CHAPTER 7

# Analytic Approach to Pattern Matching



## 7.0. Introduction

Repeated patterns and related phenomena in words are known to play a central role in many facets of computer science, telecommunications, coding, data compression, and molecular biology. One of the most fundamental questions arising in such studies is the frequency of pattern occurrences in another string known as the *text*. Applications of these results include gene finding in biology, code synchronization, user search in wireless communications, detecting signatures of an attacker in intrusion detection, and discovering repeated strings in the Lempel-Ziv schemes and other data compression algorithms.

The basic *pattern matching* is to find for a given (or random) pattern $w$ or a set of patterns $\mathcal{W}$ and text $X$ how many times $\mathcal{W}$ occurs in the text and how long it takes for $\mathcal{W}$ to occur in $X$ for the first time. These two problems are not unrelated as we have already seen in Chapter 6. Throughout this chapter we allow patterns to overlap and we count overlapping occurrences separately. For example, $w = abab$ occurs three times in the text $= bababababb$.

We consider pattern matching problems in a probabilistic framework in which the text is generated by a probabilistic source while the pattern is given. In Chapter 1 various probabilistic sources were discussed. Here we succinctly summarize assumptions adopted in this chapter. In addition, we introduce a new general source known as a *dynamic source* recently proposed by Vallée. In Chapter 2 algorithmic aspects of pattern matching and various efficient algorithms for finding patterns were discussed. In this chapter, as in Chapter 6, we focus on *analysis*. However, unlike Chapter 6, we apply here analytic tools of combinatorics and analysis of algorithms to discover general laws of pattern occurrences. An immediate consequence of our results is the possibility to set *thresholds* at which a pattern in a text starts being (statistically) meaningful.

The approach we undertake to analyze pattern matching problems is through a formal description by means of regular languages. Basically, such a description of *contexts* of one, two, or several occurrences gives access to expectation, variance, and higher moments, respectively. A systematic translation into *generating functions* of a complex variable $z$ is available by methods of analytic combinatorics deriving from the original Chomsky-Schützenberger theorem. Then, the structure of the implied generating functions at a pole, usually at $z = 1$, provides the necessary asymptotic information. In fact, there is an important phenomenon of *asymptotic simplification* where the essentials of combinatorial-probabilistic features are reflected by the singular forms of generating functions. For instance, variance coefficients come out naturally from this approach together with a suitable notion of correlation. Perhaps the originality of the present approach lies in such a joint use of combinatorial-enumerative techniques and of analytic-probabilistic methods.

There are various pattern matching problems. In its simplest form, the pattern $\mathcal{W} = w$ is a single string $w$ and one searches for some/all occurrences of $w$ as a block of consecutive symbols in the text. This problem is known as the *exact string matching* and its analysis is presented in Section 7.2 (cf. also Chapter 6). We adopt a symbolic approach, and first describe a language that contains all occurrences of $w$. Then we translate this language into a generating function that will lead to precise evaluation of the mean and the variance of the number of occurrences of the pattern. Finally, we prove the central and local limit laws, and large deviations.

In the *generalized string matching* problem the pattern $\mathcal{W}$ is a set rather than a single pattern. In its most general formulation, the pattern is a pair $(\mathcal{W}_0, \mathcal{W})$ where $\mathcal{W}_0$ is the so called *forbidden set.* If $\mathcal{W}_0 = \emptyset$, then $\mathcal{W}$ appears in the text whenever a word from $\mathcal{W}$ occurs as a string with overlapping allowed. When $\mathcal{W}_0 \neq \emptyset$ one studies the number of occurrences of strings in $\mathcal{W}$ under the condition that there is no occurrence of a string from $\mathcal{W}_0$ in the text $X$. This could be called a *restricted* string matching since one restricts the text to those strings that do not contain strings from $\mathcal{W}_0$. Finally, setting $\mathcal{W} = \emptyset$ (with $\mathcal{W}_0 \neq \emptyset$) we search for the number of text strings that do not contain any pattern from $\mathcal{W}_0$. In particular, for $\ell \leq k$ if $\mathcal{W}_0$ consists of runs of zeros of length at least $\ell$ and at most $k$, then we deal with the so called $(\ell, k)$ sequences that find application in magnetic recoding.

We shall present a complete analysis of the generalized string matching problem in Section 7.3. We first consider the so called *reduced set of patterns* in which a string in $\mathcal{W}$ cannot be a substring of another string in $\mathcal{W}$. We shall generalize our combinatorial language approach from Section 7.2 to derive the mean, variance, central and local limit laws, and large deviations. Then we analyze the generalized string pattern matching with $\mathcal{W}_0 = \emptyset$ and adopt a different approach. We shall construct an automaton to recognize the pattern $\mathcal{W}$ that turns out to be a de Bruijn graph. The generating function of the number of occurrences will have a matrix form with the main matrix representing the transition matrix of the associated de Bruijn graph. Finally, we consider the $(\ell, k)$ sequences and enumerate them leading to the Shannon capacity.

In Section 7.4 we discuss a new pattern matching problem called the *subsequence pattern matching* or the *hidden pattern matching.* In this case the pattern $\mathcal{W} = a_1 a_2 \cdots a_m$, where $a_i$ is a symbol of the underlying alphabet, is to occur as a *subsequence* rather than a string (consecutive symbols) in a text. We say that $\mathcal{W}$ is hidden in the text. For example, `date` occurs as a subsequence in the text `hidden pattern`, in fact four times, but not even once as a string. The gaps between occurrences of $\mathcal{W}$ may be bounded or unrestricted. The extreme cases are: *fully unconstrained* problem where all gaps are unbounded; and the *fully constrained* problem where all gaps are bounded. We analyze these and mixed cases.

In Section 7.5 we generalized all of the above pattern matching problems and analyze the *generalized subsequence problem.* In this case, the pattern is $\mathcal{W} = (\mathcal{W}_1, \ldots, \mathcal{W}_d)$ where $\mathcal{W}_i$ is a collection of strings (a language). We say that the generalized pattern $\mathcal{W}$ occurs in the text $X$ if $X$ contains $\mathcal{W}$ as a *subsequence* $(w_1, w_2, \ldots, w_d)$ where $w_i \in \mathcal{W}_i$. Clearly, it includes all the problems discussed so far. We shall analyze this generalized pattern matching for general probabilistic dynamic sources (which include among others Markov sources and mixing sources). The novelty of the analysis lies in translating probabilities into composition of operators. Under a mild decomposability assumption, these operators entertain spectral representations that allows us to derive precise asymptotic behavior for quantities of interest.

Finally, in the last section we study a different pattern matching, namely the one in which the pattern is part of the (random) text. We coin the term *self-repetitive pattern matching.* More precisely, we look for the longest substring of the text occurring at a given position that has another copy in the text. This new quantity, when averaged over all possible positions of the text, is actually the typical *depth* in a suffix trie (cf. Chapter 2) built over (randomly generated) text. We analyze it using analytic techniques such as generating functions and the Mellin transform. We reduce its analysis to the exact pattern matching; thus we call the technique the *string-ruler* method. In fact, we prove that the probability generating function of the depth in a suffix trie is asymptotically close to the probability generating function of the depth in a trie that is built over $n$ *independently* generated texts. Such tries have been extensively studied in the past and we have pretty good understanding of their probabilistic behaviors. This allows us to conclude that the depth in a suffix trie is asymptotically normal.

## 7.1. Probabilistic models

We study here pattern matching in a probabilistic framework in which the text is generated randomly. Let us first introduce some general probabilistic models of generating sequences. The reader is also referred to Chapter 1 for a brief introduction to probabilistic models. For the convenience of the reader, we repeat here some definitions.

Throughout we shall deal with sequences of discrete random variables. We write $(X_k)_{k=1}^{\infty}$ for a one-sided infinite sequence of random variables; however, we often abbreviate it as $X$ provided it is clear from the context that we are talking about a sequence, not a single variable. We assume that the sequence $(X_k)_{k=1}^{\infty}$ is defined over a finite alphabet $\mathcal{A} = \{a_1, \dots, a_V\}$ of size $V$. A partial sequence is denoted as $X_m^n = (X_m, \dots, X_n)$ for $m < n$. Finally, we shall always assume that a probability measure exists, and we write $P(x_1^n) = \mathbf{P}(X_k = x_k,\ 1 \leq k \leq n,\ x_k \in \mathcal{A})$ for the probability mass, where we use lowercase letters for a realization of a stochastic process.

Sequences are generated by information sources, usually satisfying some constraints. We also call them *probabilistic models*. Throughout, we assume the existence of a stationary probability distribution, that is, for any string $w$ the probability that the text $X$ contains an occurrence of $w$ at position $k$ is equal to $P(w)$ independently of the position $k$. For $P(w) > 0$, we denote by $P(u \mid w)$ the conditional probability equals $P(wu)/P(w)$.

The most elementary source is a *memoryless source* also known as the *Bernoulli source*.

(B) Memoryless or Bernoulli Source

Symbols of the alphabet $\mathcal{A} = \{a_1, \dots, a_V\}$ occur independently of one another; thus $X = X_1X_2X_3\dots$ can be described as the outcome of an infinite sequence of Bernoulli trials in which $\mathbf{P}(X_j = a_i) = p_i$ and $\sum_{i=1}^{V} p_i = 1$. Throughout, we assume that at least for one $i$ we have $0 < p_i < 1$.

In many cases, assumption (B) is not very realistic. When this is the case, assumption (B) may be replaced by:

(M) Markov Source of order one

There is a Markovian dependency between consecutive symbols in a string; that is, the probability $p_{ij} = \mathbf{P}(X_{k+1} = a_j | X_k = a_i)$ describes the conditional probability of sampling symbol $a_j$ immediately after symbol $a_i$. We denote by $\mathrm{P} = \{p_{ij}\}_{i,j=1}^{V}$ the transition matrix, and by $\mu = (\pi_1, \ldots, \pi_V)$ the stationary vector satisfying $\mu\mathrm{P} = \mu$. (Throughout, we assume that the Markov chain is irreducible and aperiodic.) A general Markov source of order $r$ is characterized by the transition matrix $V^r \times V$ with coefficients being $P(j \in \mathcal{A} \mid u)$ for $u \in \mathcal{A}^r$.

In some situations more general sources must be considered (for which one still can obtain reasonably precise analysis). Recently, Vallée introduced new sources called *dynamic sources* that we briefly describe here and use in the analysis of the generalized subsequence problem in Section 7.5. To introduce such sources we start with a description of a *dynamic system* defined by:

- A topological partition of the unit interval $\mathcal{I} := (0, 1)$ into a disjoint set of open intervals $\mathcal{I}_a, a \in \mathcal{A}$.
- An encoding mapping $\chi$ which is constant and equal to $a \in \mathcal{A}$ on each $\mathcal{I}_a$.
- A shift mapping $T:\ \mathcal{I} \to \mathcal{I}$ whose restriction to $\mathcal{I}_a$ is a bijection of class $\mathcal{C}^2$ from $\mathcal{I}_a$ to $\mathcal{I}$. The local inverse of $T$ restricted to $\mathcal{I}_a$ is denoted by $h_a$.

Observe that such a dynamic system produces infinite words of $\mathcal{A}^\infty$ through the encoding $\chi$. For an initial $x \in \mathcal{I}$ the source outputs a word, say $w(x) = (\chi x, \chi T x, \ldots)$.

(DS) Probabilistic Dynamic Source

A source is called a *probabilistic dynamic source*, if the unit interval of a dynamic system is endowed with a density $f$.

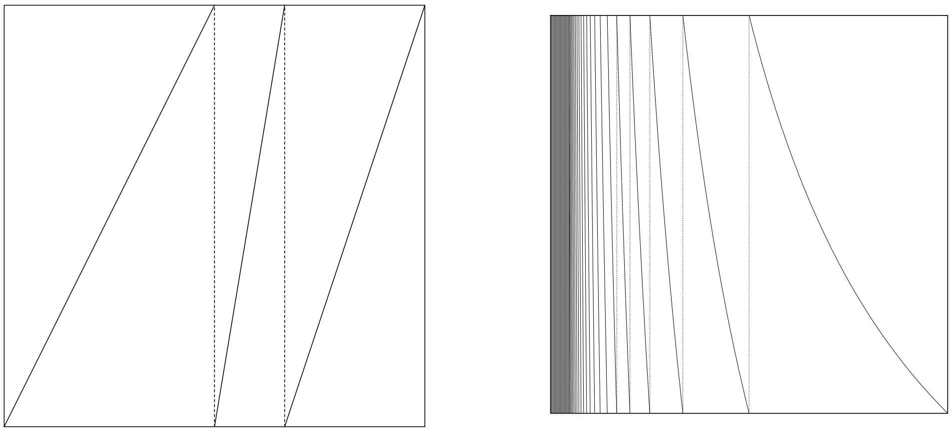
Example 7.1.1. A memoryless source associated with the probability distribution $\{p_i\}_{i=1}^{V}$ (where $V$ can be finite or infinite) is modeled by a dynamic source in which the components $w_k(x) = \chi T^k x$ are independent and the corresponding topological partition of $\mathcal{I}$ is defined as

$$\mathcal{I}_m := (q_m, q_{m+1}], \qquad q_m = \sum_{j<m} p_j.$$

In particular, a symmetric $V$-ary memoryless source can be described as

$$T(x) = \{Vx\}, \qquad \chi(x) = \lfloor Vx \rfloor,$$

where $\lfloor x \rfloor$ is the integer part of $x$ and $\{x\} = x - \lfloor x \rfloor$ is the fractional part of $x$ (cf. Figure 7.1(a)).



(a) (b)

**Figure 7.1.** Dynamic Sources discussed in Example 7.1.1: (a) memoryless with the shift mapping $T_m(x) = \langle (x - q_m)/p_{m+1} \rangle$ (b) continued fraction source with $T_m(x) = 1/x - m = \langle 1/x \rangle$.

Here is another example of a source with memory related to continued fractions. The alphabet $\mathcal{A}$ is the set of all natural numbers and the partition of $\mathcal{I}$ is defined as $\mathcal{I}_m = (1/(m+1), 1/m)$. The restriction of $T$ to $\mathcal{I}_m$ is the decreasing linear fractional transformation $T(x) = 1/x - m$, that is,

$$T(x) = \{1/x\}, \qquad \chi(x) = \lfloor 1/x \rfloor.$$

Observe that the inverse branches $h_m$ are defined as $h_m(x) = 1/(x+m)$ (cf. Figure 7.1(b)).

Let us observe that a word of length $k$, say $w = w_1 w_2 \cdots w_k$ is associated with the mapping $h_w := h_{w_1} \circ h_{w_2} \circ \cdots \circ h_{w_k}$ which is an inverse branch of $T^k$. In fact, all words that begin with the same prefix $w$ belong to the same *fundamental interval* defined as $\mathcal{I}_w = (h_w(0), h_w(1))$. Furthermore, for probabilistic dynamic sources with the density $f$, one easily computes the probability of $w$ as the measure of the interval $\mathcal{I}_w$.

The probability $P(w)$ of a word $w$ can be explicitly computed through the special *generating operator* $\mathbf{G}_w$ define as follows

$$\mathbf{G}_w[f](t) := |h'_w(t)| f \circ h_w(t). \tag{7.1.1}$$

One recognizes in $\mathbf{G}_w[f](t)$ a density mapping, that is, $\mathbf{G}_w[f](t)$ is the density of $f$ mapped over $h_w(t)$. The probability of $w$ can then be computed as

$$P(w) = \left| \int_{h_w(0)}^{h_w(1)} f(t)dt \right| = \int_0^1 |h'_w(t)| f \circ h_w(t) dt = \int_0^1 \mathbf{G}_w[f](t) dt. \tag{7.1.2}$$

Let us now consider a concatenation of two words $w$ and $u$. For memoryless sources $P(w \cdot u) = P(w)P(u)$. For Markov sources one still obtains the product of *conditional* probabilities. Dynamic sources replaces the product of probabilities by the product (composition) of generating operators. To see this, we observe that

$$\mathbf{G}_{w\cdot u} = \mathbf{G}_u \circ \mathbf{G}_w, \tag{7.1.3}$$

where we write $\mathbf{G}_w := \mathbf{G}_w[f](t)$. Indeed, $h_{wu} = h_w \circ h_u$ and $\mathbf{G}_{w\cdot u} = h'_w \circ h_u \cdot h'_u \cdot f \circ h_w \circ h_u$ while $\mathbf{G}_w = h'_w \cdot f \circ h_w$ and then $\mathbf{G}_u \circ \mathbf{G}_w = h_u \cdot h'_w \circ h_u \cdot f \circ h_w \circ h_u$, as desired.

## 7.2. Exact string matching

In the *exact string matching* problem the pattern $w = w_1 w_2 \cdots w_m$ of length $m$ is *given* while the text $X = X_1^n = X_1 \ldots X_n$ of length $n$ is generated by a random source. Observe that since the pattern $\mathcal{W}$ is given, its length $m$ will *not* vary with $n$ when $n \to \infty$ (asymptotic analysis).

There are several parameters of interest in the string matching, but two of them stand out. Namely, the number of times $w$ occurs in $X$ which we denote as $N_n := N_n(w)$ and define formally by

$$N_n(w) = \mathrm{Card}(\{i : \ X_{i-m+1}^i = w, \ m \le i \le n\}).$$

We can write $N_n(w)$ in an equivalent form as follows

$$N_n(w) = I_m + I_{m+1} + \cdots + I_n \tag{7.2.1}$$

where $I_i = 1$ if $w$ occurs at position $i$ and $I_i = 0$ otherwise.

The second parameter is the *waiting time* $T_w$ defined as the first time $w$ occurs in the text $X$, that is,

$$T_w := \min\{n : \ X_{n-m+1}^n = w\}.$$

One can also define $T_j$ as the minimum length of the text in which the pattern $w$ occurs $j$ times. Clearly, $T_w = T_1$. These parameters are not independent since

$$\{T_w > n\} = \{N_n(w) = 0\}. \tag{7.2.2}$$

More generally,

$$\{T_j \le n\} = \{N_n(w) \ge j\}. \tag{7.2.3}$$

Relation (7.2.3) is called the *duality principle* in Chapter 6.

Our goal is to estimate the frequency of pattern occurrences $N_n$ in a text generated by a Markov source. We allow patterns to overlap when counting occurrences (e.g., if $w = abab$, then it occurs twice in $X = abababb$ when overlapping is allowed; it occurs only once if overlapping is not allowed). We study probabilistic behavior of $N_n$ through two generating functions, namely:

$$N_r(z) = \sum_{n\geq 0} \mathbf{P}(N_n(w) = r) z^n,$$
$$N(z,u) = \sum_{r=1}^{\infty} N_r(z) u^r = \sum_{r=1}^{\infty}\sum_{n=0}^{\infty} \mathbf{P}(N_n(w) = r) z^n u^r$$

that are defined for $|z| \leq 1$ and $|u| \leq 1$.

Throughout this section we adopt a combinatorial approach to string matching, that is, we use combinatorial calculus to find combinatorial relationships between sets of words satisfying certain properties (i.e., languages). Alternatively, we could start with the representation (7.2.1) and use probabilistic tools along the lines already discussed in Chapter 6.

### 7.2.1. Languages representations

We start our combinatorial analysis with some definitions. For any language $\mathcal{L}$ we define its *generating function* $L(z)$ as

$$L(z) = \sum_{u\in\mathcal{L}} P(u) z^{|u|},$$

where $P(u)$ is the stationary probability of $u$ occurrence, $|u|$ is the length of $u$, and we assume that $P(\varepsilon) = 1$. Notice that $L(z)$ is defined for all complex $z$ such that $|z| < 1$. In addition, we define the *$w$-conditional generating function* of $\mathcal{L}$ as

$$L_w(z) = \sum_{u\in\mathcal{L}} P(u|w) z^{|u|} = \sum_{u\in\mathcal{L}} \frac{P(wu)}{P(w)} z^{|u|},$$

Since we allow overlaps, the structure of the pattern has a profound impact on the number of occurrences. To capture this, we introduce the autocorrelation language and the autocorrelation polynomial. Given a string $w$, we define the *autocorrelation set* $\mathcal{S}$ as:

$$\mathcal{S} = \{w_{k+1}^m : \ w_1^k = w_{m-k+1}^m\}. \tag{7.2.4}$$

By $\mathcal{P}(w)$ we denote the set of positions $k \geq 1$ satisfying $w_1^k = w_{m-k+1}^m$. In other words, if $w = vu$ and $w = ux$ for some words $v$, $x$ and $u$, then $x$ belongs to $\mathcal{S}$ and $|u| \in \mathcal{P}(w)$. Notice that $\varepsilon \in \mathcal{S}$. The generating function of the language $\mathcal{S}$ is denoted as $S(z)$ and we call it the *autocorrelation polynomial*. Its $w$-conditional generating function is denoted $S_w(z)$. In particular, for Markov sources (of order one)

$$S_w(z) = \sum_{k\in\mathcal{P}(w)} P(w_{k+1}^m \mid w_k^k) z^{m-k}. \tag{7.2.5}$$

Before we proceed, let us present a simple example illustrating the definitions introduced so far.

Example 7.2.1. Let us assume that $w = aba$ over a binary alphabet $\mathcal{A} = \{a, b\}$. Observe that $\mathcal{P}(w) = \{1, 3\}$ and $\mathcal{S} = \{\varepsilon, ba\}$, where $\varepsilon$ is the empty word. Thus, for the unbiased memoryless source we have $S(z) = 1 + \frac{z^2}{4}$, while for the Markovian model of order one, we obtain $S_{aba}(z) = 1 + p_{ab}p_{ba}z^2$.

Our goal is to estimate the number of pattern occurrences in a text. Alternatively, we can seek the generating function of a language that consists of all words containing some occurrences of $w$. Given a pattern $w$, we introduce the following languages:

(i) $\mathcal{T}_r$ as a set of words containing exactly $r$ occurrences of $w$.

(ii) $\mathcal{R}$ as a set of words containing only one occurrence of $w$, located at the right end.

(iii) $\mathcal{U}$ defined as

$$\mathcal{U} = \{u : \ w \cdot u \in \mathcal{T}_1\}, \tag{7.2.6}$$

that is, a word $u \in \mathcal{U}$ if $w \cdot u$ has exactly one occurrence of $w$ at the left end of $w \cdot u$.

(iv) $\mathcal{M}$ defined as

$$\mathcal{M} = \{v : \ w \cdot v \in \mathcal{T}_2 \text{ and } w \text{ occurs at the right end of } w \cdot v\},$$

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

Example 7.2.2. Let $\mathcal{A} = \{a, b\}$ and $w = abab$. Then $r = aa\mathtt{abab} \in \mathcal{R}$ since there is only one occurrence of $w$ at the right end of $r$. Also, $u = bbbb \in \mathcal{U}$ since $wu$ has only one occurrence of $w$ at the left end; but $v = abbbb \notin \mathcal{U}$ since $wv = ababbbbb$ has two occurrences of $w$. Furthermore, $ab \in \mathcal{M}$ since $wm = \mathtt{ababab} \in \mathcal{T}_2$ has two occurrences of $w$ at the left and the right ends. Finally, $t = bb\mathtt{ababab}bb\mathtt{abab}bb \in \mathcal{T}_3$ and one observes that $t = rm_1m_2u$ where $r = bbabab \in \mathcal{R}$, $m_1 = ab \in \mathcal{M}$, $m_2 = bbabab \in \mathcal{M}$, and $u = bb \in \mathcal{U}$.

We now describe languages $\mathcal{T}_{\geq 1} = \bigcup_{r \geq 1} \mathcal{T}_r$ (set of words containing at least once occurrence of $w$) and $\mathcal{T}_r$ in terms of $\mathcal{R}$, $\mathcal{M}$, and $\mathcal{U}$. Recall that $\mathcal{M}^r$ denotes the concatenation of $r$ languages $\mathcal{M}$, and $\mathcal{M}^0 = \{\varepsilon\}$. Also, $\mathcal{M}^+ = \cup_{r\geq 1} M^r$ and $\mathcal{M}^* = \cup_{r \geq 0} M^r$.

Theorem 7.2.3. *The languages $\mathcal{T}_r$ for $r \geq 1$ and $\mathcal{T}_{\geq 1}$ satisfy the relations*

$$\mathcal{T}_r = \mathcal{R} \cdot \mathcal{M}^{r-1} \cdot \mathcal{U} \ , \tag{7.2.7}$$

*and therefore*

$$\mathcal{T}_{\geq 1} = \mathcal{R} \cdot \mathcal{M}^* \cdot \mathcal{U}. \tag{7.2.8}$$

*In addition, we have:*

$$\mathcal{T}_0 \cdot w = \mathcal{R} \cdot \mathcal{S}. \tag{7.2.9}$$

*Proof.* To prove (7.2.7), we obtain our decomposition of $\mathcal{T}_r$ as follows: The first occurrence of $w$ in a word belonging to $\mathcal{T}_r$ determines a prefix $p \in \mathcal{T}_r$ that is in $\mathcal{R}$. After concatenating a nonempty word $v$ we create the second occurrence of $w$ provided $v \in \mathcal{M}$. This process is repeated $r-1$ times. Finally, after the last $w$ occurrence we add a suffix $u$ that does not create a new occurrence of $w$, that is, $wu$ is such that $u \in \mathcal{U}$. Clearly, a word belongs to $\mathcal{T}_{\geq 1}$ if for some $1 \leq r < \infty$ it is in $\mathcal{T}_r$.

The derivation of (7.2.9) is left to the reader as Exercise 7.2.1. ∎

EXAMPLE 7.2.4. Let $w = TAT$. The following string belongs to $\mathcal{T}_3$:

$$\overbrace{CCTAT}^{\mathcal{R}}\underbrace{AT}_{\mathcal{M}}\underbrace{GATAT}_{\mathcal{M}}\overbrace{GGA}^{\mathcal{U}}.$$

We now prove the following result that summarizes relationships between the languages $\mathcal{R}$, $\mathcal{M}$, and $\mathcal{U}$.

THEOREM 7.2.5. *The languages $\mathcal{M}$, $\mathcal{R}$, and $\mathcal{U}$ satisfy*

$$\mathcal{M}^* = \mathcal{A}^* \cdot w + \mathcal{S}, \tag{7.2.10}$$

$$\mathcal{U} \cdot \mathcal{A} = \mathcal{M} + \mathcal{U} - \{\varepsilon\}, \tag{7.2.11}$$

$$w(\mathcal{M} - \varepsilon) = \mathcal{A} \cdot \mathcal{R} - \mathcal{R}. \tag{7.2.12}$$

*Proof.* We first deal with (7.2.10). Clearly, $\mathcal{A}^* w$ contains at least one occurrence of $w$ on the right, hence $\mathcal{A}^* w \subset \mathcal{M}^*$. Furthermore, a word $v$ in $\mathcal{M}^*$ is not in $\mathcal{A}^* \cdot w$ if and only if its size $|v|$ is smaller than $|w|$ (e.g., think of $v = ab \in \mathcal{M}$ for $w = abab$). Then the second $w$ occurrence in $wv$ overlaps with $w$, which means that $v$ is in $\mathcal{S}$.

Let us turn now to (7.2.11). When one adds a character $a \in \mathcal{A}$ right after a word $u$ from $\mathcal{U}$, two cases may occur. Either $wua$ still does not contain a second occurrence of $w$ (which means that $ua$ is a nonempty word of $\mathcal{U}$) or a new $w$ appears, clearly at the right end. Hence $\mathcal{U} \cdot \mathcal{A} \subseteq \mathcal{M} + \mathcal{U} - \varepsilon$. Let now $v \in \mathcal{M} - \varepsilon$, then by definition $wv \in \mathcal{T}_2 \subseteq \mathcal{U}\mathcal{A} - \mathcal{U}$ which proves (7.2.11).

We now prove (7.2.12). Let now $x = ar$ be a word in $w \cdot (\mathcal{M} - \varepsilon)$ where $a \in \mathcal{A}$. As $x$ contains exactly two occurrences of $w$ located at its left and right ends, $r$ is in $\mathcal{R}$ and $x$ is in $\mathcal{A} \cdot \mathcal{R} - \mathcal{R}$, hence $w(\mathcal{M} - \varepsilon) \subseteq \mathcal{A} \cdot \mathcal{R} - \mathcal{R}$. To prove $\mathcal{A} \cdot \mathcal{R} - \mathcal{R} \subseteq w(\mathcal{M} - \varepsilon)$, we take a word $arw$ from $\mathcal{A} \cdot \mathcal{R}$ that is not in $\mathcal{R}$. Then $arw$ contains a second $w$ occurrence starting in $ar$. As $rw$ is in $\mathcal{R}$, the only possible position is at the left end, and then $x$ is in $w(\mathcal{M} - \varepsilon)$. This proves (7.2.12). ∎

### 7.2.2. Generating functions

The next step is to translate the relationships between languages into the associated generating functions. Therefore, we must now select the probabilistic model according to which the text is generated. We derive our results for a Markov model of order one. We adopt the following notation: To extract a particular element, say with index $(i,j)$, from a matrix, say P, we shall write $[\mathrm{P}]_{i,j} = p_{i,j}$. We also recall that $(\mathrm{I}-\mathrm{P})^{-1} = \sum_{k\geq 0} \mathrm{P}^k$ provided $||\mathrm{P}|| < 1$ for a matrix norm $||\cdot||$. We also write $\Pi$ for the stationary matrix that consists of $V$ identical rows equal to $\mu$. Finally, by Z we denote the *fundamental matrix* $\mathrm{Z} = (\mathrm{I} - (\mathrm{P}-\Pi))^{-1}$ where I is the identity matrix.

The next lemma translates the relationships between languages (7.2.10)–(7.2.12) into generating functions $M_w(z)$, $U_w(z)$ and $R(z)$ of languages $\mathcal{M}$, $\mathcal{U}$ and $\mathcal{R}$ (we recall that the first two generating function are *conditioned* on $w$ appearing just before any word from $\mathcal{M}$ and $\mathcal{U}$). We define a function $F(z)$ by

$$F(z) = \frac{1}{\mu_{w_1}}[\sum_{n\geq 0}(\mathrm{P}-\Pi)^{n+1}z^n]_{w_m,w_1} = \frac{1}{\mu_{w_1}}[(\mathrm{P}-\Pi)(\mathrm{I}-(\mathrm{P}-\Pi)z)^{-1}]_{w_m,w_1} \tag{7.2.13}$$

for $|z| <\| \mathrm{P}-\Pi \|^{-1}$, where $\mu_{w_1}$ is the stationary probability of the first symbol $w_1$ of $w$. For memoryless sources $F(z) = 0$.

LEMMA 7.2.6. *For Markov sources (of order one), the generating functions associated with languages $\mathcal{M}$,$\mathcal{U}$, and $\mathcal{R}$ satisfy*

$$\frac{1}{1-M_w(z)} = S_w(z) + P(w)z^m\left(\frac{1}{1-z} + F(z)\right), \tag{7.2.14}$$

$$U_w(z) = \frac{M_w(z)-1}{z-1}, \tag{7.2.15}$$

$$R(z) = P(w)z^m \cdot U_w(z), \tag{7.2.16}$$

*provided the underlying Markov chain is aperiodic and ergodic.*

*Proof.* We first prove (7.2.15). Let us consider language relationship (7.2.11) from Theorem 7.2.5, which we rewrite as $\mathcal{U}\cdot\mathcal{A} - \mathcal{U} = \mathcal{M} - \varepsilon$. Observe that $\sum_{b\in\mathcal{A}} p_{ab}z = z$. Hence, set $\mathcal{U}\cdot\mathcal{A}$ yields (conditioning on the left occurrence of $w$)

$$\sum_{u\in\mathcal{U}}\sum_{b\in\mathcal{A}} P(ub|w)z^{|ub|} = \sum_{a\in\mathcal{A}}\sum_{u\in\mathcal{U},\ell(u)=a} P(u|w)z^{|u|}\sum_{b\in\mathcal{A}} p_{ab}z = U_w(z)\cdot z,$$

where $\ell(u)$ denotes the last symbol of the word $u$. Of course, $\mathcal{M}-\varepsilon$ and $\mathcal{U}$ translate into $M_w(z)-1$ and $\mathcal{U}_w(z)$, and (7.2.15) is proved.

We now turn our attention to (7.2.16), and we use relationship (7.2.12) $w\mathcal{M} - w = \mathcal{A}\mathcal{R} - \mathcal{R}$ of Theorem 7.2.5. In order to compute the *conditional* generating function of $\mathcal{A}\cdot\mathcal{R}$ we proceed as follows

$$\sum_{ab\in\mathcal{A}^2}\sum_{bv\in\mathcal{R}} P(abv)z^{|abv|} = z^2\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{A}}\mu_a p_{ab}\sum_{bv\in\mathcal{R}} P(v|v_{-1}=b)z^{|v|}.$$

But due to the stationarity of the underlying Markov chain $\sum_a \mu_a p_{ab} = \mu_b$. As $\mu_b P(v|v_{-1} = b) = P(bv)$, we get $zR(z)$. Furthermore, $w \cdot \mathcal{M} - w$ translates into $P(w)z^m \cdot (M_w(z) - 1)$. By just proved (7.2.15), this is $P(w)z^m \cdot \mathcal{U}_w(z)(z-1)$, and after a simplification, we obtain (7.2.16).

Finally, we deal with (7.2.14), and prove it using (7.2.10) from Theorem 7.2.5. The left-hand side of (7.2.10) involves language $\mathcal{M}$, hence we must condition on the left occurrence of $w$. In particular, $\bigcup_{r\geq 1} \mathcal{M}^r + \varepsilon$ of (7.2.10) translates into $\frac{1}{1-\mathcal{M}_w(z)}$. Now we deal with $\mathcal{A}^* \cdot w$ of the right-hand side of (7.2.10). *Conditioning* on the left occurrence of $w$, the generating function $A_w(z)$ of $\mathcal{A}^* \cdot w$ is

$$\begin{aligned} A_w(z) &= \sum_{n\geq 0} \sum_{|u|=n} z^{n+m} P(uw|u_{-1} = w_m) \\ &= \sum_{n\geq 0} \sum_{|u|=n} z^n P(uw_1|u_{-1} = w_m) P(w_2 \ldots w_m|w_1) z^m. \end{aligned}$$

We have $P(w_2 \ldots w_m|w_1)z^m = \frac{1}{\mu_{w_1}} z^m P(w)$, and for $n \geq 0$:

$$\sum_{|u|=n} P(uw_1|u_{-1} = w_m) = [\mathrm{P}^{n+1}]_{w_m,w_1}$$

where, we recall, $w_m$ is the last character of $w$. In summary, the language $\mathcal{A}^* \cdot w$ contributes $P(w)z^m \left[\frac{1}{\mu_{w_1}} \sum_{n\geq 0} \mathrm{P}^{n+1} z^n\right]_{w_m,w_1}$, while the language $\mathcal{S} - \{\varepsilon\}$ introduces $S_w(z) - 1$. Using the equality $\mathrm{P}^{n+1} - \Pi = (\mathrm{P} - \Pi)^{n+1}$ (which follows from a consecutive application of the identity $\Pi\mathrm{P} = \Pi$), and observing that for any symbols $a$ and $b$

$$\left[\frac{1}{\mu_b} \sum_{n\geq 0} \Pi z^n\right]_{ab} = \sum_{n\geq 0} z^n = \frac{1}{1-z}.$$

we finally obtain the sum in (7.2.14). This completes the proof of the theorem. ■

The lemma above together with Theorem 7.2.3 suffice to derive generating functions $N_r(z)$ and $N(z,u)$ in an explicit form.

Theorem 7.2.7. *Let $w$ be a given pattern of size $m$, and $X$ be a random text of length $n$ generated according to an ergodic and aperiodic Markov chain with the transition probability matrix* P. *Define*

$$D_w(z) = (1-z)S_w(z) + z^m P(w)(1 + (1-z)F(z)). \tag{7.2.17}$$

*Then*

$$N_0(z) = \frac{1-R(z)}{1-z} = \frac{S_w(z)}{D_w(z)}, \tag{7.2.18}$$

$$N_r(z) = R(z) M_w^{r-1}(z) U_w(z)\ , \quad r \geq 1, \tag{7.2.19}$$

$$N(z,u) = R(z) \frac{u}{1-uM_w(z)} U_w(z), \tag{7.2.20}$$

*where*

$$M_w(z) = 1 + \frac{z-1}{D_w(z)} , \tag{7.2.21}$$

$$U_w(z) = \frac{1}{D_w(z)}, \tag{7.2.22}$$

$$R(z) = z^m P(w) \frac{1}{D_w(z)}. \tag{7.2.23}$$

We recall that for memoryless sources, $F(z) = 0$, and hence

$$D(z) = (1-z)S(z) + z^m P(w). \tag{7.2.24}$$

*Proof.* We only comment on the derivation of $N_0(z)$ since the rest follows directly from our previous results. Observe that

$$N_0(z) = \sum_{n\geq 0} \mathbf{P}(N_n = 0) z^n = \sum_{n\geq 0} (1 - \mathbf{P}(N_n > 0)) z^n = \frac{1}{1-z} - \sum_{r=1}^{\infty} N_r(z),$$

thus the first expression follows form (7.2.19). The second expression is a direct translation of $\mathcal{T}_0 \cdot w = \mathcal{R} \cdot \mathcal{A}$ (cf. (7.2.9)) which reads $N_0(z)P(w)z^m = R(z)S_w(z)$ in terms of the appropriate generating functions. ∎

### 7.2.3. Moments and limit laws

In the previous section we derived an explicit formula for the generating function $N(z,u) = \sum_{n\geq 0} \mathbf{E}(u^{N_n}) z^n$ and $N_r(z)$. These formulas can be used to obtain explicit and asymptotic expressions for moments of $N_n$ (cf. Theorem 7.2.8), the central limit theorem (cf. Theorem 7.2.11), and large deviations (cf. Theorem 7.2.12). We start with derivation of the mean and the variance of $N_n$.

THEOREM 7.2.8. *Under the assumptions of Theorem 7.2.7 and $nP(w) \to \infty$, one has, for $n \geq m$:*

$$\mathrm{E}[N_n(w)] = P(w)(n-m+1) , \tag{7.2.25}$$

*and*

$$\mathrm{Var}[N_n(w)] = nc_1 + c_2 + O(R^{-n}), \text{ for } R > 1 \tag{7.2.26}$$

*where*

$$c_1 = P(w)(2S_w(1) - 1 - (2m-1)P(w) + 2P(w)E_1)), \tag{7.2.27}$$

$$\begin{aligned} c_2 = &\, P(w)((m-1)(3m-1)P(w) - (m-1)(2S_w(1)-1) - 2\mathcal{S}'_w(1)) \\ &- 2(2m-1)P(w)^2 E_1 + 2E_2 P(w)^2, \end{aligned} \tag{7.2.28}$$

*and the constants $E_1$, $E_2$ are*

$$E_1 = \frac{1}{\mu_{w_1}} [(\mathrm{P} - \Pi)\mathrm{Z}]_{w_m, w_1}, \quad E_2 = \frac{1}{\mu_{w_1}} [(\mathrm{P}^2 - \Pi)\mathrm{Z}^2]_{w_m, w_1},$$

*Proof.* Notice that first moment estimate can be derived directly from the definition of the stationary probability of $w$. In order to grasp higher moments we will use analytic tools applied to generating functions. We compute the first two moments of $N_n$ from $N(z,u)$ since $\mathbf{E}(N_n) = [z^n]N_u(z,1)$ and $\mathbf{E}(N_n(N_n - 1)) = [z^n]N_{uu}(z,1)$ where $N_u(z,1)$ and $N_{uu}(z,1)$ are the first and the second derivatives of $N(z,u)$ with respect to variable $u$ at $(z,1)$. By Theorem 7.2.7 we find

$$\begin{aligned} N_u(z,1) &= \frac{z^m P(w)}{(1-z)^2}, \\ N_{uu}(z,1) &= \frac{2z^m P(w) M_w(z) D_w(z)}{(1-z)^3}. \end{aligned}$$

Now we observe that both expressions admit as a numerator a function that is analytic beyond the unit circle. Furthermore, for a positive integer $k > 0$

$$[z^n](1-z)^{-k} = \binom{n+k-1}{k-1} = \frac{\Gamma(n+k)}{\Gamma(k)\Gamma(n+1)}, \tag{7.2.29}$$

(where $\Gamma(x)$ is the Euler gamma function), we find for $n \geq m$

$$\mathbf{E}(N_n) = [z^n]N_u(z,1) = P(w)[z^{n-m}](1-z)^{-2} = (n-m+1)P(w).$$

In order to estimate variance, we introduce

$$\Phi(z) = 2z^m P(w) M_w(z) D_w(z),$$

and observe that

$$\Phi(z) = \Phi(1) + (z-1)\Phi'(1) + \frac{(z-1)^2}{2}\Phi''(1) + (z-1)^3 f(z),$$

where $f(z)$ is the remainder of the Taylor expansion of $\Phi(z)$ up to order 3 at $z = 1$. For *memoryless sources*, $\Phi(z)$ and thus $f(z)$ are polynomials of degree $2m-2$ and $[z^n](z-1)f(z)$ is 0 for $n \geq 2m-1$. Hence, by (7.2.29) we arrive at

$$\mathbf{E}(N_n(N_n-1)) = [z^n]N_{uu}(z,1) = \Phi(1)\frac{(n+2)(n+1)}{2} - \Phi'(1)(n+1) + \frac{1}{2}\Phi''(1).$$

But $\mathcal{M}_w(z)D_w(z) = D_w(z) + (1-z)$ and taking into account formula (7.2.24) for $D(z)$, we finally obtain (7.2.26).

For *Markov sources*, $D_w(z)$ has an additional term, namely

$$[z^n]\frac{2(z^{2m}P(w)^2F(z))}{(1-z)^2},$$

where $F(z)$, defined in (7.2.13), is analytic beyond the unit circle for $|z| \leq R$, with $R > 1$. The Taylor expansion of $F(z)$ is $E_1 + (1-z)E_2$, and applying (7.2.29) again yields the result. ∎

Recall that $\mathrm{P} = \Pi$ for memoryless sources, so $E_1 = E_2 = 0$ and (7.2.26) reduces to an equality for $n \geq 2m-1$. Thus

$$\mathrm{Var}[N_n(w)] = nc_1 + c_2 \tag{7.2.30}$$

with

$$\begin{aligned} c_1 &= P(w)(2S(1) - 1 - (2m-1)P(w)), \\ c_2 &= P(w)((m-1)(3m-1)P(w) - (m-1)(2S(1)-1) - 2S'(1)). \end{aligned}$$

In passing we should notice that from the generating function $N(z,u)$ we can compute all moments of $N_n$. Instead, however, we present some limit laws for $\mathbf{P}(N_n = r)$ for different values of $r$: We consider $r = O(1)$, $r = \mathbf{E}(N_n) + x\sqrt{\mathrm{Var}(N_n)}$ (central and local limit regime), and $r = (1+\delta)\mathbf{E}(N_n)$ (large deviations). From the central limit theorem (cf. Theorem 7.2.11 below) we conclude that the normalized random variable $(N_n - \mathbf{E}(N_n))/\sqrt{\mathrm{Var}(N_n)}$ converges also in moments to the moments of the standard normal distribution. This follows from the fact that in the theorem below we prove the convergence of the normalized generating function to an analytic function, namely $e^{u^2/2}$ for $u$ complex in the vicinity of zero. Since an analytic function has well defined derivatives, convergence in moments follows. We shall leave a formal proof to the reader (cf. Exercise 7.2.3).

Theorem 7.2.9. *Under the assumptions of Theorem 7.2.8, let $\rho_w$ be the root of $D_w(z) = 0$ of the smallest modulus and multiplicity one. Then, $\rho_w$ is real such that $\rho_w > 1$, and there exists $\rho > \rho_w$ such that for $r = O(1)$*

$$\mathbf{P}(N_n(w) = r) = \sum_{j=1}^{r+1} (-1)^j a_j \binom{n}{j-1} \rho_w^{-(n+j)} + O(\rho^{-n}) \ , \tag{7.2.31}$$

*where*

$$a_{r+1} = \frac{\rho_w^m P(w)\,(\rho_w - 1)^{r-1}}{(D_w'(\rho_w))^{r+1}} \ , \tag{7.2.32}$$

*and the remaining coefficients can be computed according to*

$$a_j = \frac{1}{(r+1-j)!} \lim_{z \to \rho_w} \frac{d^{r+1-j}}{dz^{r+1-j}} \left( N_r(z)(z-\rho_w)^{r+1} \right) \tag{7.2.33}$$

*with $j = 1, 2, \ldots, r$.*

In order to prove Theorem 7.2.9, we need the following simple result.

Lemma 7.2.10. *The equation $D_w(z) = 0$ has at least one root, and all its roots are of modulus greater than 1.*

*Proof.* Poles of $D_w(z) = (1-z)/(1-M_w(z))$ are clearly poles of $\frac{1}{1-M_w(z)}$. As $\frac{1}{1-M_w(z)}$ is the generating function of a language, it converges for $|z| < 1$ and has no pole of modulus smaller than 1. Since $D_w(1) \neq 0$, then $z = 1$ is a simple pole of $1/(1 - M_w(z))$. As all its coefficients are real and non negative, there is no other pole of modulus $|z| = 1$. It follows that all roots of $D_w(z)$ are of modulus greater than 1. The existence of a root is guaranteed since $D_w(z)$ is either a polynomial (Bernoulli model) or a ratio of polynomials (Markov model). ■

*Proof of Theorem 7.2.9.* We first re-write the formula on $N_r(z)$ as follows

$$N_r(z) = \frac{z^m P(w)(D_w(z) + z - 1)^{r-1}}{D_w^{r+1}(z)} \ . \tag{7.2.34}$$

Observe that $\mathbf{P}(N_n(w) = r)$ is the coefficient at $z^n$ of $N_r(z)$. By Hadamard's theorem, asymptotics of the coefficients of a generating function depend on the singularities of the underlying generating function. In our case, the generating function $N_r(z)$ is a rational function, thus we can only expect poles (for which the denominator $D_w(z)$ vanishes). Lemma 7.2.10 above establishes the existence and properties of such a pole. Therefore, the generating function $N_r(z)$ can be expanded around its root of smallest modulus, let $\rho_w$ be this smallest modulus, in Laurent's series:

$$N_r(z) = \sum_{j=1}^{r+1} \frac{a_j}{(z-\rho_w)^j} + \widetilde{N}_r(z) \tag{7.2.35}$$

where $\widetilde{N}_r(z)$ is analytical in $|z| < \rho'$ and $\rho'$ is defined as $\rho' = \inf\{|\rho| : \ \rho > \rho_w \text{ and } D_w(\rho) = 0\}$. The constants $a_j$ satisfy (7.2.33). This formula simplifies into (7.2.32) for the leading constant $a_{-r-1}$. As a consequence of analyticity we have for $1 < \rho_w < \rho < \rho'$: $[z^n]\widetilde{N}^{(r)}(z) = O(\rho^{-n})$. Hence, the term $\widetilde{N}_r(z)$ contributes only to the lower terms in the asymptotic expansion of $N_r(z)$. After some algebra, and noting that $[z^n]1/(1-z)^{k+1} = \binom{n+k}{n}$, we prove Theorem 7.2.9. ■

In the next theorem we establish the central limit theorem in its strong form (i.e., local limit theorem).

THEOREM 7.2.11. *Under the same assumption as in Theorem 7.2.8 we have*

$$\mathbf{P}(N_n(w) \le \mathbf{E}(N_n) + x\sqrt{\mathrm{Var}(N_n)}) = \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{7.2.36}$$

*If, in addition, $p_{ij} > 0$ for all $i, j \in \mathcal{A}$, then for any bounded real interval $B$*

$$\sup_{x\in B} \left| \mathbf{P}(N_n(w) = \lfloor \mathbf{E}(N_n) + x\sqrt{\mathrm{Var}(N_n)} \rfloor) - \frac{1}{\sqrt{2\pi \mathrm{Var}(N_n)}} e^{-\frac{1}{2}x^2} \right| = o\left(\frac{1}{\sqrt{n}}\right) \tag{7.2.37}$$

*as $n \to \infty$.*

*Proof.* Let $r = \lfloor \mathbf{E}(N_n) + x\sqrt{\mathrm{Var}(N_n)} \rfloor$ with $x = O(1)$. We compute $\mathbf{P}(N_n(w) \le r)$ (central limit theorem) and $\mathbf{P}(N_n(w) = r)$ (local limit theorem) for $r = \mathbf{E}(N_n) + x\sqrt{\mathrm{Var}(N_n)}$ when $x = O(1)$. Let $\nu_n = \mathbf{E}(N_n(w)) = (n-m+1)P(w)$ and $\sigma_n^2 = \mathrm{Var}(N_n(w)) = c_1 n + O(1)$. To establish normality of $(N_n(w) - \nu_n)/\sigma_n$, it suffices, according to Lévy's continuity theorem, to prove the following

$$\lim_{n\to\infty} e^{-\tau\nu_n/\sigma_n} N_n(e^{\tau/\sigma_n}) = e^{\tau^2/2} \tag{7.2.38}$$

for complex $\tau$ (actually, $\tau = iv$ suffices). Again, by Cauchy's theorem

$$N_n(u) = \frac{1}{2\pi i}\oint \frac{N(z,u)}{z^{n+1}} dz = \frac{1}{2\pi i}\oint \frac{uP(w)}{D_w^2(z)(1-uM_w(z))z^{n+1-m}} dz \ ,$$

where the integration is along a circle around the origin. The evaluation of this integral is standard and it appeals to the Cauchy residue theorem. Namely, we enlarge the circle of integration to a bigger one, say $R > 1$, such that the bigger circle contains the dominant pole of the integrand function. Observe that the Cauchy integral over the bigger circle is $O(R^{-n})$. Let us now substitute $u = e^t$ and $z = e^\rho$. Then, the poles of the integrand are the roots of the equation

$$1 - e^t M_w(e^\rho) = 0. \tag{7.2.39}$$

This equation implicitly defines in some neighborhood of $t = 0$ a unique $C^\infty$ function $\rho(t)$, satisfying $\rho(0) = 0$. Notably, all other roots $\rho$ satisfy $\inf|\rho| = \rho' > 0$. Then, the residue theorem with $e^{\rho'} > R > e^\rho > 1$ leads to

$$N_n(e^t) = C(t)e^{-(n+1-m)\rho(t)} + O(R^{-n}) \tag{7.2.40}$$

where

$$C(t) = \frac{P(w)}{D_w^2(\rho(t))M_w'(\rho(t))}.$$

To study properties of $\rho(t)$, we observe that the cumulant formula implies $\mathbf{E}(N_n(w)) = [t]\log N_n(e^t)$ and $\sigma_n^2 = [t^2]\log N_n(e^t)$ where, we recall, $[t^r]f(t)$ denotes the coefficient of $f(t)$ at $t^r$. In our case, $\nu_n \sim -n\rho'(0)$ as well as $\sigma_n^2 \sim -n\rho''(0)$. Set now in (7.2.40) $t = \tau/\sigma_n \to 0$ for some complex $\tau$. Since uniformly in $t$ we have $\rho(t) = t\rho'(0) + \rho''(0)t^2/2 + O(t^3)$ for $t \to 0$, our estimate (7.2.40) leads to

$$\begin{aligned} e^{-\tau\nu_n/\sigma_n} N_n(e^{\tau/\sigma_n}) &= \exp\left(\frac{\tau^2}{2} + O(n\tau^3/\sigma_n^3)\right) \\ &= e^{\tau^2/2}\left(1 + O(1/\sqrt{n})\right), \end{aligned}$$

which proves (7.2.36) after applying the Berry-Essen inequality that allows to derive the error term $O(1/\sqrt{n})$ for the probability distribution.

To establish the local limit theorem, we observe that if $p_{ij} > 0$ for all $i, j \in \mathcal{A}$, then $\rho(t) > 0$ for $t \neq 0$ (cf. Exercise 7.2.4). We can obtain much more refined local limit result. Indeed, we find for $x = o(n^{1/6})$

$$\mathbf{P}(N_n = \mathbf{E}(N_n) + x\sqrt{nc_1}) = \frac{1}{\sqrt{2\pi nc_1}} e^{-\frac{1}{2}x^2}\left(1 - \frac{\kappa_3}{2c_1^{3/2}\sqrt{n}}\left(x - \frac{x^3}{3}\right)\right) + O(n^{-3/2}) \ , \tag{7.2.41}$$

where $\kappa_3$ a constant (i.e., the third cumulant). This completes the proof of Theorem 7.2.11. ■

Finally, we establish precise large deviations for $N_n$. Large deviations play a central role in many applications, most notably in data mining and molecular biology, since it allows to establish a threshold for overrepresented and underrepresented patterns.

Theorem 7.2.12. *Let $r = a\mathrm{E}[N_n]$ with $a = (1+\delta)P(w)$ for $\delta \neq 0$. For complex $t$, define $\rho(t)$ to be the root of*

$$1 - e^t M_w(e^\rho) = 0 \ , \tag{7.2.42}$$

*and define $\omega_a$ and $\sigma_a$ by*

$$-\rho'(\omega_a) = a, \quad -\rho''(\omega_a) = \sigma_a^2.$$

*Then*

$$\mathbf{P}(N_n(w) = (1+\delta)\mathbf{E}(N_n)) \sim \frac{1}{\sigma_a\sqrt{2\pi(n-m+1)}} e^{-(n-m+1)I(a)+\theta_a} \tag{7.2.43}$$

*where $I(a) = a\omega_a + \rho(\omega_a)$ and*

$$\theta_a = \log \frac{P(w)e^{m\rho(\omega_a)}}{D_w(e^{\rho(\omega_a)}) + (1 - e^{\rho(\omega_a)})D'_w(e^{\rho(\omega_a)})}, \tag{7.2.44}$$

*and $D_w(z)$ is defined in (7.2.17).*

*Proof.* From (7.2.40) we conclude that

$$\lim_{n\to\infty} \frac{\log N_n(e^t)}{n} = -\rho(t) \ .$$

By the Gärtner-Ellis theorem we find

$$\lim_{n\to\infty} \frac{\log \mathbf{P}(N_n > na)}{n} = -I(a) \ ,$$

where

$$I(a) = a\omega_a + \rho(\omega_a)$$

with $\omega_a$ being a solution of $-\rho'(t) = a$, A stronger version of the above result is possible and we derive it in the sequel. In fact, we use (7.2.41) and the "shift of mean" technique.

As in the local limit regime, we could use Cauchy's formula to compute the probability $\mathbf{P}(N_n = r)$ for $r = \mathbf{E}(N_n) + xO(\sqrt{n})$. But, formula (7.2.41) is only good for $x = O(1)$ while we need $x = O(\sqrt{n})$ for the large deviations. To expand its validity, we shift the mean of the generating function $N_n(u)$ to a new value, say $m = an = (1+\delta)P(w)(n-m+1)$, so we can again apply the central limit formula (7.2.41) around the new mean. To accomplish this, let us re-write (7.2.40) as for any $R > 0$

$$N_n(e^t) = C(t)[g(t)]^{n-m+1} + O(R^{-n})$$

where $g(t) = e^{-\rho(t)}$. (In the derivation below, for simplicity we dropped $O(R^{-n})$ term.) The above suggests that $N_n(e^t)$ is the moment generating function of a sum $S_n$ of $n-m+1$ "almost" independent random variables $X_1, \ldots, X_{n-m+1}$ having moment generating function equal to $g(t)$ and $Y$ whose moment generating function is $C(t)$. Observe that $\mathbf{E}(S_n) = (n-m+1)P(w)$ while we need to estimate the tail of $S_n$ around $(1+\delta)(n-m+1)P(w)$. To achieve it, we introduce a new random variable $\widetilde{X}_i$ whose moment generating function $\widetilde{g}(t)$ is

$$\widetilde{g}(t) = \frac{g(t+\omega)}{g(\omega)}$$

where $\omega$ will be chosen later. Then, the mean and the variance of the new variable $\widetilde{X}$ is

$$\mathbf{E}(\widetilde{X}) = \frac{g'(\omega)}{g(\omega)} = -\rho'(\omega) \ ,$$

$$\mathrm{Var}(\widetilde{X}) = \frac{g''(\omega)}{g(\omega)} - \left(\frac{g'(\omega)}{g(\omega)}\right)^2 = -\rho''(\omega) \ .$$

Let us now choose $\omega_a$ such that

$$-\rho'(\omega_a) = \frac{g'(\omega_a)}{g(\omega_a)} = a = P(w)(1+\delta) \ .$$

Then, the new sum $\widetilde{S}_n - Y = \widetilde{X}_1 + \ldots + \widetilde{X}_{n-m+1}$ has a new mean $(1+\delta)P(w)(n-m+1) = a(n-m+1)$, and hence we can apply to $\widetilde{S}_n - Y$ the central limit result (7.2.41). To translate from $\widetilde{S_n} - Y$ to $S_n$ we use the following simple formula

$$[e^{tM}]\left(g^n(t)\right) = \frac{g^n(\omega)}{e^{\omega M}}[e^{tM}]\left(\frac{g^M(t+\omega)}{g^M(\omega)}\right) \tag{7.2.45}$$

where $M = a(n-m+1)$ and $[e^{tn}]g(t)$ denotes the coefficient of $g(t)$ at $z^n = e^{tn}$ (where $z = e^t$). Now, we can apply (7.2.41) to the right-hand side of the above to obtain

$$[e^{tM}]\left(\frac{g^M(t+\omega)}{g^M(\omega)}\right) \sim \frac{1}{\sigma_a\sqrt{2\pi(n-m+1)}}.$$

To obtain the final result we must take into account the effect of $Y$ whose moment generating function is $C(t)$. This leads to replacing $a = 1+\delta$ by $a =$

**Table 7.1.** $Z$ score vs $p$-value of tandem repeats in *A.thaliana*.

| Oligomer | Obs. | $p$-val (large dev.) | $Z$-sc. |
|---|---|---|---|
| AATTGGCGG | 2 | $8.059 \times 10^{-4}$ | 48.71 |
| TTTGTACCA | 3 | $4.350 \times 10^{-5}$ | 22.96 |
| ACGGTTCAC | 3 | $2.265 \times 10^{-6}$ | 55.49 |
| AAGACGGTT | 3 | $2.186 \times 10^{-6}$ | 48.95 |
| ACGACGCTT | 4 | $1.604 \times 10^{-9}$ | 74.01 |
| ACGCTTGG | 4 | $5.374 \times 10^{-10}$ | 84.93 |
| GAGAAGACG | 5 | $0.687 \times 10^{-14}$ | 151.10 |

$1+\delta+C'(0)/n$ resulting the the correction term $e^{\theta_a} = e^{C'(0)\omega_a}$. Theorem 7.2.12 is proved. ■

We illustrate the above results on an example taken from molecular biology.

Example 7.2.13. Biologists apply the so called $Z$-score and $p$-value to determine whether biological sequences such as DNA or protein contain a biological signal, that is, an underrepresented or overrepresented patterns. These quantities are defined as

$$Z(w) = \frac{\mathbf{E}(N_n) - N_n(w)}{\sqrt{\mathrm{Var}(N_n(w))}},$$
$$pval(r) = P(N_n(w) > r).$$

$Z$-score indicates by how many standard deviations the observed value $N_n(w)$ is away from the mean. Clearly, this score makes sense only if one can prove, as we did in Theorem 7.2.11, that $Z$ satisfies (at least asymptotically) the Central Limit Theorem (CLT). On the other hand, $p$-value is used for rare occurrences, far away from the mean where one needs to apply the large deviations as in Theorem 7.2.12.

The range of validity of $Z$-score and $p$-value are important as illustrated in Table 7.2.13 where results for 2008 nucleotides long fragments of *A.thaliana* (a plant genome) are presented. In the table for each 9-mer the number of observations is presented in the first column following by the large deviations probability computed from Theorem 7.2.12 and $Z$-score. We observe that for *AATTGGCGG* and *AAGACGGTT* the $Z$-scores are about 48 while $p$-values differ by two order of magnitudes. In fact, occurrences of these 9-mers are very rare, and therefore $Z$-score is not an adequate measure.

### 7.2.4. Waiting times

We shall now discuss the waiting times $T_w$ and $T_j$, where $T_w = T_1$ is the first time $w$ occurs in the text, while $T_j$ is the minimum length of the text in which

$w$ occurs $j$ times. Fortunately, we do not need re-derive generating function of $T_j$ since, as we have already indicated in (7.2.3), the following *duality principle* holds

$$\{N_n \geq j\} = \{T_j \leq n\},$$

and in particular, $\{T_w > n\} = \{N_n = 0\}$. Therefore, if

$$T(u,z) = \sum_{n\geq 0}\sum_{j\geq 0} \mathbf{P}(T_j = n) z^n u^j,$$

then by the duality principle we have

$$(1-u)T(u,z) + u(1-z)N(z,u) = 1,$$

and one obtains $T(u,z)$ from Theorem 7.2.7. Waiting times were analyzed in depth in Chapter 6.

Finally, observe that the above duality principle implies

$$\mathbf{E}(T_w) = \sum_{n\geq 0} \mathbf{P}(N_n = 0) = N_0(1).$$

In particular, for *memoryless sources*, from Theorem 7.2.7 we conclude that

$$N_0(z) = \frac{z^m S(z)}{(1-z)S(z) + z^m P(w)}.$$

Hence

$$\begin{aligned}\mathbf{E}(T_w) &= \sum_{n\geq 0} \mathbf{P}(N_n(w) = 0) = N_0(1) = \frac{S(1)}{P(w)} \\ &= \sum_{k\in\mathcal{P}(w)} \frac{1}{P(w_1^k)} = \frac{1}{P(w)} + \sum_{k\in\mathcal{P}(w)-\{m\}} \frac{1}{P(w_1^k)} \qquad (7.2.46)\end{aligned}$$

## 7.3. Generalized string matching

In this section we consider generalized pattern matching in which a set of patterns (rather than a single pattern) is given. We assume that the pattern is a pair of sets of words $(\mathcal{W}_0, \mathcal{W})$ where $\mathcal{W} = \bigcup_{i=1}^d \mathcal{W}_i$ consists of sets $\mathcal{W}_i \subset \mathcal{A}^{m_i}$ (i.e., all words in $\mathcal{W}_i$ have a fixed length equal to $m_i$). The set $\mathcal{W}_0$ is called the *forbidden set*. For $\mathcal{W}_0 = \emptyset$ one is interested in the number of pattern occurrences, $N_n(\mathcal{W})$, defined as the *number of patterns* from $\mathcal{W}$ occurring in the text $X_1^n$ generated by a (random) source. Another parameter of interest may be the *number of positions* in $X_1^n$ where *a pattern* from $\mathcal{W}$ appears (clearly, some patterns may occur more than once at some positions). The latter quantity we denote as $\Pi_n$. If we define $\Pi_n^{(i)}$ as the number of positions where a word from $\mathcal{W}_i$ occurs, then

$$N_n(\mathcal{W}) = \Pi_n^{(1)} + \cdots + \Pi_n^{(d)}.$$

Notice that at any given position of the text and for a given $i$ only one word from $\mathcal{W}_i$ can occur.

For $\mathcal{W}_0 \neq \emptyset$ one studies the number of occurrences $N_n(\mathcal{W})$ under the condition that $N_n(\mathcal{W}_0) := \Pi_n^{(0)} = 0$, that is, there is no occurrence of a pattern from $\mathcal{W}_0$ in the text $X_1^n$. This could be called a *restricted* pattern matching since one restricts the text to those strings that do not contain strings from $\mathcal{W}_0$.

Finally, we may set $\mathcal{W}_i = \emptyset$ for $i = 1, \ldots, d$ with $\mathcal{W}_0 \neq \emptyset$ and count the number of text strings that do not contain any pattern from $\mathcal{W}_0$. (Alternatively, we can estimate the probability that a randomly selected text $X_1^n$ does not contain any pattern from $\mathcal{W}_0$.) In particular, define for $\ell \leq k$

$$\mathcal{W}_0 = \{\underbrace{0\ldots0}_{\ell}, \ldots, \underbrace{0\ldots0}_{k}\}, \tag{7.3.1}$$

that is, $\mathcal{W}_0$ consists of runs of zeros of length at least $\ell$ and at most $k$. A text satisfying the property that no pattern from $\mathcal{W}_0$ defined in (7.3.1) occurs in it is called a $(\ell, k)$ sequence. Such sequences are used for magnetic coding.

In this section, we first present an analysis of the generalized pattern matching with $\mathcal{W}_0 = \emptyset$ and $d = 1$ that we call the *reduced pattern set* (i.e., no pattern is a substring of another pattern) followed by a detailed analysis of the generalized pattern matching. We describe two methods of analysis. First, we generalize our language approach from the previous section, and then for the general pattern matching case we apply de Bruijn's automaton and spectral analysis of matrices. Finally, we enumerate $(\ell, k)$ sequences and compute the so called Shannon capacity for such sequences.

Throughout this section we assume that the text is generated by a (non-degenerate) memoryless source (B), as defined in Section 7.1.

### 7.3.1. String matching over reduced set of patterns

We analyze here a special case of the generalized pattern matching with $\mathcal{W}_0 = \emptyset$ and $d = 1$. In this case we shall write $\mathcal{W}_1 := \mathcal{W} = \{w_1, \ldots, w_K\}$ where $w_i$ ($1 \leq i \leq K$) are given patterns with fixed length $|w_i| = m$. We shall generalize the results from the exact pattern matching section, but we omit most of the proofs or move them to exercises.

As before, let $\mathcal{T}_{\geq 1}$ be a language of words containing at least one occurrence from the set $\mathcal{W}$, and for any nonnegative integer $r$, let $\mathcal{T}_r$ be the language of words containing exactly $r$ occurrences from $\mathcal{W}$. In order to characterize $\mathcal{T}_r$ we introduce some additional languages for any $1 \leq i, j \leq K$:

- $\mathcal{M}_{ij} = \{v : \ w_i v \in \mathcal{T}_2$ *and* $w_j$ occurs at the right end of $v\}$;
- $\mathcal{R}_i$ defined as the set of words containing only one occurrence of $w_i$, located at the right end;
- $\mathcal{U}_i = \{u : \ w_i u \in \mathcal{T}_1\}$, that is, a set of words $u$ such that the only occurrence of $w_i \in \mathcal{W}$ in $w_i u$ is on the left.

We also need to generalize the autocorrelation set and the autocorrelation polynomial to a set of patterns. For any given two strings $w$ and $u$, let

$$\mathcal{S}_{w,u} = \{u_{k+1}^m : w_{m-k+1}^m = u_1^k\}$$

be the *correlation set*. The set of positions $k$ satisfying $u_1^k = w_{m-k+1}^m$ is denoted as $\mathcal{P}(w,u)$. If $w = x \cdot v$ and $u = v \cdot y$ for some words $x$, $y$, $v$, then $y \in \mathcal{S}_{w,u}$ and $|v| \in \mathcal{P}(w,u)$. The correlation polynomial, $S_{w,u}(z)$, of $w$ and $u$ is the associated generating function of $\mathcal{S}_{w,u}$, that is,

$$S_{w,u}(z) = \sum_{k \in \mathcal{P}(w,u)} P(u_{k+1}^m) z^{m-k}.$$

In particular, for $w_i, w_j \in \mathcal{W}$ we define $\mathcal{S}_{i,j} := \mathcal{S}_{w_i,w_j}$. The *correlation matrix* of $\mathcal{W}$ is denoted as $\mathrm{S}(z) = \{S_{w_i w_j}(z)\}_{i,j=1,K}$.

Example 7.3.1. Consider a DNA sequence over the alphabet $\mathcal{A} = \{A, C, G, T\}$ generated by a memoryless source with $P(A) = \frac{1}{5}$, $P(C) = \frac{3}{10}$, $P(G) = \frac{3}{10}$ and $P(T) = \frac{1}{5}$. Let $w_1 = ATT$ and $w_2 = TAT$. Then the correlation matrix $\mathrm{S}(z)$ is

$$\mathrm{S}(z) = \begin{pmatrix} 1 & 1 + \frac{z^2}{25} \\ 1 + \frac{z}{5} & 1 + \frac{z^2}{25} \end{pmatrix}.$$

In order to analyze the number of occurrences $N_n(\mathcal{W})$ and its generating functions we first generalize the language relationships discussed in Theorem 7.2.3. Observe that

$$\mathcal{T}_r = \sum_{1 \le i,j \le K} \mathcal{R}_i \mathcal{M}_{ij}^{r-1} \mathcal{U}_j,$$
$$\mathcal{T}_{\ge 1} = \sum_{r \ge 1} \sum_{1 \le i,j \le K} \mathcal{R}_i \mathcal{M}_{ij}^{r-1} \mathcal{U}_j,$$

where $\sum$ denotes disjoint union of sets. As in Theorem 7.2.5, one finds the following relationships between just introduced languages

$$\bigcup_{k \ge 1} \mathcal{M}_{i,j}^k = \mathcal{A}^* \cdot w_j + \mathcal{S}_{ij} - \varepsilon \qquad 1 \le i,j \le K,$$
$$\mathcal{U}_i \cdot \mathcal{A} = \bigcup_j \mathcal{M}_{ij} + \mathcal{U}_i - \varepsilon, \qquad 1 \le i \le K,$$
$$\mathcal{A} \cdot \mathcal{R}_j - (\mathcal{R}_j - w_j) = \bigcup_i w_i \mathcal{M}_{ij}, \qquad 1 \le j \le K,$$
$$\mathcal{T}_0 \cdot w_j = \mathcal{R}_j + \mathcal{R}_i(\mathcal{S}_{ij} - \varepsilon), \qquad 1 \le i,j \le K.$$

Let us now analyze $N_n(\mathcal{W})$ in a probabilistic framework. To simplify our presentation, we assume that the text is generated by a memoryless source. Then the above language relationships translate directly into generating functions, as discussed the last section.

Before we proceed, we adopt the following notations. Lower-case letters are reserved for vectors which are assumed to be column vectors (e.g., $\mathrm{x}^t = (x_1, \ldots, x_K)$) except for vectors of generating functions which we denote by uppercase letters (e.g., $\mathrm{U}^t(z) = (U_1(z), \ldots, U_K(z))$ where $U_i(z)$ is the generating function of a language $\mathcal{U}_{w_i}$). In the above the upper index "$t$" denotes transpose. We shall use upper-case letters for matrices (e.g., $\mathrm{S}(z) = \{S_{w_i w_j}(z)\}_{i,j=1,K}$). In particular, we write I for the identity matrix, and $\vec{1}^t = (1, \ldots, 1)$ for the vector of all ones.

Now we are ready to present exact formulas for the generating function $N_r(z) = \sum_{n\geq 0} \mathbf{P}(N_n(\mathcal{W}) = r)z^n$ and $N(z,u) = \sum_{k\geq 0} N_r(z)u^r$. The following theorem is a direct consequences of our definitions and language relationships.

THEOREM 7.3.2. *Let $\mathcal{W} = \{w_1, \ldots, w_K\}$ be a given set of reduced patterns each of length $m$, and $X$ be a random text of length $n$ generated by a memoryless source. The generating functions $N_r(z)$ and $N(z,u)$ can be computed as follows:*

$$N_r(z) = \mathrm{R}^t(z)\mathrm{M}^{r-1}(z)\mathrm{U}(z) \tag{7.3.2}$$

$$N(z,u) = \mathrm{R}^t(z)u(\mathrm{I} - u\mathrm{M}(z))^{-1}\mathrm{U}(z) \ , \tag{7.3.3}$$

*where, denoting $\mathrm{w}^t = (P(w_1), \ldots, P(w_K))$ and $\vec{1}^t = (1, 1, \ldots, 1)$, we have*

$$\mathrm{M}(z) = (\mathrm{D}(z) + (z-1)\mathrm{I})\mathrm{D}(z)^{-1}, \tag{7.3.4}$$

$$(\mathrm{I} - \mathrm{M}(z))^{-1} = \mathrm{S}(z) + \frac{z^m}{1-z}\vec{1}\cdot \mathrm{w}^t, \tag{7.3.5}$$

$$\mathrm{U}(z) = \frac{1}{1-z}(\mathrm{I} - \mathrm{M}(z))\cdot\vec{1}, \tag{7.3.6}$$

$$\mathrm{R}^t(z) = \frac{z^m}{1-z}\mathrm{w}^t\cdot(\mathrm{I} - \mathrm{M}(z)), \tag{7.3.7}$$

*and*

$$\mathrm{D}(z) = (1-z)\mathrm{S}(z) + z^m\vec{1}\cdot\mathrm{w}^t.$$

Using these results and following footsteps of our analysis for the exact pattern matching, we arrive at the following asymptotic results.

THEOREM 7.3.3. *Let the text $X$ be generated by a memoryless source with $P(w_i) > 0$ for $i = 1, \ldots, K$ and $P(\mathcal{W}) = \sum_{w_i\in\mathcal{W}} P(w_i) = \mathrm{w}^t\cdot\vec{1}$.*

(i) *The following holds*

$$\mathbf{E}(N_n(\mathcal{W})) = (n-m+1)P(\mathcal{W}),$$

$$\mathrm{Var}(N_n(\mathcal{W})) = (n-m+1)\left(P(\mathcal{W}) + P^2(\mathcal{W}) - 2mP^2(\mathcal{W}) + 2\mathrm{w}^t(\mathrm{S}(1)-\mathrm{I})\vec{1}\right) + m(m-1)P^2(\mathcal{W}) - 2\mathrm{w}^t\dot{\mathrm{S}}(1)\cdot\vec{1} \ ,$$

*where $\dot{\mathrm{S}}(1)$ denotes the derivative of the matrix $\mathrm{S}(z)$ at $z=1$.*

(ii) *Let* $\rho_{\mathcal{W}}$ *be the smallest root of multiplicity one of* $\det \mathrm{D}(z) = 0$ *outside the unit circle* $|z| \leq 1$. *There exists* $\rho > \rho_{\mathcal{W}}$ *such that for* $r = O(1)$

$$\mathbf{P}(N_n(\mathcal{W}) = r) = (-1)^{r+1}\frac{a_{r+1}}{r!}(n)_r \rho_{\mathcal{W}}^{-(n-m+r+1)} + \sum_{j=1}^{r}(-1)^j a_j \binom{n}{j-1}\rho_{\mathcal{W}}^{-(n+j)} + O(\rho^{-n}) ,$$

*where* $a_r$ *are computable constants.*

(iii) *Let* $B$ *be a bounded real interval and* $r = \lfloor \mathbf{E}(N_n) + x\sqrt{\mathrm{Var}(N_n)} \rfloor$. *Then*

$$\sup_{x\in B}\left|\mathbf{P}(N_n(\mathcal{W}) = r) - \frac{1}{\sqrt{2\pi \mathrm{Var}(N_n)}}e^{-\frac{1}{2}x^2}\right| = o\left(\frac{1}{\sqrt{n}}\right),$$

*as* $n \to \infty$.

(iv) *Let* $r = (1+\delta)\mathbf{E}(N_n)$ *with* $\delta \neq 0$, *and let* $a = (1+\delta)P(\mathcal{W})$. *Define* $\tau(t)$ *to be the root of*

$$\det(\mathrm{I} - e^t \mathrm{M}(e^\tau)) = 0 ,$$

*and* $\omega_a$ *and* $\sigma_a$ *to be*

$$-\tau'(\omega_a) = -a, \qquad -\tau''(\omega_a) = \sigma_a^2.$$

*Then*

$$\mathbf{P}(N_n(\mathcal{W}) = r) \sim \frac{1}{\sigma_a\sqrt{2\pi(n-m+1)}}e^{-(n-m+1)I(a)+\theta_a}$$

*where* $I(a) = a\omega_a + \tau(\omega_a)$ *and* $\theta_a$ *is a computable constant (cf. Exercise 7.3.3).*

*Proof.* We only sketch the derivation of part (iii) but we present two proofs. Our starting point is

$$N(z,u) = \mathrm{R}^t(z)u(\mathrm{I} - u\mathrm{M}(z))^{-1}\mathrm{U}(z)$$

shown in Theorem 7.3.2 to hold for $|z| < 1$ and $|u| < 1$. We may proceed in two different ways.

Method A: Determinant Approach.

Observe that

$$(\mathrm{I} - u\mathrm{M}(z))^{-1} = \frac{\mathrm{B}(z,u)}{\det(\mathrm{I} - u\mathrm{M}(z))}$$

where $\mathrm{B}(z,u)$ is a complex matrix. Let

$$Q(z,u) := \det(\mathrm{I} - u\mathrm{M}(z)),$$

and let $z_0 := \rho(u)$ be the smallest root of

$$Q(z,u) = \det(\mathrm{I} - u\mathrm{M}(z)) = 0.$$

Observe that $\rho(1) = 1$ by (7.3.5).

For our central limit result, we restrict out interest to $\rho(u)$ in a vicinity of $u = 1$. Such a root exists and is unique since for real $z$ the matrix $\mathrm{M}(z)$ has all positive coefficients. The Perron–Frobenius theorem implies that all other roots $\rho_i(u)$ are of smaller modulus. Finally, one can analytically continue $\rho(u)$ to a complex neighborhood of $u$. Thus Cauchy's formula yields for some $A < 1$

$$\begin{aligned} N_n(u) := [z^n]N(z,u) &= \frac{1}{2\pi i}\oint \frac{\mathrm{R}^t(z)\mathrm{B}(z,u)\mathrm{U}(z)}{Q(z,u)}\frac{dz}{z^{n+1}} \\ &= C(u)\rho^{-n}(u)(1 + O(A^n)) \end{aligned}$$

where $C(u) = -\mathrm{R}^t(\rho(u))\mathrm{B}(\rho(u),u)\mathrm{U}(\rho(u))\rho^{-1}(u)/Q'(\rho(u),u)$. As in the proof of Theorem 7.2.11, we recognize a quasi-power form for $N_n(u)$ that directly leads to the central limit theorem. An application of a saddle point method completes the proof of the local limit theorem.

Method B: Eigenvalue Approach

We apply now the Perron–Frobenius theorem for positive matrices together with a matrix spectral representation to obtain even more precise asymptotics. Our starting point is the following formula

$$[\mathrm{I} - u\mathrm{M}(z)]^{-1} = \sum_{k=0}^{\infty} u^k \mathrm{M}^k(z). \tag{7.3.8}$$

Now, observe that $\mathrm{M}(z)$ for real $z$, say $x$, is a positive matrix since each element $M_{ij}(x)$ is the generating function of the language $\mathcal{M}_{ij}$ and for any $v \in \mathcal{M}_{ij}$ we have $P(v) > 0$ for memoryless sources. Let then $\lambda_1(x), \lambda_2(x), \ldots, \lambda_K(x)$ are eigenvalues of $\mathrm{M}(x)$. By Perron–Frobenius result we know that $\lambda_1(x)$ is simple, real and $\lambda_1(x) > |\lambda_i(x)|$ for $i \geq 2$. (To simplify our further derivation, we also assume that $\lambda_i(x)$ are simple but this assumption will not have any significant impact on our asymptotics, as we shall see below.) Let $\mathrm{l}_i$ and $\mathrm{r}_i$, $i = 1, \ldots, K$ are left and right eigenvectors corresponding to $\lambda_1(x), \lambda_2(x), \ldots, \lambda_K(x)$ eigenvalues, respectively. We set $\langle \mathrm{l}_1, \mathrm{r}_1 \rangle = 1$ where $\langle \mathrm{x}, \mathrm{y} \rangle$ is the scalar product of the vectors x and y. Since $\mathrm{r}_i$ is orthogonal to the left eigenvector $\mathrm{r}_j$ for $j \neq i$, we can write for any vector x

$$\mathrm{x} = \langle \mathrm{l}_1, \mathrm{x} \rangle \mathrm{r}_1 + \sum_{i=2}^{K} \langle \mathrm{l}_i, \mathrm{x} \rangle \mathrm{r}_i.$$

This yields

$$\mathrm{M}(x)\mathrm{x} = \langle \mathrm{l}_1, \mathrm{x} \rangle \lambda_1(x)\mathrm{r}_1 + \sum_{i=2}^{K} \langle \mathrm{l}_i, \mathrm{x} \rangle \lambda_i(x)\mathrm{r}_i.$$

Since $\mathrm{M}^k(x)$ has eigenvalues $\lambda_1^k(x), \lambda_2^k(x), \ldots, \lambda_K^k(x)$, then — dropping even the assumption about eigenvalues $\lambda_2, \ldots, \lambda_K$ being simple — we arrive at

$$\mathrm{M}^k(x)\mathrm{x} = \langle \mathrm{l}_1, \mathrm{x} \rangle \mathrm{r}_1 \lambda_1^k(x) + \sum_{i=2}^{K'} q_i(k) \langle \mathrm{l}_i, \mathrm{x} \rangle \mathrm{r}_i \lambda_i^k(x) \tag{7.3.9}$$

where $q_i(k)$ is a polynomial in $k$ ($q_i(k) \equiv 1$ when the eigenvalues $\lambda_2, \ldots, \lambda_K$ are simple). Finally, we observe that we can analytically continue $\lambda_1(x)$ to complex plane due to separation of $\lambda_1(x)$ form other eigenvalues leading to $\lambda_1(z)$.

Applying now (7.3.9) to (7.3.8) and using it in the formula for $N(z,u)$ derived in Theorem 7.3.2 we obtain

$$\begin{aligned} N(z,u) &= \mathrm{R}^t(z)u[\mathrm{I} - u\mathrm{M}(z)]^{-1}\mathrm{U}(z) \\ &= u\mathrm{R}^t(z)\left(\sum_{k=0}^{\infty} u^k \lambda_1^k(z)\langle \mathrm{l}_1(z), \mathrm{U}(z)\rangle \mathrm{r}_1(z) \right. \\ &\quad \left. + \sum_{i=2}^{K'} u^k \lambda_i^k(z)\langle \mathrm{l}_i(z), \mathrm{U}(z)\rangle \mathrm{r}_i(z)\right) \\ &= \frac{uC_1(z)}{1-u\lambda_1(z)} + \sum_{i=2}^{K'} \frac{uC_i(z)}{1-u\lambda_i(z)} \end{aligned}$$

for some polynomials $C_i(z)$. This representation entails to apply the Cauchy formula yielding, as before, for $A < 1$ and a polynomial $B(u)$

$$N_n(u) := [z^n]N(z,u) = B(u)\rho^{-n}(u)(1 + O(A^n))$$

where $\rho(u)$ is the smallest root of $1-u\lambda(z) = 0$ which coincides with the smallest root of $\det(\mathrm{I} - u\mathrm{M}(u)) = 0$. In the above $A < 1$ since $\lambda_1(z)$ dominates all the other eigenvalues. In the next section we return to this method and discuss it in some more depth. ■

### 7.3.2. Analysis of the generalized string matching

In this section we deal with a general pattern matching problem where words in $\mathcal{W}$ are not of the same length, that is, $\mathcal{W} = \bigcup_{i=1}^{d} \mathcal{W}_i$ such that $\mathcal{W}_i$ is a subset of $\mathcal{A}^{m_i}$ with all $m_i$ being different. We still keep $\mathcal{W}_0 = \emptyset$ (i.e., there are no forbidden words). In the next section, we consider the case $\mathcal{W}_0 \neq \emptyset$. We present here a powerful method based on a finite automata (i.e., de Bruijn graph). This approach is very versatile, but unfortunately is not as insightful as the combinatorial approach discussed so far.

Our goal is to derive the probability generating function $N_n(u) = \mathbf{E}(u^{N_n(\mathcal{W})})$ of the number of pattern $\mathcal{W}$ occurrences in the text. We start with building an automaton that scans the text $X_1X_2\cdots X_n$ and recognizes occurrences of patterns from the set $\mathcal{W}$. As a matter of fact, our automaton is a de Bruijn graph that we describe in the sequel: Let $M = \max\{m_1, \ldots, m_d\} - 1$ and $\mathcal{B} = \mathcal{A}^M$. The de Bruijn automaton is built over the state space $\mathcal{B}$. Let $b \in \mathcal{B}$ and $a \in \mathcal{A}$. Then a transition from a state $b$ upon scanning symbol $a$ of the text is to $\hat{b} \in \mathcal{B}$ such that

$$\hat{b} = b_2b_3\cdots b_M a,$$

that is, the leftmost symbol of $b$ is erased and symbol $a$ is appended on the right. We shall denote such a transition as $ba \mapsto \hat{b}$ or $ba \in \mathcal{A}\hat{b}$ since the first symbol of $b$ has been deleted when scanning symbol $a$. When scanning a text of length $n-M$ one constructs an associated path of length $n-M$ in the de Bruijn automaton that begins at a state formed by the first $M$ symbols of the text, that is, $b = X_1X_2\cdots X_M$.

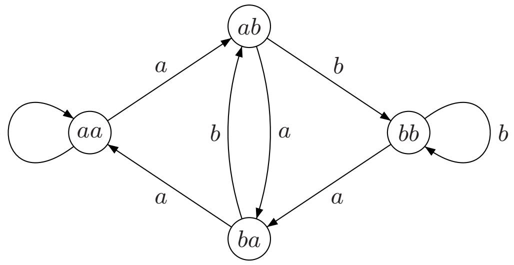



**Figure 7.2.** The de Bruijn graph for $\mathcal{W} = \{ab, aab, aba\}$.

To record the *number* of pattern occurrences we equip the automaton with a counter $\phi(b,a)$. When a transition occurs, we increment $\phi(b,a)$ by the number of occurrences of patterns from $\mathcal{W}$ in the text $ba$. Since all occurrences of patterns from $\mathcal{W}$ that end at $a$ are contained in the text of the form $ba$, we realize that

$$\phi(b,a) = N_{M+1}(\mathcal{W}, ba) - N_M(\mathcal{W}, b)$$

where $N_k(\mathcal{W}, x)$ is the number of pattern occurrences in the text $x$ of length $k$. Having built such an automaton, we construct a transition $V^M \times V^M$ matrix $\mathrm{T}(u)$ as a function of a complex variable $u$ and indexed by $\mathcal{B}\times\mathcal{B}$ such that

$$\begin{aligned}[\mathrm{T}(u)]_{b,\hat{b}} &:= P(a)u^{\phi(b,a)}[\![\, ba \in \mathcal{A}\hat{b}\,]\!] \\ &= P(a)u^{N_{M+1}(\mathcal{W},ba)-N_M(\mathcal{W},b)}[\![\,\hat{b} = b_2b_3\cdots b_M a\,]\!]\end{aligned} \tag{7.3.10}$$

where Iverson's bracket convention is used:

$$[\![B]\!] = \begin{cases} 1 \text{ if the property } B \text{ holds,} \\ 0 \text{ otherwise.}\end{cases}$$

Example 7.3.4. Let $\mathcal{W} = \{ab, aab, aba\}$. Then $M = 2$, the de Bruijn graph is presented in Figure 7.2, and the matrix $\mathrm{T}(u)$ is shown below

$$\mathrm{T}(u) = \begin{array}{c} \\ aa \\ ab \\ ba \\ bb \end{array}\!\!\begin{array}{c} \begin{array}{cccc} aa & ab & ba & bb \end{array} \\ \begin{pmatrix} P(a) & P(b)\,u & 0 & 0 \\ 0 & 0 & P(a)\,u^2 & P(b) \\ P(a) & P(b) & 0 & 0 \\ 0 & 0 & P(a) & P(b) \end{pmatrix}\end{array}.$$

Next, we extend the above construction to scan a text of length $k \ge M$. By combinatorial properties of matrix products, the entry of index $b, \hat{b}$ of the power $\mathrm{T}^k(u)$ cumulates all terms corresponding to starting in state $b$, ending in state $\hat{b}$, and recording the total number of occurrences of patterns $\mathcal{W}$ found upon scanning the last $k$ letters of the text. Therefore,

$$\left[\mathrm{T}^k(u)\right]_{b,\hat{b}} = \sum_{v\in\mathcal{A}^k} P(v)\, u^{N_{M+k}(\mathcal{W},bv)-N_M(\mathcal{W},b)}. \tag{7.3.11}$$

Define now a vector $\mathrm{x}(u)$ indexed by $b$ as

$$[\mathrm{x}(u)]_b = P(b)\, u^{N_M(\mathcal{W},b)}.$$

Then, the summation of all the entries of the row vector $\mathrm{x}(u)^{\mathbf{t}}\mathrm{T}^k(u)$ is achieved by means of the vector $\vec{1} = (1,\ldots,1)$ so that the quantity $\mathrm{x}(u)^{\mathbf{t}}\mathrm{T}(u)^k\vec{1}$ represents the probability generating function of $N_{k+M}(\mathcal{W})$ taken over all texts of length $M+k$. By setting $n = M+k$ we prove the following theorem.

Theorem 7.3.5. *Consider a general pattern* $\mathcal{W} = (\mathcal{W}_1,\ldots,\mathcal{W}_d)$ *with* $M = \max\{m_1,\ldots,m_d\}-1$. *Let* $\mathrm{T}(u)$ *be the transition matrix defined as*

$$[\mathrm{T}(u)]_{b,\hat{b}} := P(a)u^{N_{M+1}(\mathcal{W},ba)-N_M(\mathcal{W},b)}[\![\, \hat{b} = b_2b_3\cdots b_Ma \,]\!]$$

*where* $b,\hat{b}\in\mathcal{A}^M$ *and* $a\in\mathcal{A}$. *Then*

$$N_n(u) = \mathbf{E}(u^{N_n(\mathcal{W})}) = \mathrm{b}^t(u)\mathrm{T}^n(u)\vec{1} \tag{7.3.12}$$

*where* $\mathrm{b}^t(u) = \mathrm{x}^t(u)\mathrm{T}^{-M}(u)$. *Also,*

$$N(z,u) = \sum_{n\geq 0} N_n(z)z^n = \mathrm{b}^t(u)(\mathrm{I}-z\mathrm{T}(u))^{-1}\vec{1} \tag{7.3.13}$$

*for* $|z|<1$.

Let us now return for a moment to the reduced pattern case discussed in the previous section and compare expression (7.3.13) derived here with (7.3.3) of Theorem 7.3.2 that we repeat below

$$N(z,u) = \mathrm{R}^t(z)u(\mathrm{I}-u\mathrm{M}(z))^{-1}\mathrm{U}(z).$$

Although there is a striking resemblance of these formulas they are quite different. In (7.3.3) $\mathrm{M}(z)$ is a matrix of $z$ representing generating functions of languages $\mathcal{M}_{ij}$, while $\mathrm{T}(u)$ is a function of $u$ and it is the transition matrix of the associated de Bruijn graph. Nevertheless, the eigenvalue method discussed in the proof of Theorem 7.3.3 can be directly applied to derive limit laws of $N_n(\mathcal{W})$ for general set of patterns $\mathcal{W}$. We shall discuss it next.

To study asymptotics of $N_n(\mathcal{W})$ we need to estimate the growth of $T^n(u)$ which is governed by the growth of the largest eigenvalue, as we have already seen in the previous sections. Here, however, the situation is a little more complicated since the matrix $\mathrm{T}(u)$ is irreducible but not necessary primitive (cf. Chapter 1 for in depth discussion). To be more precise, $\mathrm{T}(u)$ is *irreducible* if its associated de Bruijn graph is strongly connected, while for *primitivity* of $\mathrm{T}(u)$ we require that the greatest common divisor of the cycle weights of the de Bruijn graph is equal to one.

Let us first verify irreducibility of $\mathrm{T}(u)$. As easy to check the matrix is irreducible since for any $g \geq M$ and $b, \hat{b} \in \mathcal{A}^M$ there are two words $w, v \in \mathcal{A}^g$ such that $bw = v\hat{b}$ (e.g., for $g = M$ one can take $w = \hat{b}$ and $v = b$). Thus $\mathrm{T}^g(u) > 0$ for $u > 0$ which is sufficient for irreducibility.

Let us now have a closer look at the primitivity of $\mathrm{T}(u)$. We start with a precise definition. Let $\psi(b, \hat{b}) := \phi(ba)$ where $ba \mapsto \hat{b}$ be the counter value when transitioned form $b$ to $\hat{b}$. Let also $\mathcal{C}$ be a cycle in the associated de Bruijn graph. Define the total weight of the cycle $\mathcal{C}$ as

$$\psi(\mathcal{C}) = \sum_{b, \hat{b} \in \mathcal{C}} \psi(b, \hat{b}).$$

Finally, we set $\psi_{\mathcal{W}} = \gcd(\psi(\mathcal{C}) : \ \mathcal{C} \text{ cycle})$. If $\psi_{\mathcal{W}} = 1$, then we say $\mathrm{T}(u)$ is primitive.

EXAMPLE 7.3.4 (*continued*). Consider again the matrix $\mathrm{T}(u)$ and its associated graph shown in Figure 7.2. There are six cycles of respective weights $0, 3, 2, 0, 0, 1$, therefore $\psi_{\mathcal{W}} = 1$ and $\mathrm{T}(u)$ is primitive.

Consider now another matrix

$$\mathrm{T}(u) = \begin{pmatrix} P(a) & P(b)\, u^4 \\ P(a)\, u^2 & P(b)\, u^3 \end{pmatrix}.$$

This time there are three cycles of weights $0, 6$ and $3$ and $\psi_{\mathcal{W}} = 3$. The matrix is not primitive. Observe that the characteristic polynomial $\lambda(u)$ of this matrix is a polynomial in $u^3$.

Observe that the diagonal elements of $\mathrm{T}(u)^k$ (i.e., its trace) are polynomials in $u^\ell$ if and only if $\ell$ divides $\psi_{\mathcal{W}}$; therefore, the characteristic polynomial $\det(z\mathrm{I} - \mathrm{T}(u))$ of $\mathrm{T}(u)$ is a polynomial in $u^{\psi_{\mathcal{W}}}$. Indeed, it is known that for any matrix $A$

$$\det(\mathrm{I} - A) = \exp\left( \sum_{k \geq 0} -\frac{\mathrm{Tr}[A^k]}{k} \right)$$

where $\mathrm{Tr}[A]$ is the trace of $A$.

Asymptotic behavior of the generating function $N_n(u) = \mathbf{E}(u^{N_n(\mathcal{W})})$, hence $N_n(\mathcal{W})$, depends on the growth of $\mathrm{T}^n(u)$. The next lemma summarizes some useful properties of $\mathrm{T}(u)$ and its eigenvalues. For the matrix $\mathrm{T}(u)$ of dimension $|\mathcal{A}|^M \times |\mathcal{A}|^M$ we denote by $\lambda_j(u)$ for $j = 1, \ldots, R = |\mathcal{A}^M|$ its eigenvalues and we assume that $|\lambda_1(u)| \geq |\lambda_2(u)| \geq \cdots \geq |\lambda_R(u)|$. To simplify notation, we often drop the index of the largest eigenvalue, that is, $\lambda(u) := \lambda_1(u)$. Observe that $\varrho(u) = |\lambda(u)|$ is known as the *spectral radius* and it is equal to

$$\varrho(u) = \lim_{n \to \infty} ||T^n(u)||^{1/n}$$

where $||\cdot||$ is any matrix norm.

LEMMA 7.3.6. *Let $\mathcal{G}_M(\mathcal{W})$ and $\mathrm{T}(u)$ denote, respectively, the de Bruijn graph and its associated matrix defined in (7.3.10) for general pattern $\mathcal{W}$. Assume $P(\mathcal{W}) > 0$.*

(i) *For $u > 0$ the matrix $\mathrm{T}(u)$ has a unique dominant eigenvalue $\lambda(u)$ ($> \lambda_j(u)$ for $j = 2, \dots, |\mathcal{A}|^M$) that is strictly positive and a dominant eigenvector $\mathrm{r}(u)$ whose all entries are strictly positive. Furthermore, there exists a complex neighborhood of the real positive axis on which the mappings $u \to \lambda(u)$ and $u \to \mathrm{r}(u)$ are well-defined and analytic.*

(ii) *Define $\Lambda(s) := \log \lambda(e^s)$ for $s$ complex. For real $s$ the function $s \to \Lambda(s)$ is strictly increasing and strictly convex. In addition,*

$$\Lambda(0) = 1, \qquad \Lambda'(0) = P(\mathcal{W}) > 0, \qquad \Lambda''(0) := \sigma^2(\mathcal{W}) > 0.$$

(iii) *For any $\theta \in (0, 2\pi)$ and $x$ real $\varrho(xe^{i\theta}) \le \varrho(x)$.*

(iv) *For any $\theta \in (0, 2\pi)$, if $\psi_{\mathcal{W}} = 1$, then for $x$ real $\varrho(xe^{i\theta}) < \varrho(x)$; otherwise $\psi_{\mathcal{W}} = d > 1$ and $\varrho(xe^{i\theta}) = \varrho(x)$ if and only if $\theta = 2k\pi/d$.*

*Proof.* We first prove (i). Take $u > 0$ real positive. Then the matrix $\mathrm{T}(u)$ has positive entries, and for any exponent $g \ge M$ the $g$th power of $\mathrm{T}(u)$ has strictly positive entries, as shown above (see irreducibility of $\mathrm{T}(u)$). Therefore, by the Perron–Frobenius theorem (cf. also Chapter 1) there exists an eigenvalue $\lambda(u)$ that dominates strictly all the others. Moreover, it is simple and strictly positive. In other words, one has

$$\lambda(u) := \lambda_1(u) > |\lambda_2(u)| \ge |\lambda_3(u)| \ge \cdots.$$

Furthermore, the corresponding eigenvector $r(u)$ has all its components strictly positive. Since the dominant eigenvalue is separated from other eigenvalues, by perturbation theory there exists a complex neighborhood of the real positive axis where the functions $u \to \lambda(u)$ and $u \to \mathrm{r}(u)$ are well-defined and analytic. Moreover, $\lambda(u)$ is an algebraic function since it satisfies the characteristic equation $\det(\lambda \mathrm{I} - \mathrm{T}(u)) = 0$.

We now prove part (ii). The increasing property for $\lambda(u)$ (and thus for $\Lambda(s)$) is a consequence of the fact that if $A$ and B are nonnegative irreducible matrices such that $A_{i,j} \ge B_{i,j}$ for all $(i, j)$, then the spectral radius of $A$ is larger than the spectral radius of B.

For convexity of $\Lambda(s)$, it is sufficient to prove that for $u, v > 0$

$$\lambda(\sqrt{uv}) \le \sqrt{\lambda(u)}\sqrt{\lambda(v)}.$$

Since eigenvectors are defined up to a constant, one can always choose the eigenvectors $\mathrm{r}(\sqrt{uv})$, $\mathrm{r}(u)$, and $\mathrm{r}(v)$ such that

$$\max_i \frac{r_i(\sqrt{uv})}{\sqrt{r_i(u)\, r_i(v)}} = 1.$$

Suppose that this maximum is attained at some index $i$. We denote by $P_{ij}$ the coefficient at $u$ in $\mathrm{T}(u)$, that is, $P_{ij} = [u^{\psi}][\mathrm{T}(u)]_{ij}$. By the Cauchy-Schwarz inequality we have

$$\begin{aligned}
\lambda(\sqrt{uv})r_i(\sqrt{uv}) &= \sum_j P_{ij}\,(\sqrt{uv})^{\psi(i,j)}\,r_j(\sqrt{uv})\\
&\leq \sum_j P_{ij}(\sqrt{uv})^{\psi(i,j)}\,\sqrt{r_j(u)\,r_j(v)}\\
&\leq \left(\sum_j P_{ij}\,u^{\psi(i,j)}\,r_j(u)\right)^{1/2}\left(\sum_j P_{ij}\,v^{\psi(i,j)}\,r_j(v)\right)^{1/2}\\
&= \sqrt{\lambda(u)}\sqrt{\lambda(v)}\,\sqrt{r_i(u)\,r_i(v)},
\end{aligned}$$

which implies convexity of $\Lambda(s)$. To show that $\Lambda(s)$ is strictly convex, we argue as follows: Observe that for $u = 1$ the matrix $\mathrm{T}(u)$ is stochastic, hence $\lambda(1) = 1$ and $\Lambda(0) = 0$. As we shall see below, the mean and the variance of $N_n(\mathcal{W})$ are equal asymptotically to $n\Lambda'(0)$ and $n\Lambda''(0)$, respectively. From the problem formulation, we conclude that $\Lambda'(0) = P(\mathcal{W}) > 0$ and $\Lambda''(0) = \sigma^2(\mathcal{W}) > 0$. Therefore, $\Lambda'(s)$ and $\Lambda''(s)$ cannot be always 0 and (since they are analytic) they cannot be zero on any interval. This implies that $\Lambda(s)$ is strictly increasing and strictly convex.

We now establish part (iii). For $|u| = 1$, and $x$ real positive, consider two matrices $\mathrm{T}(x)$ and $\mathrm{T}(xu)$. From (i) we know that for $\mathrm{T}(x)$ there exist a dominant strictly positive eigenvalue $\lambda := \lambda(x)$ and a dominant eigenvector $\mathrm{r} := \mathrm{r}(x)$ whose all entries $r_j$ are strictly positive. Consider an eigenvalue $\nu$ of $\mathrm{T}(xu)$ and its corresponding eigenvector $s := s(u)$. Denote by $v_j$ the ratio $s_j/r_j$. One can always choose r and $s$ such that $\max_{1\leq j\leq R}|v_j| = 1$. Suppose that this maximum is attained for some index $i$. Then

$$|\nu s_i| = |\sum_j P_{ij}\,(xu)^{\psi(i,j)}\,s_j| \leq \sum_j P_{ij}\,x^{\psi(i,j)}\,r_j = \lambda r_i. \tag{7.3.14}$$

We conclude that $|\nu| \leq \lambda$, and part (iii) is proven.

Finally we deal with part (iv). Suppose now that the equality $|\nu| = \lambda$ holds. Then, all the previous inequalities in (7.3.14) become equalities. First, for all indices $\ell$ such that $P_{i,\ell} \neq 0$, we deduce that $|s_\ell| = r_\ell$, and $v_\ell$ has modulus 1. For these indices $\ell$, we have the same equalities in (7.3.14) as for $i$. Finally, the transitivity of the de Bruijn graph entails that that each complex $v_j$ is of modulus 1. Now, the converse of the triangular inequality shows that for every edge $(i,j) \in \mathcal{G}_M(\mathcal{W})$ we have

$$u^{\psi(i,j)}v_j = \frac{\nu}{\lambda}v_i,$$

and for any cycle of length $L$ we conclude that

$$\left(\frac{\nu}{\lambda}\right)^L = u^{\psi(\mathcal{C})}.$$

However, for any pattern $\mathcal{W}$ there exists a cycle $\mathcal{C}$ of length one with weight $\psi(\mathcal{C}) = 0$, as easy to see. This proves that $\nu = \lambda$ and that $u^{\psi(\mathcal{C})} = 1$ for any cycle $\mathcal{C}$. If $\psi_{\mathcal{W}} = \gcd(\psi(\mathcal{C}),\ \mathcal{C} \text{ cycle}) = 1$, then $u = 1$ and $\varrho(xe^{i\theta}) < \varrho(x)$ for $\theta \in (0, 2\pi)$.

Suppose now that $\psi_{\mathcal{W}} = d > 1$. Then, the characteristic polynomial and the dominant eigenvalue $\lambda(v)$ are functions of $v^d$. The lemma is proved. ∎

Lemma 7.3.6 provides the main technical support to prove the forthcoming results; in particular, to establish asymptotic behavior of $\mathrm{T}^n(u)$ for large $n$. Indeed, our starting point is (7.3.13) to which we apply the spectral decomposition as in (7.3.9) to conclude that

$$N(z,u) = \frac{c(u)}{1 - z\lambda(u)} + \sum_{i \geq 2} \frac{c_i(u)}{(1 - z\lambda_i(u))^{\alpha_i}}.$$

where $\alpha_i \geq 1$ are some integers. In the above, $\lambda(u)$ is the dominant eigenvalue, while $\lambda_i(u) < \lambda(u)$ are other eigenvalues. The numerator has the expression $c(u) = \mathrm{b}^t(u)\langle \mathrm{l}(u), \vec{1}\rangle \mathrm{r}(u)$ where $\mathrm{l}(u)$ and $\mathrm{r}(u)$ are the left and the right dominant eigenvectors and $\mathrm{b}^t(u)$ is defined after (7.3.12). Then Cauchy's coefficient formula implies

$$N_n(u) = c(u)\lambda^n(u)(1 + O(A^n)) \tag{7.3.15}$$

for some $A < 1$. Equivalently, the moment generating function for $N_n(\mathcal{W})$ is given by the following uniform approximation in a neighborhood of $s = 0$

$$\mathbf{E}(e^{sN_n(\mathcal{W})}) = d(s)\lambda^n(e^s)(1 + O(A^n)) = d(s)\exp\left(n\Lambda(s)\right)\left(1 + O(A^n)\right) \tag{7.3.16}$$

where $d(s) = c(e^s)$ and $\Lambda(s) = \log \lambda(e^s)$.

There is another, more general, derivation of (7.3.15). Observe that the spectral decomposition of $\mathrm{T}(u)$ when $u$ lies in a sufficiently small complex neighborhood of any compact subinterval of $(0, +\infty)$ is of the form

$$\mathrm{T}(u) = \lambda(u)\mathrm{Q}(u) + \mathrm{R}(u) \tag{7.3.17}$$

where $\mathrm{Q}(u)$ is the projection under the dominant eigensubspace and $\mathrm{R}(u)$ a matrix whose spectral radius equals $|\lambda_2(u)|$. Therefore,

$$\mathrm{T}(u)^n = \lambda(u)^n Q(u) + R(u)^n,$$

entails the estimate (7.3.15). The next results follows immediately from (7.3.16).

THEOREM 7.3.7. *Let $\mathcal{W} = (\mathcal{W}_0, \mathcal{W}_1, \ldots, \mathcal{W}_d)$ be a generalized pattern with $\mathcal{W}_0 = \emptyset$ generated by a memoryless source. For large $n$*

$$\mathbf{E}(N_n(\mathcal{W})) = n\Lambda'(0) + O(1) = nP(\mathcal{W}) + O(1), \tag{7.3.18}$$

$$\mathrm{Var}(N_n(\mathcal{W})) = n\Lambda''(0) + O(1) = n\sigma^2(\mathcal{W}) + O(1) \tag{7.3.19}$$

*where $\Lambda(s) = \log \lambda(e^s)$ and $\lambda(u)$ is the largest eigenvalue of $\mathrm{T}(u)$. Furthermore,*

$$\mathbf{P}(N_n(\mathcal{W}) = 0) = C\lambda^n(0)(1 + O(A^n))$$

*where $C > 0$ is a constant and $A < 1$.*

Now we establish limit laws, starting with the central limit law and its local limit law.

THEOREM 7.3.8. *Under the same assumption as for Theorem 7.3.7, the following holds*

$$\sup_{x \in B} \left| \mathbf{P}\left( \frac{N_n(\mathcal{W}) - nP(\mathcal{W})}{\sigma(\mathcal{W})\sqrt{n}} \le x \right) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \right| = O\left(\frac{1}{\sqrt{n}}\right) \quad (7.3.20)$$

*where $B$ is a bounded real interval.*

*Proof.* The uniform asymptotic expansion (7.3.16) of a sequence of moment generating functions is known as a "quasi-powers approximation". Then an application of the classical Levy continuity theorem leads to the Gaussian limit law. An application of the Berry-Essen inequality provides the speed of convergence which is $O(1/\sqrt{n})$. This proves the theorem. ∎

Finally, we deal with the large deviations.

THEOREM 7.3.9. *Under the same assumption as before, Let $\omega_a$ be a solution of*

$$\omega\lambda'(\omega) = a\lambda(\omega)$$

*for some $a \neq P(\mathcal{W})$, where $\lambda(u)$ is the largest eigenvalue of $\mathrm{T}(u)$. Define*

$$I(a) = a \log \omega_a - \log \lambda(\omega_a). \quad (7.3.21)$$

*Then there exists a constant $C > 0$ such that $I(a) > 0$ for $a \neq P(\mathcal{W})$ and*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(N_n(\mathcal{W}) \le an\right) = -I(x) \qquad \text{if } \ 0 < x < P(\mathcal{W}) \quad (7.3.22)$$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(N_n(\mathcal{W}) \ge na\right) = -I(x) \qquad \text{if } \ P(\mathcal{W}) < x < C. \quad (7.3.23)$$

*Proof.* We consider now large deviations and establish (7.3.22). The variable $N_n(\mathcal{W})$ is by definition of at most linear growth, and there exists a a constant $C$ such that $N_n(\mathcal{W}) \le Cn + O(1)$. Let $0 < x < P(\mathcal{W})$. Cauchy's coefficient formula provides

$$\mathbf{P}\left(N_n(\mathcal{W}) \le k\right) = \frac{1}{2i\pi} \int_{|u|=r} \frac{N_n(u)}{u^k} \frac{du}{u(1-u)}.$$

For ease of exposition, we first discuss the case of a primitive pattern. We recall that a pattern is primitive if $\psi_{\mathcal{W}} = \gcd(\psi(\mathcal{C}),\ \mathcal{C} \text{ cycle}) = 1$. The strict domination property expressed in Lemma 7.3.6(iv) for primitive patterns implies that the above integrand is strictly maximal at the intersection of the circle $|u| = r$ and the positive real axis. Near the positive real axis, where the contribution of the integrand is concentrated, the following uniform approximation holds, with $k = na$:

$$\frac{N_n(u)}{u^k} = \exp\left(n\left(\log\lambda(u) - a\log u\right)\right)(1 + o(1)) \quad (7.3.24)$$

The saddle point equation is then obtained by cancelling the first derivative yielding

$$F(\omega) := \frac{\omega\lambda'(\omega)}{\lambda(\omega)} = a. \tag{7.3.25}$$

Note that the function $F$ is exactly the derivative of $\Lambda(s)$ at point $s := \log\omega$. Since $\Lambda(s)$ is strictly convex, the left side is an increasing function of its argument as proved in Lemma 7.3.6(ii). Also, we know form this lemma that the value $F(0) = 0$, $F(1) = P(\mathcal{W})$ while we set $F(\infty) = C$. Thus, for any real $a$ in $(0, C)$, equation (7.3.25) always admits a unique positive solution that we denote by $\omega \equiv \omega_a$. Moreover, for $a \neq P(\mathcal{W})$, one has $\omega_a \neq 1$. Since the function

$$u \to -\log\frac{\lambda(u)}{u^a}$$

admits a strict maximum at $u = \omega_a$, hence this maximum $I(a)$ is strictly positive. Finally, the usual saddle point approximation applies and one finds

$$\mathbf{P}\left(\frac{N_n(\mathcal{W})}{n} \le a\right) = \left(\frac{\lambda(\omega_a)}{\omega_a^a}\right)^n \Theta(n),$$

where $\Theta(n)$ is of the order of $n^{-1/2}$. In summary, the large deviation rate is

$$I(a) = -\log\frac{\lambda(\omega_a)}{\omega_a^a} \qquad \text{with} \qquad \frac{\omega_a\lambda'(\omega_a)}{\lambda(\omega_a)} = a.$$

as shown in the theorem.

In the general case when the pattern is not primitive, the strict inequality of Lemma 7.3.6(iv) is not satisfied, and several saddle points may be present on the circle $|u| = r$, which will lead to some oscillations. We must, in this case, use the weaker inequality of Lemma 7.3.6, namely, $\varrho(xe^{i\theta}) \le \varrho(x)$, which replaces the strict inequality. However, the factor $(1-u)^{-1}$ present in the integrand of (7.3.24) attains its maximum modulus on $|u| = r$ solely at $u = r$. Thus, the contribution of possible saddle points can only affect a fraction of the contribution from $u = r$. Consequently, (7.3.22) and (7.3.21) continue to be valid. A similar reasoning provides the right tail estimate, with $I(a)$ still given by (7.3.21). This completes the proof of (7.3.22). ∎

We complete this analysis with a local limit law.

Theorem 7.3.10. *If* $\mathrm{T}(u)$ *is primitive, then*

$$\sup_{x\in B}\left|\mathbf{P}\left(N_n = nP(\mathcal{W}) + x\sigma(\mathcal{W})\sqrt{n}\right) - \frac{1}{\sigma(\mathcal{W})\sqrt{n}}\frac{e^{x^2/2}}{\sqrt{2\pi}}\right| = o\left(\frac{1}{\sqrt{n}}\right) \tag{7.3.26}$$

*where $B$ is a bounded real interval. Furthermore, under the above additional assumption, one can find constants $\sigma_a$ and $\delta_a$ such that*

$$\mathbf{P}(N_n(\mathcal{W}) = a\mathbf{E}(N_n)) \sim \frac{1}{\sigma_a\sqrt{2\pi n}}e^{-nI(a)+\theta_a} \tag{7.3.27}$$

*where $I(a)$ is defined in (7.3.21) above.*

*Proof* Stronger "regularity conditions" are needed in order to obtain local limit estimates. Roughly, one wants to exclude the possibility that the discrete distribution is of a lattice type, being supported by a nontrivial sublattice of the integers. (For instance, we need to exclude the possibility for $N_n(\mathcal{W})$ to be always odd, or of the parity of $n$, and so on.) Observe first that positivity and irreducibility of the matrix $\mathrm{T}(u)$ is not enough as shown in Example 7.3.4.

By Lemma 7.3.6, one can estimate the probability distribution of $N_n(\mathcal{W})$ by the classical saddle point method in the case when $\mathcal{W}$ is primitive. Again, one starts from Cauchy's coefficient integral,

$$\mathbf{P}(N_n(\mathcal{W}) = k) = \frac{1}{2i\pi} \int_{|u|=1} N_n(u) \, \frac{du}{u^{k+1}}, \tag{7.3.28}$$

where $k$ is of the form $k = nP(\mathcal{W})n + x\sigma(\mathcal{W})\sqrt{n}$. Property (iv) of Lemma 7.3.6 grants us precisely the fact that any closed arc of the unit circle not containing $u = 1$ brings an exponentially negligible contribution. A standard application of the saddle point technique does the job. In this way, the proof of the local limit law of Theorem 7.3.10 is completed. Finally, the precise large deviations follows from the local limit result and an application of the method of shift discussed in the proof of Theorem 7.2.12. ■

### 7.3.3. Forbidden words and $(\ell, k)$ sequences

Finally, consider the general pattern $\mathcal{W} = (\mathcal{W}_0, \mathcal{W}_1, \ldots, \mathcal{W}_d)$ with nonempty forbidden set $\mathcal{W}_0$. In this case, we study the number of occurrences $N_n(\mathcal{W}|\mathcal{W}_0 = 0)$ of patterns $\mathcal{W}_1, \ldots \mathcal{W}_d$ under the condition that there is no occurrence in the text of any pattern from $\mathcal{W}_0$.

Fortunately, we can recover almost all results from our previous analysis after re-defining the matrix $\mathrm{T}(u)$ and its de Bruijn graph. We now change (7.3.10) to

$$[\mathrm{T}(u)]_{b,\hat{b}} := P(a)u^{\phi(b,a)} [\![\, ba \in \mathcal{A}\hat{b} \text{ and } ba \not\subset \mathcal{W}_0 ]\!] \tag{7.3.29}$$

where $ba \subset \mathcal{W}_0$ means that any subword of $ba$ belongs to $\mathcal{W}_0$. In words, we force the matrix $\mathrm{T}(u)$ to be zero at any position that leads to a word containing patterns from $\mathcal{W}_0$, that is, we eliminate from the de Bruijn graph any transition that contains a forbidden word. Having matrix $\mathrm{T}(u)$ constructed, we can repeat all previous results except that it is much harder to find explicit formulas even for the mean and the variance (cf. Exercise 7.3.4)

Finally, we consider a degenerated general pattern in which $\mathcal{W}_i = \emptyset$ for all $i = 1, \ldots, d$ except nonempty $\mathcal{W}_0$. In this case, we count the number of sequences that do not contain a pattern from $\mathcal{W}_0$. We only consider the special case of this problem, that of $(\ell, k)$ sequences for which $\mathcal{W}_0$ is defined in (7.3.1). In particular, we compute the so called *Shannon capacity* $C_{\ell,k}$ defined as

$$C_{\ell,k} = \lim_{n\to\infty} \frac{\log(\text{number of } (\ell,k) \text{ sequence of length } n)}{n}.$$

We first compute the ordinary generating function $T_{\ell,k}(z) = \sum_{w \in \mathcal{T}_{\ell,k}} z^{|w|}$ of all $(\ell, k)$ words denoted as $\mathcal{T}_{\ell,k}$. To enumerate $\mathcal{T}_{\ell,k}$ we define $\mathcal{D}_{\ell,k}$ as the set of all words consisting only of runs of 0's whose length is between $\ell$ and $k$. The generating function $D(z)$ is clearly equal to

$$D(z) = z^\ell + z^{\ell+1} + \cdots + z^k = z^\ell \frac{1 - z^{k-\ell+1}}{1-z}.$$

We now observe that $\mathcal{T}_{\ell,k}$ can be symbolically written as

$$\mathcal{T}_{\ell,k} = \mathcal{D}_{\ell,k} \left( \{1\} \times \varepsilon + \bar{\mathcal{D}}_{\ell,k} + \bar{\mathcal{D}}_{\ell,k} \times \bar{\mathcal{D}}_{\ell,k} + \cdots + \bar{\mathcal{D}}_{\ell,k}^k + \cdots \right), \tag{7.3.30}$$

where $\bar{\mathcal{D}}_{\ell,k} = \{1\} \times \mathcal{D}_{\ell,k}$. Above basically says that the collection of $(\ell, k)$ sequences, $\mathcal{T}_{\ell,k}$, is a concatenation of $\{1\} \times \mathcal{D}_{\ell,k}$. Thus (7.3.30) translates into the generating functions $T_{\ell,k}(z)$ as follows

$$\begin{aligned} T_{\ell,k}(z) = D(z) \frac{1}{1 - zD(z)} &= \frac{z^\ell (1 - z^{k+1-\ell})}{1 - z - z^{\ell+1} + z^{k+2}} \\ &= \frac{z^\ell + z^{\ell+1} + \cdots + z^k}{1 - z^{\ell+1} - z^{\ell+2} - \cdots - z^{k+1}}. \end{aligned} \tag{7.3.31}$$

Then Shannon capacity $C_{\ell,k}$ is

$$C_{\ell,k} = \lim_{n\to\infty} \frac{\log [z^n] T_{\ell,k}(z)}{n}.$$

If $\rho$ is the smallest root in absolute value of $1 - z^{\ell+1} - z^{\ell+2} - \cdots - z^{k+1} = 0$, then clearly

$$C_{\ell,k} = -\log \rho.$$

Example 7.3.11. In this example, we show that one can enumerate more precisely $(\ell, k)$ sequences. In fact, since the function $T_{\ell,k}(z)$ is rational we can compute $[z^n]T_{\ell,k}(z)$ exactly. Let us consider a particular case, namely, $\ell = 1$ and $k = 3$. Then the denumerator in (7.3.31) becomes $1 - z^2 - z^3 - z^4$, and its roots are

$$\rho_{-1} = -1, \quad \rho_0 = 0.682327\ldots, \quad \rho_1 = -0.341164\ldots + i1.161541\ldots, \quad \rho_2 = \bar{\rho}_1.$$

Computing residues we obtain

$$\begin{aligned} [z^n]T_{1,3}(z) = &\frac{\rho_0 + \rho_0^2 + \rho_0^3}{(\rho_1 + 1)(\rho_0 - \rho_1)(\rho_0 - \bar{\rho}_1)} \rho_0^{-n-1} \\ &+ (-1)^{n+1} \frac{1}{(\rho_0 + 1)(\rho_1 + 1)(\bar{\rho}_1 + 1)} + O(r^{-n}), \end{aligned}$$

where $r \approx 0.68$. More specifically,

$$[z^n]T_{1,3}(z) = 0.594(1.465)^{n+1} + 0.189(-1)^{n+1} + O(0.68^n).$$

for large $n$.

## 7.4. Subsequence pattern matching

In string matching problem, given a pattern $\mathcal{W}$ one searches for some/all occurrences of $\mathcal{W}$ as a block of consecutive symbols in a text. We analyzed various string matching problems in the previous sections. Here we concentrate on *subsequence pattern matching.* In this case we search for a given pattern $\mathcal{W} = w_1 w_2 \dots w_m$ in the text $X = x_1 x_2 \dots x_n$ as a *subsequence*, that is, we look for indices $1 \le i_1 < i_2 < \cdots < i_m \le n$ such that $x_{i_1} = w_1$, $x_{i_2} = w_2$, $\cdots$, $x_{i_m} = w_m$. We also say that the word $\mathcal{W}$ is *"hidden"* in the text; thus we call this the *hidden pattern* problem. For example, `date` occurs as a subsequence in the text `hidden pattern`, in fact four times, but not even once as a string.

More specifically, we allow the possibility of imposing an additional set of constraints $\mathcal{D}$ on the indices $i_1, i_2, \dots, i_m$ to record a valid subsequence occurrence. For a given family of integers $d_j$ ($d_j \ge 1$, possibly $d_j = \infty$), one should have $(i_{j+1} - i_j) \le d_j$. More formally, the hidden pattern specification is determined by a pair $(\mathcal{W}, \mathcal{D})$ where $\mathcal{W} = w_1 \cdots w_m$ is a word of length $m$ and the *constraint* $\mathcal{D} = (d_1, \dots, d_{m-1})$ is an element of $(\mathbf{N}^+ \cup \{\infty\})^{m-1}$.

Example 7.4.1. With # representing a 'don't-care-symbol' and the subscript denoting a strict upper bound on the length of the associated gap, a typical pattern may look like

$$\mathtt{ab}\#_2\mathtt{r}\#\mathtt{ac}\#\mathtt{a}\#\mathtt{d}\#_4\mathtt{a}\#\mathtt{br}\#\mathtt{a} \tag{7.4.1}$$

where $\# = \#_\infty$ and $\#_1$ is omitted; That is '`ab`' should occur first contiguously, followed by '`r`' with a gap of $< 2$ symbols, followed anywhere later in the text by '`ac`', etc.

The case when all the $d_j$'s are infinite is called the (fully) *unconstrained problem.* When all the $d_j$'s are finite, then we speak of the (fully) *constrained* problem. In particular, the case where all $d_j$ are equal to one reduces to the exact string matching problem. Furthermore, observe that when all $d_j < \infty$ (fully constrained pattern), the problem can be treated as the generalized string matching discussed in Section 7.3. In this case, the general pattern $\mathcal{W}$ is a set consisting of all words satisfying the constraint $\mathcal{D}$. However, if at least one $d_j$ is infinite, then the techniques discussed so far are not well suited to handle it. Therefore, in this section, we develop new methods that make the analysis possible.

If an $m$-tuple $I = (i_1, i_2, \dots, i_m)$ $(1 \le i_1 < i_2 < \cdots < i_m)$ satisfies the constraint $\mathcal{D}$ with $i_{j+1} - i_j \le d_j$, then it is called a *position tuple.* Let $\mathcal{P}_n(\mathcal{D})$ be the set of all positions subject to the separation constraint $\mathcal{D}$, satisfying furthermore $i_m \le n$. Let also $\mathcal{P}(\mathcal{D}) = \bigcup_n \mathcal{P}_n(\mathcal{D})$. An *occurrence* of pattern $\mathcal{W}$ subject to the constraint $\mathcal{D}$ is a pair $(I, X)$ formed with a position $I = (i_1, i_2, \dots, i_m)$ of $\mathcal{P}_n(\mathcal{D})$ and a text $X = x_1 x_2 \cdots x_n$ for which $x_{i_1} = w_1, x_{i_2} = w_2, \dots, x_{i_m} = w_m$. Thus, what we call an occurrence is a text augmented with the distinguished positions at which the pattern occurs. The number $\Omega$ of occurrences of pattern

$\mathcal{W}$ in text $X$ as a subsequence subject to the constraint $\mathcal{D}$ is then a sum of characteristic variables

$$\Omega(X) = \sum_{I \in \mathcal{P}_{|X|}(\mathcal{D})} Z_I(X), \tag{7.4.2}$$

where $Z_I(X) := [\![\mathcal{W}$ occurs at position $I$ in $X]\!]$. When the text $X$ is of length $n$, then we often write $\Omega_n := \Omega(X)$.

In order to proceed we need to introduce important notion of *blocks* and *aggregates*. In the general case, we assume that the subset $\mathcal{F}$ of indices $j$ for which $d_j$ is finite ($d_j < \infty$) has cardinality $m - b$ with $1 \le b \le m$. The two extreme values of $b$, namely, $b = m$ and $b = 1$, describe the (fully) unconstrained and the (fully) constrained problem, respectively. Thus, the subset $\mathcal{U}$ of indices $j$ for which $d_j$ is unbounded ($d_j = \infty$) has cardinality $b - 1$. It then separates the pattern $\mathcal{W}$ into $b$ independent subpatterns that are called the *blocks* and are denoted by $\mathcal{W}_1, \mathcal{W}_2, \ldots \mathcal{W}_b$. All the possible $d_j$ "inside" any $\mathcal{W}_r$ are finite and form the subconstraint $\mathcal{D}_r$, so that a general hidden pattern specification $(\mathcal{W}, \mathcal{D})$ is equivalently described as a $b$-tuple of fully constrained hidden patterns $((\mathcal{W}_1, \mathcal{D}_1), (\mathcal{W}_2, \mathcal{D}_2), \ldots, (\mathcal{W}_b, \mathcal{D}_b))$.

Example 7.4.1 (*continued*). Consider again

$$\mathtt{ab}\#_2\mathtt{r}\#\mathtt{ac}\#\mathtt{a}\#\mathtt{d}\#_4\mathtt{a}\#\mathtt{br}\#\mathtt{a},$$

in which one has $b = 6$, the six blocks being

$$\mathcal{W}_1 = \mathtt{a}\#_1\mathtt{b}\#_2\mathtt{r},\ \mathcal{W}_2 = \mathtt{a}\#_1\mathtt{c},\ \ \mathcal{W}_3 = \mathtt{a},\ \mathcal{W}_4 = \mathtt{d}\#_4\mathtt{a},\ \ \mathcal{W}_5 = \mathtt{b}\#_1\mathtt{r},\ \ \mathcal{W}_6 = \mathtt{a}.$$

In the same way, an occurrence position $I = (i_1, i_2, \ldots, i_m)$ of $\mathcal{W}$ subject to constraint $\mathcal{D}$ gives rise to $b$ suboccurrences, $I^{[1]}, I^{[2]}, \ldots I^{[b]}$, the $r$th term $I^{[r]}$ representing an occurrence of $\mathcal{W}_r$ subject to constraint $\mathcal{D}_r$. The $r$th *block* $B^{[r]}$ is the closed segment whose end points are the extremal elements of $I^{[r]}$, and the *aggregate* of position $I$, denoted by $\alpha(I)$, is the collection of these $b$ blocks.

Example 7.4.1 (*continued*). Taking the pattern of Example 7.4.1, the position tuple

$$I = (6, 7, 9, 18, 19, 22, 30, 33, 50, 51, 60)$$

satisfies the constraint $\mathcal{D}$ and gives rise to six subpositions,

$$\overbrace{(6,7,9)}^{I^{[1]}},\quad \overbrace{(18,19)}^{I^{[2]}},\quad \overbrace{(22)}^{I^{[3]}},\quad \overbrace{(30,33)}^{I^{[4]}},\quad \overbrace{(50,51)}^{I^{[5]}},\quad \overbrace{(60)}^{I^{[6]}}\ ;$$

accordingly, the resulting aggregate $\alpha(I)$,

$$\overbrace{[6,9]}^{B^{[1]}},\quad \overbrace{[18,19]}^{B^{[2]}},\quad \overbrace{[22]}^{B^{[3]}},\quad \overbrace{[30,33]}^{B^{[4]}},\quad \overbrace{[50,51]}^{B^{[5]}},\quad \overbrace{[60]}^{B^{[6]}}\ ,$$

is formed with six blocks.

### 7.4.1. Mean and variance analysis

Hereafter, we assume that $\mathcal{W}$ is given and the text $X$ is generated by a (non-degenerate) *memoryless* source. The first moment of the number of occurrences, $\Omega(X)$, is easily obtained by describing the collection of all occurrences in terms of formal languages, as already discussed in previous sections. We consider the collection of position-text pairs

$$\mathcal{O} := \{(I, X) \ \ ; \ \ I \in \mathcal{P}_{|X|}(\mathcal{D})\},$$

with the size of an element being by definition the length $n$ of the text $X$. The weight of an element of $\mathcal{O}$ is taken to be equal to $Z_I(X)P(X)$, where $P(X)$ is the probability of the text. In this way, $\mathcal{O}$ can also be regarded as the collection of all occurrences weighted by probabilities of the text. The corresponding generating function of $\mathcal{O}$ equipped with this weight is

$$O(z) = \sum_{(I,X)\in\mathcal{O}} Z_I(X)P(X)\, z^{|X|} = \sum_X \left( \sum_{I\in\mathcal{P}_{|X|}(\mathcal{D})} Z_I(X) \right) P(X) z^{|X|}, \quad (7.4.3)$$

and, with the definition of $\Omega$,

$$O(z) = \sum_X \Omega(X)P(X)\, z^{|X|} \ = \sum_n \mathbf{E}(\Omega_n) z^n. \quad (7.4.4)$$

As a consequence, one has $[z^n]O(z) = \mathbf{E}(\Omega_n)$, so that $O(z)$ serves as the generating function of the sequence of expectations $\mathbf{E}(\Omega_n)$.

On the other hand, each occurrence can be viewed as a "context" with an initial string, then the first letter of the pattern, then a separating string, then the second letter, etc. The collection $\mathcal{O}$ is therefore described combinatorially by

$$\mathcal{O} \ = \mathcal{A}^\star \times \{w_1\} \times \mathcal{A}^{<d_1} \times \{w_2\} \times \mathcal{A}^{<d_2} \times \ldots \times \{w_{m-1}\} \times \mathcal{A}^{<d_{m-1}} \times \{w_m\} \times \mathcal{A}^\star. \quad (7.4.5)$$

There, for $d < \infty$, $\mathcal{A}^{<d}$ denotes the collection of all words of length strictly less $d$, i.e., $\mathcal{A}^{<d} := \bigcup_{i<d} \mathcal{A}^i$, whereas, for $d = \infty$, $\mathcal{A}^{<\infty}$ denotes the collection of all finite words, i.e., $\mathcal{A}^{<\infty} := \mathcal{A}^\star = \bigcup_{i<\infty} \mathcal{A}^i$. Since the source is memoryless, the rules discussed at the end of the last section can be applied, and they give access to $O(z)$ from the description (7.4.5). The generating function functions associated to $\mathcal{A}^{<d}$ and $\mathcal{A}^{<\infty}$ are

$$A^{<d}(z) = 1+z+z^2+\cdots+z^{d-1} = \frac{1-z^d}{1-z}, \qquad A^{<\infty}(z) = 1+z+z^2+\cdots = \frac{1}{1-z}.$$

Thus, the description (7.4.5) of occurrences automatically translates into

$$O(z) \equiv \sum_{n\geq 0} \mathrm{E}[\Omega_n]\ z^n = \left(\frac{1}{1-z}\right)^{b+1} \times \left(\prod_{i=1}^m p_{w_i} z\right) \times \left(\prod_{i\in\mathcal{F}} \frac{1-z^{d_i}}{1-z}\right). \quad (7.4.6)$$

One finally finds

$$\mathbf{E}(\Omega_n) = [z^n]O(z) = \frac{n^b}{b!}\left(\prod_{i\in\mathcal{F}} d_i\right) P(\mathcal{W})\left(1+O\left(\frac{1}{n}\right)\right), \tag{7.4.7}$$

and a complete asymptotic expansion could be easily obtained.

For the analysis of variance and especially of higher moments, it is essential to work with a centered random variable $\Xi$ defined, for each $n$, as

$$\Xi_n := \Omega_n - \mathbf{E}(\Omega_n) = \sum_{I\in\mathcal{P}_n(\mathcal{D})} Y_I, \tag{7.4.8}$$

where $Y_I := Z_I - \mathbf{E}(Z_I) = Z_I - P(\mathcal{W})$. The second moment of the centered variable $\Xi$ equals the variance of $\Omega_n$ and with the centered variables defined above by (7.4.8), one has

$$\mathbf{E}(\Xi_n^2) = \sum_{I,J\in\mathcal{P}_n(\mathcal{D})} \mathbf{E}(Y_I Y_J). \tag{7.4.9}$$

From this last equation, we need to analyze *pairs of positions* $(I, X), (J, X) = (I, J, X)$ relative to a common text $X$. We denote by $\mathcal{O}_2$ this set, that is,

$$\mathcal{O}_2 := \{(I, J, X) \ \ ; \ \ I, J \in \mathcal{P}_{|X|}(\mathcal{D})\},$$

and we weight each element $(I, J, X)$ by $Y_I(X)Y_J(X)P(X)$. The corresponding generating function, which enumerates pairs of occurrences, is

$$\begin{aligned} O_2(z) &:= \sum_{(I,J,X)\in\mathcal{O}_2} Y_I(X)Y_J(X)P(X)\, z^{|X|} \\ &= \sum_X \left(\sum_{I,J\in\mathcal{P}_{|X|}(\mathcal{D})} Y_I(X)Y_J(X)\right) P(X)z^{|X|} \end{aligned}$$

and, with (7.4.9),

$$O_2(z) = \sum_{n\geq 0}\ \sum_{I,J\in\mathcal{P}_n(\mathcal{D})} \mathbf{E}(Y_I Y_J)\, z^n = \sum_{n\geq 0} \mathbf{E}(\Xi_n^2)\, z^n.$$

The process entirely parallels the derivation of (7.4.3) and (7.4.4), and, one has $[z^n]O_2(z) = \mathbf{E}(\Xi_n^2)$, so that $O_2(z)$ serves as the generating function (in the usual sense) of the sequence of moments $\mathbf{E}(\Xi_n^2)$.

There are two kinds of pairs $(I, J)$ depending whether they intersect or not. When $I$ and $J$ do not intersect, the corresponding random variables $Y_I$ and $Y_J$ are independent, and the corresponding covariance $E[Y_I Y_J]$ reduces to 0. As a consequence, one may restrict attention to pairs of occurrences $I, J$ that intersect at one place at least. Suppose that there exist two occurrences of pattern $\mathcal{W}$ at positions $I$ and $J$ which intersect at $\ell$ distinct places. We then

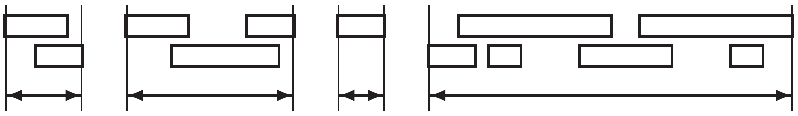


**Figure 7.3.** A pair of position tuples $I, J$ with $b = 6$ blocks each and the joint aggregates; the number of degrees of freedom is here $\beta(I, J) = 4$.

denote by $\mathcal{W}_{I\cap J}$ the subpattern of $\mathcal{W}$ that occurs at position $I \cap J$, and by $P(\mathcal{W}_{I\cap J})$ the probability of this subpattern. Since the expectation $\mathbf{E}(Z_I Z_J)$ equals $P(\mathcal{W})^2/P(\mathcal{W}_{I\cap J})$ provided that $\mathcal{W}$ agrees on every position of $I \cap J$, the expectation $\mathbf{E}(Y_I Y_J) = P(\mathcal{W})^2 e(I, J)$ involves a *correlation number* $e(I, J)$

$$e(I, J) = \frac{[\![\mathcal{W} \text{ agree } I \cap J]\!]}{P(\mathcal{W}_{I\cap J})} - 1. \tag{7.4.10}$$

Remark that this relation remains true even if the pair $(I, J)$ is not intersecting, since, in this case, one has $P(\mathcal{W}_{I\cap J}) = P(\varepsilon) = 1$.

The asymptotic behavior of variance is driven by the overlapping of blocks involved in $I$ and $J$, rather than plainly by the cardinality of $I \cap J$. In order to formalize this, define first the (joint) *aggregate* $\alpha(I, J)$ to be the system of blocks obtained by merging together all intersecting blocks of the two aggregates $\alpha(I)$ and $\alpha(J)$. The number of blocks $\beta(I, J)$ of $\alpha(I, J)$ plays a fundamental rôle here, since it measures the *degree of freedom* of pairs; we also call $\beta(I, J)$ the degree of pair $(I, J)$. Figure 7.3 illustrates graphically this notion.

EXAMPLE 7.4.2. Consider the pattern $\mathcal{W} = \boxed{\mathtt{a}\#_3\mathtt{b}\#_4\mathtt{r}}\,\#\,\boxed{\mathtt{a}\#_4\mathtt{c}}$ composed of two blocks. Then the text `aarbarbccaracc` contains several valid occurrences of $\mathcal{W}$ including two at positions $I = (2, 4, 6, 10, 13)$ and $J = (5, 7, 11, 12, 13)$. The individual aggregates are $\alpha(I) = \{[2, 6], [10, 13]\}, \alpha(J) = \{[5, 11], [12, 13]\}$ so that the joint quantities are: $\alpha(I, J) = [2, 13]$ and $\beta(I, J) = 1$. This pair has exactly degree 1.

When $I$ and $J$ intersect, there exists at least one block of $\alpha(I)$ that intersects a block of $\alpha(J)$, so that the degree $\beta(I, J)$ is at most equal to $2b - 1$. Next, we partition $\mathcal{O}_2$ according to the value of $\beta(I, J)$ and write

$$\mathcal{O}_2^{[p]} := \{(I, J, X) \in \mathcal{O}_2 \quad ; \quad \beta(I, J) = 2b - p\}$$

for the collection of intersecting pairs $(I, J, X)$ of occurrences for which the degree of freedom equals $2b - p$. From the preceding discussion, only $p \geq 1$ needs to be considered and

$$O_2(z) = O_2^{[1]}(z) + O_2^{[2]}(z) + O_2^{[3]}(z) + \cdots + O_2^{[2b]}(z).$$

As we see next, it is only the first term of this sum that matters asymptotically.

In order to conclude the discussion, we need the notion of *full pairs*: a pair $(I, J)$ of $\mathcal{P}_q(\mathcal{D}) \times \mathcal{P}_q(\mathcal{D})$ is *full* if the joint aggregate $\alpha(I, J)$ completely covers the interval $[1, q]$; see Figure 7.4. (Clearly, the possible values of length $q$ are finite, since $q$ is at most equal to $2\ell$, where $\ell$ is the length of the constraint $\mathcal{D}$.)

Example 7.4.3. Consider the pattern $\mathcal{W} = \mathtt{a}\#_3\mathtt{b}\#_4\mathtt{r}\#\mathtt{a}\#_4\mathtt{c}$. The text `-aarbarbccaracc` also contains two other occurrences of $\mathcal{W}$, at positions $I' = (1, 4, 6, 12, 13)$ and $J' = (5, 7, 11, 12, 14)$. Now, $I'$ and $J'$ are intersecting, and the aggregates are $\alpha(I') = \{[1, 6], [12, 13]\}, \alpha(J') = \{[5, 11], [12, 14]\}$ so that $\alpha(I', J') = \{[1, 11], [12, 14]$. We have here an example of a full pair of occurrences with a number of blocks $\beta(I', J') = 2$.

There is a fundamental translation invariance due to the independence of symbols in the Bernoulli model that entails a combinatorial isomorphism ($\cong$ represents combinatorial isomorphism)

$$\mathcal{O}_2^{[p]} \cong (\mathcal{A}^\star)^{2b-p+1} \times \mathcal{B}_2^{[p]},$$

where $\mathcal{B}_2^{[p]}$ is the subset of $\mathcal{O}_2$ formed of full p airs such that $\beta(I, J)$ equals $2b-p$. In essence, the gaps can be all grouped together (their number is $2b-p+1$, which is translated by the prefactor $(\mathcal{A}^\star)^{2b-p+1}$), while what remains constitutes a full occurrence. The generating function of $\mathcal{O}_2^{[p]}$ is accordingly

$$O_2^{[p]}(z) = \left(\frac{1}{1-z}\right)^{2b-p+1} \times B_2^{[p]}(z)$$

where $B_2^{[p]}(z)$ is the generating function of the collection $\mathcal{B}_2^{[p]}$. From our earlier discussion, it is a *polynomial*. Now, an easy dominant pole analysis entails that $[z^n]O_2^{[p]} = O(n^{2b-p})$. This proves that the dominant contribution to the variance is given by $[z^n]O_2^{[1]}$, which is of order $O(n^{2b-1})$.

The variance $\mathbf{E}(\Xi_n^2)$ involves the constant $B_2^{[1]}(1)$ that is the total weight of the collection $\mathcal{B}_2^{[1]}$. Recall that this collection is formed of intersecting full pairs of occurrences of degree $2b-1$. The polynomial $B_2^{[1]}(z)$ is itself the generating function of the collection $\mathcal{B}_2^{[1]}$, and it is conceptually an extension of Guibas and Odlyzko's autocorrelation polynomial. We shall later make precise the relation between both polynomials.

We summarize our findings in the following theorem.

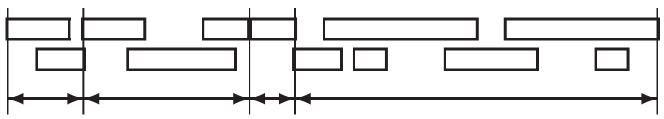


**Figure 7.4.** A full pair of position tuples $I, J$ with $b = 6$ blocks each.

THEOREM 7.4.4. *Consider a general constraint $\mathcal{D}$ with a number of blocks equal to $b$. The mean and the variance of the number of occurrences $\Omega_n$ of a pattern $\mathcal{W}$ subject to constraint $\mathcal{D}$ satisfy*

$$\begin{aligned}\mathbf{E}(\Omega_n) &= \frac{P(\mathcal{W})}{b!}\left(\prod_{j\,:\,d_j<\infty} d_j\right) n^b\left(1+O(n^{-1})\right),\\ \mathrm{Var}(\Omega_n) &= \sigma^2(\mathcal{W})n^{2b-1}\left(1+O(n^{-1})\right),\end{aligned}$$

*where the "variance coefficient" $\sigma^2(\mathcal{W})$ involves the autocorrelation $\kappa(\mathcal{W})$*

$$\sigma^2(\mathcal{W}) = \frac{P^2(\mathcal{W})}{(2b-1)!}\,\kappa^2(\mathcal{W}) \qquad \textit{with} \quad \kappa^2(\mathcal{W}) := \sum_{(I,J)\in\mathcal{B}_2^{[1]}} e(I,J) \tag{7.4.11}$$

*The set $\mathcal{B}_2^{[1]}$ is the collection of all pairs of position tuple $(I,J)$ that satisfy three conditions: $(i)$ they are full; $(ii)$ they are intersecting; $(iii)$ there is a single pair $(r,s)$ with $1\le r,s\le b$ for which the $r$th block $B^{[r]}$ of $\alpha(I)$ and the $s$th block $C^{[s]}$ of $\alpha(J)$ intersect.*

The computation of the autocorrelation $\kappa(\mathcal{W})$ reduces to $b^2$ computations of correlations $\kappa(\mathcal{W}_r,\mathcal{W}_s)$, relative to pairs $(\mathcal{W}_r,\mathcal{W}_s)$ of blocks. Note that each correlation of the form $\kappa(\mathcal{W}_r,\mathcal{W}_s)$ involves a totally constrained problem and is discussed below. Let $D(\mathcal{D}) := \prod_{i:\ d_i<\infty} d_i$. Then, one has

$$\kappa^2(\mathcal{W}) = D^2(\mathcal{D}) \sum_{1\le r,s\le b} \frac{1}{D(\mathcal{D}_r)D(\mathcal{D}_s)}\binom{r+s-2}{r-1}\binom{2b-r-s}{b-r}\,\kappa(\mathcal{W}_r,\mathcal{W}_s), \tag{7.4.12}$$

where $\kappa(\mathcal{W}_r,\mathcal{W}_s)$ is the sum of the $e(I,J)$ taken over all full intersecting pairs $(I,J)$ formed with an position tuple $I$ of block $\mathcal{W}_r$ subject to constraint $\mathcal{D}_r$ and an position tuple $J$ of block $\mathcal{W}_s$ subject to constraint $\mathcal{D}_s$. Let us explain the formula (7.4.12) in words: for a pair $(I,J)$ of the set $\mathcal{B}_2^{[1]}$, there is a single pair $(r,s)$ of indices with $1\le r,s\le b$ for which the $r$th block $B^{[r]}$ of $\alpha(I)$ and the $s$th block $C^{[s]}$ of $\alpha(J)$ intersect. Then, there exist $r+s-2$ blocks before the block $\alpha(B^{[r]},C^{[s]})$ and $2b-r-s$ blocks after it. We then have three different degrees of freedom: $(i)$ the relative order of blocks $B^{[i]}(i<r)$ and blocks $C^{[j]}(j<s)$, and similarly the relative order of blocks $B^{[i]}(i>r)$ and blocks $C^{[j]}(j>s)$; $(ii)$ the lengths of the blocks (there are $D_j$ possible lengths for the $j$th block); $(iii)$ finally the relative positions of the blocks $B^{[r]}$ and $C^{[s]}$.

In particular, in the unconstrained case, the parameter $b$ equals $m$, and each block $\mathcal{W}_r$ is reduced to the symbol $w_r$. Then the "correlation coefficient" $\kappa^2(\mathcal{W})$ simplifies to

$$\kappa^2(\mathcal{W}) := \sum_{1\le r,s\le m}\binom{r+s-2}{r-1}\binom{2m-r-s}{m-r}[\![w_r=w_s]\!]\left(\frac{1}{p_{w_r}}-1\right). \tag{7.4.13}$$

In words, once you fix the position of the intersection, called pivot, then amongst the $r+s-2$ elements smaller than the pivot one assigns freely $r-1$ to the first occurrence and the remaining $s-1$ to the second. One proceeds similarly for the $2m-r-s$ elements larger than the pivot.

### 7.4.2. Autocorrelation polynomial revisited

Finally, we compare the autocorrelation coefficient $\kappa(\mathcal{W})$ with the autocorrelation polynomial $S_w(z)$ introduced in the last section for the exact string matching problem. Let now $w = w_1 w_2 \dots w_m$ be again a string of length $m$, and all the symbols of $w$ must occur at consecutive places, so that a valid position $I$ is an interval of length $m$. We recall that the autocorrelation set $\mathcal{P}(w) \subset [1..m]$ involves all indices $k$ such that the prefix $w_1^k$ coincides with the suffix $w_{m-k+1}^m$. Here, an index $k \in \mathcal{P}(w)$ is relative to a intersecting pair of positions $(I, J)$ and one has $w_1^k = w_{I\cap J}$.

In the previous section, we introduced the autocorrelation polynomial $S_w(z)$ as

$$S_w(z) = \sum_{k\in\mathcal{P}_w} P(w_{k+1}^m) z^{m-k} = P(w) \sum_{k\in\mathcal{P}(w)} \frac{1}{P(w_1^k)} z^{m-k}.$$

We also define

$$C_w(z) = \sum_{k\in\mathcal{P}(w)} z^{m-k}.$$

Since the polynomial $B_2^{[1]}$ involves coefficients $e(I, J)$ this polynomial can be written as function of the two autocorrelations polynomials $A_w$ and $C_w$,

$$B_2^{[1]}(z) = P(w) z^m \; [A_w(z) - P(w)\, C_w(z)].$$

Put simply, the variance coefficient of the hidden pattern problem extends the classical autocorrelation quantities associated with strings.

### 7.4.3. Central limit laws

Our goal is to prove that the sequence $\Omega_n$ appropriately centered and scaled tends to the normal distribution. We consider the following standardized random variable $\tilde{\Xi}_n$ which is defined for each $n$ by

$$\tilde{\Xi}_n := \frac{\Xi_n}{n^{b-1/2}} = \frac{\Omega_n - \mathbf{E}(\Omega_n)}{n^{b-1/2}}, \qquad (7.4.14)$$

where $b$ is the number of blocks of the constraint $\mathcal{D}$. We shall show that $\tilde{\Xi}_n$ behaves asymptotically as a normal variable with mean 0 and standard deviation $\sigma$. By the classical *moment convergence theorem* this is established once all moments of $\tilde{\Xi}_n$ are known to converge to the appropriate moments of the standard normal distribution.

We remind the reader that if $G$ is a standard normal variable (i.e., a Gaussian distributed variable with mean 0 and standard deviation 1), then for any integral $s \geq 0$

$$\mathbf{E}(G^{2s}) = 1 \cdot 3 \cdots (2s-1), \qquad \mathbf{E}(G^{2s+1}) = 0. \qquad (7.4.15)$$

We shall accordingly distinguish two cases based on the parity of $r$, $r = 2s$ and $r = 2s+1$, and prove that

$$\mathrm{E}[\Xi_n^{2s+1}] = o(n^{(2s+1)(b-1/2)}), \qquad \mathbf{E}(\Xi_n^{2s}) \sim \sigma^{2s}\,(1\cdot 3\cdots(2s-1))\,n^{2sb-s}, \tag{7.4.16}$$

which implies Gaussian convergence of $\tilde{\Xi}_n$.

Theorem 7.4.5. *The random variable $\Omega_n$ over a random text of length $n$ generated by a memoryless source asymptotically obeys a* Central Limit Law *in the sense that its distribution is asymptotically normal: for all $x = O(1)$, one has*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\Omega_n - \mathbf{E}(\Omega_n)}{\sqrt{\mathrm{Var}(\Omega_n)}} \le x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt. \tag{7.4.17}$$

*Proof.* The proof below is combinatorial; it basically reduces to grouping and enumerating adequately the various combinations of indices in the sum that expresses $\mathbf{E}(\Xi_n^r)$. Once more, $\mathcal{P}_n(\mathcal{D})$ is formed of all the sets of positions in $[1,n]$ subject to the constraint $\mathcal{D}$ and we set $\mathcal{P}(\mathcal{D}) := \bigcup_n \mathcal{P}_n(\mathcal{D})$. Then totally distributing the terms in $\Xi^r$ yields

$$\mathbf{E}(\Xi_n^r) = \sum_{(I_1,\ldots,I_r)\in\mathcal{P}_n^r(\mathcal{D})} \mathbf{E}(Y_{I_1}\cdots Y_{I_r}). \tag{7.4.18}$$

An $r$-tuple of sets $(I_1,\ldots,I_r)$ in $\mathcal{P}^r(\mathcal{D})$ is said to be *friendly* if each $I_k$ intersects at least one other $I_\ell$, with $\ell \neq k$ and we let $\mathcal{Q}^{(r)}(\mathcal{D})$ be the set of all friendly collections in $\mathcal{P}^r(\mathcal{D})$. For $\mathcal{P}^r$, $\mathcal{Q}^{(r)}$, and their derivatives below, we add the subscript $n$ each time the situation is particularized to texts of length $n$. If $(I_1,\ldots,I_r)$ does not lie in $\mathcal{Q}^{(r)}(\mathcal{D})$, then $\mathbf{E}(Y_{I_1}\cdots Y_{I_r}) = 0$, since at least one of the $Y_I$'s is independent of the other factors in the product and the $Y_I$'s have been centered, $\mathbf{E}(Y_I) = 0$. One can thus restrict attention to friendly families and get the basic formula

$$\mathbf{E}(\Xi_n^r) = \sum_{(I_1,\ldots,I_r)\in\mathcal{Q}_n^{(r)}(\mathcal{D})} \mathbf{E}(Y_{I_1}\cdots Y_{I_r}), \tag{7.4.19}$$

where the expression involves fewer terms than in (7.4.18). From there, we proceed in two stages. First, restrict attention to friendly families that give rise to the dominant contribution and introduce a suitable subfamily $\mathcal{Q}_\star^{(r)} \subset \mathcal{Q}^{(r)}$; in so doing, moments of odd order appear to be negligible. Next, for even order $r$, the family $\mathcal{Q}_\star^{(r)}$ involves a symmetry and it suffices to consider another smaller subfamily $\mathcal{Q}_{\star\star}^{(r)} \subset \mathcal{Q}_\star^{(r)}$ that corresponds to a "standard" form of position tuple intersection; this last reduction precisely gives rise to the even Gaussian moments.

**Odd moments.** Given $(I_1,\ldots,I_r) \in \mathcal{Q}^{(r)}$, the aggregate $\alpha(I_1,I_2,\ldots,I_r)$ is defined as the aggregation (in the sense of the variance calculation above) of $\alpha(I_1)\cup\cdots\cup\alpha(I_r)$. Next, the *number of blocks* of $(I_1,\ldots,I_r)$ is the number of blocks of the aggregate $\alpha(I_1, \ldots, I_r)$; if $p$ is the total number of intersecting blocks of the aggregate $\alpha(I_1, \ldots, I_r)$, the aggregate $\alpha(I_1, I_2, \ldots I_r)$ has $rb - p$ blocks. Like previously, we say that the family $(I_1, \ldots, I_r)$ of $\mathcal{Q}_q^{(r)}$ is *full* if the aggregate $\alpha(I_1, I_2, \ldots I_r)$ completely covers the interval $[1, q]$. In this case, the length of the aggregate is at most $rd(m-1)+1$, and the generating function of full families is a polynomial $P_r(z)$ of degree at most $rd(m-1)+1$ with $d = \max_{j \in \mathcal{F}} d_j$. Then, the generating function of families of $\mathcal{Q}^{(r)}$ whose block number equals $k$ is of the form

$$\left(\frac{1}{1-z}\right)^{k+1} \times P_r(z),$$

so that the number of families of $\mathcal{Q}_n^{(r)}$ whose block number equals $k$ is $O(n^k)$. This observation proves that the dominant contribution to (7.4.19) arises from friendly families with a maximal block number. It is clear that the minimum number of intersecting blocks of any element of $\mathcal{Q}^{(r)}$ equals $\lceil r/2 \rceil$, since it coincides exactly with the minimum number of edges of a graph with $r$ vertices which contains no isolated vertex. Then the maximum block number of a f friendly family equals $rb - \lceil r/2 \rceil$. In view of this fact and the remarks above regarding cardinalities, we immediately have

$$\mathrm{E}\left[\Xi_n^{2s+1}\right] = O\left(n^{(2s+1)b-s-1}\right) = o\left(n^{(2s+1)(b-1/2)}\right)$$

which establishes the limit form of odd moments in (7.4.16).

**Even Moments**. We are thus left with estimating the even moments. The dominant term is relative to friendly families of $\mathcal{Q}^{(2s)}$ with an intersecting block number equal to $s$, whose set we denote by $\mathcal{Q}_\star^{(2s)}$. In such a family, each subset $I_k$ intersects one and only one other subset $I_\ell$. Furthermore, if the blocks of $\alpha(I_h)$ are denoted by $B_h^{[u]}, 1 \le u \le b$, there exists only one block $B_k^{[u_k]}$ of $\alpha(I_k)$ and only one block $B_\ell^{[u_\ell]}$ that contains the points of $I_k \cap I_\ell$. This defines an involution $\tau$ such that $\tau(k) = \ell$ and $\tau(\ell) = k$ for all pairs of indices $(\ell, k)$ for which $I_k$ and $I_\ell$ intersect. Furthermore, given the symmetry relation $\mathbf{E}(Y_{I_1} \cdots Y_{I_{2s}}) = \mathbf{E}(Y_{I_{\rho(1)}} \cdots Y_{I_{\rho(2s)}})$ it suffices to restrict attention to friendly families of $\mathcal{Q}_\star^{(2s)}$ for which the involution $\tau$ is the standard one with cycles $(1,2)$, $(3,4)$, etc; for such "standard" families whose set is denoted by $\mathcal{Q}_{\star\star}^{(2s)}$, the pairs that intersect are thus $(I_1, I_2), \ldots, (I_{2s-1}, I_{2s})$. Since the set $\mathcal{K}_{2s}$ of involutions of $2s$ elements has cardinality $K_{2s} = 1 \cdot 3 \cdot 5 \cdots (2s-1)$ , the equality

$$\sum_{\mathcal{Q}_{\star n}^{(2s)}} \mathbf{E}(Y_{I_1} \cdots Y_{I_{2s}}) = K_{2s} \sum_{\mathcal{Q}_{\star\star n}^{(2s)}} \mathbf{E}(Y_{I_1} \cdots Y_{I_{2s}}), \tag{7.4.20}$$

entails that we can work now solely with standard families.

The class of position tuples relative to standard families is $\mathcal{A}^\star \times (\mathcal{A}^\star)^{2sb-s-1} \times \mathcal{B}_{2s}^{[s]} \times \mathcal{A}^\star$; this class involves the collection $\mathcal{B}_{2s}^{[s]}$ of all full friendly $2s$-tuples of position tuples with a number of blocks equal to $s$. Since $\mathcal{B}_{2s}^{[s]}$ is exactly a shuffle of $s$ copies of $\mathcal{B}_2^{[1]}$ (as introduced in the study of the variance), the associated generating function is

$$\left(\frac{1}{1-z}\right)^{2sb-s+1} \quad (2sb-s)! \left(\frac{B_2^{[1]}(z)}{(2b-1)!}\right)^s,$$

where $B_2^{[1]}(z)$ is the already introduced autocorrelation polynomial. Upon taking coefficients, we obtain the estimate

$$\sum_{\mathcal{Q}_{\star\star n}^{(2s)}} \mathbf{E}(Y_{I_1} \cdots Y_{I_{2s}}) \sim n^{(2b-1)s} \sigma^{2s}. \tag{7.4.21}$$

In view of the formulæ (7.4.18), (7.4.19), (7.4.20), and (7.4.21) above, this yields the estimate of even moments and leads to the second relation of (7.4.16). This completes the proof of Theorem 7.4.5. ■

The even Gaussian moments eventually come out as the number of involutions, which corresponds to a fundamental asymptotic symmetry present in the problem. In this perspective the specialization of the proof to the fully unconstrained case is reminiscent of the derivation of the usual central limit theorem (dealing with sums of independent variables) by moments methods.

### 7.4.4. Limit laws for fully constrained pattern

In this section, we strengthen our results for *fully constrained pattern* in which all gaps $d_j$ are finite. We set $D = \prod_j d_j$, and $\ell = \sum_j d_j$. Observe that in this case, we can reduced the subsequence problem to a generalized string matching problem with the generalized pattern $\mathcal{W}$ consisting of all words that satisfy $(\mathcal{W}, \mathcal{D})$. Thus our previous results apply, in particular, Theorems 7.3.8 and 7.3.10. This leads to the following result.

Theorem 7.4.6. *Consider a fully constrained pattern with mean and variance found in Theorem 7.4.4 for $b = 1$.*

(i) *The random variable $\Omega_n$ satisfies a* Central Limit Law *with speed of convergence $1/\sqrt{n}$:*

$$\sup_x \left| \mathbf{P}\left(\frac{\Omega_n - DP(\mathcal{W})n}{\sigma(\mathcal{W})\sqrt{n}} \le x\right) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \right| = O\left(\frac{1}{\sqrt{n}}\right). \tag{7.4.22}$$

(ii) Large deviations *from the mean value have exponentially small probability: there exist a constant $\eta > 0$ and a nonnegative function $I(x)$ defined throughout $(0, \eta)$ such that $I(x) > 0$ for $x \neq DP(\mathcal{W})$ and*

$$\begin{cases} \lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\frac{\Omega_n}{n} \le x\right) = -I(x) \text{ if } 0 < x < DP(\mathcal{W}) \\ \lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\frac{\Omega_n}{n} \ge x\right) = -I(x) \text{ if } DP(\mathcal{W}) < x < \eta \end{cases}, \tag{7.4.23}$$

*except for at most a finite number of exceptional values of $x$. More precisely,*

$$I(x) = -\log \frac{\lambda(\zeta)}{\zeta^x} \qquad \text{with } \zeta \equiv \zeta(x) \text{ defined by } \quad \frac{\zeta\lambda'(\zeta)}{\lambda(\zeta)} = x \tag{7.4.24}$$

*where $\lambda(u)$ is the largest eigenvalue of the matrix* $\mathrm{T}(u)$ *of the associate de Bruijn graph constructed for $\mathcal{W} = \{v: \ v = w_1u_1w_2\cdots w_{m-1}u_{m-1}w_m$, where $u_i \in \mathcal{A}^{d_i-1}, \ 1 \le i \le m-1\}$.*

(iii) *For primitive patterns (cf. Section 7.3.2)* a Local Limit Law *holds:*

$$\sup_k \left| \mathbf{P}\left(\Omega_n = k\right) - \frac{1}{\sigma(\mathcal{W})\sqrt{n}} \frac{e^{x(k)^2/2}}{\sqrt{2\pi}} \right| = o\left(\frac{1}{\sqrt{n}}\right), \tag{7.4.25}$$

*where*

$$x(k) = \frac{k - DP(\mathcal{W})n}{\sigma(\mathcal{W})\sqrt{n}}$$

*for $n \to \infty$.*

Example 7.4.7. We motivated our desire to study the subsequence problem by an example form computer security. Indeed, if one wants to detect "suspicious" activities (e.g., signatures viewed as subsequences in an audit file), it is important to set up a threshold in order to avoid false alarms. This problem can be rephrased as one of finding a threshold $\alpha_0 = \alpha_0(\mathcal{W}; n, \beta)$ such that

$$\mathbf{P}(\Omega_n > \alpha_0) \le \beta,$$

for small given $\beta$ (say $\beta = 10^{-5}$). Based on frequencies of letters and the assumption that a memoryless model is (at least roughly) relevant, one can estimate the mean value and the standard deviation coefficients $P(\mathcal{W}), \sigma(\mathcal{W})$ as discussed above. The Gaussian limits granted by Theorems 7.4.5 and 7.3.8 then reduce the problem to solving an approximate system, which in the (fully) constrained case reads

$$\alpha_0 = nP(\mathcal{W}) + x_0\sigma(\mathcal{W})\sqrt{n}, \qquad \beta = \frac{1}{\sqrt{2\pi}} \int_{x_0}^{\infty} e^{-t^2/2}\, dt.$$

This system admits the approximate solution

$$\alpha_0 \approx n\pi(\omega) + \sigma(\mathcal{W})\sqrt{2n\log(1/\beta)}. \tag{7.4.26}$$

for small $\beta$.

## 7.5. Generalized subsequence problem

In the *generalized subsequence problem* the pattern is $\mathcal{W} = (\mathcal{W}_1, \ldots, \mathcal{W}_d)$ where $\mathcal{W}_i$ is a set of strings (a language). We say that the generalized pattern $\mathcal{W}$

occurs in the text $X$ if $X$ contains $\mathcal{W}$ as a *subsequence* $(w_1, w_2, \ldots, w_d)$ where $w_i \in \mathcal{W}_i$. An occurrence of the pattern in $X$ is a sequence

$$(u_0, w_1, u_1, \ldots, w_d, u_d)$$

such that $X = u_0 w_1 u_1 \cdots w_d u_d$. We shall study the associated language $\mathcal{L}$ that can be described as

$$\mathcal{L} = \mathcal{A}^* \cdot \mathcal{W}_1 \cdot \mathcal{A}^* \cdots \mathcal{W}_d \cdot \mathcal{A}^*. \tag{7.5.1}$$

More precisely, an occurrence of $\mathcal{W}$ is a sequence of $d$ disjoint intervals $I = (I_1, I_2, \ldots I_d)$ such that $I_j := [k_j^1, k_j^2]$ where $1 \le k_j^1 \le k_j^2 \le n$ is a portion of text $X_1^n$ where $w_j \in \mathcal{W}_j$ occurs. We denote by $\mathcal{P}_n := \mathcal{P}_n(\mathcal{W})$ the set of all valid occurrences $I$. The number of occurrences $\Omega_n$ of $\mathcal{W}$ in the text $X$ of size $n$ is then

$$\Omega_n = \sum_{I \in \mathcal{P}_n(\mathcal{L})} Z_I, \tag{7.5.2}$$

where $Z_I(X) := [\![\mathcal{W} \text{ occurs at position } I \text{ in } X]\!]$.

In passing, we observe that the generalized subsequence problem is the most general pattern matching considered so far. It contains the exact string matching (cf. Section 7.2), generalized string matching (cf. Section 7.3), and the subsequence pattern matching known also as hidden patterns (cf. Section 7.4). In this section we present an analysis of the first two moments of $\Omega_n$ for the generalized subsequence pattern matching problem for dynamic sources discussed in Section 7.1.

### 7.5.1. Generating operators for dynamic sources

In Section 7.1 we have introduced a general probabilistic source known as a dynamic source. In this section we analyze the generalized subsequence model for such sources.

We start with a brief description of the methodology of generating operators that are used in the analysis of dynamic sources. We recall from Section 7.1 that the *generating operator* $\mathbf{G}_w$ is defined as $\mathbf{G}_w[f](t) := |h'_w(t)| f \circ h_w(t)$ for a density function $f$ and a word $w$. In particular, in (7.1.2) we proved that $P(w) \int_0^1 f(t)dt = \int_0^1 \mathbf{G}_w[f](t)dt$ for any function $f(t)$, which implies that $P(w)$ is an eigenvalue of the operator $\mathbf{G}_w$. Furthermore, the generating operator for $w \cdot u$ is $\mathbf{G}_{w \cdot u} = \mathbf{G}_u \circ \mathbf{G}_w$, where $w$ and $u$ are words (cf. (7.1.3)) and $\circ$ is the composition of operators.

Consider now a language $\mathcal{B} \subset \mathcal{A}^*$. Its *generating operator* $\mathbf{B}(z)$ is then defined as

$$\mathbf{B}(z) := \sum_{w \in \mathcal{B}} z^{|w|} \; \mathbf{G}_w.$$

We observe that the ordinary generating function of a language $\mathcal{B}$ is related to the generating operators. Indeed,

$$B(z) := \sum_{w \in \mathcal{B}} z^{|w|} P(w) = \sum_{w \in \mathcal{B}} z^{|w|} \int_0^1 \mathbf{G}_w[f](t)dt = \int_0^1 \mathbf{B}(z)[f](t)dt. \tag{7.5.3}$$

If $\mathbf{B}(z)$ is well defined at $z = 1$, then $\mathbf{B}(1)$ is called the *normalized operator* of $\mathcal{B}$. In particular, using (7.1.1) we can compute

$$P(\mathcal{B}) = \sum_{w\in\mathcal{B}} P(w) = \int_0^1 \mathbf{B}(1)dt.$$

Furthermore, the operator

$$\mathbf{G} := \sum_{a\in\mathcal{A}} \mathbf{G}_a, \tag{7.5.4}$$

is the normalized operator of the alphabet $\mathcal{A}$ and plays a fundamental role in the analysis.

From the product formula (7.1.3) of the generating operators $\mathbf{G}_w$ we conclude that unions and Cartesian products of languages translates into sums and compositions of the associated operators. For instance, the operator associated with $\mathcal{A}^\star$ is

$$(I - z\mathbf{G})^{-1} := \sum_{i\geq 0} z^i \mathbf{G}^i.$$

where $\mathbf{G}^i = \mathbf{G} \circ \mathbf{G}^{i-1}$.

In order to proceed, we must restrict our attention to a class of dynamic sources called *decomposable* that satisfy two properties: (i) there exists a unique positive dominant eigenvalue $\lambda$ and a dominant eigenvector denoted as $\varphi$ (which is unique under the normalization $\int \varphi(t)dt = 1$); (ii) there is a spectral gap between the dominant eignevalue and other eignevalues. These properties entail the separation of the operator $\mathbf{G}$ into two parts

$$\mathbf{G} = \lambda\mathbf{P} + \mathbf{N} \tag{7.5.5}$$

such that the operator $\mathbf{P}$ is the projection relative to the dominant eigenvalue $\lambda$ while $\mathbf{N}$ is the operator relative to the remainder of the spectrum (cf. Section 7.3). Furthermore (cf. Exercise 7.5.1)

$$\mathbf{P} \circ \mathbf{P} = \mathbf{P}, \tag{7.5.6}$$

$$\mathbf{P} \circ \mathbf{N} = \mathbf{N} \circ \mathbf{P} = 0. \tag{7.5.7}$$

The last property implies that for any $i \geq 1$

$$\mathbf{G}^i = \lambda^i \mathbf{P} + \mathbf{N}^i. \tag{7.5.8}$$

In particular, for the *density* operator $\mathbf{G}$ the dominant eigenvalue $\lambda = P(\mathcal{A}) = 1$ and $\varphi$ is the unique stationary distribution. The function 1 is the left eigenvector. Then using (7.5.8) we arrive at

$$(I - z\mathbf{G})^{-1} = \frac{1}{1-z}\mathbf{P} + \mathbf{R}(z), \tag{7.5.9}$$

where

$$\mathbf{R}(z) := (I - z\mathbf{N})^{-1} - \mathbf{P} = \sum_{k\geq 0} z^k(\mathbf{G}^k - \mathbf{P}). \tag{7.5.10}$$

Observe that the first part of (7.5.9) has a pole at $z = 1$ and due to the spectral gap the operator $\mathbf{N}$ has spectral radius $\nu < \lambda = 1$. Furthermore, the operator $\mathbf{R}(z)$ is analytic in $|z| < (1/\nu)$ and again thanks to the existence of the spectral gap, the series $\mathbf{R}(1)$ is of geometric type. We shall point out below that the speed of convergence of $\mathbf{R}(z)$ is closely related to the decay of the correlation between two consecutive symbols. Finally, we list some additional properties of just introduced operators (cf. Exercise 7.5.2) true for any function $g(t)$ defined between 0 and 1.

$$\mathbf{N}[\varphi] = 0, \qquad \mathbf{P}[g](t) = \varphi(t)\int_0^1 g(t')dt' \tag{7.5.11}$$

$$\int_0^1 \mathbf{P}[g](t)dt = \int_0^1 g(t)dt, \qquad \int_0^1 \mathbf{N}[g](t)dt = 0, \tag{7.5.12}$$

where $\varphi$ is the stationary density.

Theory built so far allows us, among others, to define precisely the correlation between languages in terms of the generating operators. From now on we restrict our analysis to the so called *nondense* languages $\mathcal{B}$ for which the associated generating operator $\mathbf{B}(z)$ is analytic in a disk $|z| > 1$. First, observe that for a nondense language $\mathcal{B}$, the normalized generating operator $\mathbf{B}$ satisfies

$$\int_0^1 \mathbf{P} \circ \mathbf{B} \circ \mathbf{P}[g](t) = P(\mathcal{B}) \left(\int_0^1 g(t)dt\right). \tag{7.5.13}$$

Let us now define the correlation coefficient between two languages, say $\mathcal{B}$ with the generating operator $\mathbf{B}$ and $\mathcal{C}$ with generating operator $\mathbf{C}$. Two types of correlations may occur between such languages. If $\mathcal{B}$ and $\mathcal{C}$ do not overlap, then $\mathcal{B}$ may be before $\mathcal{C}$, or after $\mathcal{C}$. We define the *correlation coefficient* $c(\mathcal{B}, \mathcal{C})$ (and in an analogous way $c(\mathcal{C}, \mathcal{B})$) as

$$\begin{aligned} P(\mathcal{B})P(\mathcal{C})c(\mathcal{B}, \mathcal{C}) &:= \sum_{k\geq 0} \left[P(\mathcal{B} \times \mathcal{A}^k \times \mathcal{C}) - P(\mathcal{B})P(\mathcal{C})\right] \\ &= \int_0^1 \mathbf{C} \circ \mathbf{R}(1) \circ \mathbf{B}[\varphi](t). \end{aligned} \tag{7.5.14}$$

To see this we observe, using (7.5.5)–(7.5.13),

$$\begin{aligned} \int_0^1 \mathbf{C} \circ \mathbf{R}(1) \circ \mathbf{B}[\varphi](t)dt &= \int_0^1 \mathbf{C} \circ \left(\sum_{k\geq 0}(\mathbf{G}^k - \mathbf{P})\right) \circ \mathbf{B}[\varphi](t)dt \\ &= \sum_{k\geq 0}\left(\int_0^1 \mathbf{C} \circ \mathbf{G}^k \circ \mathbf{B}[\varphi](t)dt - \int_0^1 \mathbf{C} \circ \mathbf{P} \circ \mathbf{B}[\varphi](t)\right) \\ &= \sum_{k\geq 0}\left(P(\mathcal{B} \times \mathcal{A}^k \times \mathcal{B}) - P(\mathcal{B})P(\mathcal{C})\right). \end{aligned}$$

We say that $\mathcal{B}$ and $\mathcal{C}$ overlap if there exist words $b$, $u$ and $c$ such that $u \neq \varepsilon$ and $(bu, uc) \in (\mathcal{B} \times \mathcal{C}) \cup (\mathcal{C} \times \mathcal{B})$. Then we denote by $\mathcal{B} \uparrow \mathcal{C}$ the set of words that be obtained by overlapping words from $\mathcal{B}$ and $\mathcal{C}$. The correlation coefficient of the overlapping languages $\mathcal{B}$ and $\mathcal{C}$ is defined as

$$d(\mathcal{B},\mathcal{C}) := \frac{P(\mathcal{B} \uparrow \mathcal{C})}{P(\mathcal{B})P(\mathcal{C})} \tag{7.5.15}$$

Finally, the total correlation coefficient $m(\mathcal{B},\mathcal{C})$ between $\mathcal{B}$ and $\mathcal{C}$ is defined as

$$m(\mathcal{B},\mathcal{C}) = c(\mathcal{B},\mathcal{C}) + c(\mathcal{C},\mathcal{B}) + d(\mathcal{B},\mathcal{C}), \tag{7.5.16}$$

that is,

$$\begin{aligned} P(\mathcal{B})P(\mathcal{C})m(\mathcal{B},\mathcal{C}) &= P(\mathcal{B} \uparrow \mathcal{C}) \\ &+ \sum_{k\geq 0} \left[ P(\mathcal{B} \times \mathcal{A}^k \times \mathcal{C}) + P(\mathcal{C} \times \mathcal{A}^k \times \mathcal{B}) - 2P(\mathcal{B})P(\mathcal{C}) \right]. \end{aligned}$$

We shall need these coefficients in the analysis of the generalized subsequence problem for dynamic sources.

### 7.5.2. Mean and variance

In this section we shall derive the mean and the variance of the number of occurrences $\Omega_n(\mathcal{W})$ of the generalized pattern as a subsequence for a dynamic source.

We first give a sketch of the forthcoming proof:

- We first describe the generating operators of the language $\mathcal{L}$ defined in (7.5.1) that we repeat it here

$$\mathcal{L} = \mathcal{A}^* \times \mathcal{W}_1 \times \mathcal{A}^* \cdots \mathcal{W}_d \times \mathcal{A}^*.$$

  It turns out that the quasi-inverse $(I - z\mathbf{G})^{-1}$ operator is involved in such a generating operator.
- We then decompose the operator with the help of (7.5.9). We obtain a term related to $(1-z)^{-1}\mathbf{P}$ that gives the main contribution to the asymptotics, and another term coming from the operator $\mathbf{R}(z)$.
- We then compute the generating function of $\mathcal{L}$ using (7.5.3).
- Finally, we extract asymptotic behavior from the generating function.

The main finding of this section is summarized in the next theorem.

THEOREM 7.5.1. *Consider a decomposable dynamical source endowed in the stationary density $\varphi$ and let $\mathcal{W} = (\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_d)$ be a generalized nondense pattern.*
(i) *The expectation $\mathbf{E}(\Omega_n)$ of the number of occurrences of the generalized pattern $\mathcal{W}$ in a text of length $n$ satisfies asymptotically*

$$\mathbf{E}(\Omega_n(\mathcal{W})) = \binom{n+d}{d} P(\mathcal{W}) + \binom{n+d-1}{d-1} P(\mathcal{W}) \left[ C(\mathcal{W}) - T(\mathcal{W}) \right] + O(n^{d-2}),$$

*where*

$$T(\mathcal{W}) = \sum_{i=1}^{d} \sum_{w\in\mathcal{W}_i} \frac{|w|P(w)}{P(\mathcal{W}_i)} \tag{7.5.17}$$

*is the average length, and the correlation coefficient $C(\mathcal{W})$ is the sum of the correlations $c(\mathcal{W}_{i-1}, \mathcal{W}_i)$ between languages $\mathcal{W}_i$ and $\mathcal{W}_{i+1}$ as defined in (7.5.14).*
(ii) *The variance of $\Omega_n$ is asymptotically equal to*

$$\mathrm{Var}(\Omega_n(\mathcal{W})) = \sigma^2(\mathcal{W})\ n^{2d-1}\left(1+O(n^{-1})\right), \tag{7.5.18}$$

*where the coefficient*

$$\sigma^2(\mathcal{W}) = P^2(\mathcal{W})\left[\frac{d-2T(\mathcal{W})}{d!(d-1)!} + \frac{m(\mathcal{W})}{(2d-1)!}\right]$$

*and the total correlation–coefficient $m(\mathcal{W})$ can be computed as*

$$m(\mathcal{W}) := \sum_{1\le i,j\le d} \binom{i+j-2}{i-1}\binom{2d-i-j}{d-i} m(\mathcal{W}_i, \mathcal{W}_j).$$

*where $m(\mathcal{W}_i, \mathcal{W}_{i+1})$ are defined in (7.5.16).*

*Proof.* We only prove part (i) leaving the proof of part (ii) as an exercise (cf. Exercise 7.5.3). We shall start with the language representation $\mathcal{L}$ defined in (7.5.1) that we recalled above. Its generating operator is

$$\mathbf{L}(z) = (I - z\mathbf{G})^{-1} \circ \mathbf{L}_r(z) \circ (I - z\mathbf{G})^{-1} \circ \cdots \circ \mathbf{L}_1(z) \circ (I - z\mathbf{G})^{-1}. \tag{7.5.19}$$

After applying the transformation (7.5.8) to $\mathbf{L}(z)$, we obtain an operator $\mathbf{M}_1(z)$

$$\mathbf{M}_1(z) = \left(\frac{1}{1-z}\right)^{d+1} \mathbf{P} \circ \mathbf{L}_r(z) \circ \mathbf{P} \circ \cdots \circ \mathbf{P} \circ \mathbf{L}_1(z) \circ \mathbf{P}$$

that has a pole of order $r+1$ at $z=1$. Near $z=1$, each operator $\mathbf{L}_i(z)$ is analytic and admits the expansion

$$\mathbf{L}_i(z) = \mathbf{L}_i + (z-1)\mathbf{L}_i'(1) + O(z-1)^2.$$

Therefore, the leading term of the expansion is

$$\left(\frac{1}{1-z}\right)^{d+1} \mathbf{P} \circ \mathbf{L}_r \circ \mathbf{P} \circ \cdots \circ \mathbf{P} \circ \mathbf{L}_1 \circ \mathbf{P}. \tag{7.5.20}$$

The second main term is a sum of $r$ terms, each of them obtained by replacing the operator $\mathbf{L}_i(z)$ by its derivative $\mathbf{L}_i'(1)$ at $z=1$. The corresponding generating function $M_1(z)$ satisfies near $z=1$

$$M_1(z) = \left(\frac{1}{1-z}\right)^{d+1} P(\mathcal{W}) - \left(\frac{1}{1-z}\right)^{d} P(\mathcal{W})T(\mathcal{W}) + O\left(\frac{1}{1-z}\right)^{d-1}. \tag{7.5.21}$$

where the average length $T(\mathcal{W})$ is defined in (7.5.17).

After applying (7.5.8) in $\mathbf{L}(z)$, we obtain an operator $\mathbf{M}_2(z)$ that has a pole of order $r$ at $z = 1$. This is a sum of $d+1$ terms, each of the term containing an occurrence of the operator $\mathbf{R}(z)$ between two generating operators of languages $\mathcal{W}_{i-1}, \mathcal{W}_i$. The corresponding generating function $M_2(z)$ has also a pole of order $r$ at $z = 1$ and satisfies near $z = 1$

$$M_2(z) = \left(\frac{1}{1-z}\right)^d P(\mathcal{W}) \sum_{i=2}^{d} c(\mathcal{L}_{i-1}, \mathcal{L}_i) + O\left(\frac{1}{1-z}\right)^{d-1}.$$

Here, the correlation number $c(\mathcal{B},\mathcal{C})$ between $\mathcal{B}$ and $\mathcal{C}$ is defined in (7.5.14). To complete the proof we need only to extract the coefficients of $P(z)/(1-z)^d$, as already discussed in previous sections. ∎

## 7.6. Self-repetitive pattern matching

In this last section of the chapter, we change the model. So far we postulated the pattern $w$ is given. Hereafter, we make the pattern part of the text, which is still randomly generated. To simplify our presentation, we assume that the text is emitted by a memoryless source. We should point out that the quantity analyzed here is in fact the typical depth in a (compact) suffix trie built over the suffixes of a randomly generated text.

### 7.6.1. Formulation of the problem

Let $i$ be an arbitrary integer smaller than or equal to $n$. We define $D_n(i)$ to be the largest value of $k \le n$ such that $X_i^{i+k-1}$ occurs at least twice in the text $X_1^n$ of length $n$; in other words, such that $N_n(X_i^{i+k-1}) \ge 2$. We recall that $N_n(w)$ is the number of times pattern $w$ occurs in the text $X_1^n$. Clearly, $N_n(X_i^{i+k-1}) \ge 1$. Our goal is to determine probabilistic behavior of a "typical" $D_n(i)$, that is, we define $D_n$ to be equal to $D_n(i)$ when $i$ is randomly and uniformly selected between 1 and $n$. More precisely,

$$\mathbf{P}(D_n = \ell) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{P}(D_n(i) = \ell)$$

for any $1 \le \ell \le n$.

Let $w \in \mathcal{A}^k$ be an arbitrary word of size $k$. Observe that

$$\mathbf{P}(D_n(i) \ge k \ \&\ X_i^{i+k-1} = w) = \mathbf{P}(N_n(w) \ge 2 \ \&\ X_i^{i+k-1} = w),$$

and

$$\sum_{i=1}^{n} \mathbf{P}(N_n(w) = r \ \&\ X_i^{i+k-1} = w) = r\mathbf{P}(N_n(w) = r).$$

Recall that $N_n(u) = \mathbf{E}(u^{N_n(w)}) = \sum_{r\geq 0}\mathbf{P}(N_n(w) = r)u^r$ is the probability generating function of $N_n(w)$. We sometimes shall write $N_{n,w}(u)$ to underline the fact that the pattern $w$ is given. From above we conclude that

$$\begin{aligned}
\mathbf{P}(D_n \geq k) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}(D_n(i) \geq k) \\
&= \sum_{w\in\mathcal{A}^k}\frac{1}{n}\sum_{i=1}^{n}\mathbf{P}(D_n(i) \geq k \ \&\ X_i^{i+k-1} = w) \\
&= \frac{1}{n}\sum_{w\in\mathcal{A}^k}\sum_{r\geq 2} r\mathbf{P}(N_n(w) = r) \\
&= \sum_{w\in\mathcal{A}^k}\left(\mathbf{P}(w) - \frac{1}{n}N'_{n,w}(0)\right) \\
&= 1 - \frac{1}{n}\sum_{w\in\mathcal{A}^k} N'_{n,w}(0),
\end{aligned}$$

where $N'_{n,w}(0)$ denotes the derivative of $N_n(u)$ at $u = 0$

Let now $D_n(u) = \mathbf{E}(u^{D_n}) = \sum_k \mathbf{P}(D_n = k)u^k$ be the probability generating function of $D_n$. Then above implies

$$D_n(u) = \frac{1}{n}\frac{(1-u)}{u}\sum_{w\in\mathcal{A}^*} u^{|w|}N'_{n,w}(0),$$

and the bivariate generating function $D(z,u) = \sum_n nD_n(u)z^n$ becomes

$$D(z,u) = \frac{1-u}{u}\sum_{w\in\mathcal{A}^*} u^{|w|}\frac{\partial}{\partial u}N_w(z,0) \tag{7.6.1}$$

where $N_w(z,u) = \sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\mathbf{P}(N_n(w) = r)z^n u^r$. In Section 7.2 we worked with $\sum_{n=0}^{\infty}\sum_{r=1}^{\infty}\mathbf{P}(N_n(w) = r)z^n u^r$ and in (7.2.20) of Theorem 7.2.7 we provided a formula for it. Adding the term $N_0(z) = S_w(z)/D_w(z)$ we finally arrive at

$$N_w(z,u) = \frac{z^{|w|}\mathbf{P}(w)}{D_w^2(z)}\frac{u}{1-uM_w(z)} + \frac{S_w(z)}{D_w(z)},$$

where $M_w(z)$ is defined in (7.2.21) and $D_w(z) = (1-z)S_w(z) + z^{|w|}\mathbf{P}(w)$ (cf. 7.2.24) with $S_w(z)$ being the autocorrelation polynomial for $w$. Since

$$\frac{\partial}{\partial u}N_w(z,0) = z^{|w|}\frac{\mathbf{P}(w)}{D_w^2(z)},$$

we finally arrive at the following lemma that is the starting point of the subsequent analysis.

LEMMA 7.6.1. *The bivariate generating function for $D_n$ is*

$$D(z,u) = \frac{1-u}{u} \sum_{w \in \mathcal{A}^*} (zu)^{|w|} \frac{\mathbf{P}(w)}{((1-z)S_w(z) + z^{|w|}\mathbf{P}(w))^2} \tag{7.6.2}$$

*for $|u| < 1$ and $|z| < 1$, where $S_w(z)$ is the autocorrelation polynomial for $w$.*

In this section, we prove the following result for a random text generated by a memoryless source over a finite alphabet $\mathcal{A}$ of size $V$ with $p_i$ being the probability of emitting symbol $i \in \mathcal{A}$. We denote by $h = -\sum_{i=1}^{V} p_i \log p_i$ the entropy rate of the source, and $h_2 = \sum_{i=1}^{V} p_i \log^2 p_i$. The reader is asked in Exercise 7.6.1 to extend below theorem to Markov sources.

THEOREM 7.6.2. (i) *For a biased memoryless source (i.e., $p_i \neq p_j$ for some $i \neq j$) and any $\varepsilon > 0$*

$$\mathbf{E}(D_n) = \frac{1}{h}\log n + \frac{\gamma}{h} + \frac{h_2}{h^2} + P_1(\log n) + O(n^{-\varepsilon}), \tag{7.6.3}$$

$$\mathrm{Var}(D_n) = \frac{h_2 - h^2}{h^3}\log n + O(1) \tag{7.6.4}$$

*where $P_1(\cdot)$ is a periodic function with small amplitude in the case where the tuple $(\log p_1, \ldots,$*
*$\log p_V)$, is collinear with a rational tuple (i.e., $\log p_j / \log p_1 = r/s$ for some integers $r$ and $s$) and converges to zero otherwise.*
*Furthermore, $(D_n - \mathbf{E}(D_n))/\mathrm{Var}(D_n)$ is asymptotically normal with mean zero and variance one that is, for fixed $x \in R$*

$$\lim_{n\to\infty} \mathbf{P}\{D_n \leq \mathbf{E}(D_n) + x\sqrt{\mathrm{Var}(D_n)}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \ ,$$

*and for all integer $m$*

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{D_n - \mathbf{E}(D_n)}{\sqrt{\mathrm{Var} D_n}}\right]^m = \begin{cases} 0 & \text{when } m \text{ is odd} \\ \frac{m!}{2^{m/2}(\frac{m}{2})!} & \text{when } m \text{ is even.} \end{cases}$$

(ii) *For the unbiased source (i.e., $p_1 = \cdots = p_V = 1/V$), $h_2 = h^2$, the expected value $\mathbf{E}(D_n)$ is given by (7.6.3) above, and for any $\varepsilon > 0$*

$$\mathrm{Var}(D_n) = \frac{\pi^2}{6\log^2 V} + \frac{1}{12} + P_2(\log n) + O(n^{-\varepsilon})$$

*where $P_2(\log n)$ is a periodic function with small amplitude The limiting distribution of $D_n$ does not exist, but one finds*

$$\lim_{n\to\infty} \sup_x \mid \mathbf{P}(D_n \leq x) - \exp(-nV^{-x}) \mid = 0$$

*for any fixed real $x$.*

In passing we observe that the quantity $D_n$ is also the depth of a randomly selected suffix in a compact suffix trie. Such a trie is a compacted version of suffix tries defined in Chapter 2. In a compact suffix trie one deletes all unary nodes at the *bottom* of the non-compact suffix trie. Observe that in a compact suffix trie, which we further call simply a suffix trie, the path from the root to node $i$ (representing the $i$th suffix) is the shortest suffix that distinguishes it from all other suffixes. The quantity $D_n(i)$ defined above represents the depth of the $i$ suffix in the associated suffix trie, while $D_n$ is the *typical depth*, that is, the depth of a randomly selected terminal node in the suffix trie. Theorem 7.6.2 tells us that the typical depth is normally distributed with the average depth asymptotically equal to $\frac{1}{h}\log n$ and variance $\Theta(\log n)$ for biased memoryless source. In the unbiased case variance is $O(1)$ and the (asymptotic) distribution is of the extreme distribution type. Interestingly, as proved below, the depth in a suffix trie (built over *one* sequence generated by a memoryless source) is asymptotically equivalent to the depth in a trie built over $n$ *independently* generated strings. Thus suffix tries resemble tries!

### 7.6.2. Random tries resemble suffix tries

The proof of Theorem 7.6.2 hinges on establishing asymptotic equivalence between $D_n$ introduced above and a new random variable $D_n^T$ defined as follows: First, for $n$ *independently* generated texts (by the same memoryless source as for $D_n$) we denote by $D_n^T(i)$ for an integer $i \le n$ the length of the longest prefix of the $i$th text that is also a prefix of another text, say the $j$th text, $j \ne i$. Then the random variable $D_n^T$ is defined by selecting integer $i$ uniformly between 1 and $n$. We also define $D_n^T(u) = \sum_k \mathbf{P}(D_n^T = k)u^k$ and $D^T(z,u) = \sum_n nD_n^T(u)z^n$. Observe that $D_n^T$ is in fact the typical depth in a trie built over these $n$ independent texts.

It is relatively easy to derive the generating function of $D_n^T$, as shown below.

Lemma 7.6.3. *For all* $n \ge 1$

$$D_n^T(u) = \frac{1-u}{u}\sum_{w\in\mathcal{A}^*} u^{|w|}\mathbf{P}(w)(1-\mathbf{P}(w))^{n-1},$$

$$D^T(z,u) = \frac{1-u}{u}\sum_{w\in\mathcal{A}^*} u^{|w|}\frac{z\mathbf{P}(w)}{(1-z+\mathbf{P}(w)z)^2}$$

*for all* $|u| \le 1$ *and* $|z| < 1$.

*Proof.* It suffices to observe that

$$\mathbf{P}(D_n^T(i) < k) = \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)(1-\mathbf{P}(w))^{n-1}.$$

Indeed, $D_n^T(i) < k$ if there is a word $w \in \mathcal{A}^k$ such the a prefix of the $i$th string is equal to $w$ and none of the other text prefixes are equal to $w$. ∎

Our goal now is to prove that $D_n(u)$ and $D_n^T(u)$ are asymptotically close as $n \to \infty$. This requires several preparatory steps outlined below that will lead to

$$D_n^T(u) - D_n(u) = (1-u)O(n^{-\varepsilon}) \tag{7.6.5}$$

for some $\varepsilon > 0$ and all $|u| < \beta$ for $\beta > 1$. Consequently,

$$|\mathbf{P}(D_n \leq k) - \mathbf{P}(D_n^T \leq k)| = O(n^{-\varepsilon}\beta^{-k})$$

for all positive integers $k$. In Lemma 7.6.11 we shall prove that $D_n^T$ is asymptotically normal, hence $D_n$ is normal. This will prove Theorem 7.6.2.

We start with a lemma indicating that for most words $w$ the autocorrelation polynomial $S_w(z)$ is very close to 1 for $z$ non-negative. This lemma provides information about analytical properties of the autocorrelation polynomial.

Lemma 7.6.4. *There exist $\delta < 1$, $\theta > 0$ and $\rho > 1$ such that $\rho\delta < 1$ and*

$$\sum_{w \in \mathcal{A}^k} [\![ |S_w(\rho) - 1| \leq (\rho\delta)^k\theta ]\!] \mathbf{P}(w) \geq 1 - \theta\delta^k. \tag{7.6.6}$$

*Proof.* To simply notations, let $P_k$ be the probability measure on $\mathcal{A}^k$ such that $P_k(A) = \sum_{w\in\mathcal{A}^k} [\![ w \in A ]\!] \mathbf{P}(w)$. Thus we need to prove that $P_k(S_w(\rho) \leq 1 + (\rho\delta)^k\theta) \geq 1 - \theta\delta^k$.

Let $i$ be an integer smaller than $k \in \mathcal{P}(w)$, where $\mathcal{P}(w)$ is the autocorrelation set for $w$. It is easy to see that (cf. Exercise 7.6.2)

$$P_k(k - i \in \mathcal{P}(w)) = \left(\sum_{j=1}^{V} p_j^{\lfloor k/i \rfloor + 1}\right)^r \left(\sum_{j=1}^{V} p_j^{\lfloor k/i \rfloor}\right)^{i-r} \tag{7.6.7}$$

where $r = k - \lfloor k/i \rfloor i$. Denoting $p = \max_i p_i$ we have

$$P_k(k - i \in \mathcal{P}(w)) \leq p^{k-i}.$$

Thus $P_k(\max(\mathcal{P}(w) - \{k\}) \geq k/2) \leq \sum_{i=1}^{\lfloor k/2 \rfloor} P_k(i+1 \in \mathcal{P}(w)) \leq \frac{p^{k/2}}{1-p}$. Now, if the word $w$ is such that $\max(\mathcal{P}(w) - \{k\}) < k/2$, then $S_w(\rho) \leq \sum_{i > \lfloor k/2 \rfloor}^{k} \rho^i p^i \leq \rho^k \frac{p^{k/2}}{1-p}$. Therefore, it suffices for (7.6.6) to select $\delta = \sqrt{p}$, $\theta = (1-p)^{-1}$ and $\rho > 1$ such that $\rho\delta < 1$. ■

In the next lemma we show that $D(z,u)$ can be analytically continued above the unit disk, that is, for $|u| > 1$.

Lemma 7.6.5. *The generating function $D(z,u)$ can be analytically continued for all $|u| < \delta^{-1}$ and $|z| < 1$ where $\delta < 1$.*

*Proof.* Let $|u| < 1$ and $|z| < 1$. Consider the following identity

$$\sum_w (uz)^{|w|} \frac{\mathbf{P}(w)}{(1-z)^2} = \frac{1}{(1-uz)(1-z)^2}.$$

Therefore, for $|z| < 1$

$$uD(z,u) - \frac{(1-u)}{(1-uz)(1-z)^2} = (1-u)\sum_w (zu)^{|w|}\mathbf{P}(w)\left(\frac{1}{D_w^2(z)} - \frac{1}{(1-z)^2}\right)$$
$$= (u-1)\sum_w (zu)^{|w|}\mathbf{P}(w)\frac{1}{D_w^2(z)(1-z)^2}(D_w(z)-(1-z))(D_w(z)+(1-z)),$$

where we recall $D_w(z)-(1-z) = (1-z)(S_w(z)-1)+\mathbf{P}(w)z^{|w|}$. By Lemma 7.6.4

$$P_k(|D_w(z)-(1-z)| \leq (|1-z|+1)\delta^{|w|}) \geq 1 - O(\delta^{|w|})$$

for all $w$ such that $|w| = k$. Moreover, for any bounded function $f(w)$ such that $f(w) \leq f_{max}$ for all $w$ with $|w| = k$, we also have the following estimate for all $y$:

$$\sum_{|w|=k} \mathbf{P}(w)f(w) \leq y + f_{\max}P_k(f(w) > y)\ . \tag{7.6.8}$$

In particular, we take $f(w) = D_w(z) - (1-z)$ and we have $f_{\max} = O(1)$ since $|S_w(z)| < (1-p)^{-1}$ ($p$ as defined as in proof of lemma 7.6.4). Now taking $y = (|1-z|+1)\delta^k$, using the above we obtain

$$uD(z,u) - \frac{1-u}{(1-uz)(1-z)^2} = (u-1)\sum_{k=0}^{\infty}(zu)^k O((|1-z|+1)\delta^k + \delta^k)$$

for all $w$. In conclusion,

$$uD(z,u) - \frac{(1-u)}{(1-uz)(1-z)^2} = O\left(\frac{u-1}{1-\delta|u|}\right)$$

for $\delta < 1$ and $|z| < 1$, which completes the proof. ■

Before we proceed, we need two technical lemmas.

LEMMA 7.6.6. *There exists $K$, a constant $\rho' > 1$ and $\alpha > 0$ such that for all $w$ with $|w| \geq K$ we have*

$$|S_w(z)| \geq \alpha$$

*for $|z| \leq \rho'$ with $\rho' > 1$ such that $p\rho' < 1$.*

*Proof.* Let $\ell$ be an integer and $\rho' > 1$ such that $p\rho' + (p\rho')^\ell < 1$. Let $k > \ell$ and let $w$ such that $|w| > k$. Let $i = \max(\mathcal{P} - \{k\})$. If $i \leq \ell$, then for all $z$ such that $|z| \leq \rho'$ we have

$$|S_w(z)| \geq 1 - \frac{(p\rho')^\ell}{1-p\rho'}.$$

If $i > \ell$, let $q = \lfloor k/i \rfloor$, then $w = u^q v$ where $u$ is the prefix of length $i$ of word $w$, and $v$ is the suffix of length $k - iq$. Thus

$$S_w(z) = \frac{1 - (\mathbf{P}(u)z^i)^q}{1 - \mathbf{P}(u)z^i} + (\mathbf{P}(u)z^i)^q S_{uv}(z),$$

where $S_{uv}(z)$ is the autocorrelation polynomial of $uv$. This implies

$$|S_w(z)| \geq \frac{1 - (p\rho')^{qi}}{1 + (p\rho')^i} - \frac{(p\rho')^{iq} - (p\rho')^k}{1 - p\rho'}.$$

But since $i > \ell$, we obtain

$$|S_w(z)| \geq \frac{1 - (p\rho') - 3(p\rho')^{k-\ell}}{1 + p\rho'} > 0,$$

which completes the proof. ∎

Lemma 7.6.7. *There exists an integer $K'$ such that for $|w| \geq K'$ there is only one root of $D_w(z)$ in the disk $|z| \leq \rho'$ for $\rho' > 1$.*

*Proof.* Let $K_1$ be such that $(p\rho')^{K_1} < \alpha(\rho' - 1)$ holds for the $\alpha$ and $\rho'$ as in Lemma 7.6.6. Denote $K' = \max\{K, K_1\}$, where $K$ is defined above. Note also that the above condition implies that for all $w$ such that $|w| = k > K'$ we have $\mathbf{P}(w)(\rho')^k < \alpha(\rho - 1)$. Hence, for $|w| > K'$ we have $|\mathbf{P}(w)z^k| < |(z-1)S_w(z)|$ on the circle $|z| = \rho' > 1$. Therefore, by Rouché's theorem the polynomial $D_w(z)$ has the same number of roots as $(1 - z)S_w(z)$ in the disk $|z| \leq \rho'$. But, the polynomial $(1-z)S_w(z)$ has only a single root in this disk since by Lemma 7.6.6 we have $|S_w(z)| > 0$ in $|z| \leq \rho'$. ∎

We just establish that there exists the smallest root of $D_w(z) = 0$ that we denote as $A_w$. Let also $C_w$ and $D_w$ be the first and the second derivatives of $D_w(z)$ at $z = A_w$, respectively. Using bootstrapping, one easily obtains the following expansions

$$\begin{aligned}
A_w &= 1 + \frac{1}{S_w(1)}\mathbf{P}(w) + O(\mathbf{P}(w)^2),\\
C_w &= -S_w(1) + \left(k - \frac{2S_w'(1)}{S_w(1)}\right)\mathbf{P}(w) + O(\mathbf{P}(w)^2),\\
E_w &= -2S_w'(1) + \left(k(k-1) - \frac{3S_w''(1)}{S_w(1)}\right)\mathbf{P}(w) + O(\mathbf{P}(w)^2),
\end{aligned}$$

where $S_w'(1)$ and $S_w''(1)$, respectively, denote the first and the second derivatives of $S_w(z)$ at $z = 1$.

Finally, we are ready to compare $D_n(u)$ with $D_n^T(u)$ to conclude that they do not differ too much as $n \to \infty$. Let us define two new generating functions

$Q_n(u)$ and $Q(z,u)$ that represent the difference between $D_n(u)$ and $D_n^T(u)$, that is,

$$Q_n(u) = \frac{u}{1-u}\left(D_n(u) - D_n^T(u)\right),$$

and

$$Q(z,u) = \sum_{n=0}^{\infty} n\, Q_n(u) z^n = \frac{u}{1-u}\left(D(z,u) - D^T(z,u)\right).$$

Then

$$Q(z,u) = \sum_w u^{|w|}\mathbf{P}(w)\left(\frac{z^{|w|}}{D_w(z)^2} - \frac{z}{(1-z+\mathbf{P}(w)z)^2}\right).$$

It is not difficult to establish asymptotics of $Q_n(u)$ by appealing to the Cauchy theorem. This is done in the following lemma.

Lemma 7.6.8. *There exists $B > 1$ such that for all $|u| \le \beta$ the following evaluation holds*

$$Q_n(u) = \frac{1}{n}\sum_w u^{|w|}\mathbf{P}(w)\left(A_w^{|w|-n-1}\left(\frac{n+1-|\mathcal{W}|}{C_w^2 A_w} + \frac{E_w}{C_w^3}\right) - n(1-\mathbf{P}(w))^{n-1}\right) + O(B^{-n})$$

*for some $\beta > 1$.*

*Proof.* By Cauchy's formula

$$nQ_n(u) = \frac{1}{2i\pi}\oint Q(z,u)\frac{dz}{z^{n+1}},$$

where the integration is along a loop contained in the unit disk that encircles the origin. Let $w$ be such that $|w| \ge K'$, where $K'$ is defined in Lemma 7.6.7. From the proof of Lemma 7.6.7 we conclude that $D_w(z)$ and $(1-z+\mathbf{P}(w)z)$ have only one root in $|z| \le \rho$ for some $\rho > 1$. Applying Cauchy's residue theorem we obtain

$$\frac{1}{2i\pi}\oint u^{|w|}\mathbf{P}(w)\frac{dz}{z^{n+1}}\left(\frac{z^{|w|}}{D_w(z)^2} - \frac{z}{(1-z+\mathbf{P}(w)z)^2}\right) =$$

$$= u^{|w|}\mathbf{P}(w)\left(\frac{A_w^{|w|-n-1}}{u}\left(\frac{n+1-|w|}{C_w^2 A_w} + \frac{E_w}{C_w^3}\right) - n(1-\mathbf{P}(w))^{n-1}\right) + I_w(\rho,u),$$

where

$$I_w(\rho,u) = \frac{\mathbf{P}(w)}{2i\pi}\int_{|z|=\rho} u^{|w|}\frac{dz}{z^{n+1}}\left(\frac{z^{|w|}}{D_w(z)^2} - \frac{z}{(1-z+\mathbf{P}(w)z)^2}\right).$$

To establish a bound for $I_w(\rho,u)$ we argue exactly in the same manner as in the proof of Theorem 7.6.5. This leads for $|w| > K'$ to

$$\sum_{|w|=k} I_w(\rho,u) = O((\delta\rho u)^k \rho^{-n})$$

since for all $w$ we also have $S_w(\rho) \leq 1/(1-p\rho)$ and $D_w(z) = O(\rho^k)$ in the circle $|z| \leq \rho$. Set now $\beta = (\delta\rho)^{-1} > 1$. Then, for $|u| < \beta$ we have

$$\sum_{\{w:\ |w|>K'\}} I_w(\rho, u) = O(\sum_w \mathbf{P}(w)\rho^{|w|-n}) = O(\rho^{-n}).$$

This proves our bound since the other terms ($|w| < K'$) contribute only $B^{-n}$ for some $B > 1$ due to the fact that all roots of $D_w(z)$ have magnitudes greater than 1. ∎

In the next lemma we show that $Q_n(u) \to 0$ as $n \to \infty$.

LEMMA 7.6.9. *For all* $1 < \beta < \delta^{-1}$, *there exists* $\varepsilon > 0$ *such that* $Q_n(u) = (1-u)O(n^{-\varepsilon})$ *uniformly for* $|u| \leq \beta$.

*Proof.* The expansion of $E_w$ with respect to $\mathbf{P}(w)$, and Lemma 7.6.4 show that as $n \to \infty$ the following holds $\sum_w u^{|w|}\mathbf{P}(w)A_w^{-n}E_w/C_w^3 = O(1)$. Therefore, by Lemma 7.6.8 we have

$$Q_n(u) = \sum_w u^{|w|}\mathbf{P}(w)\left(\frac{A_w^{|w|-n-2}}{C_w^2} - (1-\mathbf{P}(w))^{n-1}\right) + O(1/n)\ .$$

Let now $f_w(x)$ be a function defined for $x$ real by

$$f_w(x) = \frac{A_w^{|w|-x-2}}{C_w^2} - (1-\mathbf{P}(w))^{x-1}.$$

By the same arguments as used in proving (7.6.8) in Theorem 7.6.5, we note that $\sum_w u^{|w|}\mathbf{P}(w)f_w(x)$ is absolutely convergent for all $x$ and $u$ such that $|u| \leq \beta$. The function $\bar{f}_w(x) = f_w(x) - f_w(0)e^{-x}$ is exponentially decreasing when $x \to +\infty$ and is $O(x)$ when $x \to 0$; therefore its Mellin transform defined as

$$\bar{f}_w^*(s) = \int_0^\infty \bar{f}_w(x)x^{s-1}dx$$

is well defined for $\Re(s) > -1$. In this region we obtain

$$\bar{f}_w^*(s) = \Gamma(s)\left(A_w^{|w|-1}\frac{(\log A_w)^{-s}-1}{A_wC_w^2} - \frac{(-\log(1-\mathbf{P}(w))^{-s}-1}{1-\mathbf{P}(w)}\right)\ ,$$

where $\Gamma(s)$ is the gamma function. Let $g^*(s,u)$ be the Mellin transform of the series $\sum_w u^{|w|}\mathbf{P}(w)\bar{f}_w(x)$ which exists at least in the strip $(-1,0)$. Formally, we have

$$g^*(s,u) = \sum_w u^{|w|}\mathbf{P}(w)\bar{f}_w^*(s)\ .$$

We can reverse the Mellin transform $g^*(s,u)$ provided that the following holds.

LEMMA 7.6.10. *The function* $g^*(s,u)$ *is analytical in* $\Re(s) \in (-1,c)$ *for some* $c > 0$.

Assuming Lemma 7.6.10 is granted, we have

$$Q_n(u) = \frac{1}{2i\pi}\int_{\varepsilon-i\infty}^{\varepsilon+i\infty} g^*(s,U)n^{-s}ds + O(1/n) \ + \sum_w u^{|w|}\mathbf{P}(w)f_w(0)e^{-n},$$

for some $\varepsilon \in (0,c)$. Notice that the last term of the above contributes $O(e^{-n})$, and can be safely ignored. Furthermore, a simple majorization under the integral gives the evaluation $Q_n(u) = O(n^{-\varepsilon})$ which completes the proof. ∎

*Proof of Lemma 7.6.10*: We establish the absolute convergence of $g^*(s,u)$ for all $s$ such that $\Re(s) \in (-1,c)$ and $|u| \le \beta$. Let us define $h^*(s,u) = \frac{g^*(s,u)}{\Gamma(s)}$. Note that for any fixed $s$ we have the following

$$(\log A_w)^{-s} = \left(\frac{\mathbf{P}(w)}{1+S_w(1)}\right)^{-s}(1+O(\mathbf{P}(w))) \ ,$$
$$(-\log(1-\mathbf{P}(w)))^{-s} = \mathbf{P}(w)^{-s}(1+O(\mathbf{P}(w))) \ .$$

Thus

$$\frac{(\log A_w)^{-s}-1}{A_w^{2-|w|}C_w^2} - \frac{(-\log(1-\mathbf{P}(w)))^{-s}-1}{1-\mathbf{P}(w)} =$$
$$= \mathbf{P}(w)^{-s}\left[(1+a_w(1))^s(1+O(|w|\mathbf{P}(w)) - (1+O(\mathbf{P}(w))\right] + O(|w|\mathbf{P}(w)) \ .$$

By Lemma 7.6.4, $P_k(S_w(1) \le 1+\theta\delta^k) \ge 1 - O(\delta^k)$, and hence

$$h^*(s,u) = \sum_{k=0}^{\infty}\left(\sup\{p^{-\Re(s)}, q^{-\Re(s)}\}|u|\delta\right)^k O(1)$$

that absolutely converges for all values of $s$ such that $\Re(s) < c$ where $c$ satisfies $\sup\{p^{-c}, q^{-c}\} < (\delta\beta)^{-1}$. Since $h^*(0,u) = 0$ by definition, the pole of $\Gamma(s)$ at $s = 0$ is canceled in $g^*(s,u)$, and therefore $h^*(s,u)$ does not show any singularities in the strip $\Re(s) \in (-1,c)$. ∎

To complete the proof of our main Theorem 7.6.2, we need an asymptotic analysis of $D_n^T(u)$ which is presented next. We recall that $D_n^T$ represents also the typical depth in a trie built from $n$ independently generated strings.

Lemma 7.6.11. *There exists $\varepsilon > 0$ such that*

$$D_n^T(u) = (1-u)n^{\kappa(u)}(\Gamma(\kappa(u)) + P(\log n, u))) + O(n^{\varepsilon}),$$

*where*

$$u\sum_{i=1}^{V} p_i^{1-\kappa(u)} = 1$$

*and $P(\log n, u)$ is periodic function with small amplitude in the case where the vector $(\log p_1,$
$\ldots, \log p_V)$ is collinear with a rational tuple, and converges to zero when $n \to \infty$ otherwise.*

*Proof.* We begin with the identity

$$D_n^T(u) = \frac{1-u}{u} \sum_{w \in \mathcal{A}^*} u^{|w|} \mathbf{P}(w)(1-\mathbf{P}(w))^{n-1}.$$

We argue exactly in the same manner as in the proof of Lemma 7.6.8. We find the Mellin transform $T^*(s,u) = \int_0^\infty x^{s-1} dx u/(1-u) D_x^T(u) dx$ to be

$$T^*(s,u) = \sum_{w \in \mathcal{A}^*} u^{|w|} \mathbf{P}(w)(-\log(1-\mathbf{P}(w)))^{-s} \Gamma(s).$$

Using the fact that for $s$ bounded $(-\log(1-\mathbf{P}(w)))^{-s} = \mathbf{P}(w)^{-s}(1+O(s\mathbf{P}(w)))$, we conclude

$$T^*(s,u) = \Gamma(s) \left( \frac{u \sum_{i=1}^V p_i^{1-s}}{1 - u\sum_{i=1}^V p_i^{1-s}} + g(s,u) \right),$$

where

$$g(s,u) = O\left( \frac{us \sum_{i=1}^V p_i^{2-\Re(s)}}{1-|u|\sum_{i=1}^V p_i^{2-\Re(s)}} \right).$$

Let $\kappa(u)$ be the main root of $1 = u\sum_{i=1}^V p_i^{1-s}$. The other roots of $1 = u\sum_{i=1}^V p_i^{1-s}$, are countable and we denote them as $\kappa_k(u)$ for $k \neq 0$ integer. For all integers $k$ we have $\Re(\kappa_k(u)) \geq \kappa(u)$. Using the inverse Mellin we find

$$D_n^T = \frac{1-u}{2i\pi u} \int_{-i\infty}^{+i\infty} T^*(s,u) n^{-s} ds.$$

We now consider $|u| < \delta^{-1}$ for $\delta < 1$. Then there exists $\varepsilon$ such that for $\Re(s) \leq \varepsilon$ the function $g(s,u)$ has no singularity. Moving the integration path to the left of $\Re(s) = \varepsilon$, and applying the reside theorem we find the following estimate

$$D_n^T(u) = (1-u)\frac{\Gamma(\kappa(u))}{h(u)} n^{\kappa(u)} + (1-u)\sum_k \frac{\Gamma(\kappa_k(u))}{h_k(u)} n^{\kappa_k(u)} + O(n^{-\varepsilon}) \quad (7.6.9)$$

with $h(u) = -\sum_i p_i^{1-\kappa(u)} \log p_i$ and $h_k(u) = -\sum_i p_i^{1-\kappa_k(u)} \log p_i$. When $\log p_i$'s are collinear with a rational vector, then there is subset of $\kappa_k(u)$ that have the same real part as $\kappa(u)$ and also equally spaced on the vertical line $\Re(s) = \Re(\kappa(u))$. In this case their contribution to (7.6.9) is

$$n^{\kappa(u)} \sum_k \frac{\Gamma(\kappa_k(u))}{h(u)} \exp((\kappa_k(u) - \kappa(u)) i \log n).$$

When the $\log p_i$'s are not collinear with a rational vector the contribution of the $\kappa_k(u)$ divided by $n^{\kappa(u)}$ converges to zero when $n \to \infty$. ∎

The last lemma completes the proof of Theorem 7.6.2. Indeed, it suffices to observe that for $t \to 0$

$$\kappa(e^t) = c_1 t + \frac{c_2}{2} t^2 + O(t^3) \quad (7.6.10)$$

where $c_1 = 1/h$ and $c_2 = (h_2 - h^2)/h^3$. We concentrate first on the asymmetric case. From the expression of $D_n^T(u)$ we find immediately the first and the second moments via the first and the second derivatives of $D_n^T(u)$ at $u = 1$ with the appropriate asymptotic expansion in $c_1 \log n$ and in $c_2 \log n$. In order to obtain the limiting normal distribution we prove

$$e^{-tc_1 \log n/\sqrt{c_2 \log n}} D_n^T \left(e^{t/\sqrt{c_2 \log n}}\right) \to e^{t^2/2}$$

using $n^{\kappa(u)} = \exp(\kappa(u) \log n)$ and referring to expansion (7.6.10).

For the symmetric case there is no normal limiting distribution since variance is $O(1)$. However, there are oscillation due to the fact that all $\kappa_k(u)$ are aligned on a vertical line. This completes the proof of Theorem 7.6.2.

## Problems

*Section 7.2*

7.2.1 Prove (7.2.9).

7.2.2 In Theorem 7.2.8 we prove that for irreducible aperiodic Markov chain the variance $\mathrm{Var}(N_n) = nc_1 + c_2$ (cf. (7.2.26)). Prove that $c_1 > 0$.

7.2.3 Prove that $(N_n - \mathbf{E}(N_n))/\sqrt{\mathrm{Var}(N_n)}$ converges in moments to the appropriate moments of the standard normal distribution.

7.2.4 Let $\rho(t)$ be a root of $1 - e^t M_{\mathcal{W}}(e^\rho) = 0$. Observe that $\rho(0) = 0$. Prove that $\rho(t) > 0$ for $t \neq 0$ for $p_{ij} > 0$ for all $i, j \in \mathcal{A}$.

7.2.5 Prove the expression (7.2.44) for $\theta_a$ of Theorem 7.2.12 (cf. Denise and Régnier (2004)).

*Section 7.3*

7.3.1 Extend the analysis of Section 7.3 to multisets $\mathcal{W}$, that is, a word $w_i$ may occur several times in $\mathcal{W}$.

7.3.2 Prove language relationships (7.3.2)–(7.3.2).

7.3.3 Derive explicit formulas for $\theta_a$ appearing in Theorem 7.3.3(iv).

7.3.4 Find explicit formulas for the values of the mean $\mathbf{E}(N_n(\mathcal{W}))$ and of the variance $\mathrm{Var}(N_n(\mathcal{W}))$ for the generalized pattern matching discussed in Section 7.3.2 for $\mathcal{W}_0 = \emptyset$ and $\mathcal{W}_0 \neq \emptyset$.

7.3.5 Derive explicit formulas for $\sigma_a$ and $\theta_a$ in (7.3.27) appearing in Theorem 7.3.10.

7.3.6 Enumerate $(\ell, k)$ sequences over a non binary alphabet (i.e., generalize the analysis of Section 7.3.3).

*Section 7.4*

7.4.1 Find an explicit formula for the generating function $B_2^{[p]}(z)$ of the collection $\mathcal{B}_2^{[p]}$.

7.4.2 Design a dynamic programming algorithm to compute the correlation algorithm, $\kappa^2(\mathcal{W})$.

7.4.3 Establish the rate of convergence for the Gaussian law from Theorem 7.4.5.

7.4.4 For the fully unconstrained subsequence problem establish the large deviations (cf. Janson (pear)).

7.4.5 Provide details of the proof for Theorem 7.4.6.

7.4.6 Let $\mathcal{W} = \{w_1, \ldots, w_d\}$ be a set of patterns $w_i$. The pattern $\mathcal{W}$ occurs as a subsequence in the text if any of $w_i$ occurs as a subsequence. Analyze this generalization of the subsequence pattern matching.

7.4.7 Let $w$ be a pattern. Set $W$ to be a window size with $|w| \leq W \leq n$. Consider the *windowed subsequence pattern matching* in which $w$ must appear as a subsequence within the window $W$. Analyze the number of windows that has at least one occurrence of $w$ as a subsequence within the window (cf. Gwadera, Atallah, and Szpankowski 2003).

*Section 7.5*

7.5.1 Prove the generating operators identities (7.5.5)–(7.5.8).

7.5.2 Prove (7.5.11)–(7.5.13).

7.5.3 Prove the second part of Theorem 7.5.1, that is, derive formula (7.5.18) for variance of $\Omega_n(\mathcal{W})$.

7.5.4 Does the central limit theorem holds for the generalized subsequence problem discussed in Section 7.5? What about large deviations?

*Section 7.6*

7.6.1 Extend Theorem 7.6.2 for Markov sources.

7.6.2 Prove (7.6.7) and extend it to Markov sources (cf. Apostolico and Szpankowski 1992).

7.6.3 Let $\kappa(u)$ be the main root of $1 = u\sum_{i=1}^{V} p_i^{1-s}$, and $\kappa_k(u)$ for $k \neq 0$ integer are other roots of $1 = u\sum_{i=1}^{V} p_i^{1-s}$. Prove that for all integers $k$ we have $\Re(\kappa_k(u)) \geq \kappa(u)$.

## Notes

Algorithmic aspects of pattern matching are presented in numerous books. We mention here Crochemore and Rytter (1994) and Gusfield (1997) (cf. also Apostolico (1985)). Public domain utilities like `agrep`, `grappe`, `webglimpse` for finding general patters were recently developed by Wu and Manber (1995), Kucherov and Rusinowitch (1997), and others. Various data compression schemes are studied in Wyner and Ziv (1989), Wyner (1997), Yang and Kieffer (1998), Ziv and Lempel (1978), Ziv and Merhav (1993)). Prediction based on pattern matching is discussed in Jacquet, Szpankowski, and Apostol (2002). Algorithmic aspect of pattern matching can also be found in Chapter 2 and Chapter 8 of this book.

In this chapter the emphasis is on analysis of pattern matching problems by analytic methods in a probabilistic framework. Probabilistic models are discussed in Section 7.1 and Chapter 1. Markov models are presented in many standard books (cf. Karlin and Ost (1987)). Dynamic sources were introduced by Vallée (2001) (cf. also Clement, Flajolet, and Vallée (2001), Bourdon and Vallée (2002)). General stationary ergodic sources are discussed in Shields (1969).

In this chapter analytic tools are used to investigate combinatorial pattern matching problems. The reader is referred to Alon and Spencer (1992), Szpankowski (2001), Waterman (1995) (cf. also Arratia and Waterman (1989, 1994)) for in-depth discussion of probabilistic tools. Analytic techniques are thoroughly explained in Sedgewick and Flajolet (1995) and Szpankowski (2001). The reader may also consult Atallah, Jacquet, and Szpankowski (1993), Bender (1973), Clement et al. (2001), Hwang (1996), Jacquet and Szpankowski (1994, 1998). The Perron–Frobenius theory and the spectral decomposition of matrices can be found in Gantmacher (1959), Karlin and Taylor (1975), Kato (1980), Szpankowski (2001). Operator theory is discussed in Kato (1980).

Exact string matching is presented in Section 7.2. There are numerous references. Our approach is founded in the work of Guibas and Odlyzko (1981a) and Guibas and Odlyzko (1981b). The presentation of this section follows very closely recent work of Régnier and Szpankowski (1998a) and Régnier (2000). More probabilistic approach is adopted in Chapter 2 and in Prum et al. (1995). Example 7.2.13 is taken from Denise, Régnier, and Vandenbogaert (2001).

Generalized string matching problem discussed in Section 7.3 was introduced in Bender and Kochman (1993). The analysis of string matching over reduced set of patterns appears in Régnier and Szpankowski (1998b) (cf. also Guibas and Odlyzko (1981b)). An automaton approach to motif finding was proposed in Nicodème et al. (2002). The general string matching was first dealt with in Bender and Kochman (1993), however, our presentation follows a different path simplifying previous analyses. It is closely related to the subsequence pattern matching analysis presented in Flajolet, Guivarc'h, Szpankowski, and Vallée (2001). The $(\ell, k)$ sequence analysis is taken from Szpankowski (2001).

The subsequence pattern matching or the hidden pattern matching discussed in Section 7.4 is based on Flajolet et al. (2001). Proceeding along different tracks, Janson (pear) has related this particular case to his treatment of $U$–statistics via Gaussian Hilbert spaces; see Chapter XI of Janson's book Janson (1997) for the type of method employed. Example 7.4.7 was fully developed in Gwadera et al. (2003).

The generalized subsequence pattern matching discussed in Section 7.5 is taken from Bourdon and Vallée (2002). The operator generating function approach for dynamic sources was developed by Vallée (2001).

In Section 7.6 we present some results for the self-repetitive pattern matching. Theorem 7.6.2 was proved in Jacquet and Szpankowski (1994), however, our proof in this section is somewhat simplified. In particular, proof of the crucial Lemma 7.6.1 is new and based on results presented in Section 7.2. Lemma 7.6.11 is due to Jacquet and Régnier (1986) (for an extension to Markov sources see Jacquet and Szpankowski (1991)). Mellin transform is explained in depth in Flajolet, Gourdon, and Dumas (1995), Szpankowski (2001). Tries are treated in depth in Mahmoud (1992) and Szpankowski (2001). As mentioned, the quantity $D_n$ analyzed in the section is also the typical depth in a suffix tries introduced in Chapter 2 (cf. also Apostolico (1985)). Probabilistic analysis of suffix tries can be found in Apostolico and Szpankowski (1992), Devroye, Szpankowski, and Rais (1992), Szpankowski (1993a, 1993b). As discussed in the section, suffix tries are often appear in analysis of data compression schemes (cf. Wyner and Ziv (1989), Wyner (1997), Yang and Kieffer (1998), Ziv and Lempel (1978), Ziv and Merhav (1993)).