$

By Lemma 4, we conclude from (13) that for all p_j and for all non-zero integer vectors $(k_1, \dots, k_m)$

$$\|k_1 \log_2 p_1 + \cdots + k_m \log_2 p_m\| \ge C \left(\max_{1 \le j \le m} |k_j| \right)^{-m-\varepsilon},$$

and therefore

$$\| \log_2 P(d) \| \ge \left(\max_{1 \le j \le m} |k_j| \right)^{-m-\varepsilon} \ge \left(\sum_{1 \le j \le m} k_j \right)^{-m-\varepsilon} = |d|^{-m-\varepsilon}.$$

Consequently, by Jensen's inequality, we obtain

$$\begin{aligned} \overline{r} &\ge \frac{1}{2D} \sum_{d \in D} P(d) |d|^{-2m-2\varepsilon} \\ &\ge \frac{1}{2D} \left(\sum_{d \in D} P(d) |d| \right)^{-2m-2\varepsilon} \\ &\ge D^{-2m-1-2\varepsilon}. \end{aligned}$$

This completes the proof of Theorem 3. ■

Note that Theorem 4 of [14] states a lower bound for the redundancy rate in the form $\overline{r} \ge D^{-9} (\log D)^{-8}$ (for almost all memoryless sources). In view of Theorem 2 this cannot be true for large m.

2.6 Markov sources

Finally, we state corresponding properties for Markov sources. The proof is almost the same as for memoryless sources except that it is technically more challenging. In section 5 we shortly comment on the differences.

Theorem 4. *Let $m \ge 2$ and $\mathcal{S}$ be a Markov source of order 1 on an alphabet of size m with transition matrix $\mathbf{P} = (p_{ij})_{1 \le i,j \le m}$ with $p_{ij} > 0$ ($1 \le i, j \le m$). Furthermore let $D_0 > 1$ be an arbitrary real number.*

(i) *Then there exists a VV code with average delay $D \ge D_0$ such that its average redundancy rate satisfies*

$$\overline{r} = O(D^{-\frac{m+4}{m+2}}), \tag{18}$$

and maximal length is $O(D \log D)$. There also exists a VV code with average delay $D \ge D_0$ for which worst case redundancy rate satisfies

$$r^* = O(D^{-\frac{m+3}{m+2}}), \tag{19}$$

however, the maximal length might be infinite.

(ii) *For almost all source parameters, and for every sufficiently large D_0, there exists a VV code with the average delay D satisfying $D_0 \le D \le 2D_0$ such that its average redundancy rate is bounded by*

$$\overline{r} \le D^{-\frac{m^2+4}{m+2} + \varepsilon}, \tag{20}$$

where $\varepsilon > 0$ and the maximal length is $O(D \log D)$. There also exists a VV code with the average delay D satisfying $D_0 \le D \le 2D_0$ such that the maximal redundancy is bounded by

$$r^* \le D^{-\frac{m^2+3}{m+2}+\varepsilon}. \tag{21}$$

for any $\varepsilon > 0$.
(iii) *Finally, for almost all source parameters, and for every VV code with average delay $D \ge D_0$ (where D_0 is sufficiently large) we have*

$$r^* \ge \overline{r} \ge D^{-2m^2+2m-3-\varepsilon}, \tag{22}$$

where $\varepsilon > 0$.

3 Proof of Lemma 3

This section is devoted to the proof of our crucial Lemma 3 We shall use techniques similar to those already presented in [14].

The main thrust of the proof is to construct a complete prefix free set $\mathcal{C}$ of words (i.e., a dictionary) on an alphabet of size m such that $\log_2 P(d)$ is *very close* to an integer $\ell(d)$ with high probability. This is accomplished by constructing an m-ary tree $\mathcal{T}$ in which edges are labeled from left to right by the symbol of the alphabet $\mathcal{A} = \{a_1, \dots, a_m\}$. Leaves of such an m-ary tree can be identified with a complete prefix free set $\mathcal{C}$. Furthermore, the sequence of labels on a path from the root to a leaf translates into symbols of the corresponding word d in the complete prefix free set $\mathcal{C}$. Finally, we apply Kraft's inequality (cf. Lemma 6 below) to conclude that there exists a (VV) code C with $|C(d)| = \ell(d)$ and small average redundancy rate.

In the first step, we set $k_i^0 := \lfloor p_i N^2 \rfloor$ $(1 \le i \le m)$ and

$$x = k_1^0 \log_2 p_1 + \cdots + k_m^0 \log_2 p_m.$$

By our assumption (9) of Lemma 3, there exist integers $0 \le k_j^1 < N$ such that

$$\left\langle x + k_1^1 \log_2 p_1 + \cdots + k_m^1 \log_2 p_m \right\rangle = \left\langle (k_1^0 + k_1^1) \log_2 p_1 + \cdots + (k_m^0 + k_m^1) \log_2 p_m \right\rangle < \frac{4}{N^\eta}.$$

Now consider all paths in a (potentially) infinite m-ary tree starting at the root with $k_1^0 + k_1^1$ edges of type a_1, $k_2^0 + k_2^1$ edges of type a_2,..., and $k_m^0 + k_m^1$ edges of type a_m. Let $\mathcal{C}_1$ denote the set of the corresponding words over the input alphabet. (These are the first words of our prefix free set we are going to construct.) By an application of Stirling's formula it follows that there are two positive constants c', c'' with

$$\frac{c'}{N} \le P(\mathcal{C}_1) = \binom{(k_1^0 + k_1^1) + \cdots + (k_m^0 + k_m^1)}{k_1^0 + k_1^1, \dots, k_m^0 + k_m^1} p_1^{k_1^0+k_1^1} \cdots p_m^{k_m^0+k_m^1} \le \frac{c''}{N} \tag{23}$$

uniformly for all k_j^1 with $0 \le k_j^1 < N$. In summary, by construction all words $d \in \mathcal{C}_1$ have the property that

$$\langle \log_2 P(d) \rangle < \frac{4}{N^\eta},$$

that is, $\log_2 P(d)$ is very close to an integer. Note further that all words in $d \in \mathcal{C}_1$ have about the same length

$$n_1 = (k_1^0 + k_1') + \cdots + (k_m^0 + k_m') = N^2 + O(N),$$

and words in $\mathcal{C}_1$ constitute the first crop of "good words". Finally, let $\mathcal{B}_1 = \mathcal{A}^{n_1} \setminus \mathcal{C}_1$ denote all words of length n_1 not in $\mathcal{C}_1$ (cf. the first full tree in Figure 1). Then

$$1 - \frac{c''}{N} \le P(\mathcal{B}_1) \le 1 - \frac{c'}{N}.$$

In the second step, we consider all words $r \in \mathcal{B}_1$ and concatenate them with appropriately chosen words d_2 of length $\sim N^2$ such that $\log_2 P(rd_2)$ is close to an integer *with high probability.* The construction is almost the same as in the first step. For every word $r \in \mathcal{B}_1$ we set

$$x(r) = \log_2 P(r) + k_1^0 \log_2 p_1 + \cdots + k_m^0 \log_2 p_m.$$

By (9) there exist integers $0 \le k_j^2(r) < N$ $(1 \le j \le m)$ such that

$$\left\langle x(r) + k_1^2(r)\log_2 p_1 + \cdots + k_m^2(r)\log_2 p_m \right\rangle < \frac{4}{N^\eta}.$$

Now consider all paths (in the infinite tree $\mathcal{T}$) starting at $r \in \mathcal{B}_1$ with $k_1^0 + k_1^2(r)$ edges of type a_1, $k_2^0 + k_2^2(r)$ edges of type a_2, $\ldots$, and $k_m^0 + k_m^2(r)$ edges of type a_m (that is, we concatenated r with properly chosen words d_2) and denote this set by $\mathcal{C}_2^+(r)$. We again have that the total probability of these words is bounded from below and above by

$$\begin{aligned} P(r)\frac{c'}{N} &\le P(\mathcal{C}_2(r)) = P(r)\binom{(k_1^0 + k_1^2(r)) + \cdots + (k_m^0 + k_m^2(r))}{k_1^0 + k_1^2(r), \ldots, k_m^0 + k_m^2(r)} p_1^{k_1^0 + k_1^2(r)} \cdots p_m^{k_m^0 + k_m^2(r)} \\ &\le P(r)\frac{c''}{N}. \end{aligned}$$

Furthermore, by construction we have

$$\langle \log_2 P(d) \rangle < \frac{4}{N^\eta}$$

for all $d \in \mathcal{C}_2^+(r)$.

Similarly, we can construct a set $\mathcal{C}_2^-(r)$ instead of $\mathcal{C}_2^+(r)$ for which we have $1 - \langle \log_2 P(d) \rangle < 4/N^\eta$. We will indicate in the sequel whether we will use $\mathcal{C}_2^+(r)$ or $\mathcal{C}_2^-(r)$.

Let $\mathcal{C}_2 = \bigcup(\mathcal{C}_2^+(r) : r \in \mathcal{B}_1)$ (or $\mathcal{C}_2 = \bigcup(\mathcal{C}_2^-(r) : r \in \mathcal{B}_1)$). Then all words $d \in \mathcal{C}_2$ have almost the same length

$$|d| = 2N^2 + O(2N),$$

their probabilities satisfy

$$\langle \log_2 P(d) \rangle < \frac{4}{N^\eta} \quad \left(\text{or} \quad 1 - \langle \log_2 P(d) \rangle < \frac{4}{N^\eta} \right)$$

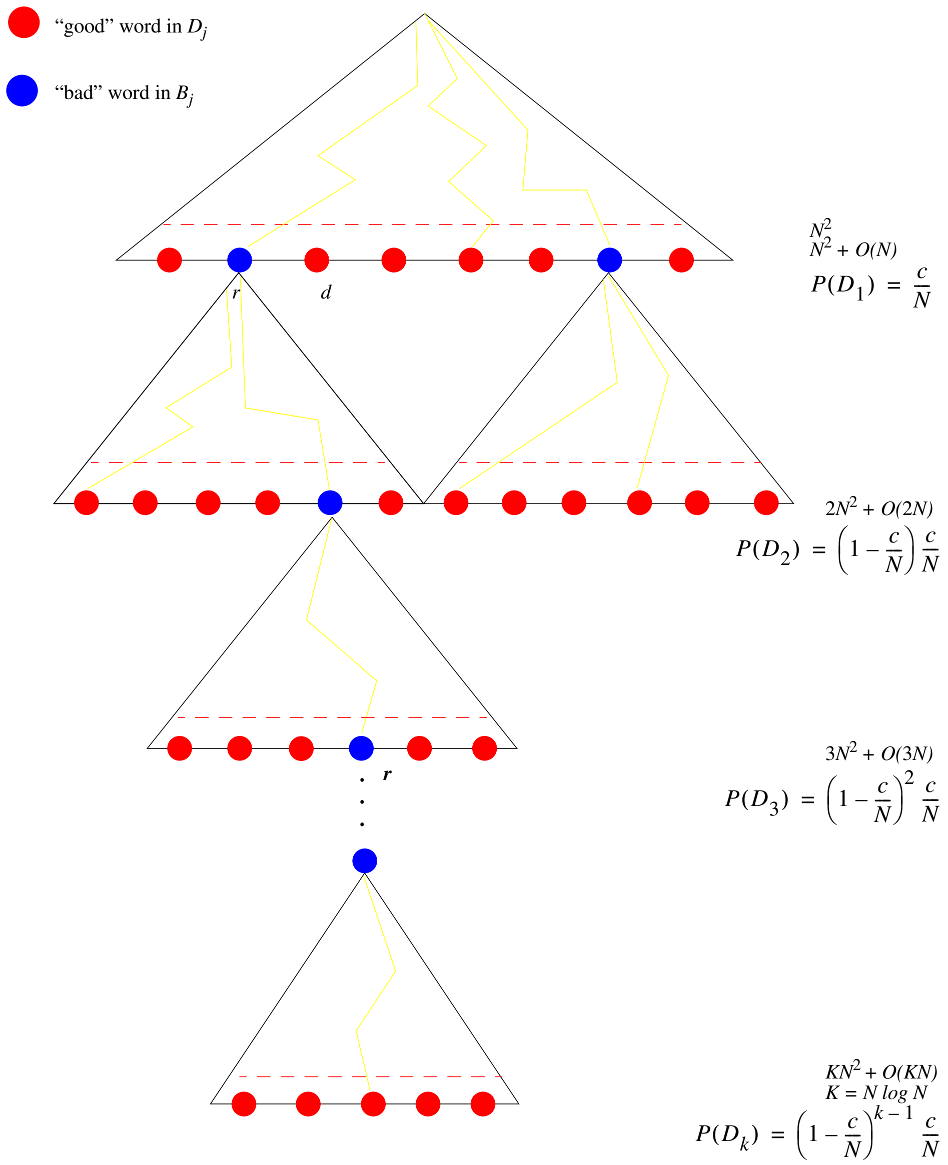
“good” word in D_j
“bad” word in B_j
r
d
N^2
$N^2 + O(N)$
$P(D_1) = \frac{c}{N}$
$2N^2 + O(2N)$
$P(D_2) = \left(1 - \frac{c}{N}\right)\frac{c}{N}$
r
$3N^2 + O(3N)$
$P(D_3) = \left(1 - \frac{c}{N}\right)^2\frac{c}{N}$
$KN^2 + O(KN)$
$K = N \log N$
$P(D_k) = \left(1 - \frac{c}{N}\right)^{k-1}\frac{c}{N}$

Figure 1: Illustration to the construction of the VV code.

and the total probability is bounded by

$$\frac{c'}{N}\left(1-\frac{c''}{N}\right) \leq P(\mathcal{C}_2) \leq \frac{c''}{N}\left(1-\frac{c'}{N}\right).$$

The variant of the Shannon code to which we alluded in several places above, is now constructed. For every $r \in \mathcal{B}_1$, let $\mathcal{B}^+(r)$ (or $\mathcal{B}^-(r)$) denote the set of paths (resp. words) starting with r of length $2(k_1^0 + \cdots + k_m^0) + (k_1^1 + k_1^2(r) + \cdots + k_m^1 + k_m^2(r))$ that are not contained in $\mathcal{C}_2^+(r)$ (or $\mathcal{C}_2^-(r)$) and set $\mathcal{B}_2 = \bigcup(\mathcal{B}_2^+(r) : r \in \mathcal{B}_1)$ (or $\mathcal{B}_2 = \bigcup(\mathcal{B}_2^-(r) : r \in \mathcal{B}_1)$). Observe that the probability of $\mathcal{B}_2$ is bounded by

$$\left(1-\frac{c''}{N}\right)^2 \leq P(\mathcal{B}_2) \leq \left(1-\frac{c'}{N}\right)^2.$$

We continue this construction, and in step j we define sets of words $\mathcal{C}_j$ and $\mathcal{B}_j$ such that all words $d \in \mathcal{C}_j$ satisfy

$$\langle \log_2 P(d) \rangle < \frac{4}{N^\eta} \quad \left(\text{or} \quad 1 - \langle \log_2 P(d) \rangle < \frac{4}{N^\eta}\right)$$

and the length of $d \in \mathcal{C}_j \cup \mathcal{B}_j$ is given by

$$|d| = jN^2 + \mathcal{O}(jN).$$

The probabilities of $\mathcal{C}_j$ and $\mathcal{B}_j$ are bounded by

$$\frac{c'}{N}\left(1-\frac{c''}{N}\right)^{j-1} \leq P(\mathcal{C}_j) \leq \frac{c''}{N}\left(1-\frac{c'}{N}\right)^{j-1},$$

and

$$\left(1-\frac{c''}{N}\right)^j \leq P(\mathcal{B}_j) \leq \left(1-\frac{c'}{N}\right)^j.$$

This construction is terminated after $K = O(N \log N)$ steps so that

$$P(\mathcal{B}_K) \leq c''\left(1-\frac{c'}{N}\right)^K \leq \frac{1}{N^\beta}$$

for some $\beta > 0$. This also ensures that

$$P(\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_K) > 1 - \frac{1}{N^\beta}.$$

The complete prefix free set $\mathcal{C}$ on the m-ary alphabet is given by

$$\mathcal{C} = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_K \cup \mathcal{B}_K.$$

By the above construction, it is also clear that the average delay of $\mathcal{C}$ is bounded by

$$c_1 N^3 \leq D = \sum_{d \in \mathcal{C}} P(d)\,|d| \leq c_2 N^3$$

for certain constants $c_1, c_2 > 0$. Notice further that the maximal code length satisfies

$$\max_{d\in\mathcal{C}} |d| = \mathcal{O}\left(N^3 \log N\right) = \mathcal{O}\left(D \log D\right).$$

For every $d \in \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_K$ we can choose a non-negative integer $\ell(d)$ with

$$|\ell(d) + \log_2 P(d)| < \frac{2}{N^\eta}.$$

In particular, we have

$$0 \le \ell(d) + \log_2 P(d) < \frac{2}{N^\eta}$$

if $\langle \log_2 P(d)\rangle < 2/N^\eta$ and

$$-\frac{2}{N^\eta} < \ell(d) + \log_2 P(d) \le 0$$

if $1 - \langle \log_2 P(d)\rangle < 2/N^\eta$. For $d \in \mathcal{B}_K$ we simply set $\ell(d) = \lceil -\log_2 P(d)\rceil$.

The (final) problem is now to *adjust* the choices of "+" resp. "−" in the above construction so that Kraft's inequality is satisfied. For this purpose we use the following easy property (that we adopt from Khodak [14]).

Lemma 6 (Khodak, 1972)**.** *Let $\mathcal{C}$ be a finite set with probability distribution P and suppose that for every $d \in \mathcal{C}$ we have $|\ell(d) + \log_2 P(d)| \le 1$ for a nonnegative integer $\ell(d)$. If*

$$\sum_{d\in\mathcal{C}} P(d)(\ell(d) + \log_2 P(d)) \ge 2 \sum_{d\in\mathcal{C}} P(d)(\ell(d) + \log_2 P(d))^2, \tag{24}$$

then there exists an injective mapping $C : \mathcal{C} \to \{0,1\}^*$ such that C is a prefix free set and $|C(d)| = \ell(d)$ for all $d \in \mathcal{C}$.

Proof. We again use the local expansion $2^{-x} = 1 - x\log 2 + \eta(x)$ for $|x| \le 1$, where $((\log 4)/4)x^2 \le \eta(x) \le (\log 4)x^2$. Hence

$$\begin{aligned}
\sum_{d\in\mathcal{C}|} 2^{-\ell(d)} &= \sum_{d\in\mathcal{C}|} P(d) 2^{-(\ell(d)+\log_2 P(d))} \\
&= 1 - \log 2 \sum_{d\in\mathcal{C}} P(d)(\ell(d) + \log_2 P(d)) + \sum_{d\in\mathcal{C}} P(d)\eta\left(\ell(d) + \log_2 P(d)\right) \\
&\le 1 - \log 2 \sum_{d\in\mathcal{C}} P(d)(\ell(d) + \log_2 P(d)) + 2\log 2 \sum_{d\in\mathcal{C}} P(d)(\ell(d) + \log_2 P(d))^2 \\
&\overset{(24)}{\le} 1
\end{aligned}$$

If (24) is satisfied, then Kraft's inequality follows, and there exists an injective mapping $C : \mathcal{C} \to \{0,1\}^*$ such that C is a prefix free set and $|C(d)| = \ell(d)$ for all $d \in \mathcal{C}$. ∎

We set

$$E_j = \sum_{d\in\mathcal{C}_j} P(d)(\ell(d) + \log_2 P(d)).$$

Then $E_j > 0$ if we have chosen "+" in the above construction and $E_j < 0$ if we have chosen "−". In any case we have

$$|E_j| \le P(\mathcal{C}_j)\frac{2}{N^\eta} \le \frac{2c''}{N^{1+\eta}}\left(1-\frac{c'}{N}\right)^{j-1} \le \frac{2c''}{N^{1+\eta}}.$$

Suppose for a moment that we have always chosen "+", that is $E_j > 0$ for all $j \ge 1$, and that

$$\sum_{j=1}^{K} E_j \le \frac{8+2c''}{N^{1+\eta}}. \tag{25}$$

We can assume that N is large enough that $2/N^\eta \le 1/2$. Hence, the assumptions of Lemma 6 are trivially satisfied since $0 \le \ell(d) + \log_2 P(d) < 1/2$ implies $2(\ell(d) + \log_2 P(d))^2 < \ell(d) + \log_2 P(d)$ for all $d \in \mathcal{C}$. If (25) does not hold (if we have chosen always "+"), then one can select "+" and "−" so that

$$\frac{8}{N^{1+\eta}} \le \sum_{j=1}^{K} E_j \le \frac{8+4c''}{N^{1+\eta}}.$$

Indeed, if the partial sum $\sum_{j=i}^{K} E_i \le (8+2c'')N^{-1-\eta}$, then the sign of E_j is chosen to be "+" and if $\sum_{j=i}^{K} E_i > (8+2c'')N^{-1-\eta}$ then the sign of E_j is chosen to be "−". Since

$$\sum_{d\in\mathcal{C}} P(d)(\ell(d)+\log_2 P(d))^2 \le \frac{4}{N^{2\eta}} \le \frac{4}{N^{1+\eta}} \le \sum_{d\in\mathcal{C}} P(d)(\ell(d)+\log_2 P(d))$$

the assumption of Lemma 6 is satisfied. Thus, there exists a prefix free coding map $C : \mathcal{C} \to \{0,1\}^*$ with $|C(d)| = \ell(d)$ for all $d \in \mathcal{C}$. Furthermore, the average redundancy rate is bounded by

$$\overline{r} \le \frac{1}{D}\sum_{d\in\mathcal{C}} P(d)(|C(d)| + \log_2 P(d)) \le (8+4c'')\frac{1}{DN^{1+\eta}}.$$

Since the average code length D is of order N^3 we have

$$\overline{r} = \mathcal{O}\left(D^{-1-\frac{1+\eta}{3}}\right) = \mathcal{O}\left(D^{-\frac{4+\eta}{3}}\right).$$

This proves the upper bound for $\overline{r}$ of Lemma 3.

The proof of the upper bound for r^* is very similar. The only difference is that we always use the "+" in the above construction and do not stop. We set

$$\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots .$$

By construction, every word $d \in \mathcal{C}$ satisfies

$$\langle \log_2 P(d) \rangle \le \frac{4}{N^\eta}$$

and the average delay of $\mathcal{C}$ is bounded by

$$c_1 N^3 \le D = \sum_{d\in\mathcal{C}} P(d)\,|d| \le c_2 N^3.$$

Consequently, if we set $\ell(d) = \lceil -\log_2 P(d) \rceil$, then Kraft's inequality is trivially satisfied and there exists a code C with $|C(d)| = \ell(d)$ for all $d \in \mathcal{C}$ (the Shannon code). Furthermore, we have

$$r^* = \frac{1}{D} \sup_{d\in\mathcal{C}} (|C(d)| + \log_2 P(d)) \le \frac{2}{DN^\eta} = \mathcal{O}\left(D^{-1-\frac{\eta}{3}}\right)$$

as proposed. This completes the proof of Theorem 3.

Remark. If all $\log_2 p_j$ are rational, then the above construction is (almost) trivial. There are *lots* of integers k_j such that

$$P(d) = \sum_{j=1}^{k} k_j \log_2 p_j$$

is an integer. Thus, the redundancy can be estimated by the probability of the *remaining set* $\mathcal{B}_K$.

4 Proof of Lemma 4

Lemma 4 states that for almost all $p_j > 0$ (with $p_1 + \cdots + p_m = 1$) the set

$$X = \{\langle k_1 \log_2 p_1 + \cdots + k_m \log_2 p_m \rangle : 0 \le k_j < N \ (1 \le j \le m)\}$$

has dispersion

$$\delta(X) \le N^{-m+\varepsilon} \tag{26}$$

for all sufficiently large N and for all non-zero integer vectors $(k_1, \ldots, k_m)$ we have

$$\|k_1 \log_2 p_1 + \cdots + k_m \log_2 p_m\| \ge C \left(\max_{1\le j\le m} |k_j| \right)^{-m-\varepsilon} \tag{27}$$

for some constant $C > 0$.

In view of the above, we just have to show (26) and (27) for almost all p_j. These kind of problems fall into the field of *metric Diophantine approximation* that is well established in number theory (see [4, 5, 19, 23]). One of the problems in this field is to obtain some information about the following linear forms

$$L = k_0 + k_1\gamma_1 + \cdots + k_m\gamma_m,$$

where k_j are integers and γ_j are randomly chosen real numbers. In fact, one is usually interested in lower bounds for $|L|$ in terms of $\max |k_j|$.

In our context, we have $\gamma_j = \log_2 p_j$ so that the γ_j's are related by $2^{\gamma_1} + \cdots + 2^{\gamma_m} = 1$. This means that they cannot be chosen independently. They are situated on a proper submanifold of the m-dimensional space. It has turned out that metric Diophantine approximation in this case is much more complicated than in the independent case. Fortunately, there exist now proper results that we can use for our purpose.

Theorem 5 (Dickinson and Dodson [6]). *Suppose that $m \geq 2$ and $1 \leq k < m$. Let U be an open set in $\mathbf{R}^k$ and, for $1 \leq j \leq m$, let $\Psi_j : U \to \mathbf{R}$ be C^1 real functions. Let $\eta > 0$ be real. Then for almost all $u = (u_1, \dots, u_k) \in U$, there exists $N_0(u)$ such that for all $N \geq N_0(u)$ we have*

$$|k_0 + k_1\Psi_1(u) + \cdots + k_m\Psi_m(u)| \geq N^{-m+(m-k)\eta}(\log N)^{m-k}$$

for all non-zero integer vectors $(k_0, k_1, \dots, k_m)$ with

$$\max_{1 \leq j \leq k} |k_j| \leq N \quad \text{and} \quad \max_{k < j \leq m} |k_j| \leq N^{1-\eta}/(\log N).$$

Remark. More precisely, let us define a convex body consisting of all real vectors $(y_1, \dots, y_m)$ with

$$\begin{aligned} |y_0 + y_1\Psi_1(u) + \ldots + y_m\Psi_m(u)| &\leq N^{-m+(m-k)\eta}(\log N)^{m-k}, \\ |y_j| &\leq N, \quad (j = 1, \dots, k), \\ |y_j| &\leq N^{1-\eta}(\log N)^{-1}, \quad (j = k+1, \dots, m). \end{aligned} \tag{28}$$

Dickinson and Dodson [6, p. 278] showed in the course of the proof of their Theorem 2 that the set

$$S(N) := \left\{ u \in U : \exists\ (k_0, k_1, \dots, k_m) \in \mathbf{Z}^{m+1} \text{ with } 0 < \max_{1 \leq j \leq m} |k_j| < N^{1-\eta} \text{ satisfying (28)} \right\}$$

satisfies

$$\left| \limsup_{N \to \infty} S(N) \right| = 0,$$

where $|\cdot|$ denotes the Lebesgue measure. This means that almost no u belongs to infinitely many sets $S(N)$. In other words, for almost every u, there exists $N_0(u)$ such that $u \notin S(N)$ for every $N \geq N_0(u)$. And this is stated in Theorem 5.

For $m = 2$, Theorem 5 can be improved as shown below.

Theorem 6 (R.C. Baker [3]). *Let Ψ_1 and Ψ_2 be C^3 real functions defined on an interval $[a, b]$. For x in $[a, b]$, set*

$$k(x) = \Psi_1'(x)\Psi_2''(x) - \Psi_1''(x)\Psi_2'(x).$$

Assume that $k(x)$ is non-zero almost everywhere and that $|k(x)| \leq M$ for all x in $[a, b]$ and set $\kappa = \min\{10^{-3}, 10^{-8}M^{-1/3}\}$. Then for almost all x in $[a, b]$, there are infinitely many positive integers N such that

$$|k_0 + k_1\Psi_1(x) + k_2\Psi_2(x)| \geq \kappa N^{-2}$$

for all integers k_0, k_1, k_2 with $0 < \max\{|k_1|, |k_2|\} \leq N$.

Using Theorem 5 and Theorem 6 we are now in a position to prove (26) and (27).

Proof of (27). For this purpose we can directly apply Theorem 5, where $k = m-1$ and U is an open set contained in $\Delta = \{u = (u_1, \dots, u_{m-1}) \in \mathbf{R}^{m-1} : u_1 \ge 0, \dots, u_{m-1} \ge 0, u_1 + \cdots + u_{m-1} \le 1\}$ and $\Psi_j(u) = \log_2(u_j)$ $(1 \le j \le m-1)$, resp. $\Psi_m(u) = \log_2(1 - u_1 - \cdots - u_{m-1})$. We also know that, for almost all u, the numbers $1, \Psi_1(u), \dots, \Psi_m(u)$ are linearly independent over the rationals, hence,

$$L := k_0 + k_1\Psi_1(u) + \cdots + k_m\Psi_m(u) \ne 0$$

for all non-zero integer vectors $(k_0, k_1, \dots, k_m)$.

Set $J = \max_{1\le j\le m} |k_j|$ and define N by $N^{1-\eta} = J \log N$. Assume that J is large enough to give $N \ge N_0(u)$. We then have (for suitable constants $c_1, c_2 > 0$)

$$|L| \ge N^{-m+\eta}(\log N) \ge c_1 J^{-m-(m-1)\eta/(1-\eta)} (\log J)^{(1-m)/(1-\eta)} \ge c_2 J^{-m-\varepsilon}$$

for $\varepsilon = 2(m-1)\eta/(1-\eta)$ and J large enough. This completes the proof of (27).

Proof of (26). To simplify our presentation, we first apply Theorem 6 in the case of $m = 2$ and then briefly indicate how it generalizes. First of all we want to point out that Theorems 5 and 6 are lower bounds for the homogeneous linear form $L = k_0 + k_1\Psi_1(u) + \cdots + k_m\Psi_m(u)$ in terms of $\max |k_j|$. Using techniques from "Geometry of Numbers" (see below) these lower bounds can be transformed into upper bounds for the dispersion of the set $X = \{\langle k_1\Psi_1(u) + \cdots + k_m\Psi_m(u)\rangle : 0 \le k_1, \dots, k_m < N\}$.

In particular we will use the notion of successive minima of convex bodies. Let $B \subseteq \mathbf{R}^d$ be a 0-symmetric convex body. Then the successive minima λ_j are defined by $\lambda_j = \inf\{\lambda > 0 : \lambda B$ contains j linearly independent integer vectors$\}$. One of the first main results of "Geometry of Numbers" is *Minkowski's Second Theorem* saying that $2^d/d! \le \lambda_1 \cdots \lambda_d \mathrm{Vol}_d(B) \le 2^d$, see [5, 19].

Let x and N be the same as Theorem 6 and consider the convex body $B \subseteq \mathbf{R}^3$ that is defined by the inequalities

$$\begin{aligned} |y_0 + y_1\Psi_1(x) + y_2\Psi_2(x)| &\le \kappa N^{-2}, \\ |y_1| &\le N, \\ |y_2| &\le N. \end{aligned}$$

By Theorem 6 the set B does not contain a non-zero integer point. Thus, the first minimum λ_1 of B is ≥ 1. Note that $\mathrm{Vol}_3(B) = 8\kappa$. Then from Minkowski's Second Theorem we conclude that the three minima of this convex body satisfy $\lambda_1\lambda_2\lambda_3 \le 1/\kappa$. Since $1 \le \lambda_1 \le \lambda_2$ we thus get $\lambda_3 \le \lambda_1\lambda_2\lambda_3 \le 1/\kappa$ and consequently $\lambda_1 \le \lambda_2 \le \lambda_3 \le 1/\kappa$. In other words, there exist constants κ_2 and κ_3, and three linearly independent integer vectors (a_0, a_1, a_2), (b_0, b_1, b_2) and (c_0, c_1, c_2) such that

$$\begin{aligned} |a_0 + a_1\Psi_1(x) + a_2\Psi_2(x)| &\le \kappa_2 N^{-2}, \\ |b_0 + b_1\Psi_1(x) + b_2\Psi_2(x)| &\le \kappa_2 N^{-2}, \\ |c_0 + c_1\Psi_1(x) + c_2\Psi_2(x)| &\le \kappa_2 N^{-2}, \\ \max\{|a_i|, |b_i|, |c_i|\} &\le \kappa_3 N. \end{aligned}$$

Using these linearly independent integer vectors, we can show that the dispersion of

$$X = \{\langle k_1\Psi_1(x) + k_2 \log_2 \Psi_2(x)\rangle : 0 \le k_1, k_2 \le 7\kappa_3 N\}$$

is small.

Let ξ be a real number (that we want to approximate by an element of X) and consider the (regular) system of linear equations

$$\begin{aligned}
-\xi + \theta_a(a_0 + a_1\Psi_1(x) + a_2\Psi_2(x)) + & \\
+\theta_b(b_0 + b_1\Psi_1(x) + b_2\Psi_2(x)) + \theta_c(c_0 + c_1\Psi_1(x) + c_2\Psi_2(x)) &= 4\kappa_2 N^{-2}, \\
\theta_a a_1 + \theta_b b_1 + \theta_c c_1 &= 4\kappa_3 N, \qquad (29) \\
\theta_a a_2 + \theta_b b_2 + \theta_c c_2 &= 4\kappa_3 N.
\end{aligned}$$

Denote by $(\theta_a, \theta_b, \theta_c)$ its unique solution and set

$$t_a = \lfloor\theta_a\rfloor, \quad t_b = \lfloor\theta_b\rfloor, \quad t_c = \lfloor\theta_c\rfloor,$$

and

$$k_j = t_a a_j + t_b b_j + t_c c_j \quad (j = 0, 1, 2).$$

Of course, k_0, k_1, k_2 are integers and from the second and third equation of (29) combined with $\max\{|a_i|, |b_i|, |c_i|\} \le \kappa_3 N$ it follows that

$$\kappa_3 N \le \min\{k_1, k_2\} \le \max\{k_1, k_2\} \le 7\kappa_3 N,$$

in particular, k_1 and k_2 are positive integers. Moreover, by considering the first equation of (29) we see that

$$\kappa_2 N^{-2} \le -\xi + k_0 + k_1\Psi_1(x) + k_2\Psi_2(x) \le 7\kappa_2 N^{-2}.$$

Since this estimate is independent of the choice of ξ this implies

$$\delta(X) \le 7\kappa_2 N^{-2}.$$

Clearly, we can apply this procedure for the functions $\Psi_1(x) = \log_2 x$ and $\Psi_2(x) = \log_2(1 - x)$ and for any interval $[a, b]$ with $0 < a < b < 1$.

This also shows that we can choose $\varepsilon = 0$ in the case $m = 2$ for infinitely many N in Lemma 3, provided that we introduce an (absolute) numerical constant.

Finally, we discuss the general case $m \ge 2$ (and prove Lemma 4). We consider the convex body $B \subseteq \mathbf{R}^{m+1}$ that has volume 2^{m+1} and is defined by (28):

$$\begin{aligned}
|y_0 + y_1\Psi_1(u) + \ldots + y_m\Psi_m(u)| &\le N^{-m+(m-k)\eta}(\log N)^{m-k}, \\
|y_j| &\le N, \quad (j = 1, \ldots, k), \\
|y_j| &\le N^{1-\eta}(\log N)^{-1}, \quad (j = k+1, \ldots, m).
\end{aligned}$$

By assumption, the first minimum λ_1 of B satisfies $\lambda_1 \geq N^{-\eta}$, thus, by Minkowski's Second Theorem, its last minimum λ_m is bounded by $\lambda_m \leq N^{n\eta}$. Consequently, we have $n+1$ linearly independent vectors $\mathbf{q}^{(i)}$, $i = 0, \ldots, m$, such that

$$\|\mathbf{q}^{(i)} \cdot \Psi(u)\| \leq N^{-m+(m-k)\eta+m\eta}(\log N)^k, \qquad \|\mathbf{q}^{(i)}\|_\infty \leq N^{1+m\eta}.$$

We now argue as above, and consider a system of linear equations analogous to (29). Hence, for any real number ξ, there are positive integers $k_1, \ldots, k_m$ such that

$$\| -\xi + k_1\Psi_1(u) + \ldots + k_m\Psi_m(u)\| < \frac{1}{N^{m-\varepsilon}}, \qquad \max k_j \leq N,$$

where $\varepsilon > 0$ can be made arbitrarily small by taking sufficiently small values of η. Applied to the functions $\Psi_j(u) = \log_2(u_j)$ $(1 \leq j \leq m-1)$ and $\Psi_m(u) = \log_2(1 - u_1 - \cdots - u_{m-1})$, this proves (26). This completes the proof of Lemma 4.

5 Proof for Markov Sources

In this section, we extend our results to Markov sources of order 1 (Theorem 4) by indicating necessary changes in our previous proofs.

We assume that the transition matrix of the Markov source is given by

$$\mathbf{P} = (p_{ij})_{1 \leq i,j \leq m},$$

where $p_{ij} = \Pr\{X_{k+1} = j \,|X_k = i\} > 0$. The stationary distribution $p_1, \ldots, p_m$ is then uniquely defined by $p_j = \sum_{i=1}^m p_i p_{ij}$. For example, for $m = 2$ we have

$$p_1 = \frac{p_{21}}{p_{21} + p_{12}} \quad \text{and} \quad p_2 = \frac{p_{12}}{p_{21} + p_{12}}.$$

The probability of a message x_1^n becomes

$$P(x_1^n) = \hat{p} \prod_{i,j=1}^m p_{ij}^{k_{ij}},$$

where $\hat{p} = p_\ell$ if $x_0 = \ell$ and k_{ij} is the size of the set $\{k \in \{1, \ldots, n-1\} : (x_k, x_{k+1}) = (i,j)\}$. Note that there are some consistency conditions:

$$\sum_{i,j=1}^m k_{ij} = n - 1,$$
$$\sum_{i=1}^m k_{ij} = \sum_{i=1}^m k_{ji} + \nu_j(x_1^n) \quad (1 \leq j \leq m),$$

where $\nu = \nu_j(x_1^n) \in \{0, 1, -1\}$ depending on x_1 and x_n. For example, if $x_1 = x_n$ then $\nu = 0$. We call a vector $\mathbf{k} = (k_{ij})$ of integers *admissible* if it satisfies these conditions. This means that if n is not fixed then we can only vary $m^2 - m + 1$ of the m^2 "parameters" k_{ij} "independently". For example, if $m = 2$ then we can represent $\log_2 P(x_1^n)$ by

$$\log_2 P(x_1^n) = c_0 + k_{11} \log_2 p_{11} + k_{12} \log_2 (p_{12} p_{21}) + k_{22} \log_2 p_{22}, \tag{30}$$

where $c_0 = c_0(x_1, x_n)$ attains finitely many possible values.

We will further need the following asymptotic expansions which can be found in [12, Theorem 5] and Whittle [25]. For $a, b \in \{1, \ldots, m\}$ and an admissible integer vector $\mathbf{k} = (k_{ij})$ let $N_{\mathbf{k}}^{a,b}$ denote the number of sequences of length $n = \sum_{i,j=1}^m k_{ij} + 1$, where $x_0 = a$, $x_n = b$. Then

$$N_{\mathbf{k}}^{a,b} \sim \frac{k_{ba}}{k_b} \cdot \det{}_{bb}(\mathbf{I} - \mathbf{k}^*) \cdot \binom{k_1}{k_{11}, \ldots, k_{1m}} \cdots \binom{k_m}{k_{m1}, \ldots, k_{mm}}, \tag{31}$$

where $k_j = \sum_{i=1}^m k_{ij}$, $\mathbf{k}^* = (k_{ij}/k_i)_{1 \le i,j \le m}$ and $\det_{bb}(\mathbf{I} - \mathbf{k}^*)$ is the determinant of $\mathbf{I} - \mathbf{k}^*$ in which row b and column b are deleted.

With the help of these formulae, we can prove corresponding properties for Markov sources. In particular, we get a slightly modified **Lemma 2**. Instead of $X = \{\langle k_1\gamma_1 + \cdots + k_m\gamma_m\rangle : 0 \le k_j < N \ (1 \le j \le m)\}$ we must work with

$$X = \left\{ \left\langle c_0 + \sum_{i,j=1}^m k_{ij}\gamma_{ij} \right\rangle : \mathbf{k} \text{ admissible and } 0 \le k_{ij} < N \ (1 \le i, j \le m) \right\} \tag{32}$$

for some c_0. In particular, for $m = 2$ such a set can be represented as

$$X = \{\langle c_0 + k_{11}\gamma_1 + k_{12}(\gamma_{12} + \gamma_{21}) + k_{22}\gamma_{22}\rangle : 0 \le k_{11}, k_{12}, k_{22} < N\}$$

Clearly, we get the same result for this modified set X.

Next we have to get an analogue to **Lemma 3**. We assume that the dispersion of the set as in 32) is bounded by $\delta(X) \le 2/N^\eta$ and show that there exist codes with average code length $D = \Theta(N^{m+2})$, of maximal code length of order $\Theta(N^{m+2}\log N)$ and of average redundancy rate $\overline{r} = O(D^{-1-\frac{\eta+1}{m+2}})$. Furthermore there exist codes with average code length $D = \Theta(N^{m+2})$ and worst case redundancy $r^* = O(D^{-1-\frac{\eta}{m+2}})$.

The only difference in the proof is that (23) has to be replaced by a similar inequality. Suppose that $p_{ij} > 0$ constitute the transition probabilities and let p_j be the stationary distribution. Set $k_{ij} = \lfloor p_i p_{ij} N^2 \rfloor$ $(i, j \in \{1, \ldots, m\})$ and suppose that $0 \le k'_{ij} \le N$ $(i, j \in \{1, \ldots, m\})$ with $k'_{01} = k'_{1,0}$. Then we have for some constants c', c''.

$$\frac{c'}{N^m} \le N_{\mathbf{k}+\mathbf{k}'}^{a,b}\, p_a \prod_{i,j=1}^m p_{ij}^{k_{ij}+k'_{ij}} \le \frac{c''}{N^m}$$

where $N_{\mathbf{k}}^{a,b}$ is defined above (31). As in the proof of (23), this follows from (31) and Stirling's formula.

Now the (modified) proof of Lemma 3 follows the same footsteps as in the memoryless case. Instead of $k_i^0 = \lfloor p_i N^2 \rfloor$ we use $k_{ij} = \lfloor p_i p_{ij} N^2 \rfloor$ and so on.

Now part (i) of **Theorem 4** follows immediately. We just have to set $\eta = 1$.

There is even a modified **Lemma 4**. We have to apply Theorem 5 for properly chosen $\Psi_j(u)$ $(1 \le j \le m^2 - m + 1)$ with $k = m^2 - m$. Hence the upper bound of (ii) of **Theorem 4** holds by applying the modified Lemma 3 with $\eta = m^2 - m + 1 - \varepsilon$.

There is only one slight change in the proof of part (iii) of **Theorem 4**. Since the linear form in (30) is not homogeneous in k_{ij} we have to add an additional variable that is always set to 1 and apply the above procedure. This results in showing that for almost all Markov sources we have for all probabilities $P(x_1^n)$

$$\| \log_2 P(x_1^n) \| \ge C \, (\max k_{ij})^{-(m^2-m+2)-\varepsilon} \, .$$

This is the reason why the exponent $m^2 - m + 2$ appears instead of "expected exponent" $m^2 - m + 1$.

References

[1] J. Abrahams, Code and parse trees for lossless source encoding, *Proc., Compression and Complexity of SEQUENCES '97*, Positano, Italy, 1997.

[2] J Allouche and J. Shallit, *Automatic Sequences*, Cambridge University Press, Cambridge, 2003.

[3] R. C. Baker, Dirichlet's theorem on Diophantine approximation, *Math. Proc. Cambridge Philos. Soc.* 83, 37–59, 1978.

[4] V. I. Bernik and M. M. Dodson, *Metric Diophantine approximation on manifolds*, Cambridge Tracts in Mathematics, 137, Cambridge University Press, Cambridge, 1999.

[5] J. W. S. Cassels, *An Introduction to Diophantine Approximation*, Cambridge University Press, 1957.

[6] H. Dickinson and M. M. Dodson, Extremal manifolds and Hausdorff dimension, *Duke Math. J.* 101, 271–281, 2000.

[7] M. Drmota and R. Tichy, *Sequences, Discrepancies, and Applications*, Springer Verlag, Berlin Heidelberg, 1997.

[8] M. Drmota, Y. Reznik, S. Savari, and W. Szpankowski, Precise Asymptotic Analysis of the Tunstall Code *IEEE Intern. Symposium on Information Theory*, 2334-2337, Seattle, 2006.

[9] F. Fabris, Variable-length-to-variable-length source coding: A greedy step-by-step algorithm (Corresp.), *IEEE Trans. Info. Theory*, 38, 1609 - 1617, 1992.

[10] Freeman, G.H.; Divergence and the construction of VV-length lossless codes by source-word extensions Data Compression Conference, 1993. DCC '93., 79-88, 1993

[11] R. G. Gallager, *Discrete Stochastic Processes*, Kluwer, Boston 1996.

[12] P. Jacquet and W. Szpankowski, Markov Types and Minimax Redundancy for Markov Sources, *IEEE Trans. Information Theory*, 50, 1393-1402, 2004.

[13] F. Jelinek and K. S. Schneider, On variable-length-to-block coding, *Trans. Information Theory* IT-18, 765-774, 1972.

[14] G.L. Khodak, Bounds of redundancy estimates for word-based encoding of sequences produced by a Bernoulli source (Russian), *Problemy Peredachi Informacii* 8, 21–32, 1972.

[15] R. Krichevsky, *Universal Compression and Retrieval,* Kluwer, Dordrecht, 1994.

[16] S. A. Savari, Variable-to-Fixed Length Codes and the Conservation of Entropy, *Trans. Information Theory* 45, 1612-1620, 1999.

[17] S. A. Savari and R. G. Gallager, Generalized Tunstall codes for sources with memory, *IEEE Trans. Inform. Theory*, 43, 658-668, 1997.

[18] S. Savari and W. Szpankowski, On the Analysis of Variable-to-Variable Length Codes Bell Labs Technical Memorandum (10009642-011025-01TM), 2002 (see also *2002 International Symposium on Information Theory*, Lausanne 2002).

[19] W. M. Schmidt, *Diophantine Approximation*, Lecture Notes Math. 785, Springer, Berlin, 1980.

[20] V. M. Sidel'nikov, Statistical Properties of Transformations Realized by Finite Automata, *Kibernetika* 6, 1-14, 1965. (In Russian)

[21] W. Szpankowski, Asymptotic Average Redundancy of Huffman (and other) Block Codes, *Trans. Information Theory* 46, 2434-2443, 2000.

[22] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001.

[23] V. G. Sprindžuk, *Metric Theory of Diophantine Approximations*, Scripta Ser. Math., Wiley, New York, 1979.

[24] B. P. Tunstall, "Synthesis of Noiseless Compression Codes," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, 1967.

[25] P. Whittle, Some Distribution and Moment Formulæ for Markov Chain, *J. Roy. Stat. Soc.,* Ser. B., 17, 235–242, 1955.