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

October 22, 2004

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 rate bounded by $D^{-5/3}$, where D is the average delay (e.g., the average length of a phrase). 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 extend the above result to Markov sources and prove that the worst case redundancy (i.e., for individual sequences) does not exceed $D^{-4/3}$. Finally, we consider bounds that are valid for *almost all* memoryless and Markov sources, that is, for almost all sources there exists a variable-to-variable code such that its average redundancy rate is bounded by $D^{-4/3-m/3+\varepsilon}$ while its worst case redundancy rate 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-variables codes the average and the worst case redundancy rates are at least $D^{-2m-1-\varepsilon}$ for almost all sources. Regarding proof techniques, we use throughout Diophantine approximation; in particular, we investigate Diophantine approximation problem on m-ary trees.

Index Terms: Variable-to-Variable length codes, average and maximal redundancy rates, metric Diophantine approximation, geometry of numbers, Minkowski's theorem.

[*] 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 first author was supported by the Austrian Science Foundation FWF, grant M822-N12.

[‡] The work of the third author was supported by NSF Grant CCR-0208709, the NIH grant R01 GM068959-01, and AFOSR Grant FA8655-04-1-3074.

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 [8] 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 (average and maximal) redundancy rates of VV codes by re-examining and expanding a thirtysome years old paper by Khodak [13] who in 1972 claimed to provide asymptotic upper and lower bounds for the achievable redundancy rate of variable-to-variable length codes. However, Khodak did not offer explicit variable-to-variable length codes that satisfy these bounds. Here, we present a transparent (and simplified) proof, generalize Khodak's results (e.g., we analyze maximal redundancy, almost all sources, Markov sources), 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 variable-to-variable length 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 $\mathcal{D}$, that is, $x^i \in \mathcal{D}$. We shall write d or d_i for a dictionary entry, and by D we denote the average dictionary (phrase) length (average delay). A convenient way of representing the dictionary $\mathcal{D}$ is by a complete tree that we shall call the *parsing tree*. Next, the string encoder in a variable-to-variable 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 basically the Shannon code and we concentrate on building a parsing tree for which $\log P(d)$ ($d \in \mathcal{D}$) is close to an integer. This allows us to construct a VV code with redundancy approaching zero as the average delay increases.

More precisely, for large delay D we shall show in Theorem 1 that there exists a variable-to-variable code such that for memoryless sources the average redundancy rate *decays* as $D^{-5/3}$. This result basically belongs to Khodak [13] except we present here a transparent proof and an efficient algorithm of a constructible variable-to-variable code. Next, we extend this result in several directions. First, we show that for such codes the worst case redundancy rate decays as $D^{-4/3}$, and both bounds hold also for Markov sources. More importantly, we study new bounds for *almost all* sources (i.e., in our case for almost all symbol probabilities). In particular, we show that for almost all sources there exists a variable-to-variable 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 and $\varepsilon > 0$. Finally, we conclude our analysis with a lower bound showing that for all variable-to-variables codes and for almost all 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 code and Huffman code) and variable-to-fixed code lengths (e.g., Tunstall codes). Abrahams [1] discusses literature on fixed-to-variable length codes. For a memoryless source, Szpankowski [20] provides a precise asymptotic analysis of the Huffman and other codes for fixed length blocks of source symbols. While it was known since Shannon that the redundancy rate for such codes is $O(1/D)$ (in this case D is fixed and equal to the block length), in [20] it is shown that the average redundancy rate either converges to a $constant/D$ (e.g., $0.5/D$ in the case of the Shannon code) or it exhibits very erratic behavior fluctuating between 0 and $constant/D$.

An important example of a variable-to-fixed code is the Tunstall code [23]. Savari and Gallager [16] 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 $O(1/D)$. From this brief discussion, we conclude that while FV and VF codes waste a fraction of a bit per source symbol, the VV code presented in this paper is losing a negligible information per phrase when the phase length increases.

There is scarcely any literature on VV codes with a few exceptions such as [8, 9, 13, 17]. The most interesting, as already mentioned, is an old work of Khodak [13]. To the best of our knowledge not much was done since then, except that Fabris [8] (cf. also [9, 17]) analyzed Tunstall–Huffman variable-to-variable code and provided a simple bound on its redundancy rate.

Finally, we say a word about our proof techniques. The main tools come from Diophantine approximation [5, 18]. The basic thrust of this theory is the study of small values of linear forms like $k_0 + k_1\gamma_1 + \cdots + k_m\gamma_m$ 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 $k_1 \log p_1 + \cdots + k_m \log p_m$ where p_i for $1 \leq i \leq m$ is the probability of generating the ith symbol of an m-ary alphabet. Therefore, it is natural to apply techniques from Diophantine approximation. For our almost sure results we also need non-trivial results on metric Diophantine approximation on manifolds since $\log p_1, \ldots, \log p_m$ are not independent; indeed, $p_1 + \cdots + p_m = 1$.

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, following by the presentation of our main results. We first consider redundancy rates for all sources (cf. Theorem 1) and then for *almost all* 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 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). In Section 3 we prove Theorem 1 and in Section 4 we present more challenging proof of Theorem 2 for almost all sources. Finally, in the last Section 5 we show how to extend our findings to Markov sources.

2 Main Results and Their Consequences

In this section we first define the average and the maximal redundancy rates for variable-to-variable (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. Finally, we consider *almost* all sources and describe our findings (cf. Theorem 2). We complete our analysis with a lower bound for the redundancy rates (cf. Theorem 3).

2.1 Redundancy Rates for VV Codes

Let us first introduce formally redundancy rates for variable-to-variable codes. Actually, this needs some care and we shall discuss it in depth below. We first define (asymptotic) *average* redundancy rate and then extend it to the *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 variable-to-variable codes.

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

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

where $|d|$ is the length of $d \in \mathcal{D}$. 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 [17] we arrive at

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

An application of the *Conservation of Entropy Theorem* [14, 15, 19], as in [17], leads to

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

which we adopt as our definition of the average redundancy rate.[1] Observe that (4) decomposes the redundancy rate of the variable-to-variable 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"

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

probabilities $\{P_{\mathcal{D}}(d),\ d \in \mathcal{D}\}$. Hereafter, we mostly deal with the probability induced by the dictionary, so we shall denote it as $P = P_{\mathcal{D}}$.

We shall also study (asymptotic) redundancy rate r^* for *individual sequences* defined as

$$r^* = \lim_{|x|\to\infty} \frac{\max_x [L(x) + \log P_{\mathcal{S}}(x)]}{|x|}. \tag{5}$$

Recall that the source sequence is a concatenation of phrases $x^1, x^2, \ldots, x^k$, where $k \to \infty$ and let $\ell(x^i)$ be the code length assigned to x^i. But $|x| = \sum_{i=1}^k |x^i|$, $L(x) = \sum_{i=1}^k \ell(x^i)$, and $\log P_{\mathcal{S}}(x) = \sum_{i=1}^k \log P_{\mathcal{S}}(x^i)$, hence

$$\begin{aligned}
r^* &= \lim_{k\to\infty} \frac{\max_{x^1,\ldots,x^k} \sum_{i=1}^k [\ell(x^i) + \log P_{\mathcal{S}}(x^i)]}{\sum_{i=1}^k |x^i|} \\
&= \lim_{k\to\infty} \frac{\sum_{i=1}^k \max_{x^i} [\ell(x^i) + \log P_{\mathcal{S}}(x^i)]}{\sum_{i=1}^k |x^i|} \\
&= \lim_{k\to\infty} \frac{\sum_{i=1}^k \max_{d\in\mathcal{D}} [\ell(d) + \log P(d)]}{\sum_{i=1}^k |x^i|} \\
&= \lim_{k\to\infty} \frac{k \max_{d\in\mathcal{D}} [\ell(d) + \log P(d)]}{\sum_{i=1}^k |x^i|} \\
&= \frac{\max_{d\in\mathcal{D}} [\ell(d) + \log P(d)]}{D} \qquad (a.s.).
\end{aligned}$$

In conclusion, we adopt the following definition of the *maximal* or the *worst case* redundancy

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

The main purpose of this work is to construct a (complete) prefix free set (dictionary) $\mathcal{D}$ (i.e., a complete tree) on the input alphabet $\mathcal{A}$ and a bijective mapping C (a variable-to-variable code) to another prefix free set on the binary alphabet $\{0, 1\}$ with 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 basically the Shannon code that assigns $\lceil -\log P(d) \rceil$ code length to the dictionary word $d \in \mathcal{D}$. Our goal is to build a dictionary (i.e., a complete parsing tree) such that the code length $\lceil -\log P(d) \rceil$ of the word d (a leaf in the parsing tree) is close to the ideal code length $-\log P(d)$.

Let k_i for $i = 1, \ldots, m$ be the number of symbols $a_i \in \mathcal{A}$ in $d \in \mathcal{D}$, and let p_i be the probability of generating symbol a_i. We assume throughout that p_i are *given.* Then $P(d) = p_1^{k_1} \cdots p_m^{k_m}$ and the numerator of the average redundancy rate becomes

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

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 $k_1, \dots k_m$ such that the linear form $k_1\gamma_1 + k_2\gamma_2 + \cdots + k_m\gamma_m$ is close to an integer. In the sequel we shall study properties 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.

To analyze $\langle k_1\gamma_1 + k_2\gamma_2 + \cdots + k_m\gamma_m \rangle$ we need to introduce the notion of *dispersion* and recall some properties of *continued fraction*.

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

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

where c_0 is an integer and c_j are *positive* integers for $j \geq 1$. We denote this rational number as $[c_0, c_1, \dots, c_n]$. With help of the Euclidean algorithm, it is easy to see that every rational number has a finite continued fraction expansion. Furthermore, if c_j is a given sequence of integers (that are positive for $j > 0$), then the limit $\gamma = \lim_{n\to\infty}[c_0, c_1, \dots, c_n]$ exists and is denoted by the infinite continued fraction expansion $\gamma = [c_0, c_1, c_2 \dots]$. Conversely, if $\gamma = \gamma_0$ is a real irrational number and if we recursively set

$$c_j = \lfloor \gamma_j \rfloor, \qquad \gamma_{j+1} = 1/(\gamma_j - c_j),$$

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

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

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

where integers p_n, q_n are coprime. Setting $p_{-1} = 1$, $q_{-1} = 0$, $p_0 = c_0$ and $q_0 = 1$, 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 fast. Furthermore, the convergent $\frac{p_n}{q_n}$ is (in some sense) the best rational approximations of γ, that is,

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

The denominator q_n is called the *best approximation denominator* [5].

Dispersion. Let $\|x\| = \min(\langle x \rangle, 1 - \langle x \rangle)$ 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 $\|y - x\| \le \delta(X)$. 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. Therefore, there exists $x_1 \in X$ with $\langle y - x_1 \rangle \le 2\delta(X)$ and $x_2 \in X$ with $\langle x_2 - y \rangle \le 2\delta(X)$.

The following property will be used throughout this paper.

Lemma 1. *Suppose that γ is an irrational number and $N = q_n$ is 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 $\gamma = [c_0, c_1, c_2, \ldots]$). Then*

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

Proof. For $N = q_n$ we obtain from continued fraction theory (cf. [5, 7])

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

or

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

for some $|\theta| < 1$. Consequently, the numbers $\langle k\gamma \rangle$ $(0 \le k < N = q_n)$ are quite close to the numbers $0, 1/N, 2/N, \ldots, (N-1)/N$, in particular, for every $k < N$ there exists $l < N$ with $|k\gamma - l/N| < (N-1)/N^2 < 1/N$. Furthermore, if y is an arbitrary number in $[0,1)$, then there is a number l/N with $\|y - l/N\| \le (1/2)/N$ and another number $k\gamma$ such that $\|k\gamma - l/N\| < 1/N$. Indeed, the numbers l/N mod 1 are spaced equally $1/N$ apart, hence the former inequality must hold. Thus

$$\|y - k\gamma\| \le \|y - l/N\| \ + \|k\gamma - l/N\| \ < 2/N.$$

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

Corollary 1. *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 a best approximation denominator of one of the irrational numbers. Then the dispersion of the set*

$$X = \{\langle k_1\gamma_1 + \cdots + k_m\gamma_m \rangle : 0 \le k_j < N \ (1 \le j \le m)\}$$

is bounded by

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

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

$$\left|\gamma - \frac{M}{N}\right| < \frac{1}{N^2} \tag{7}$$

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 (7) 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, that satisfies (7), can be effectively computed.

Existence of a VV Code. The existence of a VV code will be establishing by invoking the following lemma, already proved in [13].

Lemma 2. *Let $\mathcal{D}$ be a finite set with probability distribution P and suppose that for every $d \in \mathcal{D}$ we have $|\ell(d) + \log_2 P(d)| \leq 1$ for a nonnegative integer $\ell(d)$. If*

$$\sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d)) \geq 2 \sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d))^2, \tag{8}$$

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

Proof. Suppose that $|x| \leq 1$. Then we have $2^{-x} = 1 - x\log 2 + \eta(x)$ with $((\log 4)/4)x^2 \leq \eta(x) \leq (\log 4)x^2$. Hence

$$\begin{aligned}
\sum_{d\in\mathcal{D}|} 2^{-\ell(d)} &= \sum_{d\in\mathcal{D}|} P(d) 2^{-(\ell(d)+\log_2 P(d))} \\
&= 1 - \log 2 \sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d)) + \sum_{d\in\mathcal{D}} P(d)\eta\left(\ell(d) + \log_2 P(d)\right) \\
&\leq 1 - \log 2 \sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d)) + 2\log 2 \sum_{d\in\mathcal{D}} P(d)(\ell(d) + \log_2 P(d))^2 \\
&\overset{(8)}{\leq} 1.
\end{aligned}$$

If (8) is satisfied, then Kraft's inequality follows, and there exists an injective mapping $C : \mathcal{D} \to \{0,1\}^*$ such that $C(\mathcal{D})$ is a prefix free set and $|C(d)| = l_d$ for all $d \in \mathcal{D}$. ■

We are now finally ready to formulate our first main result concerning average and maximal redundancy rates for *all* $0 < p_i < 1$, $i = 1, \ldots, m$. The proof for memoryless sources is presented in Section 3, while Markov source is treated in Section 5.

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

$$\overline{r} = O(D^{-5/3}), \tag{9}$$

and the maximal phrase length is $O(D \log D)$. Furthermore, there also exists a variable-to-variable code with the average delay $D \geq D_0$ such that the worst case redundancy rate satisfies

$$r^* = O(D^{-4/3}), \tag{10}$$

however, the maximal phrase length might be infinite.

The estimate (9) for $\overline{r}$ is the same as in Khodak [13]. However, the proof presented in [13] is rather sketchy and complicated. Our method uses similar ideas to those in [13] 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 effective algorithm for an explicit VV-code with arbitrarily large average dictionary length D and small redundancy rate for any given probability distribution $p_1, \ldots, p_m > 0$ on an m-ary alphabet (and a memoryless source). 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.

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.

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) a positive real number $\varepsilon < 1$.

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

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

1. **Calculate** a 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** $k_j^0 = \lfloor p_j N^2 \rfloor$ $(1 \leq j \leq m)$, $x = \sum_{j=1}^m k_j^0 \log_2 p_j$, and $n_0 = \sum_{j=1}^m k_j^0$.
3. **Set** $\mathcal{D} = \emptyset$, $\mathcal{B} = \{\omega\}$, and $p = 0$

```
while $p < 1 - \varepsilon/4$ do
        For all $r \in \mathcal{B}$ of minimal length
        $b \leftarrow \log_2 P(r)$
        Find $0 \leq k < N$ that solves the congruence $kM \equiv 1 - \lfloor (x+b)N \rfloor \bmod N$
        $n \leftarrow n_0 + k$
        $\mathcal{D}' \leftarrow \{d \in \mathcal{A}^n : \text{type}(d) = (k_1^0, \ldots, k_{m-1}^0, k_m^0 + k)\}$
        $\mathcal{D} \leftarrow \mathcal{D} \cup r \cdot \mathcal{D}'$
        $\mathcal{B} \leftarrow (\mathcal{B} \setminus \{r\}) \cup r \cdot (\mathcal{A}^n \setminus \mathcal{D}')$
        $p \leftarrow p + P(r)P(\mathcal{D}')$, where
                $P(\mathcal{D}') = \frac{n!}{k_1^0! \cdots k_{m-1}^0!(k_m^0 + k)!} p_1^{k_1^0} \cdots p_{m-1}^{k_{m-1}^0} p_m^{k_m^0 + k}.$
end while.
```

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

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

The main step of the algorithm is a loop of the same subroutine. The input of the subroutine is a word $r \in \mathcal{B}$. The first output is an integer k with $0 \le k < N$ that satisfies $1/N \le \langle kM/N + x + b \rangle \le 2/N$. This can be realized by solving the congruence $kM \equiv 1 - \lfloor (x+b)N \rfloor \bmod N$ (e.g., with help of the Euclidean algorithm). This choice ensures that

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

For this k we can calculate the next output that consists of the set $\mathcal{D}'$ of all words d of an m-ary alphabet of type$(d) = (k_1^0, \ldots, k_{m-1}^0, k_m^0 + k)$. By construction all $d \in \mathcal{D}'$ satisfy

$$\langle \log_2 P(rd) \rangle = \langle k \log_2 p_m + x + b \rangle \le \frac{3}{4}\varepsilon.$$

Hence for all words in $d \in \mathcal{D}$ that are constructed in the above algorithm we have

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

Furthermore, in each step of the algorithm we have $P(\mathcal{D}) + P(\mathcal{B}) = 1$ and the union $\mathcal{D} \cup \mathcal{B}$ constitute a complete prefix free set on the alphabet $\mathcal{A}$. The algorithm terminates when $p = P(\mathcal{D}) > 1 - \varepsilon/4$. (The proof of Theorem 1 shows that it definitely terminates when the average dictionary length D is of order $O(N^3)$.) The redundancy can be estimated by

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

Thus we constructed a parsing tree and a VV code with a small redundancy rate.

2.4 Redundancy Rates for Almost All Sources

In this section we present better estimates for the redundancy rates but valid only for *almost all* sources, that is, almost all p_j such that $\sum_{j=1}^m p_j = 1$ and $p_j > 0$ for all $1 \le j \le m$. From mathematical point of view, these results are more challenging.

While Lemma 1 laid foundation for Theorem 1, the next lemma, which we prove in Section 4, is crucial for present results.

Lemma 3. *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

$$\delta(X) \le \frac{1}{N^{m-\varepsilon}} \tag{11}$$

for sufficiently large N. In addition, for almost all $p_j > 0$ we have

$$\|k_1 \log_2 p_1 + \cdots + k_m \log_2 p_m\| \gg \left(\max_{1 \le j \le m} |k_j| \right)^{-m-\varepsilon} \tag{12}$$

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 $p_1 > 0$ and $p_2 > 0$ 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

$$\delta(X) \le \frac{\kappa}{N^2}. \tag{13}$$

The estimate (13) is a little bit sharper than (11). However, it is only valid for infinitely many N and not for all but finitely many.[2]

Lemma 3, proved in Section 4, is used to establish our second main result valid for almost all sources.

Theorem 2. *Let $m \ge 2$ and $\mathcal{S}$ be a memoryless or a Markov source on an alphabet of size m. For memoryless source $\mathcal{S}$ we set $m' = m$, while for Markov source $\mathcal{S}$ we assume $m' = m^2 - m + 1$. Then for almost all source parameters, and for every sufficiently large D_0, there exists a variable-to-variable code with the average delay D satisfying $D_0 \le D \le 2D_0$ such that its average redundancy rate is bounded by*

$$\overline{r} \le D^{-\frac{4}{3} - \frac{m'}{3} + \varepsilon}, \tag{14}$$

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

$$r^* \le D^{-1 - \frac{m'}{3} + \varepsilon}. \tag{15}$$

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.

[2] We point out that (11) and (13) are close from being optimal. Since the set X consists of N^m points the dispersion must satisfy $\delta(X) \ge \frac{1}{2} N^{-m}$.

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.

Finally, we present one lower bound for redundancy rates that is valid for almost all sources. It will follow from (12) of Lemma 3 and the following simple lower bound (cf. Corollary 1 in [13]).

Lemma 4. *Let $\mathcal{D}$ be a finite set with probability distribution P. Then*

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

for a certain constant $c > 0$.

Proof. We have as in the proof of Lemma 2

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

where $((\log 4)/4)x^2 \le \eta(x) \le (\log 4)x^2$. Then

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

Hence, by Kraft's inequality and by the observation

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

the result follows immediately. ■

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

Theorem 3. *Let $m \ge 2$ and $\mathcal{S}$ be a memoryless or a Markov source on an alphabet of size m. For memoryless source $\mathcal{S}$ we set $m'' = m$ while for Markov source $\mathcal{S}$ we assume $m'' = m^2 - m + 2$. Then for almost all source parameters, and for every variable-to-variable code with average delay $D \ge D_0$ (where D_0 is sufficiently large) we have*

$$r^* \ge \overline{r} \ge D^{-2m''-1-\varepsilon}, \tag{16}$$

where $\varepsilon > 0$.

Proof. By Lemma 4 we have

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

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

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

Provided Lemma 3 is granted, we conclude from (12) that for all p_j and for all non-zero integer vectors $(k_1, \ldots, k_m)$

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

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

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

This completes the proof of Theorem 3. ■

Note that Theorem 4 of [13] 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 3 this cannot be true for large m.

3 Proof of Theorem 1

In this section we will prove Theorem 1 for memoryless sources. In fact, we will prove a property that is a little more general and relies on approximation properties. Note that our method is in some sense close to the ideas presented in [13].

Theorem 4. *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

$$\delta(X) \le \frac{2}{N^\eta}. \tag{17}$$

Then there exists a variable-to-variable code with the average code length D of order $\Theta(N^3)$, 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 exits another variable-to-variable code with the average code length D of order 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, Corollary 1 and Theorem 4 directly imply Theorem 1 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 below is much simpler (cf. Remark following the proof of Theorem 4).

Throughout the proof we shall use the following result that follows directly from Stirling's formula, hence its proof is omitted for brevity.

Lemma 5. *Let p_j $(1 \le j \le m)$ with $p_j > 0$ and $p_1 + p_2 + \cdots + p_m = 1$ be given. Set $k_j = \lfloor p_j N^2 \rceil$ $(1 \le j \le m)$ and suppose that $0 \le k'_j \le N$ $(1 \le j \le m)$. Then*

$$\frac{c'}{N} \le \binom{k_1 + k'_1 + \cdots + k_m + k'_m}{k_1 + k'_1, \ldots, k_m + k'_m} p_1^{k_1+k'_1} \cdots p_m^{k_m+k'_m} \le \frac{c''}{N}$$

for certain constants c', c''.

Proof of Theorem 4. The idea of the proof is to construct a complete prefix free set $\mathcal{D}$ 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 $\mathcal{A} = \{a_1, \ldots, a_m\}$. Leaves of such an m-ary tree can be identified with a complete prefix free set $\mathcal{D}$. 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 $\mathcal{D}$. By checking (8) of Lemma 2, we conclude, providing (17) holds, that there exists a (variable-to-variable) 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 (17), there exist integers $0 \le k_j^1 < N$ such that

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

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 $\mathcal{D}_1$ 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 Lemma 5 we have

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

for certain positive constants c', c''. In summary, by construction all words $d \in \mathcal{D}_1$ 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 $d \in \mathcal{D}_1$ have about the same length

$$n_1 = (k_1^0 + k_1^1) + \cdots + (k_m^0 + k_m^1) = N^2 + O(N).$$

Let finally $\mathcal{B}_1 = \mathcal{A}^{n_1} \setminus \mathcal{D}_1$ denote all words of length n_1 not in $\mathcal{D}_1$ (cf. 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 (17) 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, ..., 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 $\mathcal{D}_2^+(r)$ (cf. Figure 1). We again have that the total probability of these words is bounded from below and above by

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

Furthermore, by construction we have

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

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

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

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

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

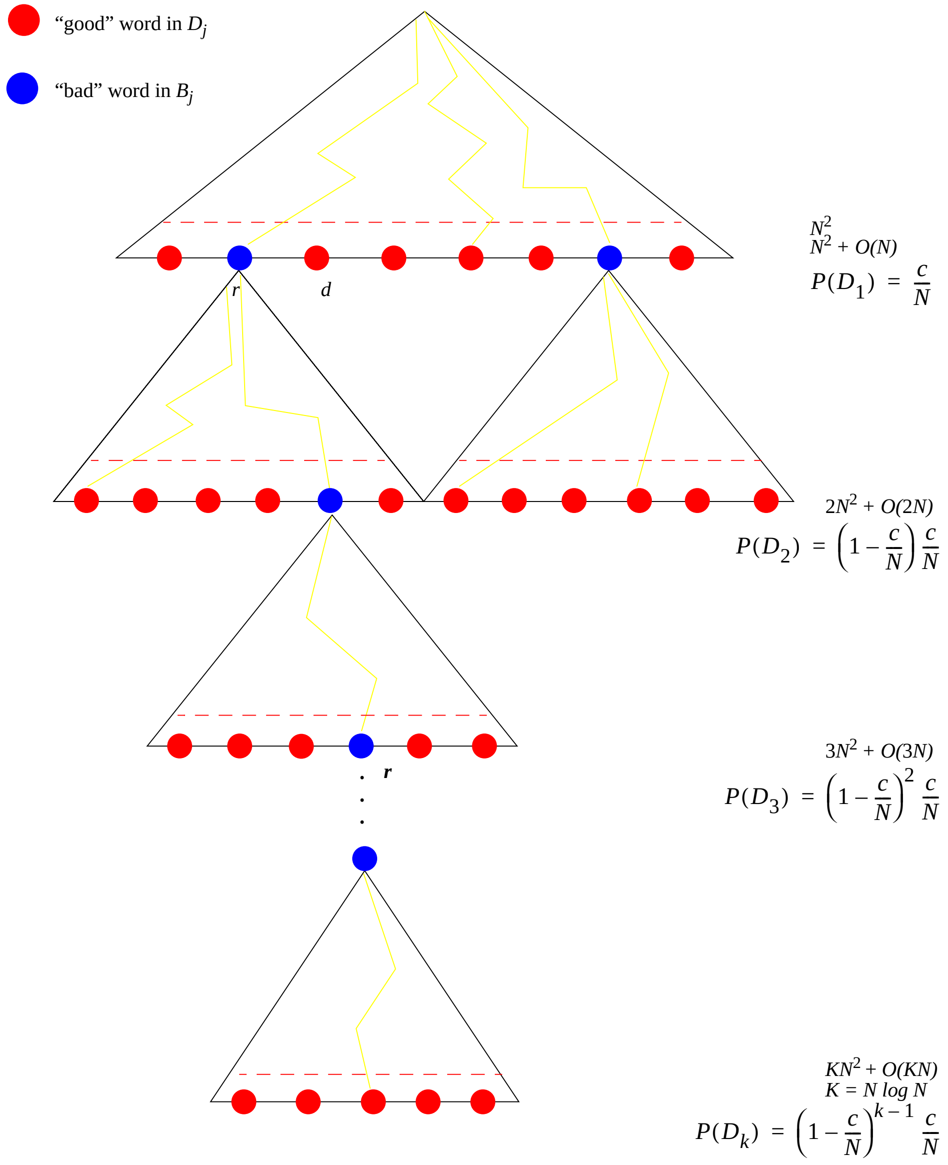
● “good” word in D_j
● “bad” word in B_j
r
d
N^2
$N^2 + O(N)$
$P(D_1) = \frac{c}{N}$
$2N^2 + O(2N)$
$P(D_2) = \left(1 - \frac{c}{N}\right) \frac{c}{N}$
r
$3N^2 + O(3N)$
$P(D_3) = \left(1 - \frac{c}{N}\right)^2 \frac{c}{N}$
$KN^2 + O(KN)$
$K = N \log N$
$P(D_k) = \left(1 - \frac{c}{N}\right)^{k-1} \frac{c}{N}$

Figure 1: Illustration to the construction of the variable-to-variable code.

their probabilities satisfy

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

and the total probability is bounded by

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

For every $r \in \mathcal{B}_1$, let $\mathcal{B}_2^+(r)$ (or $\mathcal{B}_2^-(r)$) 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 $\mathcal{D}_2^+(r)$ (or $\mathcal{D}_2^-(r)$) 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 $\mathcal{D}_j$ and $\mathcal{B}_j$ such that all words $d \in \mathcal{D}_j$ 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 $d \in \mathcal{D}_j \cup \mathcal{B}_j$ is given by

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

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

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

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

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

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

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

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

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

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

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

For every $d \in \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K$ 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 how to *adjust* the choices of "+" resp. "−" in the above construction so that we can apply Lemma 2. We set

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

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

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

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

$$\sum_{j=1}^{K} E_j \le \frac{8 + 2c''}{N^{1+\eta}}. \tag{18}$$

We can assume that N is large enough that $2/N^\eta \le 1/2$. Hence, the assumptions of Lemma 2 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 $d \in \mathcal{D}$. If (18) does not hold (if we have chosen always "+"), then it is rather clear that 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 $\sum_{j=i}^{K} E_j \le (8 + 2c'')N^{-1-\eta}$, then the sign of E_i is to be "+" and if $\sum_{j=i}^{K} E_j > (8 + 2c'')N^{-1-\eta}$ then the sign of E_i is to be "−". Since

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

the assumption of Lemma 2 is satisfied. Thus, there exists a prefix free coding map $C : \mathcal{D} \to \{0,1\}^*$ with $|C(d)| = \ell(d)$ for all $d \in \mathcal{D}$. Furthermore, the average redundancy rate is bounded by

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

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 Theorem 4.

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 never terminate. We set

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

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

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

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

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

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 $d \in \mathcal{D}$ (the Shannon code). Furthermore, we have

$$r^* = \frac{1}{D} \sup_{d\in\mathcal{D}} (\ell(d) + \log_2 P(d)) \le \frac{4}{DN^\eta} = \mathcal{O}\left(D^{-1-\frac{\eta}{3}}\right)$$

as proposed. Since k_0 is arbitrary and k_j may be changed to $-k_j$, the bound also holds for $\|L\|$. This completes the proof of Theorem 4.

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 Theorem 2

In order to prove Theorem 2 (for memoryless sources) we just have to combine Theorem 4 and Lemma 3. We recall that Lemma 3 says 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

$$\delta(X) \leq N^{-m+\varepsilon} \tag{19}$$

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

$$\|k_1 \log_2 p_1 + \cdots + k_m \log_2 p_m\| \gg \left(\max_{1 \leq j \leq m} |k_j|\right)^{-m-\varepsilon}. \tag{20}$$

In view of the above, we just have to show (19) and (20) 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, 18, 22]). One of the problems in this field is to obtain some information on 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 are not independent. They are strongly correlated, and they are situated on a proper submanifold of the m-dimensional space. It turns out that metric Diophantine approximation in this case is much more complicated than in the independent case. Fortunately, there are results that we can use for our purpose.

Theorem 5 (Dickinson and Dodson [6]). *Let $m \geq 2$ and $1 \leq k < m$ be integers. 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 a C^1 real function. Let $\eta > 0$ be real. Then for almost all $u = (u_1, \ldots, u_k) \in U$, 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 $(k_0, k_1, \ldots, k_m)$ with

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

Remark. Theorem 5 is not written explicitly in [6], however, it easily follows from the proof of Theorem 2 from [6], as we show now. Let us consider the convex body consisting of all real vectors $(y_1, \ldots, y_m)$ satisfying

$$\begin{aligned} |y_0 + y_1\Psi_1(u) + \ldots + y_m\Psi_m(u)| &\leq N^{-m+(m-k)\eta}(\log N)^{m-k}, \\ |y_j| &\leq N, \quad (j = 1, \ldots, k), \\ |y_j| &\leq N^{1-\eta}(\log N)^{-1}, \quad (j = k+1, \ldots, m). \end{aligned} \tag{21}$$

Dickinson and Dodson [6, p. 278] showed that the set

$$S(N) := \left\{ u \in U : \exists\ (k_0, k_1, \ldots, k_m) \in \mathbf{Z}^{m+1} \text{ with } 0 < \max_{1 \leq j \leq m} |k_j| < N^{1-\eta} \text{ satisfying (21)} \right\}$$

satisfies

$$\operatorname{meas}\left(\bigcap_{M\geq 0}\bigcup_{N\geq M} S(N)\right)=0,$$

where meas 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 $N \geq N_0(u)$. This is precisely the content of 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 $0 < \max\{|k_1|, |k_2|\} \leq N$.

Now we are in position to prove (19) and (20). We start with the easier proof of (20).

Proof of (20). For this purpose we can directly apply Theorem 5 with $k = m-1$, the open set U being contained in $\Delta = \{u = (u_1, \ldots, u_{m-1}) \in \mathbf{R}^{m-1} : u_1 \geq 0, \ldots, u_{m-1} \geq 0, u_1+\cdots+u_{m-1} \leq 1\}$ and $\Psi_j(u) = \log_2(u_j)$ $(1 \leq j \leq m-1)$, resp. $\Psi_m(u) = \log_2(1 - u_1 - \cdots - u_{m-1})$. Let u be in U for which there exists $N_0(u)$ as in the statement of Theorem 5. Let $k_0, \ldots, k_m$ be integers and put

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

Set $J = \max_{1\leq j\leq m} |k_j|$ and define N by $N^{1-\eta} = J \log N$. Assume that J is large enough in order that $N \geq N_0(u)$. We then have from Theorem 5 with $k = m-1$

$$|L| \geq N^{-m+\eta}(\log N) \gg J^{-m-(m-1)\eta/(1-\eta)}(\log J)^{(1-m)/(1-\eta)} \gg J^{-m-\varepsilon},$$

for $\varepsilon = 2(m-1)\eta/(1-\eta)$ and J large enough. This completes the proof of (20).

Proof of (19). To simplify our presentation, we first apply Theorem 6 for the case $m = 2$. Then, we briefly indicate how to generalize our argument. First of all, we want to point out that Theorems 5 and 6 give 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” (cf. [5, 18]) these lower bounds can be transformed into upper bounds for the dispersion of the set $X = \{\langle k_1\Psi_1(u) + \cdots + k_m\Psi_m(u)\rangle : 0 \leq k_1, \ldots, k_m < N\}$.

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$\}$, where λB is the λ-scaled version of B. One of the (first) main results of "geometry of numbers" is *Minkowski's Second Theorem* saying that $2^d/d! \leq \lambda_1 \cdots \lambda_d \mathrm{Vol}_d(B) \leq 2^d$, (cf. [5, 18]).

In the sequel we shall keep the the notation of Theorem 6. Let x and N satisfy the conclusion of that theorem and consider the convex body $B \subseteq \mathbf{R}^3$ defined by the inequalities

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

By Theorem 6 the set B does not contain a non-zero integer point. Thus, the first minimum of B must satisfy $\lambda_1 \geq 1$. Note that $\mathrm{Vol}_3(B) = 8\kappa$. From Minkowski's Second Theorem we conclude that the three minima of this convex body satisfy $\lambda_1\lambda_2\lambda_3 \leq 1/\kappa$. Since $1 \leq \lambda_1 \leq \lambda_2$ we thus get $\lambda_3 \leq \lambda_1\lambda_2\lambda_3 \leq 1/\kappa$ and consequently $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq 1/\kappa$. In other words, by restating the definition of λ_3 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

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

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

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

is small. Indeed, let ξ be a real number (that we want to approximate by an element of X) and consider the (regular) system of linear equations

$$\begin{aligned} -\xi + \theta_a(a_0 + a_1\Psi_1(x) + a_2\Psi_2(x)) + & \\ +\theta_b(b_0 + b_1\Psi_1(x) + b_2\Psi_2(x)) + \theta_c(c_0 + c_1\Psi_1(x) + c_2\Psi_2(x)) &= 4\kappa_2 N^{-2}, \\ \theta_a a_1 + \theta_b b_1 + \theta_c c_1 &= 4\kappa_3 N, \\ \theta_a a_2 + \theta_b b_2 + \theta_c c_2 &= 4\kappa_3 N. \end{aligned} \tag{22}$$

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 (22) combined with $\max\{|a_i|, |b_i|, |c_i|\} \leq \kappa_3 N$ it follows that

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

in particular, k_1 and k_2 are positive integers. Moreover, by considering the first equation of (22) we see that

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

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

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

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 justifies the paragraph following the statement of Lemma 3 and, in particular, the estimate (13).

Finally, we discuss the general case $m \geq 2$ (and prove Lemma 3). We consider the convex body $B \subseteq \mathbf{R}^{m+1}$ that has volume $2^{m+1}N^{\eta(1-m)}$ and consists of all real vectors $(y_0, y_1, \ldots, y_m)$ satisfying

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

It follows from Theorem 5 that, for almost all u and for every sufficiently large N, the first minimum of B is ≥ 1, thus, by Minkowski's Second Theorem, its last minimum is $\leq N^{(m-1)\eta}$. Consequently, we have $m+1$ linearly independent integer vectors $\mathbf{q}^{(i)} = (q_0^{(i)}, \ldots, q_m^{(i)})$, $i = 0, \ldots, m$, such that

$$\|\mathbf{q}^{(i)} \cdot \Psi(u)\| \leq N^{-m+m\eta}(\log N), \qquad \|\mathbf{q}^{(i)}\|_\infty \leq N^{1+(m-2)\eta},$$

where $\Psi(u)$ denotes the vector $(1, \Psi_1(u), \ldots, \Psi_m(u))$. Regarding the exponents of N, this is slightly weaker than what we obtained by using Theorem 6. However, we can still argue as above, and consider a system of linear equations analogous to (22), namely the system

$$\begin{aligned} -\xi + \theta_0(q_0^{(0)} + q_1^{(0)}\Psi_1(u) + \ldots + q_m^{(0)}\Psi_m(u)) + \\ + \ldots + \theta_m(q_0^{(m)} + q_1^{(m)}\Psi_1(u) + \ldots + q_m^{(m)}\Psi_m(u)) &= (m+2)N^{-m+m\eta}(\log N), \\ \theta_0 q_i^{(0)} + \ldots + \theta_m q_i^{(m)} &= (m+2)N^{1+(m-2)\eta}, \quad (i = 1, \ldots, m). \end{aligned}$$

Let $\varepsilon > 0$ be given. Proceeding exactly as for $m = 2$, with η sufficiently small, we get that, for any real number ξ, there are positive integers $k_1, \ldots, k_m$ such that

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

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 (19). This completes the proof of Lemma 3 (and consequently the proof of Theorem 2 for memoryless sources).

5 Proof for Markov Sources

In this section, we extend our results to Markov sources by indicating necessary changes in our previous proofs. We only consider binary Markov sources of order one.

We assume that the transition matrix of the Markov source is

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix},$$

where $p_{ij} = \Pr\{X_{k+1} = j \,|X_k = i\}$. The stationary distribution is then

$$p_0 = \frac{p_{10}}{p_{10}+p_{01}} \quad \text{and} \quad p_1 = \frac{p_{01}}{p_{10}+p_{01}}.$$

The probability of a message x_1^n becomes

$$P(x_1^n) = \hat{p}\, p_{00}^{k_{00}} p_{01}^{k_{01}} p_{10}^{k_{10}} p_{11}^{k_{11}},$$

where $\hat{p} = p_0$ if $x_0 = 0$ and $\hat{p} = p_1$ if $x_0 = 1$ and k_{ij} is the number of $k \in \{1, 2, \ldots, n-1\}$ such that $(x_k, x_{k+1}) = (i, j)$. Note that $k_{00} + k_{01} + k_{10} + k_{11} = n - 1$ and that $k_{01} = k_{10}$ if $x_1 = x_n$ and $k_{01} = k_{10} \pm 1$ if $x_1 \neq x_n$.

This means that

$$\log_2 P(x_1^n) = c_0 + k_{00} \log_2 p_{00} + k_{10} \log_2(p_{01}p_{10}) + k_{11} \log_2 p_{11}, \tag{23}$$

where c_0 attains only finitely many possible values. Note that we can vary k_{00}, k_{01}, and k_{11} "independently". In general we can vary $m^2 - m + 1$ of the m^2 "parameters" k_{ij} "independently".

We will further need the following asymptotic expansions which can be found in [11, Theorem 5] and Whittle [24]. For $\mathbf{k} = (k_{00}, k_{01}, k_{10}, k_{11})$ and $a, b \in \{0, 1\}$, let $N_{\mathbf{k}}^{a,b}$ denote the number of 0-1-sequences of length $n = k_{00} + k_{01} + k_{10} + k_{11} + 1$, where $x_0 = a$, $x_n = b$, and k_{ij} is the number of $k \in \{1, 2, \ldots, n-1\}$ such that $(x_k, x_{k+1}) = (i, j)$. Then

$$\begin{aligned}
N_{\mathbf{k}}^{0,0} &\sim \frac{k_{10}}{k_{10}+k_{11}} \binom{k_{00}+k_{01}}{k_{00}} \binom{k_{10}+k_{11}}{k_{10}}, \\
N_{\mathbf{k}}^{0,1} &\sim \frac{k_{01}}{k_{00}+k_{01}} \binom{k_{00}+k_{01}}{k_{00}} \binom{k_{10}+k_{11}}{k_{10}}, \\
N_{\mathbf{k}}^{1,0} &\sim \frac{k_{10}}{k_{10}+k_{11}} \binom{k_{00}+k_{01}}{k_{00}} \binom{k_{10}+k_{11}}{k_{10}}, \\
N_{\mathbf{k}}^{0,0} &\sim \frac{k_{01}}{k_{00}+k_{01}} \binom{k_{00}+k_{01}}{k_{00}} \binom{k_{10}+k_{11}}{k_{10}}
\end{aligned}$$

for those $\mathbf{k}$ which are admissible (i.e. if $a = b$ then $k_{01} = k_{10}$ and if $a \neq b$ then $k_{01} = k_{10} \pm 1$). Similar estimates hold for an m-ary alphabet.

Now we can extend our results to Markov sources over a binary alphabet. In particular, we use a slightly modified Lemma 1. Instead of $X = \{\langle k_1\gamma_1 + \cdots + k_m\gamma_m\rangle : 0 \leq k_j < N \ (1 \leq j \leq m)\}$ we must work with

$$X = \{\langle c_0 + k_1\gamma_1 + \cdots + k_m\gamma_m\rangle : 0 \leq k_j < N \ (1 \leq j \leq m)\}$$

for some c_0. Clearly, we get the same result for this modified set X.

Next, we have to modify Lemma 5. Suppose that $p_{ij} > 0$ $(i, j \in \{0, 1\})$ are the transition probabilities and p_j $(j \in \{0, 1\})$ the stationary distribution. Set $k_{ij} = \lfloor p_i p_{ij} N^2 \rfloor$ $(i, j \in \{0, 1\})$ and suppose that $0 \le k'_{ij} \le N$ $(i, j \in \{0, 1\})$ with $k'_{01} = k'_{1,0}$. Then we have for certain constants c', c''.

$$\frac{c'}{N} \le N^{0,0}_{\mathbf{k}+\mathbf{k}'} p_0 p_{00}^{k_{00}+k'_{00}} p_{01}^{k_{01}+k'_{01}} p_{10}^{k_{10}+k'_{10}} p_{11}^{k_{11}+k'_{11}} \le \frac{c''}{N}$$

(and three further similar estimates). As in the proof of Lemma 5 this follows from the above asymptotic expansion and Stirling's formula.

The (modified) proof of Theorem 4 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. Hence, Theorem 1 follows immediately.

There is even a modified Lemma 3. 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 Theorem 2 holds with $m' = m^2 - m + 1$.

There is only one slight change in the proof of Theorem 3. Since the linear form in (23) 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 $P(x_1^n)$

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

Consequently, we have to deal with $m'' = m^2 - m + 2$ instead of m' in Theorem 3.

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

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

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

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

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

[13] 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.

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

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

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

[17] 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).

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

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

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

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

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

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

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