# On $(d, k)$ Sequences Not Containing a Given Word

Philippe Jacquet
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France
Email: philippe.jacquet@inria.fr

Wojciech Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
Email: spa@cs.purdue.edu

***Abstract*— A sequence of zeros and ones is called a $(d, k)$-sequence if it does not contain runs of zeros of length either less than $d$ or greater than $k$, where $d$ and $k$ are arbitrary, but fixed, positive integers and $d < k$. For a given pattern $w$, we enumerate exactly and asymptotically $(d, k)$ sequences of length $n$ that do *not* contain a given word $w$. We use techniques of analytic algorithms such as generating functions, combinatorial calculus, and complex asymptotics.**

## I. Introduction

In many communication systems, in particular, in magnetic and optimal recording systems, one must restrict the structure of a bit stream (binary sequence) to a class of sequences satisfying certain constraints. The simplest *constrained binary sequences* are those in which runs of zeros (between two consecutive 1's) must have length at least $d$ and at most $k$, where $0 \leq d < k$. Such sequences are called the $(d, k)$ sequences. One can easily enumerate $(d, k)$ sequences (cf. [7], [15]) that ultimately allow us to compute the so called Shannon entropy. However, as observed in [8], in some situations one needs to avoid certain *patterns* in $(d, k)$ sequences. This problem belongs to the area of pattern matching [6]. In this paper, we enumerate $(d, k)$ sequences that do *not* contain a given word $w$. We present exact and asymptotic enumerations using analytic tools such as generating functions, combinatorial calculus, and complex asymptotics.

Pattern matching is a well studied problem. In general, for a given set of patterns $\mathcal{V} = \{w_1, \ldots, w_K\}$, where $w_i$ is a string, one searches for multiple occurrences of words from $\mathcal{V}$ in a *text* of length $n$. (Hereafter, we only consider a single pattern of length $m$ that we denote as $w$.) Ultimately (cf. a forthcoming paper), we are interested in the distribution of the number of occurrences of $\mathcal{V}$ elements in the (constrained) text (i.e., $(d, k)$ sequences) generated by a random source. This is motivated by communication as well computation as biology applications in which one is looking for over-represented or under-represented patterns hoping to find useful signals.

The pattern matching problem (in a probabilistic framework) goes back at least to Feller. The number of string occurrences in a random text has been intensively studied over the last two decades, with significant progress in this area being reported [1], [4], [6], [9]. For instance Guibas and Odlyzko [4], and later Jacquet and Szpankowski [10] have revealed the fundamental role played by autocorrelation vectors and their associated polynomials in the analysis of pattern matching. Régnier and Szpankowski [11], [12] established that the number of occurrences of a string is asymptotically normal under a diversity of models that include Markov chains. Nicodème, Salvy, and Flajolet [9] showed generally that the number of places in a random text at which a 'motif' (i.e., a general regular expression pattern) terminates is asymptotically normally distributed. Bender and Kochman [1] studied a generalized pattern $\mathcal{V}$ occurrences using (in the nutshell) the deBruijn graph representation that allowed the authors to establish the central limit theorem, but without explicit mean and variance. Recent surveys on pattern matching can be found in Lothaire [6] (Chaps. 6 and 7). To the best of our knowledge none of these work (with a possible exception of [1]) deals with pattern matching in constrained sequences such as $(d, k)$ sequences.

In the information theory community, $(d, k)$ sequences were analyzed since Shannon with some recent contributions [5], [7], [14], [16]. One of the earliest problem, already solved by Shannon, was to enumerate the number of binary $(d, k)$ sequences. Typically [7], constraint sequences such as $(d, k)$ sequences are analyzed through various versions of the de Bruijn graph [1], [3]. This is an elegant and general approach, however, sometimes leads to complicated and non-explicit solutions. Here, we take the view of combinatorics on words. We first construct a language of all $(d, k)$ sequences. Using generating functions and complex asymptotics, we present a simpler and more precise enumeration of various modifications of $(d, k)$ sequences. Then we enumerate $(d, k)$ sequences *not* containing a given pattern $w$. This seems to be a harder problem that is still open to the best of our knowledge. We again apply combinatorics on words to construct a language of all $(d, k)$ words not containing a given word, and using generating function and complex analysis we provide exact enumeration and asymptotic results.

## II. Main Results

We start with a relatively easy exercise of enumerating $(d, k)$ sequences of length $n$ over a binary alphabet $\{0, 1\}$. This will allow us to establish notation and introduce our methodology. In particular, for a language $\mathcal{W}$ (a collection of words) we define its *generating functions* $W(z)$ as

$$W(z) = \sum_{u \in \mathcal{W}} z^{|u|} = \sum_{n \geq 0} z^n W_n,$$

where $W_n$ counts the number of words $u$ of length $|u| = n$. We shall also write $W_n := [z^n]W(z)$ since $W_n$ is the coefficient at $z^n$ of $W(z)$.

### A. Enumeration of $(d, k)$ Sequences

Let $\mathcal{W}_{d,k}$ denote the language (set) of all $(d, k)$ sequences. We also define the set

$$\mathcal{A}_{d,k} = \{\underbrace{0 \ldots 0}_{d}, \ldots, \underbrace{0 \ldots 0}_{k}\},$$

that is, a set of runs of zero of length between $d$ and $k$. Its generating function is

$$A(z) = z^d + z^{d+1} + \cdots + z^k = \frac{z^d - z^{k+1}}{1 - z}. \quad (1)$$

We now observe that $\mathcal{W}_{d,k}$ can be symbolically written as for $d > 0$

$$\begin{aligned} \mathcal{W}_{d,k} &= \mathcal{A}_{d,k} \cdot \left(\epsilon + (\{1\} \cdot \mathcal{A}_{d,k}) + (\{1\} \cdot \mathcal{A}_{d,k})^2 + \cdots \right. &(2)\\ &+ \left. (\{1\} \cdot \mathcal{A}_{d,k})^k + \cdots \right) (\epsilon + \{1\}) \\ &+ \{1\} \cdot \left(\epsilon + (\mathcal{A}_{d,k} \cdot \{1\}) + (\mathcal{A}_{d,k} \cdot \{1\})^2 + \cdots \right. &(3)\\ &+ \left. (\mathcal{A}_{d,k} \cdot \{1\})^k + \cdots \right) (\epsilon + \mathcal{A}_{d,k}) &(4) \end{aligned}$$

where $\cdot$ denotes Cartesian product of sets (concatenation of words). It basically says that the collection of $(k, d)$ sequences, $\mathcal{W}_{d,k}$:

- starts with a run of allowed zeros, followed by repeated 1 and runs of allowed zeros, followed by 1 or an empty symbol, or
- starts with a 1, followed by repeated 1 and runs of allowed zeros, followed by a run of allowed zeros or an empty symbol.

The above can be translated into the generating function

$$W_{d,k}(z) = \sum_{u \in \mathcal{W}_{d,k}} z^{|u|} = \sum_{n=0}^{\infty} z^n W_n$$

where $W_n := [z^n]W_{d,k}(z)$ is the number of $(d, k)$ sequences of length $n$. From (4) we directly find

$$W_{d,k}(z) = \frac{A(z) + z(1 + 2A(z))}{1 - zA(z)}, \quad (5)$$

where $A(z)$ is given by (1). Extracting the coefficient at $z^n$ (e.g., using MAPLE), one finds exactly the number of $(d, k)$ sequences of length $n$. For example, there are $W_{20} = 3151$ constraint sequences with $d = 1$ and $k = 3$.

In order to obtain asymptotics of $W_n := [z^n]W_{d,k}(z)$ we apply Cauchy's formula [15]. Let $B(z) = zA(z)$ and let roots of

$$1 - zA(z) = 0$$

be $\rho_0 < |\rho_1| \leq \ldots |\rho_k|$. The smallest root $\rho_0$ is *real* and $\rho_0 < 1$. Indeed, $B(0) = 0A(0) = 0$ and $B(1) = k+d-1 > 0$ and $B(z)$ is an increasing function in the interval $(0, 1)$. Let us also define

$$\lambda_0 = \frac{1}{\rho_0}.$$

Then from Cauchy's formula

$$W_n = \frac{1}{2\pi i} \oint \frac{W_{d,k}(z)}{z^{n+1}} dz = \frac{\lambda_0 + 1/\lambda_0 + 2}{\lambda_0 + A'(\rho_0)\rho_0} \lambda_0^{n+1} + O(r^n) \quad (6)$$

where $r < \lambda_0$. Thus Shannon's capacity $C_{d,k} = \lim_{n \to \infty} n^{-1} \log W_n$ becomes

$$C_{d,k} = \log \lambda_0 + O(1/n). \quad (7)$$

For $d = 0$ we need to modify slightly our derivation since now 1 does not separate runs of zeros. In this case we have to replace (4) by a simpler one, namely,

$$\begin{aligned} \mathcal{W}_{0,k} &= \mathcal{A}_{0,k} \cdot \left(\epsilon + (\{1\} \cdot \mathcal{A}_{0,k}) + (\{1\} \cdot \mathcal{A}_{0,k})^2 + \cdots \right. \\ &+ \left. (\{1\} \cdot \mathcal{A}_{0,k})^k + \cdots \right) \end{aligned}$$

which leads to

$$W_{0,k}(z) = \frac{A(z)}{1 - zA(z)}.$$

In particular, there must be 1024 sequences $(0, 10)$, and indeed $[z^{10}]W_{0,10}(z) = 1024$.

Finally, for our main result, we must enumerate $(d, k)$ sequences that start with 0 and end with 1, that we call restricted $(d, k)$ sequences. We denote such a set as $\mathcal{W}_{d,k}^r$. Clearly

$$\mathcal{W}_{d,k}^r = \epsilon + \mathcal{B}_{d,k} + \mathcal{B}_{d,k}^2 + \cdots + \mathcal{B}_{d,k}^k + \cdots \quad (8)$$

where $\mathcal{B}_{d,k} = \mathcal{A}_{d,k} \cdot \{1\}$. The above translates into the generating function $W_{d,k}^r(z)$ as follows

$$W_{d,k}^r(z) = \frac{1}{1 - zA(z)} = \frac{1}{1 - B(z)} \quad (9)$$

where $B(z) = zA(z)$.

It is easy to enumerate asymptotically such sequences. Let $W_n^r = [z^n]W_{d,k}^r(z)$. Then by Cauchy's formula

$$W_n^r = \frac{1}{B(1/\lambda_0)} \lambda_0^{n+1} + O(r^n) \quad (10)$$

where, as before, $\rho_0 = 1/\lambda_0$ is the smallest real root of $B(z) = 1$.

### B. Enumeration of $(d, k)$ Sequences Not Containing a Given Word

In this section we consider restricted $(d, k)$ sequences (that start with 0 and end with 1). Our goal is to enumerate such sequences not containing a given word $w$.

Let $w = w_1 \ldots w_m \in \{0, 1\}^m$ with $w_1 = 0$ and $w_m = 1$. By $\mathcal{S}$ we denote its *autocorrelation set*, that is,

$$\mathcal{S} = \{w_{k+1}^m : \ w_1^k = w_{m-k+1}^m\}.$$

By $S(z) = \sum_{u \in \mathcal{S}} z^{|u|}$ we denote the generating function of $\mathcal{S}$ that is known as the *autocorrelation polynomial* of $w$ [4], [12], [15].

We will construct a language $\mathcal{T}_r^{(d,k)}$ that contains exactly $r$ occurrences of $w$ in a $(d,k)$ sequence. In this note, we concentrate on $\mathcal{T}_0 = \mathcal{T}_0^{(d,k)}$. In order to assure that we deal only with $(d,k)$ sequences, we build them over a new alphabet

$$\mathcal{B}_{d,k} = \mathcal{A}_{d,k} \cdot \{1\} = \{\underbrace{0\ldots0}_{d}1, \ldots, \underbrace{0\ldots0}_{k}1\}.$$

To simplify our notation we shall write $\mathcal{B} := \mathcal{B}_{d,k}$. We should point out that the word $w$ is also built over $\mathcal{B}$, hence $w$ occurrences are over $\mathcal{B}$, too.

We now define three languages, $\mathcal{R}$, $\mathcal{M}$ and $\mathcal{U}$:

(i) $\mathcal{T}_r^{(d,k)}$ as a set of $(d,k)$ words (over the alphabet $\mathcal{B}_{d,k}$) containing exactly $r$ occurrences of $w$.

(ii) $\mathcal{R}^{(d,k)}$ as a set of $(d,k)$ sequences (over the alphabet $\mathcal{B}_{d,k}$) containing only one occurrence of $w$, located at the right end.

(iii) $\mathcal{U}^{(d,k)}$ defined as

$$\mathcal{U}^{(d,k)} = \{u: \ w \cdot u \in \mathcal{T}_1^{(d,k)}\}, \tag{11}$$

that is, a word $u \in \mathcal{U}^{(d,k)}$ if $u$ is a $(d,k)$ sequence and $w \cdot u$ has exactly one occurrence of $w$ at the left end of $w \cdot u$.

(iv) $\mathcal{M}^{(d,k)}$ defined as

$$\begin{aligned}\mathcal{M}^{(d,k)} \quad = \quad & \{v: \ w \cdot v \in \mathcal{T}_2^{(d,k)} \\ & \text{and } w \text{ occurs at the right end of } w \cdot v\},\end{aligned}$$

that is, $\mathcal{M}^{(d,k)}$ is a $(d,k)$ language such that any word in $w \cdot \mathcal{M}^{(d,k)}$ has exactly two occurrences of $w$ at the left and right end.

In what follows, we drop the upper index $(d,k)$ unless it is crucial to make a distinction from other languages.

It is easy to see that

$$\mathcal{T}_r = \mathcal{R} \cdot \mathcal{M}^{r-1} \cdot \mathcal{U}, \tag{12}$$
$$\mathcal{T}_0 \cdot w = \mathcal{R} \cdot \mathcal{S} \tag{13}$$

Thus it suffices to find the languages $\mathcal{R}, \mathcal{M}$ and $\mathcal{U}$. The following is true (the proof follows the footsteps of [12])

$$\mathcal{M}^* = \mathcal{W}_{d,k}^r \cdot w + \mathcal{S}, \tag{14}$$
$$\mathcal{U} \cdot \mathcal{B} = \mathcal{M} + \mathcal{U} - \{\epsilon\}, \tag{15}$$
$$w(\mathcal{M} - \epsilon) = \mathcal{B} \cdot \mathcal{R} - \mathcal{R}. \tag{16}$$

where $\mathcal{W}_{d,k}^r$ is given by (8). Observe that all these languages are built over the extended alphabet $\mathcal{B}$.

The language relationships (14)–(16) translate directly into generating functions:

$$\frac{1}{1 - M(z)} = S(z) + W_{d,k}^r(z) z^m, \tag{17}$$
$$U(z) = \frac{M_w(z) - 1}{B(z) - 1}, \tag{18}$$
$$R(z) = z^m \cdot U(z), \tag{19}$$

where $W_{d,k}^r(z)$ is given in (9).

In particular, from (13) and above we find

$$T_0^{(d,k)}(z) = \frac{S(z)}{S(z)(1 - B(z)) + z^m} \tag{20}$$

where $S(z)$ is the autocorrelation polynomial for $w$ and

$$B(z) = zA(z) = z^{d+1} + \cdots + z^{k+1}.$$

For example, for $w = 0101$ we have $S(z) = 1 + z^2$ there is one $(1,3)$-sequence of length $n = 4$ that does not contain $w$, as it should be.

One can find asymptotics by applying Cauchy's formula. Let

$$D(z) = S(z)(1 - B(z)) + z^m$$

be the denominator of (20). Define $\rho_0^w$ to be the smallest real root of $D(z) = 0$. Such a root exists since $D(0) = 1$ and $D(1) = 1 - S(1)(k-d) < 0$ provided $d < k$.

In summary, we prove the following main result.

*Theorem 1:* Let $W_n^w$ denote the number of $(d,k)$ sequences of length $n$ not containing a given pattern $w$ characterized by its autocorrelation polynomial $S(z)$. Let also $\rho_0^w$ be the smallest real root of

$$D(z) := S(z)(1 - B(z)) + z^m = 0,$$

and define $\lambda_0^w = 1/\rho_0^w$. Then

$$W_n^w := [z^n] T_0^{(d,k)}(z) = \frac{S(\rho_0^w)}{D(\rho_0^w)} (\lambda_0^w)^{n+1} + O(r^n) \tag{21}$$

where $r < \lambda_0^w$. Furthermore, the Shannon capacity for $(d,k)$ sequence not containing $w$ becomes

$$C_{d,k}^w = \log \lambda_0^w + O(1/n)$$

for large $n$.

**Proof**. By Cauchy's formula and (20)

$$W_n^w = [z^n] T_0^{(d,k)}(z) = \frac{1}{2\pi i} \oint \frac{S(z)}{D(z) z^{n+1}} dz.$$

Another application of Cauchy's residue theorem, completes the proof. ∎

It is easy to see that $\rho_0^w > \rho_0$, hence the number of $(d,k)$ strings (starting with 0 and ending with one 1) exponentially outnumbers those $(d,k)$ sequences that do not contain $w$. Certainly, $C_{d,k}^w < C_{d,k}$.

## Acknowledgment

The work of W. Szpankowski was supported in part by the NSF Grants CCR-0208709, CCF-0513636, and DMS-0503742, NIH Grant R01 GM068959-01, and the AFOSR Grant FA8655-04-1-3074.

## REFERENCES


[1] E. Bender, and F. Kochman, The distribution of subword counts is usually normal. *European Journal of Combinatorics*, **14**, 265–275, 1993.

[2] J. Fan, T. L. Poo, and B. Marcus, Constraint Gain, *Trans. Information Theory*, 50, 1989-1999, 2004.

[3] P. Flajolet, W. Szpankowski, and B. Vallée, Hidden Word Statistics, *J. the ACM*, 2006.

[4] L. Guibas and A. M. Odlyzko. Periods in strings. *J. Combinatorial Theory*, 30:19–43, 1981.

[5] V. Kolesnik and V. Krachkovsky, Generating Functions and Lower Bounds on Rates for Limited Error-Correcting Codes, *Trans. Information Theory*, 37, 778-788, 1991.

[6] M. Lothaire, (Editor) *Applied Combinatorics on Words*, Cambridge University Press, 2005.

[7] B. Marcus, R. Roth and P. Siegel, Constrained Systems and Coding for Recording Channels, Chap. 20 in *Handbook of Coding Theory* (eds. V. S. Pless and W. C. Huffman), Elsevier Science, 1998.

[8] B. Moision, A. Orlitsky and P. Siegel, On codes that avoid specific differences, *Trans. Information Theory*, 47, 433-442, 2001.

[9] P. Nicodème, B. Salvy, and P. Flajolet, Motif Statistics, *European Symposium on Algorithms*, *Lecture Notes in Computer Science*, No. 1643, 194–211, 1999.

[10] 10. Jacquet, Szpankowski, "Autocorrelation on words and its applications. Analysis of suffix trees by string-ruler approach," in J. Combinatorial Theory. Ser. A. 66, pp. 237-269, 1994.

[11] M. Régnier and W. Szpankowski, On the Approximate Pattern Occurrences in a Text, *Proc. Compression and Complexity of SEQUENCE'97*, IEEE Computer Society, 253–264, Positano, 1997.

[12] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22:631–649, 1998.

[13] G. Reinert, S. Schbath, and M. Waterman, Probabilistic and statistical properties of words: an overview. *J. Comput. Biol.*, 7, 1-46, 2000.

[14] V. Stefanov and W. Szpankowski, Waiting Time Distribution for Pattern Occurrences in a Constrained Sequence, preprint 2005.

[15] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.

[16] E. Zehavi and J. Wolf, On runlength codes, *Trans. Information Theory*, 34, 45-54, 1988.