COMBINATORIAL METHODS FOR COUNTING PATTERN OCCURRENCES IN A MARKOVIAN TEXT

by

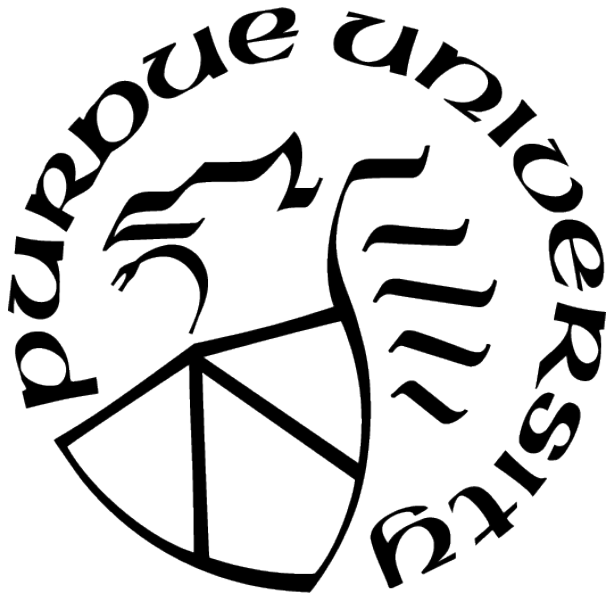
Yucong Zhang

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



Department of Statistics

West Lafayette, Indiana

December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Mark Daniel Ward, Chair

Department of Statistics

Dr. Takashi Owada

Department of Statistics

Dr. Thomas M. Sellke

Department of Statistics

Dr. Wojciech Szpankowski

Department of Computer Science

Approved by:

Dr. Jun Xie

To my loving family.

ACKNOWLEDGMENTS

First of all, I would like to send my appreciation to my academic advisor, Professor Mark Daniel Ward. You opened the door to the amazing topics of combinatorics. I have learned a lot from you, academically and personally. You are a great mentor and friend. I appreciate your tremendous support, encouragement and patience during my Ph.D. years. It has been a great honor and pleasure working with you. Thank you, Dr. Ward.

I would like to thank Professor Szpankowski, Professor Sellke, and Professor Owada, for serving on my Ph.D. committee.

My special thanks are to Professor Szpankowski, for his valuable expert advice on my research topic, and for the support from the Center for Science of Information. I would like to thank Brent Ladd and Bob Brown, for supporting me and offering me resources from the center.

Last but not the least, I would like to thank my family. I am grateful to my parents, Gang Zhang and Xiaogui Zhang, who have raised me up with endless support and love. My deepest gratitude is to my wife, Qing, whose love and trust have been continuously motivating me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

ϵ	Symbol of the empty word
$\mathcal{A}$	The set of letters in the alphabet. By convention, an alphabet set composed of ℓ letters is $\mathcal{A} = \{a_1, a_2, \cdots, a_\ell\}$; if $\ell = 2$, then $\mathcal{A} = \{a, b\}$.
$\mathcal{A}^*$	The set of all words of finite length on the alphabet $\mathcal{A}$; The set $\mathcal{A}^*$ includes the empty word ϵ
$\mathcal{U}$	A set of pattern words. A set of patterns composed of r pattern words is $\mathcal{U} = \{u_1, u_2, \cdots, u_r\}$.
w	A word (text)
$\lvert w \rvert$	The length of a word w
$\lvert w \rvert_{u_k}$	(also $\lvert w \rvert_k$) The number of occurrences of the pattern u_k in the word w, where $k \in \{1, 2, \ldots, r\}$. When no ambiguity exists, $\lvert w \rvert_{u_k}$ can be simplified to $\lvert w \rvert_k$. In particular, $\lvert w \rvert_{u_k} = 1$ if $w = u_k$.
u	The decorated word of a word u. Any words in Sans Serif font are decorated (either monocolored or bicolored, depending on the context).
$\widehat{u}$	The last letter of a word u
$\underline{\mathsf{c}}$	(also Skel(c)) The skeleton of a cluster c.
$\widetilde{\underline{\mathsf{c}}}$	(also Flip($\underline{\mathsf{c}}$)) The Flip of a skeleton $\underline{\mathsf{c}}$. It is a set of clusters which share the same skeleton $\underline{\mathsf{c}}$, and can also be represented by a fully bicolored decorated text.
$[\,\cdot\,]$	The square bracket displays the letter or word immediately before the word of interest. e.g., $[a]w$ signifies that the word w starts after a letter a.
$\pi(w)$	(also or π_w) The weight of a word w
${}^a\pi(w)$	(also ${}^a\pi_w$) The weight of a word w, which starts after the letter a. (Used in Markov processes)

p_{ij} (also $p_{i,j}$, or $p(i,j)$) The transition probability from letter i to letter j, in a Markovian stochastic process of order 1

$P_{(\ell \times \ell)}$ The transition matrix in a Markov order 1 stochastic process, where ℓ is the size of the alphabet

$\mathbf{E}(X_n)$ The expectation of a random variable X_n

$[\![\cdot]\!]$ Iverson indicator notation. $[\![S]\!] = \begin{cases} 0 & \text{if } S \text{ is false} \\ 1 & \text{if } S \text{ is true} \end{cases}$

ABSTRACT

In this dissertation, we provide combinatorial methods to obtain the probabilistic multivariate generating function that counts the occurrences of patterns in a text generated by a Markovian source. The generating function can then be expanded into the Taylor series in which the power of a term gives the size of a text and the coefficient provides the probabilities of all possible pattern occurrences with the text size. The analysis is on the basis of the inclusion-exclusion principle to pattern counting (Goulden and Jackson, 1979 and 1983) and its application that Bassino et al. (2012) used for obtaining the generating function in the context of the Bernoulli text source. We followed the notations and concepts created by Bassino et al. in the discussion of distinguished patterns and non-reduced pattern sets, with modifications to the Markovian dependence. Our result is derived in the form of a linear matrix equation in which the number of linear equations depends on the size of the alphabet. In addition, we compute the moments of pattern occurrences and discuss the impact of a Markovian text to the moments comparing to the Bernoulli case. The methodology that we use involves the inclusion-exclusion principle, stochastic recurrences, and combinatorics on words including probabilistic multivariate generating functions and moment generating functions.

1. LITERATURE REVIEW

The study of pattern occurrence problems can be traced back to the late 1970s and early 1980s, when Guibas and Odlyzko established the foundations of the analysis from a combinatorial viewpoint in their two publications in 1981 [19] [18]. They also introduced the notion of correlations between words in order to represent how a word overlaps with another. Another foundational work was done by Goulden and Jackson in 1979 [17] and 1983 [16]. They introduced a very powerful method to count pattern occurrences in a text, when the pattern set is reduced. In a reduced set of patterns, no pattern word is a substring of another pattern word. Their method is referred to as the inclusion-exclusion method (or cluster method in some publications), and characterized by counting patterns in which some occurrences are labeled.

A very detailed publication by Apostolico [2] in 1985 introduced the use of a suffix tree in computer science. A suffix tree plays a core role in pattern matching and has been used in some exceptional algorithms, such as Knuth-Morris-Pratt [22] for string searching and Lempel-Ziv 77 [33] for lossless data compression. As the research of pattern matching problems became more important and useful in computer science, more people started working on this problem.

Several authors studied the topic of no pattern occurrences during 1980s and early 1990s. In Blom et al. [6] and Breen et al. [7], this problem was considered in a probabilistic approach by using probability generating functions, in which one is more interested in the probability of certain occurrences, rather than the enumeration. Gerber et al. [14] treated the question in a inspired viewpoint through martingale argument in addition to general theory of Markov chain.

In 1991, Chrysaphinou et al. [8] developed the independent case of Guibas and Odlyzko. They assumed that the letters in a text are not generated independently. Instead, the letters are generated by a source that follows Markov of order 1 dependence. They also studied the generating function of waiting time for multiple patterns, and provided results that no pattern occurs in a text produced by a Markovian source.

In 1995, Prum et al. [28] considered the Markov model, and provided the limiting distribution of pattern occurrences but with no precise variance computation. By using a martingale approach and a conditional approach, they proposed two asymptotically standard normal statistics to find and classify words when a first-order Markov chain model is assumed.

In 1997, Flajolet et al. [12] used a bivariate generating function which follows asymptotics of singularity perturbation, and studied the number of pattern occurrences in a random binary search tree (BST). Their results demonstrate that on average, the frequency of the occurrence of any specific pattern is proportional to the size of a randomly grown BST. In contrast, the probability of BSTs with forbidden patterns is exponentially small. The small probability can be described in the form of Bessel functions.

Since the late 1990s, on the foundation of the results in nearly 20 years, three approaches were developed quite independently in the study of the pattern occurrences problem.

1.1 Inclusion-exclusion

The aforementioned inclusion-exclusion counting method was developed by Goulden and Jackson. It is usually in the form of a multivariate generating function, in which each pattern can be tracked by a formal parameter. In 1999, Noonan and Zeilberger [26] extended the inclusion-exclusion method beyond the reduced set of patterns and solved the general non-reduced case.

The method of inclusion-exclusion is then formalized by Bassino et al. [3][4][5]. They introduced new notations and concepts, and presented a state-of-the-art approach so that the general counting problem can be solved by obtaining the multivariate generating function. Their approach can be applied to Bernoulli texts with reduced or non-reduced pattern sets.

This thesis can be considered an extension of Bassino et al.'s method to Markovian sources, in a probabilistic point of view.

1.2 Pattern frequency occurrences

The univariate analysis by Guibas and Odlyzko [19][18] was further extended to the multivariate case by Szpankowski, Régnier and Jacquet [21][30][31]. In the case of the re-

duced pattern set, their studies of pattern correlation enables handling overlapping patterns precisely, and counting several patterns simultaneously.

Building on the these results, Vallée [32] applied the previous analysis to dynamical sources. Lothaire [23, Chapter 7], a pseudonym adopted to represent a group of authors, considered a Markovian source on the symbol emission.

The methodology is important in the application on data structures, and particularly useful in analyzing the complexity of tries and trees. Subsequent works include Jacquet and Szpankowski [20], Régnier and Denise [29], Fayolle and Ward [11], Gheorghiciuc and Ward [15], and Park, Hwang, Nicodème, and Szpankowski [27].

1.3 Automaton

The algorithm of automaton was originally invented by Aho and Corasick [1] in order to keep track the occurrences of finite patterns in a text. In 2002, Nicodème et al. [25] provided an algorithm to construct a "marked automata" which can identify one regular expression for either the Bernoulli or Markovian case. This algorithm was then extended by Nicodème [24] for the purpose of counting multiple patterns simultaneously and handling pattern matching with errors. The classical Aho-Corasick automaton [1] can also be used when the pattern set is finite.

For an algorithm treatment of pattern matching, see Crochemore et al. [10] [9].

2. DEFINITIONS AND NOTATIONS

In this chapter, we introduce the notations and fundamental concepts that are used throughout all chapters.

2.1 Basic definitions

Alphabet. Let $\mathcal{A}$ be the alphabet consisting of all letters under consideration. In general, an alphabet is a set of ℓ letters denoted as $\mathcal{A} = \{a_1, a_2, \ldots, a_\ell\}$. A two-element alphabet is called a binary alphabet, e.g. $\mathcal{A} = \{a, b\}$ or $\mathcal{A} = \{0, 1\}$. We will use a binary alphabet in most of our examples for simplicity.

Pattern words. The set of pattern words is denoted by $\mathcal{U} = \{u_1, u_2, \ldots, u_r\}$, where r is finite. Each pattern word in $\mathcal{U}$ is distinct. If there is only one pattern word, we omit the subscript index and refer to the pattern word as u.

The pattern set $\mathcal{U}$ is *reduced* if no pattern in $\mathcal{U}$ is a factor (i.e., a substring) of another. Otherwise the pattern set is *non-reduced*.

Occurrences. For a given pattern set $\mathcal{U} = \{u_1, u_2, \ldots, u_r\}$ and a text w of length $|w|$, we follow the definition given by Bassino et al. [5] and define a sequence of occurrences index $\mathcal{O} = (\mathcal{O}_i)_{i=1}^{|w|}$ as

$$\mathcal{O}_i := \{j \mid u_j \text{ has an occurrence ending at position } i \text{ of } w\}$$

Example 2.1.1 Given pattern words $\mathcal{U} = \{u_1 = abb, u_2 = bba, u_3 = aabb\}$ and a text $w = aabbaaabbabbbba$ (therefore, $|w| = 15$), the sequence of occurrence index $\mathcal{O}$ is

$$\mathcal{O}_i = \begin{cases} \{1\} & \text{if } i \in \{12\} \\ \{1, 3\} & \text{if } i \in \{4, 9\} \\ \{2\} & \text{if } i \in \{5, 10, 15\} \\ \emptyset & \text{otherwise} \end{cases}$$

Therefore, all occurrences are recognized by $(\mathcal{O}_i)_{i=1}^{15}$.

In Example 2.1.1, all occurrences are recognized and recorded in the non-empty sets from $(\mathcal{O}_i)_{i=1}^{15}$. However, when needed, we can select a subset of the occurrences from $(\mathcal{O}_i)_{i=1}^{n}$. The occurrences that are selected are referred to as *distinguished* occurrences. The distinguished occurrence indices are denoted by $\mathcal{D} := (\mathcal{D}_i)_{i=1}^{|w|}$ such that $\mathcal{D}_i \subseteq \mathcal{O}_i$ for all $1 \leq i \leq |w|$.

Decorated text. Given a pattern set $\mathcal{U}$ and a text w of length $|w|$, a decorated text w is a pair $\mathsf{w} = (w, \mathcal{D})$, where $\mathcal{D}$ specifies which occurrences are *distinguished* (i.e., recognized). The text w is referred to as the support of w [5].

Example 2.1.2 We select a subset of $(\mathcal{O}_i)_{i=1}^{15}$ from Example 2.1.1, and use the notation $\mathcal{D} = (\mathcal{D}_i)_{i=1}^{15}$ to denote the subset as follows

$$\mathcal{D}_i = \begin{cases} \{1\} & \text{if } i \in \{4\} \\ \{1,3\} & \text{if } i \in \{9\} \\ \{2\} & \text{if } i \in \{15\} \\ \emptyset & \text{otherwise} \end{cases}$$

A straightforward visual way to represent a decorated text is to label the pattern indices $\mathcal{D}_i$ above the letter at position i. In this way, the aforementioned decorated text $\mathsf{w} = (w, \mathcal{D})$ is then represented as

$$\overset{\textcircled{1}}{a}\overset{\textcircled{1}}{a}\overset{}{b}\overset{}{b}\overset{}{a}\overset{}{a}\overset{}{a}\overset{\textcircled{1}}{b}\overset{\textcircled{3}}{b}\overset{}{a}\overset{}{b}\overset{}{b}\overset{}{b}\overset{}{b}\overset{\textcircled{2}}{a} \tag{2.1}$$

A text is *fully decorated* when $\mathcal{D} = \mathcal{O}$. The fully decorated text in Example 2.1.1 is

$$\overset{\textcircled{1}}{a}\overset{\textcircled{1}}{a}\overset{}{b}\overset{}{b}\overset{}{a}\overset{}{a}\overset{}{a}\overset{\textcircled{1}}{b}\overset{\textcircled{3}}{b}\overset{}{a}\overset{}{b}\overset{}{b}\overset{}{b}\overset{}{b}\overset{\textcircled{2}}{a} \tag{2.2}$$

From (2.2) we can see there are 8 occurrences in the fully decorated text. In fact, there are a total of $2^8 = 256$ decorated texts sharing the same support *aabbaaabbabbbba*. The examples (2.1) and (2.2) are two of the 256 possible decorated texts.

The idea of decorated texts was first introduced in Bassino et al. [5, Section 4]. It is a very succinct and efficient way to display the distinguished occurrences of patterns by marking numbers above the corresponding letters.

We note that if a fully decorated text has exactly k occurrences of patterns, then there are 2^k decorated texts associated with it. It is because each of the k occurrences in the fully decorated text could be distinguished or not. Two decorated texts with the same support but decorated differently are considered distinct. Therefore, there are a total of 2^k decorated texts sharing the same support with the fully decorated text.

2.2 Probabilistic models of texts

We consider a text w of length $n := |w|$ a substring $\{X_k\}_{k=j}^{j+n-1}$ from a one-sided infinite sequence of random variables $\{X_k\}_{k=1}^{\infty}$, where each random variable X_k is a letter generated over a pre-defined alphabet $\mathcal{A}$, with a probability measure.

Weights. The weight of a word $w \in \mathcal{A}^*$ is denoted by $\pi(w)$ or π_w. As our focuses are the probabilities of pattern word occurrences, the weight of w is the probability of w in the model, namely, $\pi(w) = \Pr(w)$. Notice that in a enumerative model (counting the number of occurrences), we instead use $\pi(w) = 1$.

Assuming $w = x_j x_{j+1} \cdots x_{j+n-1}$, the weight of w, representing its probability mass, is given by

$$\begin{aligned} \pi(w) = \Pr(w) &= \Pr\left(\bigcap_{k=j}^{j+n-1} \{X_k = x_k\}\right) \\ &= \Pr\left(X_k = x_k;\ j \le k \le j+n-1\right) \end{aligned} \tag{2.3}$$

where the lower-case letters x_k stand for the realization of a stochastic process. Here we discuss two types of stochastic processes—a memoryless source and a Markov of order one source.

Memoryless source. If a text w is generated by a memoryless source, each random variable X_k in the sequence $\{X_k\}_{k=1}^{\infty}$ occurs independently from one another and the X_k are identically distributed. In other words, every letter a_j in the alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_\ell\}$ has a probability measure, denoted by $\pi(a_j)$ or π_{a_j}, which satisfies $\sum_{j=1}^{\ell} \pi(a_j) = 1$. Therefore, $\{X_k\}_{k=1}^{\infty}$ is an outcome of an infinite sequence of Bernoulli trials.

According to Equation (2.3), with a memoryless source, the weight of a word w is obtained by

$$\begin{aligned}\pi(w) &= \Pr(\{X_k\}_{k=j}^{j+n-1}) \\ &= \Pr\left(X_k = x_k;\ j \le k \le j+n-1\right) \\ &= \prod_{k=j}^{j+n-1} \pi(x_k)\end{aligned} \tag{2.4}$$

Markov source. When the source follows a Markov chain of order one, each random variable X_k in the sequence $\{X_k\}_{k=1}^{\infty}$ occurs with a conditional probability based on its previous letter (except the first one, X_1, which can be a designated letter or a random letter with a given probability).

An $\ell \times \ell$ stochastic matrix $P_{(\ell\times\ell)}$ should be pre-defined, providing the transition probabilities from any letter to another, as follows:

$$P_{(\ell\times\ell)} := \begin{pmatrix} p_{a_1,a_1} & p_{a_1,a_2} & \cdots & p_{a_1,a_\ell} \\ p_{a_2,a_1} & p_{a_2,a_2} & \cdots & p_{a_2,a_\ell} \\ \vdots & \vdots & \ddots & \vdots \\ p_{a_\ell,a_1} & p_{a_\ell,a_2} & \cdots & p_{a_\ell,a_\ell} \end{pmatrix} \tag{2.5}$$

where $p_{i,j}$ (or $p(i,j)$, or p_{ij} when there is no ambiguity) describes the conditional probability of the occurrence of letter j immediately after letter i:

$$p_{i,j} = \Pr(X_{k+1} = j \mid X_k = i), \quad \text{where } i, j \in \mathcal{A}. \tag{2.6}$$

If a word $w = x_j x_{j+1} \cdots x_{j+n-1}$ is generated by a Markov source of order 1, then $\pi(w)$, the weight of the word, must be dependent on the letter X_{j-1}. In order to specify this letter, we add a left superscript on the weight of w.

$$\begin{aligned}{}^{\alpha}\pi(w) &= \Pr\left(X_k = x_k;\ j \le k \le j+n-1 \mid X_{j-1} = \alpha\right) \\ &= p(\alpha, x_j) \cdot p(x_j, x_{j+1}) \cdot p(x_{j+1}, x_{j+2}) \ \cdots \ p(x_{j+n-2}, x_{j+n-1})\end{aligned} \tag{2.7}$$

where α and $x_k \in \mathcal{A}$.

2.3 Cluster

When a text is decorated by distinguished occurrences, we can define *clusters* to help identify the parts of a text which are covered by the decorated distinguished occurrences. The concept of cluster was created by Goulden and Jackson [17][16], and adopted by Bassino, Clément, Fayolle and Nicodème [4][5].

Cluster. A cluster c with respect to a pattern set $\mathcal{U}$ is a decorated text such that

- every letter of c must be covered by at least one distinguished occurrence;
- it is not possible to split the word into two parts without splitting a distinguished occurrence.

We use $\mathsf{C}_{\mathcal{U}}$ to denote the class of all clusters with respect to the pattern set $\mathcal{U}$. The subscript can be skipped when no ambiguity occurs.

In order to clarify the definition of a cluster, we have the next two examples.

Example 2.3.1 Consider a patter set $\mathcal{U} = \{u_1 = ab, u_2 = abab\}$. The decorated pattern words are

$$\mathsf{u}_1 = a\overset{\bullet 1}{b}, \text{ and } \mathsf{u}_2 = aba\overset{\bullet 2}{b}$$

By definition, we can name a few clusters, such as

$$\overset{\bullet 1}{a}b,\ a\overset{\bullet 2}{b}a\overset{\bullet 2}{b},\ \overset{\bullet 1}{a}b\overset{\bullet 2}{a}\overset{\bullet 2}{b},\ \overset{\bullet 1}{a}\overset{\bullet 2}{b}a\overset{\bullet 2}{b},\ \overset{\bullet 1}{a}\overset{\bullet 2}{b}\overset{\bullet 1}{a}\overset{\bullet 2}{b},\ a\overset{\bullet 2}{b}a\overset{\bullet 2}{b}a\overset{\bullet 2}{b},\ \overset{\bullet 1}{a}\overset{\bullet 2}{b}\overset{\bullet 1}{a}\overset{\bullet 2}{b}\overset{\bullet 1}{a}\overset{\bullet 2}{b},\ a\overset{\bullet 2}{b}a\overset{\bullet 2}{b}\overset{\bullet 1}{a}\overset{\bullet 2}{b},\ a\overset{\bullet 2}{b}a\overset{\bullet 2}{b}a\overset{\bullet 2}{b}a\overset{\bullet 2}{b},\ \text{etc.} \tag{2.8}$$

We also list some examples of decorated texts that violate the definition, therefore, do not belong to clusters:

$$ab,\ a\overset{\bullet 1}{a}b,\ a\overset{\bullet 1}{a}b\overset{\bullet 1}{b},\ \overset{\bullet 1}{a}b\overset{\bullet 1}{a}b,\ \overset{\bullet 1}{a}b\overset{\bullet 1}{a}b,\ aba\overset{\bullet 2}{b}ab,\ ababa\overset{\bullet 2}{b},\ \overset{\bullet 1}{a}bab\overset{\bullet 2}{a}b,\ \overset{\bullet 1}{a}b\overset{\bullet 2}{a}b\overset{\bullet 1}{a}\overset{\bullet 2}{b},\ \text{etc.} \tag{2.9}$$

Each item in the list (2.9) has one or more characters that conflict with the definition of a cluster. Some of the items may contain letters that are not covered by any distinguished occurrences, such as

$$\boldsymbol{ab},\ \boldsymbol{a}\overset{\bullet 1}{a}b,\ \boldsymbol{a}\overset{\bullet 1}{a}b\boldsymbol{b},\ \overset{\bullet 1}{a}b\boldsymbol{ab},\ abab\overset{\bullet 2}{\boldsymbol{ab}},\ \boldsymbol{ab}abab\overset{\bullet 2}{b}$$

where the bold letters are not covered by any occurrences.

Some other items in (2.9) have all of their letters covered by at least one cluster. However, it is possible to split the word into two parts without splitting a distinguished occurrence. Such decorated words in (2.9) are

$$\overset{\textbf{1}}{a}\overset{\textbf{1}}{b}|\overset{\textbf{1}}{a}b,\ \overset{\textbf{1}}{a}\overset{\textbf{2}}{b}|aba\overset{\textbf{2}}{b},\ \overset{\textbf{1}}{a}ba\overset{\textbf{1}}{b}|\overset{\textbf{2}}{a}\overset{\textbf{1}}{b}a\overset{\textbf{2}}{b}$$

where "|" is a splitter.

Here we make two remarks.

(1) A decorated pattern word in which only the pattern per se is distinguished, is always a cluster. For instance, $\mathsf{u}_1 = \overset{\textbf{1}}{a}b$ and $\mathsf{u}_2 = aba\overset{\textbf{2}}{b}$ are both clusters.

(2) A decorated text in which any distinguished occurrence has an overlap with another distinguished occurrence, is not necessarily a cluster, even if every letter in this decorated text is covered by at least one distinguished occurrence.[1]

The decorated text $\overset{\textbf{1}}{a}\overset{\textbf{2}}{b}\overset{\textbf{1}}{a}\overset{\textbf{2}}{b}\overset{\textbf{1}}{a}\overset{\textbf{2}}{b}$ is an example. Although every letter is covered by two distinguished occurrences (therefore, overlapped occurrences), it can still be split into two parts, $\overset{\textbf{1}}{a}\overset{\textbf{2}}{b}\overset{\textbf{1}}{a}\overset{\textbf{2}}{b}|\overset{\textbf{1}}{a}\overset{\textbf{2}}{b}$, without breaking any of its distinguished occurrences.

Example 2.3.2 We analyze the clusters of two decorated texts shown in Example 2.1.2, where the pattern set is $\mathcal{U} = \{u_1 = abb, u_2 = bba, u_3 = aabb\}$ and the support text is $w = aabbaaabbabbbba$.

(1) The fully decorated text is

$$aa\overset{\textbf{1}}{\underset{\textbf{3}}{b}}\overset{\textbf{2}}{b}aa\overset{\textbf{1}}{\underset{\textbf{3}}{b}}\overset{\textbf{2}}{b}a\overset{\textbf{1}}{b}\overset{\textbf{2}}{b}a \tag{2.10}$$

The fully decorated text (2.10) is only composed of three concatenating clusters c_1, c_2, and c_3, where

$$\mathsf{c}_1 = aa\overset{\textbf{1}}{\underset{\textbf{3}}{b}}\overset{\textbf{2}}{b}a,\quad \mathsf{c}_2 = aa\overset{\textbf{1}}{\underset{\textbf{3}}{b}}\overset{\textbf{2}}{b}a\overset{\textbf{1}}{b}b,\quad \mathsf{c}_3 = bb\overset{\textbf{2}}{a}$$

[1]The reason we make this remark is that in Bassino et al. [5, Section 4], the definition of a cluster might include this situation, which should have been avoided.

Fig. 2.1 provides details of how occurrences overlap in the three clusters. From the figure we can clearly see that (2.10) $= \mathsf{c}_1\mathsf{c}_2\mathsf{c}_3$.

(2) There are 8 occurrences in the fully decorated text. Therefore, the total number of decorated texts is $2^8 = 256$. The following decorated text is one of them.

$$a\overset{\textbf{1}}{a}bb\overset{\textbf{3}}{a}\overset{\textbf{1}}{a}bb\overset{\textbf{2}}{a}bbbba \tag{2.11}$$

The decorated text (2.11) includes both clusters and letters which do not belong to a cluster. We define the following clusters

$$\mathsf{c}_1 = a\overset{\textbf{1}}{b}b, \quad \mathsf{c}_2 = a\overset{\textbf{1}}{a}\overset{\textbf{3}}{b}b, \quad \text{and } \mathsf{c}_3 = bb\overset{\textbf{2}}{a}$$

Then (2.11) $= a\,\mathsf{c}_1\,a\,\mathsf{c}_2\,abb\,\mathsf{c}_3$, as shown in Fig. 2.2

c_1	c_2	c_3
$\overset{\textbf{1}}{a}\overset{\textbf{3}}{a}bb\overset{\textbf{2}}{a}$	$\overset{\textbf{1}}{a}\overset{\textbf{3}}{a}bb\overset{\textbf{2}}{a}bb\overset{\textbf{1}}{}$	$bb\overset{\textbf{2}}{a}$
$aabb$	$aabb$	bba
abb	abb	
bba	bba	
	abb	

Figure 2.1. $\overset{\textbf{1}}{a}\overset{\textbf{3}}{a}bb\overset{\textbf{2}}{a}\overset{\textbf{1}}{a}\overset{\textbf{3}}{b}b\overset{\textbf{2}}{a}\overset{\textbf{1}}{b}bbba = \mathsf{c}_1\mathsf{c}_2\mathsf{c}_3$. The fully decorated text (2.10) is composed of three clusters.

c_1		c_2		c_3
a	$\overset{\textbf{1}}{a}bb$	a	$\overset{\textbf{1}}{a}\overset{\textbf{3}}{a}bb$	abb	$bb\overset{\textbf{2}}{a}$
	abb		$aabb$		bba
			abb		

Figure 2.2. $\overset{\textbf{1}}{a}\overset{\textbf{3}}{a}bb\overset{\textbf{2}}{a}bbbba = a\,\mathsf{c}_1\,a\,\mathsf{c}_2\,abb\,\mathsf{c}_3$. The decorated text (2.11) contains letters (which do not belong to any clusters) and clusters.

3. GENERATING FUNCTIONS OF DECORATED TEXTS

Generating functions of decorated texts will be discussed in detail in this chapter. By applying techniques of analytic combinatorics to the multivariate generating function, we may obtain statistics such as mean, variance and covariance. The methodology of generating functions and combinatorial analysis are provided in Flajolet and Sedgewick [13].

We will start with memoryless probability models, and then take Markov sources into account.

3.1 Generating function

In general, for any set of texts $\mathcal{H}$, we define the univariate generating function

$$H(z) = \sum_{h \in \mathcal{H}} \pi(h) z^{|h|}$$

where z is a formal variable for marking the length of the texts. In particular, the alphabet set $\mathcal{A} = \{a_1, a_2, \ldots, a_\ell\}$ is a set of texts such that each text is one letter. The generating function of the alphabet is

$$A(z) = \sum_{\alpha \in \mathcal{A}} \pi(\alpha) z$$

A text is generated randomly by a source, which could follow a memoryless or a Markovian stochastic process. What we are interested in are the probabilities for the occurrences of pattern words $\mathcal{U} = \{u_1, u_2, \ldots, u_r\}$ from the set of all texts $\mathcal{A}^*$, according to

(1) the word length, and

(2) the number of occurrences (with possible overlap) of pattern words from $\mathcal{U}$.

Therefore, extra formal variables are needed to mark each $\{u_j\}_1^r$. We use x_j to serve this purpose. The multivariate generating function is

$$F_{\mathcal{U}}(z, \mathbf{x}) = F(z, \mathbf{x}) := \sum_{w \in \mathcal{A}^*} \pi(w) \cdot z^{|w|} \cdot x_1^{|w|_1} \cdot x_2^{|w|_2} \cdots x_r^{|w|_r} \tag{3.1}$$

where $\mathbf{x} := \{x_1, x_2, \cdots, x_r\}$, and $|w|_i$ is the total number of occurrences of pattern word u_i in w. The notation $|w|_i$ is standard, and was used in Bassino et al. [5] and Flajolet and Sedgewick [13, Chapter 3].

A generalized version of Equation (3.1) is to replace the set of all texts $\mathcal{A}^*$ with any set of texts $\mathcal{H}$, where $\mathcal{H}$ is a subset of $\mathcal{A}^*$.

The generating function that provides the probabilities of all possible pattern occurrences in every text length is

$$H(z, \mathbf{x}) := \sum_{w \in \mathcal{H}} \pi(w) \cdot z^{|w|} \cdot x_1^{|w|_1} \cdot x_2^{|w|_2} \cdots x_r^{|w|_r} \tag{3.2}$$

Example 3.1.1 (1) Consider the case in which all texts are generated by a memoryless source on the alphabet $\mathcal{A} = \{a, b\}$. Let the text set be $\mathcal{H} = \{aaa, aaba, abaa\}$, and the pattern set be $\mathcal{U} = \{u_1 = aa,\ u_2 = ab\}$. We have the following generating function.

$$\begin{aligned} H(z, x_1, x_2) &= \sum_{w \in \mathcal{H}} \pi(w) \cdot z^{|w|} \cdot x_1^{|w|_1} \cdot x_2^{|w|_2} \\ &= \pi_a^3 \cdot z^3 \cdot x_1^2 \ + \ \pi_a^2 \pi_b \pi_a \cdot z^4 \cdot x_1 x_2 \ + \ \pi_a \pi_b \pi_a^2 \cdot z^4 \cdot x_2 x_1 \\ &= \pi_a^3 \cdot z^3 \cdot x_1^2 \ + \ 2\pi_a^3 \pi_b \cdot z^4 \cdot x_1 x_2 \end{aligned}$$

(2) When $\mathcal{H} = \mathcal{A}^* = \{\epsilon,\ a,\ b,\ aa,\ ab,\ ba,\ bb,\ aaa,\ aab, \cdots\}$, where ϵ stands for the empty text, and if we again use $\mathcal{U} = \{u_1 = aa,\ u_2 = ab\}$, then the generating function is

$$\begin{aligned} F(z, x_1, x_2) = 1 \ &+ \ \pi_a z \ + \ \pi_b z \ + \ \pi_a^2 z^2 x_1 \ + \ \pi_a \pi_b z^2 x_2 \ + \ \pi_b \pi_a z^2 \ + \ \pi_b^2 z^2 \\ &+ \ \pi_a^3 z^3 x_1^2 \ + \ \pi_a^2 \pi_b z^3 x_1 x_2 \ + \ \cdots \end{aligned}$$

When $\mathcal{H}$ includes only one text $\mathcal{H} = \{w\}$, Equation (3.2) becomes

$$H(z, \mathbf{x}) = \pi(w) \cdot z^{|w|} \cdot x_1^{|w|_1} \cdot x_2^{|w|_2} \cdots x_r^{|w|_r} \tag{3.3}$$

In Equation (3.2) and (3.3), all occurrences must be distinguished. In other words, the generating function $H(z, \mathbf{x})$ considers the fully decorated text $(w, (\mathcal{O}_i)_{i=1}^{|w|})$.

Let Q denote the complete set of the decorated texts associated with a fully decorated text. Remember that when there are k occurrences in the fully decorated text, the number of decorated texts is 2^k.

Example 3.1.2 Let $w = abab$, and $\mathcal{U} = \{u_1 = ab, u_2 = ba\}$. The fully decorated text is

$$a\overset{\textbf{1}}{b}a\overset{\textbf{2}}{b}\overset{\textbf{1}}{a}b$$

There are 3 occurrences in the fully decorated text, and therefore, $2^3 = 8$ decorated texts, which are enumerated in the set Q.

$$\mathsf{Q} = \{abab,\ a\overset{\textbf{1}}{b}ab,\ ab\overset{\textbf{2}}{a}b,\ aba\overset{\textbf{1}}{b},\ a\overset{\textbf{1}}{b}a\overset{\textbf{2}}{b},\ a\overset{\textbf{1}}{b}a\overset{\textbf{1}}{b},\ ab\overset{\textbf{2}}{a}\overset{\textbf{1}}{b},\ a\overset{\textbf{1}}{b}\overset{\textbf{2}}{a}\overset{\textbf{1}}{b}\} \tag{3.4}$$

In a decorated text, not every occurrence is necessarily recognized. In such situations, we shall replace the formal variables x_j in the generating function (3.3) with t_j in order to mark the distinguished occurrences. Thus, the generating function of Q is given by

$$\begin{aligned} Q(z, \mathbf{t}) := \sum_{\mathrm{w} \in \mathsf{Q}} \pi(w) \cdot z^{|\mathrm{w}|} \cdot\ & t_1^{(\#\text{ distinguished occurrences of } u_1)} \\ & \cdot\ t_2^{(\#\text{ distinguished occurrences of } u_2)} \\ & \cdots \\ & \cdot\ t_r^{(\#\text{ distinguished occurrences of } u_r)} \end{aligned} \tag{3.5}$$

Example 3.1.3 Continuing the Example 3.1.2. Assuming the text is generated by a memoryless source, the generating function of the complete set of decorated texts, i.e., (3.4), is

$$\begin{aligned} Q(z, t_1, t_2) &= \pi_a^2\pi_b^2 z^4 + \pi_a^2\pi_b^2 z^4 t_1 + \pi_a^2\pi_b^2 z^4 t_2 + \pi_a^2\pi_b^2 z^4 t_1 \\ &\quad + \pi_a^2\pi_b^2 z^4 t_1 t_2 + \pi_a^2\pi_b^2 z^4 t_1^2 + \pi_a^2\pi_b^2 z^4 t_1 t_2 + \pi_a^2\pi_b^2 z^4 t_1^2 t_2 \\ &= \pi_a^2\pi_b^2 z^4 (1 + 2t_1 + t_2 + 2t_1 t_2 + t_1^2 + t_1^2 t_2) \end{aligned} \tag{3.6}$$

The generating function of the fully decorated text $a\overset{\textbf{1}}{b}a\overset{\textbf{2}}{b}\overset{\textbf{1}}{a}b$ is

$$H(z, x_1, x_2) = \pi_a^2\pi_b^2 z^4 x_1^2 x_2 \tag{3.7}$$

From Equation (3.6) and (3.7), we can observe that

$$Q(z, t_1, t_2) = H(z, t_1 + 1, t_2 + 1)$$

In fact, it is true for any set of texts $\mathcal{H}$, according to the following theorem. This classical theorem has been known for decades, and plays a fundamental role in [5] and [13, Chapter 3].

Theorem 3.1.1 Given the set of patterns $\mathcal{U} = \{u_1, u_2, \cdots, u_r\}$, for any set of texts $\mathcal{H}$, the two generating functions,
(1) $H(z, x_1, x_2, \cdots, x_r)$, the generating function of the fully decorated text, and
(2) $Q(z, t_1, t_2, \cdots, t_r)$, the generating function of the complete set of decorated texts,
have the following relation.

$$Q(z, t_1, t_2, \cdots, t_r) = H(z, t_1 + 1, t_2 + 1, \cdots, t_r + 1) \tag{3.8}$$

or equivalently,

$$H(z, x_1, x_2, \cdots, x_r) = Q(z, x_1 - 1, x_2 - 1, \cdots, x_r - 1) \tag{3.9}$$

Proof 1 In the generating function $H(z, x_1, x_2, \cdots, x_r)$, the operation of replacing an x_j with $t_j + 1$, stands for the fact that the occurrence of the corresponding u_j may be distinguished (represented by t_j) or not (represented by 1).

Therefore, by replacing every x_j with $t_j + 1$, we obtain the generating function of the complete set of decorated texts. ■

With Theorem 3.1.1, we have the following corollary.

Corollary 3.1.1 Given the set of patterns $\mathcal{U} = \{u_1, u_2, \cdots, u_r\}$, let T denote the complete set of decorated texts on the support of $\mathcal{A}^*$. Its generating function $T(z, \mathbf{t})$ satisfies

$$T(z, t_1, t_2, \cdots, t_r) = F(z, t_1 + 1, t_2 + 1, \cdots, t_r + 1) \tag{3.10}$$

or equivalently,

$$F(z, x_1, x_2, \cdots, x_r) = T(z, x_1 - 1, x_2 - 1, \cdots, x_r - 1) \tag{3.11}$$

Proof 2 Remember that $F(z, x_1, x_2, \cdots, x_r)$ is defined by Equation (3.1).

When $\mathcal{H} = \mathcal{A}^*$, we have

$$H(z, x_1, x_2, \cdots, x_r) = F(z, x_1, x_2, \cdots, x_r)$$

Because T, Q are the complete sets of decorated texts of $\mathcal{A}^*$, $\mathcal{H}$, respectively, we have $\mathsf{T} = \mathsf{Q}$ in this case. Thus, by Theorem 3.1.1, we obtain (3.10) and (3.11). ■

3.2 Generating function of clusters

A cluster is a substring in a decorated text. Thus, it is also considered a decorated text. When the pattern set $\mathcal{U} = \{u_1, u_2, \ldots, u_r\}$ is given, there exists the set of all clusters, C, on $\mathcal{A}^*$. The generating function of a set of clusters $\mathsf{C}_\mathcal{U}$ (or C, if no ambiguity), denoted by $\xi(z, \mathbf{t})$, is computed by

$$\begin{aligned} \xi(z, \mathbf{t}) := \sum_{\mathsf{c} \in \mathsf{C}} \pi(\mathsf{c}) \cdot z^{|\mathsf{c}|} \cdot t_1^{(\#\text{ distinguished occurrences of } u_1)} \\ \cdot t_2^{(\#\text{ distinguished occurrences of } u_2)} \\ \cdots \\ \cdot t_r^{(\#\text{ distinguished occurrences of } u_r)} \end{aligned} \tag{3.12}$$

A cluster is not necessarily fully decorated. If we are interested in fully decorated clusters, then the generating function can be directly obtained by $\mathbf{t} \to \mathbf{x} - \mathbf{1}$, according to Theorem 3.1.1.

3.3 Generating functions of decorated texts from Markov sources

When a text is generated by a Markov source, the weight ${}^{\alpha}\pi(w)$ of the text is defined in Equation (2.7). The left superscript α is the letter to the left of w, but not in w itself.

As long as the source is Markovian, a left superscript is necessary for all generating functions. Since we only consider first-order Markov dependence, one letter on the left superscript is enough. For higher-order Markov dependence, the left superscripts must be adjusted according to the order of the Markov dependence.

In the following example, we compare the generating functions of clusters in two stochastic processes—memoryless and Markovian.

Example 3.3.1 Let the alphabet $\mathcal{A} = \{a, b\}$, and consider the pattern set $\mathcal{U}$ that contains only one pattern word $\mathcal{U} = \{u\} = \{aba\}$.

(1) If the text source is memoryless, i.e., each letter, either a or b, is generated as a Bernoulli trial, with a probability π_a or π_b, respectively ($\pi_a + \pi_b = 1$).

Let the decorated pattern $\mathsf{u} = ab\overset{\bullet}{a}$ represent a distinguished pattern in a text. Then the set of all clusters is

$$\mathsf{C} = \{ab\overset{\bullet}{a}, ab\overset{\bullet}{a}b\overset{\bullet}{a}, ab\overset{\bullet}{a}b\overset{\bullet}{a}b\overset{\bullet}{a}, \ldots\} \tag{3.13}$$

Equivalently, it can be summarized as

$$\mathsf{C} = ab\overset{\bullet}{a} \cdot (b\overset{\bullet}{a})^* \tag{3.14}$$

It is straightforward to obtain the generating function of (3.14). (See Flajolet and Sedgewick [13] for the analogous combinatorics structures.)

$$\xi(z, t) = \frac{t\pi_a^2 \pi_b z^3}{1 - t\pi_a \pi_b z^2} \tag{3.15}$$

(2) If the text source is Markovian of order 1. In this case, the weight of any texts must depend on the letter immediately before the text. We introduce the convenient notation of putting the pre-word letter in a square bracket $[\,\cdot\,]$.

For instance, a decorated word $ab\overset{\bullet}{a}$ immediately after a letter a is $[a]ab\overset{\bullet}{a}$, which has a weight

$${}^{a}\pi(ab\overset{\bullet}{a}) = p_{aa}p_{ab}p_{ba}z^3 t$$

Similarly, the weight of $[b]ab\overset{\bullet}{b}a$ is

$$ {}^{b}\pi(ab\overset{\bullet}{b}a) = p_{ba}p_{ab}p_{ba}z^3t = p_{ab}p_{ba}^2z^3t $$

The generating function must consider the difference in weight. Thus, we group all the clusters with a pre-word $[a]$ as

$$ {}^{a}\mathsf{C} = \{[a]ab\overset{\bullet}{b}a, [a]ab\overset{\bullet}{b}ab\overset{\bullet}{b}a, [a]ab\overset{\bullet}{b}ab\overset{\bullet}{b}ab\overset{\bullet}{b}a, \ldots\} = [a]\, ab\overset{\bullet}{b}a \cdot (b\overset{\bullet}{b}a)^* $$

with the generating function

$$ {}^{a}\xi(z,t) = \frac{p_{aa}p_{ab}p_{ba}z^3t}{1 - p_{ab}p_{ba}z^2t} \tag{3.16} $$

For all the clusters with a pre-word $[b]$, we have

$$ {}^{b}\mathsf{C} = \{[b]ab\overset{\bullet}{b}a, [b]ab\overset{\bullet}{b}ab\overset{\bullet}{b}a, [b]ab\overset{\bullet}{b}ab\overset{\bullet}{b}ab\overset{\bullet}{b}a, \ldots\} = [b]\, ab\overset{\bullet}{b}a \cdot (b\overset{\bullet}{b}a)^* $$

and the generating function

$$ {}^{b}\xi(z,t) = \frac{p_{ba}p_{ab}p_{ba}z^3t}{1 - p_{ab}p_{ba}z^2t} = \frac{p_{ab}p_{ba}^2z^3t}{1 - p_{ab}p_{ba}z^2t} \tag{3.17} $$

In many cases in the following chapters, the single pre-word letter is determined by the last letter of the pre-word text. Hence, we use the symbol $\widehat{u}$ to denote the last letter of the text u.

4. INCLUSION-EXCLUSION METHOD FOR REDUCED PATTERNS WITH A MARKOVIAN TEXT SOURCE

In a set of pattern words, it is possible that some patterns are factors (i.e., substrings) of other patterns. For instance, in the set $\{aa, ab, aba\}$, the word ab is a factor of aba. It appears more difficult to keep track of these patterns simultaneously when some patterns are in others. Therefore, in this chapter, we start our analysis with the simpler reduced set of patterns, in which every pattern is not a factor of others. Then in the next chapter, we will discuss non-reduced pattern sets.

4.1 Inclusion-exclusion method

The inclusion-exclusion method was introduced by Goulden and Jackson [17] [16] to count occurrences of patterns from a reduced set of pattern words. The weight of words in the generating functions enables the probability models, for either memoryless or Markovian sources.

The power of inclusion-exclusion itself goes far beyond enumerating texts. In Goulden and Jackson [16], the number of derangements for permutations is obtained. Flajolet and Sedgewick [13, pp. 209] provides another application of counting rises in permutations.

In corollary 3.1.1, we have introduced T, the complete set of decorated texts on the support of $\mathcal{A}^*$. Let $\mathcal{U}$ be the set of pattern words, and $\mathsf{C}_{\mathcal{U}}$ (or C) the class of all clusters for the pattern set $\mathcal{U}$.

The inclusion-exclusion method provides an elegant relation of T, $\mathcal{A}$ and C. Note that we continue to use the notation of Bassino et al. [5], to which the following relation is attributed:

$$\mathsf{T} = (\mathcal{A} + \mathsf{C})^\star \tag{4.1}$$

By making use of the symbolic inclusion-exclusion principle and denoting by $|w|_u$ the number of occurrences of u in w, we directly get

$$F(z, \mathbf{x}) = \sum_{w \in \mathcal{A}^*} \pi(w) z^{|w|} \mathbf{x}^{|w|_u} = \frac{1}{1 - A(z) - \xi(z, \mathbf{x} - \mathbf{1})} \tag{4.2}$$

4.1.1 An alternative: recursive point of view

When the alphabet is $\mathcal{A} = \{a, b\}$, a decorated text could only start with 4 possible cases: empty (the text is empty), a letter a which does not belong to a cluster, a letter b which does not belong to a cluster, or a cluster.

$$\mathsf{T} = \begin{cases} \epsilon \\ a \cdots \\ b \cdots \\ \text{cluster} \cdots \end{cases}$$

Therefore,

$$\begin{aligned} T(z,t) &= 1 + \pi_a \cdot z \cdot T(z,t) + \pi_b \cdot z \cdot T(z,t) + \xi(z,t) \cdot T(z,t) \\ &= 1 + (\pi_a \cdot z + \pi_b \cdot z) \cdot T(z,t) + \xi(z,t) \cdot T(z,t) \\ &= 1 + A(z) \cdot T(z,t) + \xi(z,t) \cdot T(z,t) \\ &= 1 + (A(z) + \xi(z,t)) \cdot T(z,t) \end{aligned} \tag{4.3}$$

We then obtain

$$T(z,t) = \frac{1}{1 - A(z) - \xi(z,t)} \tag{4.4}$$

According to Corollary 3.1.1, (4.4) is equivalent to (4.2).

4.1.2 Markovian source

When a Markovian source of order 1 is considered, these probability generating functions do not only depend on the letters in the strings, *but also on the previous one letter before the strings.*

Consider a pattern $u = aba$. Using a square bracket to display the letter directly before the cluster, we write

$$[a]\ \mathsf{C} = [a]\ ab\overset{\bullet}{a} \cdot (b\overset{\bullet}{a})^* \tag{4.5}$$

and

$$[b]\ \mathsf{C} = [b]\ ab\overset{\bullet}{b}\overset{\bullet}{a} \cdot (b\overset{\bullet}{a})^* \tag{4.6}$$

to represent the clusters with a previous letter a or b, respectively.

The generating functions are

$${}^a\xi(z,t) = \frac{p_{aa}p_{ab}p_{ba} \cdot t \cdot z^3}{1 - p_{ab}p_{ba} \cdot t \cdot z^2} \tag{4.7}$$

and

$${}^b\xi(z,t) = \frac{p_{ba}^2 p_{ab} \cdot t \cdot z^3}{1 - p_{ab}p_{ba} \cdot t \cdot z^2} \tag{4.8}$$

As for the set of decorated texts T, we use similar notations and denote the texts that start **after** a letter a by ${}^aT(z,t)$. When the alphabet is $\mathcal{A} = \{a, b\}$, a decorated text could only start with 4 possible cases: empty (the text is empty), a letter a which does not belong to a cluster, a letter b which does not belong to a cluster, or a cluster.

$${}^a\mathsf{T} = [a] \begin{cases} \epsilon \\ a \cdots \\ b \cdots \\ \text{cluster} \cdots \end{cases}$$

Thus, the following equation holds for ${}^aT(z,t)$.

$${}^aT(z,t) = 1 + p_{aa}z \cdot {}^aT(z,t) + p_{ab}z \cdot {}^bT(z,t) + {}^a\xi(z,t) \cdot {}^aT(z,t) \tag{4.9}$$

Similarly, for texts start after a letter b,

$${}^bT(z,t) = 1 + p_{ba}z \cdot {}^aT(z,t) + p_{bb}z \cdot {}^bT(z,t) + {}^b\xi(z,t) \cdot {}^aT(z,t) \tag{4.10}$$

Note that the last term in (4.9) is ${}^a\xi(z,t) \cdot {}^aT(z,t)$ and the last term in (4.10) is ${}^b\xi(z,t) \cdot {}^aT(z,t)$. In both terms ${}^aT(z,t)$ appears. This is because the pattern word is $u = aba$ and then the clusters can be only structured in the form of (4.5) or (4.6). Therefore, no matter a cluster starts after a letter a or a letter b, all clusters must end with a letter a.

We write (4.9) and (4.10) together, and have

$$\begin{pmatrix} {}^{a}T(z,t) \\ {}^{b}T(z,t) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} p_{aa} & p_{ab} \\ p_{ba} & p_{bb} \end{pmatrix} z \cdot \begin{pmatrix} {}^{a}T(z,t) \\ {}^{b}T(z,t) \end{pmatrix} + \begin{pmatrix} {}^{a}\xi(z,t) \\ {}^{b}\xi(z,t) \end{pmatrix} \cdot {}^{a}T(z,t) \tag{4.11}$$

The equation above is the Markovian equivalent of $\mathsf{T} = (\mathcal{A} + \mathsf{C})^{\star}$ with a Bernoulli source.

Example 4.1.1 (One pattern word Markovian) The given pattern set only contains one word, i.e., $\mathcal{U} = \{u\} = \{aba\}$. Consider the transition matrix

$$P = \begin{pmatrix} p_{aa} & p_{ab} \\ p_{ba} & p_{bb} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 3/5 & 2/5 \end{pmatrix} \tag{4.12}$$

As we discussed in (4.5) and (4.6), the clusters are in the form of

$$[a]\ \mathsf{C} = [a]\ ab\overset{\bullet}{a} \cdot (b\overset{\bullet}{a})^{*}$$

$$[b]\ \mathsf{C} = [b]\ ab\overset{\bullet}{a} \cdot (b\overset{\bullet}{a})^{*}$$

Therefore we have their generating functions of clusters, as follows:

$${}^{a}\xi(z,t) = \frac{p_{aa}p_{ab}p_{ba} \cdot t \cdot z^3}{1 - p_{ab}p_{ba} \cdot t \cdot z^2} = \left(\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{5} \cdot t \cdot z^3\right) \Big/ \left(1 - \frac{1}{2} \cdot \frac{3}{5} \cdot t \cdot z^2\right) \tag{4.13}$$

$${}^{b}\xi(z,t) = \frac{p_{ba}^2 p_{ab} \cdot t \cdot z^3}{1 - p_{ab}p_{ba} \cdot t \cdot z^2} = \left(\left(\frac{3}{5}\right)^2 \cdot \frac{1}{2} \cdot t \cdot z^3\right) \Big/ \left(1 - \frac{1}{2} \cdot \frac{3}{5} \cdot t \cdot z^2\right) \tag{4.14}$$

Applying Equation (4.11) (or equivalently, Equations (4.9) and (4.10) combined), we obtain

$${}^{a}T(z,t) = 1 + \frac{1}{2} \cdot z \cdot {}^{a}T(z,t) + \frac{1}{2} \cdot z \cdot {}^{b}T(z,t) + {}^{a}\xi(z,t) \cdot {}^{a}T(z,t) \tag{4.15}$$

$${}^{b}T(z,t) = 1 + \frac{3}{5} \cdot z \cdot {}^{a}T(z,t) + \frac{2}{5} \cdot z \cdot {}^{b}T(z,t) + {}^{b}\xi(z,t) \cdot {}^{a}T(z,t) \tag{4.16}$$

It follows that

$$ {}^{a}T(z,t) = \frac{100 + 10z - 30tz^2 - 3tz^3}{2 \cdot (50 - 45z - 5z^2 - 15tz^2 + 6tz^3)} \tag{4.17} $$

$$ {}^{b}T(z,t) = \frac{5\,(10 + z - 3tz^2)}{(50 - 45z - 5z^2 - 15tz^2 + 6tz^3)} \tag{4.18} $$

The corresponding generating functions are

$$ \begin{aligned} {}^{a}F(z,x) &= {}^{a}T(z, x-1) \\ &= -\frac{3(z+10)\,(-(10/3) + (x-1)z^2)}{100 + (12x-12)z^3 + (-30x+20)z^2 - 90z} \\ &= 1 + z + z^2 + \left(\frac{17}{20} + \frac{3}{20}\cdot x\right) z^3 + \left(\frac{137}{200} + \frac{63}{200}\cdot x\right) z^4 \\ &\quad + \left(\frac{1133}{2000} + \frac{777}{2000}\cdot x + \frac{9}{200}\cdot x^2\right) z^5 + O(z^6) \end{aligned} \tag{4.19} $$

$$ \begin{aligned} {}^{b}F(z,x) &= {}^{b}T(z, x-1) \\ &= \frac{50 + (-15x+15)z^2 + 5z}{50 + (6x-6)z^3 + (-15x+10)z^2 - 45z} \\ &= 1 + z + z^2 + \left(\frac{41}{50} + \frac{9}{50}\cdot x\right) z^3 + \left(\frac{329}{500} + \frac{171}{500}\cdot x\right) z^4 \\ &\quad + \left(\frac{2741}{5000} + \frac{1989}{5000}\cdot x + \frac{27}{500}\cdot x^2\right) z^5 + O(z^6) \end{aligned} \tag{4.20} $$

The last part of Equation (4.19) and (4.20) can be further expand to any order of interest. They provide straightforward probabilities of any number of pattern occurrences, with a given text length.

For instance, in Equation (4.20), we have

$$ {}^{b}F(z,x) = \cdots + \left(\frac{2741}{5000} + \frac{1989}{5000}\cdot x + \frac{27}{500}\cdot x^2\right) z^5 + O(z^6) $$

The information it conveys is:

Assume a binary text is generated by a Markovian source of order one, with the transition matrix Eq. (4.12). In a text of length 5 following a letter b, the probability that the pattern word $u = aba$ does not occur is 2741/5000. The pattern word may occur only once, with the probability 1989/5000. Or, it may occur twice, with the probability 27/500.

4.1.3 Generalization of 1-pattern case

In a generalized case, when the alphabet is $\mathcal{A} = \{a_1, a_2, \ldots, a_\ell\}$, and we are only considering one pattern u, we have

$$\begin{pmatrix} {}^{a_1}T(z,t) \\ {}^{a_2}T(z,t) \\ \vdots \\ {}^{a_\ell}T(z,t) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + P_{(\ell\times\ell)} \cdot z \cdot \begin{pmatrix} {}^{a_1}T(z,t) \\ {}^{a_2}T(z,t) \\ \vdots \\ {}^{a_\ell}T(z,t) \end{pmatrix} + \begin{pmatrix} {}^{a_1}\xi(z,t) \\ {}^{a_2}\xi(z,t) \\ \vdots \\ {}^{a_\ell}\xi(z,t) \end{pmatrix} \cdot {}^{\widehat{u}}T(z,t) \tag{4.21}$$

where $P_{(\ell\times\ell)}$ stands for the Markovian transition matrix, and $\widehat{u}$ denotes the last letter in the pattern u. *The letter $\widehat{u} \in \mathcal{A}$ is solely determined by the pattern u.*

4.2 Reduced 2-pattern case

When two or more patterns exist, we start with the reduced case—meaning no pattern is a part of another, for simplicity.

Let us start with an example. We consider $u_1 = abb$, $u_2 = bba$, with binary alphabet $\mathcal{A} = \{a, b\}$. The set of clusters C of $\{u_1, u_2\}$ is given by

$$\mathsf{C} = (ab\overset{\bullet\!1}{b}, bb\overset{\bullet\!2}{a}) \cdot \begin{pmatrix} \emptyset & \{\overset{\bullet\!2}{a}, b\overset{\bullet\!2}{a}\} \\ \{\overset{\bullet\!1}{b}b\} & \emptyset \end{pmatrix}^* \cdot \begin{pmatrix} \varepsilon \\ \varepsilon \end{pmatrix} \tag{4.22}$$

Then the generating functions under the condition of the previous letter are as follows.

$$\begin{aligned} {}^{a}\xi(z,t_1,t_2) &= \left(z^3 t_1 \cdot p_{aa}p_{ab}p_{bb},\ z^3 t_2 \cdot p_{ab}p_{bb}p_{ba}\right) \\ &\quad \cdot \left(\mathbb{I} - \begin{pmatrix} 0 & zt_2 \cdot p_{ba} + z^2 t_2 p_{bb}p_{ba} \\ z^2 t_1 p_{ab}p_{bb} & 0 \end{pmatrix}\right)^{-1} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= {}^{a}\eta(z,t_1,t_2) \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \end{aligned}$$

where

$$
\begin{aligned}
{}^{a}\eta(z,t_1,t_2) = &\left(z^3t_1\cdot p_{aa}p_{ab}p_{bb},\ z^3t_2\cdot p_{ab}p_{bb}p_{ba}\right)\\
&\cdot\left(\mathbb{I}-\begin{pmatrix} 0 & zt_2\cdot p_{ba}+z^2t_2p_{bb}p_{ba}\\ z^2t_1p_{ab}p_{bb} & 0\end{pmatrix}\right)^{-1}
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
{}^{b}\xi(z,t_1,t_2) &= \left(z^3t_1\cdot p_{ba}p_{ab}p_{bb},\ z^3t_2\cdot p_{bb}^2p_{ba}\right)\\
&\quad\cdot\left(\mathbb{I}-\begin{pmatrix} 0 & zt_2\cdot p_{ba}+z^2t_2p_{bb}p_{ba}\\ z^2t_1p_{ab}p_{bb} & 0\end{pmatrix}\right)^{-1}\cdot\begin{pmatrix}1\\1\end{pmatrix}\\
&= {}^{b}\eta(z,t_1,t_2)\cdot\begin{pmatrix}1\\1\end{pmatrix}
\end{aligned}
$$

where

$$
\begin{aligned}
{}^{b}\eta(z,t_1,t_2) = &\left(z^3t_1\cdot p_{ba}p_{ab}p_{bb},\ z^3t_2\cdot p_{bb}^2p_{ba}\right)\\
&\cdot\left(\mathbb{I}-\begin{pmatrix} 0 & zt_2\cdot p_{ba}+z^2t_2p_{bb}p_{ba}\\ z^2t_1p_{ab}p_{bb} & 0\end{pmatrix}\right)^{-1}
\end{aligned}
$$

Both ${}^{a}\eta(z,t_1,t_2)$ and ${}^{b}\eta(z,t_1,t_2)$ are 1×2 row vectors, in the form of

$$
\left((\cdots)\overset{❶}{b},\ (\cdots)\overset{❷}{a}\right)
$$

With alphabet $\mathcal{A}=\{a,b\}$, a text (directly after a letter a) could only start with the following possible cases: empty (the text is empty), a letter a which does not belong to a cluster, a letter b which does not belong to a cluster, a cluster ending in $\overset{❶}{b}$, or a cluster ending in $\overset{❷}{a}$.

Thus, we have the following combinatorial structure for ${}^{a}\mathsf{T}$.

$$
{}^{a}\mathsf{T} = [a] \begin{cases} \epsilon \\ a \cdot {}^{a}\mathsf{T} \\ b \cdot {}^{b}\mathsf{T} \\ (\text{a cluster ending in } \overset{\textbf{❶}}{b}) \cdot {}^{b}\mathsf{T} \\ (\text{a cluster ending in } \overset{\textbf{❷}}{a}) \cdot {}^{a}\mathsf{T} \end{cases}
$$

Therefore, the set of decorated texts T has the following recurrence relations:

$$
{}^{a}T(z,t_1,t_2) = 1 + p_{aa}z \cdot {}^{a}T(z,t_1,t_2) + p_{ab}z \cdot {}^{b}T(z,t_1,t_2) + {}^{a}\eta(z,t_1,t_2) \cdot \begin{pmatrix} {}^{b}T(z,t_1,t_2) \\ {}^{a}T(z,t_1,t_2) \end{pmatrix}
$$

and

$$
{}^{b}T(z,t_1,t_2) = 1 + p_{ba}z \cdot {}^{a}T(z,t_1,t_2) + p_{bb}z \cdot {}^{b}T(z,t_1,t_2) + {}^{b}\eta(z,t_1,t_2) \cdot \begin{pmatrix} {}^{b}T(z,t_1,t_2) \\ {}^{a}T(z,t_1,t_2) \end{pmatrix}
$$

Or, in a well-organized form:

$$
\begin{pmatrix} {}^{a}T(z,t_1,t_2) \\ {}^{b}T(z,t_1,t_2) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} p_{aa} & p_{ab} \\ p_{ba} & p_{bb} \end{pmatrix} \cdot z \cdot \begin{pmatrix} {}^{a}T(z,t_1,t_2) \\ {}^{b}T(z,t_1,t_2) \end{pmatrix} + \begin{pmatrix} {}^{a}\eta(z,t_1,t_2) \\ {}^{b}\eta(z,t_1,t_2) \end{pmatrix} \cdot \begin{pmatrix} {}^{b}T(z,t_1,t_2) \\ {}^{a}T(z,t_1,t_2) \end{pmatrix}.
$$

Keep in mind that both ${}^{a}\eta(z,t_1,t_2)$ and ${}^{b}\eta(z,t_1,t_2)$ are 1×2 row vectors.

4.3 Generalization of reduced multi-pattern case

Reduced multi-pattern cases can be derived by the same approach of the reduced 2-pattern case. Hence, we have the following theorem.

Theorem 4.3.1 Consider a random text that is generated by a first-order Markovian source with the alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_\ell\}$. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_r\}$ denote the set of reduced pattern words. Then we can obtain ${}^{a_j}T(z, t_1, t_2, \ldots, t_r)$, the generating function of decorated text following the letter $a_j (j = 1, 2, \ldots, \ell)$, by the following linear equations:

$$\begin{pmatrix} {}^{a_1}T(z,\mathbf{t}) \\ {}^{a_2}T(z,\mathbf{t}) \\ \vdots \\ {}^{a_\ell}T(z,\mathbf{t}) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + P_{(\ell\times\ell)} \cdot z \cdot \begin{pmatrix} {}^{a_1}T(z,\mathbf{t}) \\ {}^{a_2}T(z,\mathbf{t}) \\ \vdots \\ {}^{a_\ell}T(z,\mathbf{t}) \end{pmatrix} + \begin{pmatrix} {}^{a_1}\eta(z,\mathbf{t}) \\ {}^{a_2}\eta(z,\mathbf{t}) \\ \vdots \\ {}^{a_\ell}\eta(z,\mathbf{t}) \end{pmatrix} \cdot \begin{pmatrix} {}^{\widehat{u_1}}T(z,\mathbf{t}) \\ {}^{\widehat{u_2}}T(z,\mathbf{t}) \\ \vdots \\ {}^{\widehat{u_r}}T(z,\mathbf{t}) \end{pmatrix}. \quad (4.23)$$

Some remarks regarding the notations:

(1) $\mathbf{t}$ stands for the set $\{t_1, t_2, \cdots, t_r\}$.

(2) Every ${}^{a_j}\eta(z, \mathbf{t})$ (where $j \in \{1, 2, \ldots, \ell\}$) is a $1 \times r$ row vector, in the form of

$$\left((\cdots)^{\textcircled{1}}\widehat{u_1},\ (\cdots)^{\textcircled{2}}\widehat{u_2},\ \ldots,\ (\cdots)^{\textcircled{r}}\widehat{u_r} \right).$$

(3) $\widehat{u_1}, \widehat{u_2}, \cdots, \widehat{u_r}$ denotes the last letter of $u_1, u_2, \ldots, u_r$, respectively,

where $\{\widehat{u_1},\ \widehat{u_2},\ \cdots,\ \widehat{u_r}\} \in \mathcal{A}$;

Then we have the generating functions ${}^{a_j}F(z, x_1, x_2, \ldots, x_\ell)$ that give the probability of occurrences for each pattern word:

$${}^{a_j}F(z, x_1, x_2, \ldots, x_r) = {}^{a_j}T(z, x_1 - 1, x_2 - 1, \ldots, x_r - 1)$$

4.4 An application of Theorem 4.3.1

Theorem 4.3.1 enables us to count the occurrences for multiple reduced pattern words in a first-order Markovian text. We make an example to present the procedures.

Example 4.4.1 (Three reduced patterns, Markovian case) We consider a set of patterns that contains three words, namely $\mathcal{U} = \{u_1, u_2, u_3\} = \{aa, bb, aba\}$. The transition matrix is

$$P = \begin{pmatrix} p_{aa} & p_{ab} \\ p_{ba} & p_{bb} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 3/5 & 2/5 \end{pmatrix} \quad (4.24)$$

As we discussed, the clusters are in the form of

$$[a]\ \mathsf{C} = [a]\ (a\overset{\bullet 1}{a}, b\overset{\bullet 2}{b}, ab\overset{\bullet 3}{a}) \cdot \begin{pmatrix} \overset{\bullet 1}{a} & \emptyset & b\overset{\bullet 3}{a} \\ \emptyset & \overset{\bullet 2}{b} & \emptyset \\ \overset{\bullet 1}{a} & \emptyset & b\overset{\bullet 3}{a} \end{pmatrix}^* \cdot \begin{pmatrix} \varepsilon \\ \varepsilon \\ \varepsilon \end{pmatrix}$$

$$[b]\ \mathsf{C} = [b]\ (a\overset{\bullet 1}{a}, b\overset{\bullet 2}{b}, ab\overset{\bullet 3}{a}) \cdot \begin{pmatrix} \overset{\bullet 1}{a} & \emptyset & b\overset{\bullet 3}{a} \\ \emptyset & \overset{\bullet 2}{b} & \emptyset \\ \overset{\bullet 1}{a} & \emptyset & b\overset{\bullet 3}{a} \end{pmatrix}^* \cdot \begin{pmatrix} \varepsilon \\ \varepsilon \\ \varepsilon \end{pmatrix}$$

Their generating functions are as follows.

$$\begin{aligned} {}^a\xi(z,t_1,t_2,t_3) &= \left(z^2t_1 \cdot p_{aa}^2,\ z^2t_2 \cdot p_{ab}p_{bb},\ z^3t_3 \cdot p_{aa}p_{ab}p_{ba}\right) \\ &\quad \cdot \left(\mathbb{I} - \begin{pmatrix} zt_1 \cdot p_{aa} & 0 & z^2t_3 \cdot p_{ab}p_{ba} \\ 0 & zt_2 \cdot p_{bb} & 0 \\ zt_1 \cdot p_{aa} & 0 & z^2t_3 \cdot p_{ab}p_{ba} \end{pmatrix}\right)^{-1} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\ &= \left(\frac{5z^2t_1}{20-10zt_1-6z^2t_3}, \frac{z^2t_2}{5-2zt_2}, \frac{3z^3t_3}{20-10zt_1-6z^2t_3}\right) \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad (4.25) \\ &= {}^a\eta(z,t_1,t_2,t_3) \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \end{aligned}$$

where

$${}^a\eta(z,t_1,t_2,t_3) = \left(\frac{5z^2t_1}{20-10zt_1-6z^2t_3}, \frac{z^2t_2}{5-2zt_2}, \frac{3z^3t_3}{20-10zt_1-6z^2t_3}\right) \qquad (4.26)$$

We also have,

$$
\begin{aligned}
{}^{b}\xi(z,t_1,t_2,t_3) &= \left(z^2t_1\cdot p_{ba}p_{aa},\ z^2t_2\cdot p_{bb}^2,\ z^3t_3\cdot p_{ba}^2p_{ab}\right)\\
&\quad\cdot\left(\mathbb{I}-\begin{pmatrix} zt_1\cdot p_{aa} & 0 & z^2t_3\cdot p_{ab}p_{ba}\\ 0 & zt_2\cdot p_{bb} & 0\\ zt_1\cdot p_{aa} & 0 & z^2t_3\cdot p_{ab}p_{ba}\end{pmatrix}\right)^{-1}\cdot\begin{pmatrix}1\\1\\1\end{pmatrix}\\
&= \left(\frac{3z^2t_1}{10-5zt_1-3z^2t_3},\frac{4z^2t_2}{25-10zt_2},\frac{9z^3t_3}{50-25zt_1-15z^2t_3}\right)\cdot\begin{pmatrix}1\\1\\1\end{pmatrix}\\
&= {}^{b}\eta(z,t_1,t_2,t_3)\cdot\begin{pmatrix}1\\1\\1\end{pmatrix}
\end{aligned}
\tag{4.27}
$$

where

$$
{}^{b}\eta(z,t_1,t_2,t_3)=\left(\frac{3z^2t_1}{10-5zt_1-3z^2t_3},\frac{4z^2t_2}{25-10zt_2},\frac{9z^3t_3}{50-25zt_1-15z^2t_3}\right)
\tag{4.28}
$$

Applying Formula (4.23) in Theorem 4.3.1, we have

$$
\begin{aligned}
\begin{pmatrix}{}^{a}T(z,t_1,t_2,t_3)\\ {}^{b}T(z,t_1,t_2,t_3)\end{pmatrix} &= \begin{pmatrix}1\\1\end{pmatrix}+\begin{pmatrix}1/2 & 1/2\\ 3/5 & 2/5\end{pmatrix}\cdot z\cdot\begin{pmatrix}{}^{a}T(z,t_1,t_2,t_3)\\ {}^{b}T(z,t_1,t_2,t_3)\end{pmatrix}\\
&\quad+\begin{pmatrix}{}^{a}\eta(z,t_1,t_2,t_3)\\ {}^{b}\eta(z,t_1,t_2,t_3)\end{pmatrix}\cdot\begin{pmatrix}{}^{a}T(z,t_1,t_2,t_3)\\ {}^{b}T(z,t_1,t_2,t_3)\\ {}^{a}T(z,t_1,t_2,t_3)\end{pmatrix}
\end{aligned}
\tag{4.29}
$$

Note that ${}^{a}\eta(z,t_1,t_2,t_3)$ and ${}^{b}\eta(z,t_1,t_2,t_3)$ are 1×3 row vectors. Substituting (4.26) and (4.28) into (4.29), and solving the two linear equations, we are able to obtain ${}^{a}T(z,t_1,t_2,t_3)$ and ${}^{b}T(z,t_1,t_2,t_3)$, namely,

$$
\begin{aligned}
&{}^{a}T(z,t_1,t_2,t_3)\\
&=\frac{(4zt_2-z-10)\left(3z^2t_3+5zt_1-10\right)}{100+12t_3(t_2+1)z^3+\left((20t_1+20)t_2+20t_1-30t_3-10\right)z^2-(50t_1+40t_2+90)z}
\end{aligned}
\tag{4.30}
$$

and

$$
\begin{aligned}
&{}^{b}T(z,t_1,t_2,t_3)\\
&=\frac{(2zt_2-5)\left(3z^2t_3+5zt_1-z-10\right)}{50+6t_3(t_2+1)z^3+\left((10t_1+10)t_2+10t_1-15t_3-5\right)z^2-(25t_1+20t_2+45)z}
\end{aligned}
\tag{4.31}
$$

The corresponding generating functions ${}^{a}F(z,x_1,x_2,x_3)$ and ${}^{b}F(z,x_1,x_2,x_3)$ are

$$
\begin{aligned}
&{}^{a}F(z,x_1,x_2,x_3)\\
&\quad={}^{a}T(z,x_1-1,x_2-1,x_3-1)\\
&\quad=\frac{(4(x_2-1)z-z-10)\left(3(x_3-1)z^2+5(x_1-1)z-10\right)}{100+12(x_3-1)x_2z^3+(20x_1(x_2-1)+20x_1-30x_3)z^2-(50x_1+40x_2)z}\\
&\quad=\frac{(4zx_2-5z-10)\left(3z^2x_3+5zx_1-3z^2-5z-10\right)}{100+12(x_3-1)x_2z^3+(20x_1x_2-30x_3)z^2-(50x_1+40x_2)z}
\end{aligned}
\tag{4.32}
$$

and

$$
\begin{aligned}
&{}^{b}F(z,x_1,x_2,x_3)\\
&={}^{b}T(z,x_1-1,x_2-1,x_3-1)\\
&=\frac{(2(x_2-1)z-5)\left(3(x_3-1)z^2+5(x_1-1)z-z-10\right)}{50+6(x_3-1)x_2z^3+(10x_1(x_2-1)+10x_1-15x_3)z^2-(25x_1+20x_2)z}\\
&=\frac{(2zx_2-2z-5)\left(3z^2x_3+5zx_1-3z^2-6z-10\right)}{50+6(x_3-1)x_2z^3+(10x_1x_2-15x_3)z^2-(25x_1+20x_2)z}
\end{aligned}
\tag{4.33}
$$

Computing the Taylor expansion at $z=0$ for (4.32) and (4.33), we have

$$
\begin{aligned}
{}^{a}F(z,x_1,x_2,x_3)=&1+z+\left(\frac{11}{20}+\frac{x_1}{4}+\frac{x_2}{5}\right)z^2\\
&+\left(\frac{11}{50}x_2+\frac{3}{20}x_3+\frac{3}{20}+\frac{11}{40}x_1+\frac{1}{8}x_1^2+\frac{2}{25}x_2^2\right)z^3+\mathrm{O}\left(z^4\right)
\end{aligned}
\tag{4.34}
$$

and

$$
\begin{aligned}
{}^{b}F(z,x_1,x_2,x_3) = 1 + z + \left(\frac{27}{50} + \frac{3x_1}{10} + \frac{4x_2}{25}\right) z^2 \\
+ \left(\frac{27}{125}x_2 + \frac{9}{50}x_3 + \frac{3}{25} + \frac{27}{100}x_1 + \frac{3}{20}x_1^2 + \frac{8}{125}x_2^2\right) z^3 + \mathrm{O}\left(z^4\right)
\end{aligned}
\tag{4.35}
$$

The Taylor series of ${}^{a}F(z,x_1,x_2,x_3)$ and ${}^{b}F(z,x_1,x_2,x_3)$ can be expanded further to any order of interest. We are able to obtain the probabilities of certain patterns from these coefficients.

For instance, consider the z^3 term

$$
{}^{b}F(z,x_1,x_2,x_3) = \cdots + \left(\frac{27}{125}x_2 + \frac{9}{50}x_3 + \frac{3}{25} + \frac{27}{100}x_1 + \frac{3}{20}x_1^2 + \frac{8}{125}x_2^2\right) z^3 + \mathrm{O}\left(z^4\right) \tag{4.36}
$$

The z^3 term indicates the following information.

Assume that a binary text is generated by a Markovian source of order one, with the transition matrix (4.24). We are interested in counting pattern words $\{u_1 = aa,\ u_2 = bb,\ u_3 = aba\}$. In a text of length 3 following a letter b, the probability that:

– none of the three patterns occurs, is 3/25;

– pattern $u_1 = aa$ occurs exactly once and no other patterns occur, is 27/100;

– pattern $u_2 = bb$ occurs exactly once and no other patterns occur, is 27/125;

– pattern $u_3 = aba$ occurs exactly once and no other patterns occur, is 9/50;

– pattern $u_1 = aa$ occurs exactly twice and no other patterns occur, is 3/20;

– pattern $u_2 = bb$ occurs exactly twice and no other patterns occur, is 8/125;

and all other situations have 0 probability.

As expected, the probability values above should sum up to 1.

$$
\frac{27}{125} + \frac{9}{50} + \frac{3}{25} + \frac{27}{100} + \frac{3}{20} + \frac{8}{125} = 1.
$$

5. INCLUSION-EXCLUSION METHOD FOR NON-REDUCED PATTERNS

5.1 Skeleton and Flip

In the previous chapter, we discussed reduced decorated texts, in which there is no distinguished occurrence being a factor of another distinguished one.

However, when there are multiple pattern words, it is possible that one or more patterns are factors of others. We continue to follow the notation of Bassino et al. [5].

Example 5.1.1 (Non-reduced case) Given a pattern set $\mathcal{U} = \{u_1 = ab, u_2 = aba, u_3 = baba\}$, a cluster text *ababa* could be labeled in a number of ways for distinguished pattern words, listed below as c_1 through c_{16}:

$$\mathsf{c}_1 = a\overset{\bullet 1}{b}a\overset{\bullet 3}{b}a \quad \mathsf{c}_2 = a\overset{\bullet 1}{b}a\overset{\bullet 1\bullet 3}{b}a \quad \mathsf{c}_3 = a\overset{\bullet 1}{b}a\overset{\bullet 2\bullet 3}{b}a \quad \mathsf{c}_4 = a\overset{\bullet 1}{b}a\overset{\bullet 1\bullet 2\bullet 3}{b}a$$

$$\mathsf{c}_5 = a\overset{\bullet 2}{b}a\overset{\bullet 3}{b}a \quad \mathsf{c}_6 = a\overset{\bullet 1\bullet 2}{b}a\overset{\bullet 3}{b}a \quad \mathsf{c}_7 = a\overset{\bullet 2\bullet 1\bullet 3}{b}a\overset{}{b}a \quad \mathsf{c}_8 = a\overset{\bullet 2}{b}a\overset{\bullet 2\bullet 3}{b}a$$

$$\mathsf{c}_9 = a\overset{\bullet 1\bullet 2}{b}a\overset{\bullet 1\bullet 3}{b}a \quad \mathsf{c}_{10} = a\overset{\bullet 1\bullet 2}{b}a\overset{\bullet 2\bullet 3}{b}a \quad \mathsf{c}_{11} = a\overset{\bullet 2\bullet 1\bullet 3}{b}a\overset{\bullet 2}{b}a \quad \mathsf{c}_{12} = a\overset{\bullet 1\bullet 2}{b}a\overset{\bullet 1\bullet 2\bullet 3}{b}a$$

$$\mathsf{c}_{13} = a\overset{\bullet 2}{b}a\overset{\bullet 2}{b}a \quad \mathsf{c}_{14} = a\overset{\bullet 1\bullet 2}{b}a\overset{\bullet 2}{b}a \quad \mathsf{c}_{15} = a\overset{\bullet 2\bullet 1\bullet 2}{b}a\overset{}{b}a \quad \mathsf{c}_{16} = a\overset{\bullet 1\bullet 2}{b}a\overset{\bullet 1\bullet 2}{b}a$$

This list could be extremely lengthy when there are more pattern words or with a longer cluster text. To help us better understand the different ways of labeling, *skeletons* are introduced (see Bassino et al. [5, Section 5]).

Definition 5.1.1 (Skeleton) A skeleton is a cluster such that no distinguished occurrence is a factor of another distinguished occurrence.

Two dual operations, denoted by Skel and Flip, were introduced by Bassino et al. [5, Section 5] to relate clusters and skeletons, defined as follows.

- Let c be a cluster. The skeleton Skel(c) (denoted by $\underline{\mathsf{c}}$) of a decorated text c is obtained from c by undistinguishing (moving the status of an occurrence from "distinguished" to "not distinguished") any occurrence that is a factor of another distinguished occurrence in c.

- Let $\underline{\mathsf{c}}$ be a skeleton. Then the Flip operation associates to $\underline{\mathsf{c}}$ the set Flip($\underline{\mathsf{c}}$) of all clusters whose Skeleton is $\underline{\mathsf{c}}$. Flip($\underline{\mathsf{c}}$) can be also written as $\widetilde{\underline{\mathsf{c}}}$.

It is easy to prove that the *skeleton* of a cluster c is uniquely defined. Just repeatedly remove distinguished occurrences that are factors of other distinguished occurrences, until none of these remain. This process is always unambiguous, as we will demonstrate.

5.2 Bicolored decorated cluster

Back to Example 5.1.1, one should observe that $\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3, \mathsf{c}_4$ share the same skeleton

$$\mathrm{Skel}(\mathsf{c}_1) = \mathrm{Skel}(\mathsf{c}_2) = \mathrm{Skel}(\mathsf{c}_3) = \mathrm{Skel}(\mathsf{c}_4) = \overset{\blacksquare 1}{a}b\overset{\blacksquare 3}{aba}$$

This skeleton is identical to c_1.

On the other hand, c_5 through c_{12} share another skeleton, which is identical to c_5,

$$\mathrm{Skel}(\mathsf{c}_5) = \cdots = \mathrm{Skel}(\mathsf{c}_{12}) = ab\overset{\blacksquare 2}{a}\overset{\blacksquare 3}{ba}$$

Lastly, clusters c_{13} through c_{16} share the third skeleton, which is identical to c_{13},

$$\mathrm{Skel}(\mathsf{c}_{13}) = \mathrm{Skel}(\mathsf{c}_{14}) = \mathrm{Skel}(\mathsf{c}_{15}) = \mathrm{Skel}(\mathsf{c}_{16}) = ab\overset{\blacksquare 2}{a}\overset{\blacksquare 2}{ba}$$

This example demonstrates that even for the same text (here *ababa*), there could be multiple distinct skeletons for different groups of clusters. Each cluster can only map to one skeleton. The Flip operation yields the set of all clusters that share a skeleton. Using $\underline{\mathsf{c}} = \overset{\blacksquare 1}{a}b\overset{\blacksquare 3}{aba}$ to represent the *skeleton* of c_1 through c_4, Figure 5.1 illustrates the mapping relations of the Skel and Flip operations.

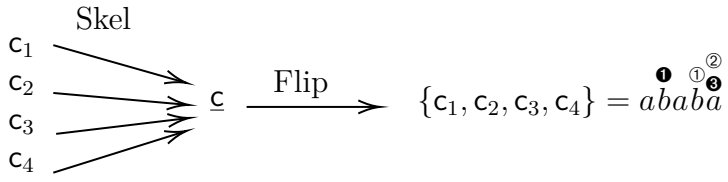


Figure 5.1. c_1 through c_4 in Example 5.1.1 share the same *skeleton*, $\overset{\blacksquare 1}{a}b\overset{\blacksquare 3}{aba}$. Flip($\underline{\mathsf{c}}$) returns the set $\{\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3, \mathsf{c}_4\}$.

In order to simplify our notation, the set of clusters given by a Flip operation can be denoted by a *bicolored decorated cluster*. Considering our Example 5.1.1, when the Flip operation is applied to $\underline{\mathsf{c}}$ (remember that $\underline{\mathsf{c}}$ is the common skeleton of $\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3$, and c_4), we have

$$\widetilde{\underline{\mathsf{c}}} = \mathrm{Flip}(\underline{\mathsf{c}}) = \{\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3, \mathsf{c}_4\} = a\overset{\bullet}{b}a\overset{\bullet}{b}\overset{\circ}{a}$$

Here we use the bicolored decorated cluster $a\overset{\bullet}{b}a\overset{\bullet}{b}\overset{\circ}{a}$ to represent the set $\{\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3, \mathsf{c}_4\}$.

It is a convenient notation adopted by Bassino et al. [5]. The black filled cirlces are located above the ending letters of those occurrences in the skeleton. Meanwhile, the white filled circles are placed above the factor occurrences. In this way, all clusters that share the same skeleton can be represented in one bicolored decorated cluster.

In a bicolored decorated cluster, each factor occurrence (labelled by a white filled circle) could be distinguished or not. Therefore, given a bicolored decorated cluster, one can easily write the full set of decorated clusters.

The full set of decorated clusters represented by $a\overset{\bullet}{b}a\overset{\bullet}{b}\overset{\circ}{a}$ includes $2^2 = 4$ decorated clusters, i.e.,

$$\{\mathsf{c}_1 = a\overset{\bullet}{b}ab\overset{\bullet}{a}, \quad \mathsf{c}_2 = a\overset{\bullet}{b}a\overset{\bullet}{b}\overset{\bullet}{a}, \quad \mathsf{c}_3 = a\overset{\bullet}{b}ab\overset{\bullet}{a}, \quad \mathsf{c}_4 = a\overset{\bullet}{b}a\overset{\bullet}{b}\overset{\bullet}{a}\}$$

They all have the same skeleton, $a\overset{\bullet}{b}ab\overset{\bullet}{a}$.

Likewise, the set of clusters

$$\begin{aligned}\{\mathsf{c}_5 &= ab\overset{\bullet}{a}b\overset{\bullet}{a}, \quad \mathsf{c}_6 = a\overset{\bullet}{b}\overset{\bullet}{a}b\overset{\bullet}{a}, \quad \mathsf{c}_7 = ab\overset{\bullet}{a}b\overset{\bullet}{a}, \quad \mathsf{c}_8 = ab\overset{\bullet}{a}b\overset{\bullet}{a},\\ \mathsf{c}_9 &= a\overset{\bullet}{b}\overset{\bullet}{a}b\overset{\bullet}{a}, \quad \mathsf{c}_{10} = a\overset{\bullet}{b}\overset{\bullet}{a}b\overset{\bullet}{a}, \quad \mathsf{c}_{11} = ab\overset{\bullet}{a}b\overset{\bullet}{a}, \quad \mathsf{c}_{12} = a\overset{\bullet}{b}\overset{\bullet}{a}b\overset{\bullet}{a}\}\end{aligned}$$

can be written as $a\overset{\circ}{b}\overset{\bullet}{a}\overset{\circ}{b}\overset{\bullet}{a}$, which includes $2^3 = 8$ decorated clusters, all listed above. These 8 clusters share the same skeleton, $ab\overset{\bullet}{a}b\overset{\bullet}{a}$.

Lastly, the set of the other four clusters

$$\{\mathsf{c}_{13} = ab\overset{\bullet}{a}b\overset{\bullet}{a}, \quad \mathsf{c}_{14} = a\overset{\bullet}{b}\overset{\bullet}{a}b\overset{\bullet}{a}, \quad \mathsf{c}_{15} = ab\overset{\bullet}{a}b\overset{\bullet}{a}, \quad \mathsf{c}_{16} = a\overset{\bullet}{b}\overset{\bullet}{a}b\overset{\bullet}{a}\}$$

is equivalent to $a\overset{①}{b}\overset{❷}{a}\overset{①}{b}\overset{❷}{a}$ (which indeed represents $2^2 = 4$ clusters). The clusters c_{13} through c_{16} share the same skeleton $ab\overset{❷}{a}b\overset{❷}{a}$.

5.3 Notations regarding a bicolored decorated cluster

A fully bicolored decorated cluster $\widetilde{\underline{\mathsf{c}}}$ is essentially a set of clusters which share the same skeleton. The bicolored decoration is composed of three parts, (1) the text of the cluster, c; (2) the distinguished occurrences defining the skeleton; and (3) the factor occurrences.

The following notations are used to represent the three parts. As an example, given $\widetilde{\underline{\mathsf{c}}} = a\overset{①}{b}\overset{❷}{a}\overset{①}{b}\overset{②}{a}$, the three parts can be denoted by $\mathsf{c}, \mathcal{D}$ and $\mathcal{F}$, where

$\mathsf{c} = ababa$

$\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5) = (\emptyset, \emptyset, 2, \emptyset, 3)$

$\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4, \mathcal{F}_5) = (\emptyset, 1, \emptyset, 1, 2)$

Definition 5.3.1 gives the general definition of $\mathcal{D}$ and $\mathcal{F}$.

Definition 5.3.1 Consider a skeleton $\underline{\mathsf{c}}$ and its fully bicolored decorated cluster Flip($\underline{\mathsf{c}}$), or $\widetilde{\underline{\mathsf{c}}}$.

We use $\mathcal{D}_i$ to denote the index above the distinguished position i of the skeleton $\underline{\mathsf{c}}$, or $\emptyset$ if position i is not labelled by a distinguished occurrence in $\underline{\mathsf{c}}$.

We let $\mathcal{F}_i$ denote the factor index above position i of the bicolored decorated cluster $\widetilde{\underline{\mathsf{c}}}$, or $\emptyset$ if position i is not labelled by a factor occurrence in $\widetilde{\underline{\mathsf{c}}}$.

Finally, we use $\mathcal{D}$ and $\mathcal{F}$ to denote the tuples of $\mathcal{D}_i$ and $\mathcal{F}_i$. i.e., $\mathcal{D} = (\mathcal{D}_i)_{1 \le i \le |\mathsf{c}|}$ and $\mathcal{F} = (\mathcal{F}_i)_{1 \le i \le |\mathsf{c}|}$.

Thus, a fully bicolored decorated cluster $\widetilde{\underline{\mathsf{c}}}$ can be denoted by the triple $(\mathsf{c}, \mathcal{D}, \mathcal{F})$.

5.4 Right extension set

Let us introduce two more concepts, *right extension set* and *bicolored right extension set*. Again, we follow the notation of Bassino et al. [5].

Definition 5.4.1 (Right extension set) The right extension set of a pair of words (u, v) is

$$\mathcal{E}_{u,v} = \Big\{ e \mid \text{there exists } e' \in \mathcal{A}^+ \text{ such that } ue = e'v \text{ with } 0 < |e| < |v| \Big\}$$

Given a right extension set, by adding the bicolored decorated numbers of distinguished occurrences in v, we obtain the *bicolored decorated right extension set*. A *right extension set* or a *bicolored decorated right extension set* is represented in a matrix.

Definition 5.4.2 (Bicolored decorated right extension set) Let u, v be two words, and v be the bicolored decorated word of v. Then the *bicolored decorated right extension set* of u, v is

$$\mathsf{E}_{u,v} = \bigcup_{e \in \mathcal{E}_{u,v}} \mathrm{Suff}_{|e|}(\mathrm{Flip}(\mathsf{v}))$$

This definition is illustrated by Fig. 5.2.

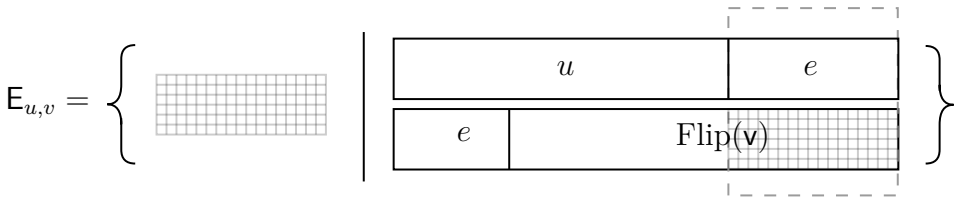


Figure 5.2. The two long rectangles ue and $e\mathsf{v}$ represent two identical strings: $ue = ev$. (Here, v is the bicolored decorated word of v.) The *bicolored decorated right extension set* $\mathsf{E}_{u,v}$ includes the part covered by the grid area, which has the same string as e and which has the bicolored decorated numbers corresponding to those in the suffix of v.

The following two examples provide more details of the two concepts.

Example 5.4.1 (Bicolored decorated right extension set) A pattern set

$$\mathcal{U} = \{ab, aba, bab, baba\}$$

gives

$$\mathsf{u}_1 = a\overset{\bullet 1}{b}, \quad \mathsf{u}_2 = ab\overset{\bullet 2}{a}, \quad \mathsf{u}_3 = ba\overset{\bullet 3}{b}, \quad \mathsf{u}_4 = bab\overset{\bullet 4}{a},$$

and

$$\mathrm{Flip}(\mathsf{u}_1) = \{a\overset{\bullet 1}{b}\}, \quad \mathrm{Flip}(\mathsf{u}_2) = \{a\overset{\circ 1}{b}\overset{\bullet 2}{a}\},$$
$$\mathrm{Flip}(\mathsf{u}_3) = \{ba\overset{\bullet 3}{b}\}, \quad \mathrm{Flip}(\mathsf{u}_4) = \{ba\overset{\circ 1}{b}\overset{\bullet 2}{a}\overset{\circ 3}{a}\overset{\bullet 4}{}\}.$$

The *right extension set* $\mathcal{E}$ (without ambiguity, we omit the subscript of $\mathcal{E}_{\mathcal{U}}$) is

$$\mathcal{E} = \begin{pmatrix} \emptyset & \emptyset & ab & aba \\ b & ba & b & ba \\ \emptyset & a & ab & aba \\ b & ba & b & ba \end{pmatrix}$$

and the *bicolored decorated right extension set* E

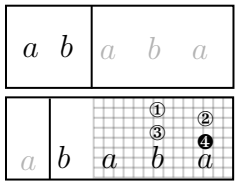
$$\mathsf{E} = \begin{pmatrix} \emptyset & \emptyset & \{a\overset{\bullet 1}{b}\} & \{a\overset{\circ 1}{b}\overset{\bullet 2}{a}\} \\ \{\overset{\bullet 1}{b}\} & \{\overset{\circ 1}{b}\overset{\bullet 2}{a}\} & \{\overset{\bullet 3}{b}\} & \{\overset{\circ 3}{b}\overset{\bullet 4}{a}\} \\ \emptyset & \{\overset{\bullet 2}{a}\} & \{a\overset{\bullet 3}{b}\} & \{a\overset{\circ 3}{b}\overset{\bullet 4}{a}\} \\ \{\overset{\bullet 1}{b}\} & \{\overset{\circ 1}{b}\overset{\bullet 2}{a}\} & \{\overset{\bullet 3}{b}\} & \{\overset{\circ 3}{b}\overset{\bullet 4}{a}\} \end{pmatrix}.$$

Here we only elaborate on the first row in E, which in turn lists E_{u_1,u_1}, E_{u_1,u_2}, E_{u_1,u_3}, and E_{u_1,u_4}.

$$\mathsf{E}_{u_1,u_1} = \mathsf{E}_{ab,ab} = \left\{ \quad \middle| \quad \begin{array}{|cc|}\hline a & b \\ \hline a & b \\ \hline \end{array}\ a\ \overset{\bullet 1}{b} \right\} = \emptyset$$

$$\mathsf{E}_{u_1,u_2} = \mathsf{E}_{ab,aba} = \left\{ \quad \middle| \quad \begin{array}{|cc|}\hline a & b \\ \hline a & b \\ \hline \end{array}\ a\ \overset{\circ 1}{b}\ \overset{\bullet 2}{a} \right\} = \emptyset$$

$$\mathsf{E}_{u_1,u_3} = \mathsf{E}_{ab,bab} = \left\{ \quad \middle| \quad \begin{array}{|cc|cc|}\hline a & b & a & b \\ \hline a & b & a & \overset{\circ 3}{b}\overset{\bullet 1}{} \\ \hline \end{array} \right\} = \{a\overset{\bullet 3}{b}\}$$

$$\mathsf{E}_{u_1,u_4} = \mathsf{E}_{ab,baba} = \left\{ \right.$$  $$\left| \right.$$  $$\left. \right\} = \{a\overset{①②}{\underset{③❹}{ba}}\}$$

Following similar ideas, the other rows in E can be obtained.

In Example 5.4.1, every non-empty right extension set only contains one element. But this is certainly not necessary. Depending on the pattern words, a right extension set could contain many elements. Example 5.4.2 provides such a case.

Example 5.4.2 For $\mathcal{U} = \{a^3, a^4\}$, we have

$$\mathsf{u}_1 = aa\overset{❶}{a}, \quad \mathsf{u}_2 = aaa\overset{❷}{a}$$

and

$$\mathrm{Flip}(\mathsf{u}_1) = \{aa\overset{❶}{a}\}, \quad \mathrm{Flip}(\mathsf{u}_2) = \{aa\overset{①}{\underset{①❷}{a}}\overset{①}{a}\}.$$

Then we have

$$\mathcal{E} = \begin{pmatrix} a + aa, & aa + aaa \\ a + aa, & a + aa + aaa \end{pmatrix}$$

and

$$\mathsf{E} = \begin{pmatrix} \{\overset{❶}{a}, a\overset{❶}{a}\} & \{\overset{①}{\underset{①❷}{aa}}, aa\overset{①}{\underset{①❷}{a}}\} \\ \{\overset{❶}{a}, a\overset{❶}{a}\} & \{\overset{❷}{a}, \overset{①}{\underset{①❷}{aa}}, aa\overset{①}{\underset{①❷}{a}}\} \end{pmatrix}.$$

5.5 Set of all clusters

The bicolored decoration enables us to write the set of clusters in a concise form similar to that in reduced cases.

Given the set of pattern words $\mathcal{U} = \{u_1, \cdots, u_r\}$, the set of all clusters C can be obtained by

$$\mathsf{C} = (\mathrm{Flip}(\mathsf{u}_1), \ldots, \mathrm{Flip}(\mathsf{u}_r)) \cdot \mathsf{E}^* \cdot \begin{pmatrix} \varepsilon \\ \vdots \\ \varepsilon \end{pmatrix} \tag{5.1}$$

This is a general result, not only for non-reduced cases. We emphasize that Equation (5.1) can be also used for reduced $\mathcal{U}$, i.e., no pattern word in $\mathcal{U}$ is a factor of another. When $\mathcal{U}$ is reduced, Equation (5.1) provides the form that we already used in previous chapters, such as Equation (4.22).

5.6 Set of all skeletons

Each pattern word in $\mathcal{U} = \{u_1, \cdots, u_r\}$ can be treated as a skeleton by itself, as we did in Example 5.4.1 and Example 5.4.2. Each of these skeletons only consists of a pattern word u_j and a monocolor label j upon $\widehat{u_j}$, the last letter of u_j. These skeletons are denoted by $\underline{\mathsf{u}}_1, \underline{\mathsf{u}}_2, \cdots, \underline{\mathsf{u}}_r$, respectively.

We also denote $\underline{\mathsf{E}} = (\underline{\mathsf{E}}_{i,j})$. For instance, in Example 5.4.2, where

$$\mathsf{E} = \begin{pmatrix} \{\overset{\bullet 1}{a}, \overset{}{a}\overset{\bullet 1}{a}\} & \{\overset{\circ 1}{\overset{\bullet 2}{a}}\overset{}{a}, \overset{\circ 1}{a}\overset{\bullet 2}{a}\overset{}{a}\} \\ \{\overset{\bullet 1}{a}, \overset{}{a}\overset{\bullet 1}{a}\} & \{\overset{\circ 1}{\overset{\bullet 2}{a}}, \overset{\circ 1}{a}\overset{\bullet 2}{a}, \overset{\circ 1}{a}\overset{\bullet 2}{a}\overset{}{a}\} \end{pmatrix}$$

We have

$$\underline{\mathsf{E}} = \begin{pmatrix} \{\overset{\bullet 1}{a}, \overset{}{a}\overset{\bullet 1}{a}\} & \{\overset{}{a}\overset{\bullet 2}{a}, \overset{}{a}\overset{}{a}\overset{\bullet 2}{a}\} \\ \{\overset{\bullet 1}{a}, \overset{}{a}\overset{\bullet 1}{a}\} & \{\overset{\bullet 2}{a}, \overset{}{a}\overset{\bullet 2}{a}, \overset{}{a}\overset{}{a}\overset{\bullet 2}{a}\} \end{pmatrix}.$$

Using these notations, we can write the set of all skeletons, denoted by $\underline{\mathsf{C}}$, in the following form.

$$\underline{\mathsf{C}} = (\underline{\mathsf{u}}_1, \underline{\mathsf{u}}_2, \cdots, \underline{\mathsf{u}}_r) \cdot \underline{\mathsf{E}}^* \cdot \begin{pmatrix} \varepsilon \\ \vdots \\ \varepsilon \end{pmatrix} \tag{5.2}$$

In fact, if a skeleton includes exactly $k+1$ occurrences, it is in the form of

$$\underline{\mathsf{C}}_{k+1} = (\underline{\mathsf{u}}_1, \underline{\mathsf{u}}_2, \cdots, \underline{\mathsf{u}}_r) \cdot \underline{\mathsf{E}}^k \cdot \begin{pmatrix} \varepsilon \\ \vdots \\ \varepsilon \end{pmatrix} \tag{5.3}$$

where $(\underline{\mathsf{u}}_1, \underline{\mathsf{u}}_2, \cdots, \underline{\mathsf{u}}_r)$ contributes the first occurrence, and $\underline{\mathsf{E}}^k$ contributes the other k occurrences. For convenience, we define the concepts of a $(k+1)$-skeleton and $(k+1)$-cluster.

Definition 5.6.1 ($(k+1)$-skeleton) A skeleton that is composed of $k+1$ occurrences is a $(k+1)$-skeleton. A $(k+1)$-skeleton is in the form of Equation (5.3).

Definition 5.6.2 ($(k+1)$-cluster) A $(k+1)$-cluster is a cluster whose skeleton is a $(k+1)$-skeleton.

When a Flip operation is applied to a $(k+1)$-skeleton $\underline{\mathsf{c}}$, every element in the set $\mathrm{Flip}(\underline{\mathsf{c}})$ (a.k.a. $\widetilde{\underline{\mathsf{c}}}$) is a $(k+1)$-cluster.

5.7 Generating functions of Flip with a Bernoulli text source

Consider a pattern set $\mathcal{U} = \{u_1, \cdots, u_r\}$ and a fully bicolored decorated cluster

$$\widetilde{\underline{\mathsf{c}}} = (\mathsf{c}, \mathcal{D}, \mathcal{F})$$

with skeleton $\underline{\mathsf{c}}$. The text of the skeleton has length $|\mathsf{c}| = \ell$, with $\mathsf{c} = \alpha_1\alpha_2 \ldots \alpha_\ell$, where $\mathcal{D} = (\mathcal{D}_i)_{1\le i\le \ell}$, $\mathcal{F} = (\mathcal{F}_i)_{1\le i\le \ell}$, and $\alpha_i \in \mathcal{A}$. Here we use the notations introduced in Definition 5.3.1.

When a text is generated by a Bernoulli source, e.g., when a letter α_i occurs with a probability π_{α_i}, then the generating function $\widetilde{\underline{c}}(z, \mathbf{t})$ of the set of clusters $\mathrm{Flip}(\underline{\mathsf{c}})$ built upon the skeleton $\underline{\mathsf{c}}$ is

$$\widetilde{\underline{c}}(z, \mathbf{t}) = \prod_{i=1}^{\ell} \left[\pi_{\alpha_i} z \times \left(\prod_{j\in\mathcal{D}_i} t_j \right) \times \left(\prod_{s\in\mathcal{F}_i} (1+t_s) \right) \right] \tag{5.4}$$

where the variable t_j (or t_s) counts the occurrences of the pattern word u_j (or u_s).

It is straightforward to verify Equation (5.4). Each $\mathcal{D}_i$ represents an occurrence in the skeleton $\underline{\mathsf{c}}$. Therefore, it must be distinguished. On the other hand, each $\mathcal{F}_i$ stands for an occurrence of a factor pattern, which could be distinguished or not.

Let us apply Equation (5.4) on the following example.

Example 5.7.1 By (5.4), we can write the generating functions of the Flips in Section 5.2 as follows.

(1) A Flip $\widetilde{\underline{c}} = a\overset{\textbf{❶}}{b}a\overset{①}{b}\overset{②}{\overset{③}{a}}$ (with the skeleton $\underline{c} = a\overset{\textbf{❶}}{b}ab\overset{③}{a}$) has the following generating function:

$$\begin{aligned}\widetilde{\underline{c}}(z,\mathbf{t}) &= \pi_a z \cdot \pi_b z t_1 \cdot \pi_a z \cdot \pi_b z(1+t_1) \cdot \pi_a z(1+t_2)t_3 \\ &= \pi_a^3\pi_b^2 z^5 \cdot t_1 \cdot (1+t_1) \cdot (1+t_2) \cdot t_3\end{aligned}$$

(2) A Flip $\widetilde{\underline{c}} = a\overset{①}{\overset{\textbf{❷}}{b}}a\overset{①}{\overset{③}{b}}\overset{②}{a}$ (with the skeleton $\underline{c} = a\overset{\textbf{❷}}{b}ab\overset{③}{a}$) has the following generating function:

$$\begin{aligned}\widetilde{\underline{c}}(z,\mathbf{t}) &= \pi_a z \cdot \pi_b z(1+t_1) \cdot \pi_a z t_2 \cdot \pi_b z(1+t_1) \cdot \pi_a z(1+t_2)t_3 \\ &= \pi_a^3\pi_b^2 z^5 \cdot (1+t_1)^2 \cdot (1+t_2) \cdot t_2 \cdot t_3\end{aligned}$$

(3) A Flip $\widetilde{\underline{c}} = a\overset{①}{\overset{\textbf{❷}}{b}}a\overset{①}{b}\overset{\textbf{❷}}{a}$ (with the skeleton $\underline{c} = a\overset{\textbf{❷}}{b}ab\overset{\textbf{❷}}{a}$) has the following generating function:

$$\begin{aligned}\widetilde{\underline{c}}(z,\mathbf{t}) &= \pi_a z \cdot \pi_b z(1+t_1) \cdot \pi_a z t_2 \cdot \pi_b z(1+t_1) \cdot \pi_a z t_2 \\ &= \pi_a^3\pi_b^2 z^5 \cdot (1+t_1)^2 \cdot t_2^2\end{aligned}$$

As we see, although a Flip can be represented in two equivalent ways—either a set of mono-color decorated clusters or a bicolored decorated cluster, the bicolored representation makes it very efficient to conclude a Flip's generating function.

5.8 Generating functions of Flip with a Markovian text source

Equation (5.4) can be readily modified when a text is generated by a Markovian source of order 1. Assuming Flip($\underline{c}$) starts after the letter α_0, the generating function ${}^{\alpha_0}\widetilde{\underline{c}}(z,\mathbf{t})$ of the set of clusters Flip($\underline{c}$) built upon the skeleton $\underline{c}$ is

$${}^{\alpha_0}\widetilde{\underline{c}}(z,\mathbf{t}) = \prod_{i=1}^{\ell}\left[p_{\alpha_{i-1},\alpha_i}\cdot z \times \left(\prod_{j\in\mathcal{D}_i} t_j\right) \times \left(\prod_{s\in\mathcal{F}_i}(1+t_s)\right)\right] \tag{5.5}$$

where $p_{\alpha_{i-1},\alpha_i}$ is the transition probability from letter α_{i-1} to letter α_i.

In a Markovian scenario, Example 5.7.1 is modified into the following one.

Example 5.8.1 Let ${}^{a}\widetilde{\underline{c}}$ denote a Flip $\widetilde{\underline{c}}$ starting after the letter a. By Equation (5.5), we write the generating functions of the Flips in Section 5.2, as follows.

(1) A Flip ${}^a\widetilde{\underline{\mathsf{c}}} = a\overset{\textcircled{1}}{b}a\overset{\textcircled{1}\textcircled{2}}{b}\overset{\textcircled{3}}{a}$ (with the skeleton $\underline{\mathsf{c}} = \overset{\textcircled{1}}{a}ba\overset{\textcircled{3}}{b}a$) has the following generating function:

$$\begin{aligned}{}^a\widetilde{\underline{c}}(z,\mathbf{t}) &= p_{a,a}z \cdot p_{a,b}zt_1 \cdot p_{b,a}z \cdot p_{a,b}z(1+t_1) \cdot p_{b,a}z(1+t_2)t_3 \\ &= p_{a,a}p_{a,b}^2p_{b,a}^2z^5 \cdot t_1 \cdot (1+t_1) \cdot (1+t_2) \cdot t_3\end{aligned}$$

(2) A Flip ${}^a\widetilde{\underline{\mathsf{c}}} = a\overset{\textcircled{1}}{b}\overset{\textcircled{2}}{a}\overset{\textcircled{1}}{b}\overset{\textcircled{3}}{a}$ (with the skeleton $\underline{\mathsf{c}} = ab\overset{\textcircled{2}}{a}\overset{\textcircled{3}}{b}a$) has the following generating function:

$$\begin{aligned}{}^a\widetilde{\underline{c}}(z,\mathbf{t}) &= p_{a,a}z \cdot p_{a,b}z(1+t_1) \cdot p_{b,a}zt_2 \cdot p_{a,b}z(1+t_1) \cdot p_{b,a}z(1+t_2)t_3 \\ &= p_{a,a}p_{a,b}^2p_{b,a}^2z^5 \cdot (1+t_1)^2 \cdot (1+t_2) \cdot t_2 \cdot t_3\end{aligned}$$

(3) A Flip ${}^a\widetilde{\underline{\mathsf{c}}} = a\overset{\textcircled{1}}{b}\overset{\textcircled{2}}{a}\overset{\textcircled{1}}{b}\overset{\textcircled{2}}{a}$ (with the skeleton $\underline{\mathsf{c}} = ab\overset{\textcircled{2}}{a}\overset{\textcircled{2}}{b}a$) has the following generating function:

$$\begin{aligned}{}^a\widetilde{\underline{c}}(z,\mathbf{t}) &= p_{a,a}z \cdot p_{a,b}z(1+t_1) \cdot p_{b,a}zt_2 \cdot p_{a,b}z(1+t_1) \cdot p_{b,a}zt_2 \\ &= p_{a,a}p_{a,b}^2p_{b,a}^2z^5 \cdot (1+t_1)^2 \cdot t_2^2\end{aligned}$$

5.9 Generating functions of clusters

Now we are able to obtain $\widetilde{\underline{\mathsf{c}}}(z,\mathbf{t})$, the generating function of a Flip. The next construction to consider is the generating functions of clusters $\xi(z,\mathbf{t})$. In this section, we come back to the non-Markov case, and continue to use the notations in Bassino et al. [5, Section 5]. The discussion of generating functions of clusters in a Markovian context will be given at the beginning of the next chapter.

According to Equation (5.1), for a non-reduced pattern set $\mathcal{U} = \{u_1, \ldots, u_r\}$, the generating function $\xi(z,\mathbf{t})$ of clusters is

$$\xi(z,\mathbf{t}) = (U_1(z,\mathbf{t}), \ldots, U_r(z,\mathbf{t})) \cdot (\mathbb{I} - \mathbb{E}(z,\mathbf{t}))^{-1} \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tag{5.6}$$

where $U_i(z,\mathbf{t})$ is the generating function of $\mathrm{Flip}(\mathsf{u}_i)$, and $\mathbb{E}(z,\mathbf{t})$ is the matrix in which every element corresponds to a generating functions of a right extension sets. Note that all $\mathrm{Flip}(\mathsf{u}_i)$ and right extension sets are bicolored [5, Section 5].

Formula (5.4) already provides a direct form of $U_i(z,\mathbf{t})$. In fact, this formula can be also applied to the elements $\mathsf{E}_{i,j}$ in the matrix E, even though they are neither clusters nor skeletons, but rather bicolored decorated right extensions.

The generating function of $\mathsf{E}_{i,j}$ is $E_{i,j}(z,\mathbf{t})$, and therefore, we use $\mathbb{E}(z,\mathbf{t})$ to denote the matrix of $(E_{i,j}(z,\mathbf{t}))$.

Example 5.9.1 Let us calculate the generating functions of clusters in the following two cases. In both cases, we assume a text is generated by a Bernoulli source.

(1) A pattern set $\mathcal{U} = \{u_1,u_2,u_3,u_4\} = \{ab, aba, bab, baba\}$. We have

$$\mathsf{u}_1 = a\overset{\bullet 1}{b},\quad \mathsf{u}_2 = ab\overset{\bullet 2}{a},\quad \mathsf{u}_3 = ba\overset{\bullet 3}{b},\quad \mathsf{u}_4 = bab\overset{\bullet 4}{a},$$

and

$$\begin{aligned}
&\mathrm{Flip}(\mathsf{u}_1) = \{a\overset{\bullet 1}{b}\}, && U_1(z,t_1,t_2,t_3,t_4) = \pi_a\pi_b z^2 t_1\\
&\mathrm{Flip}(\mathsf{u}_2) = \{a\overset{\circ 1}{b}\overset{\bullet 2}{a}\}, && U_2(z,t_1,t_2,t_3,t_4) = \pi_a^2\pi_b z^3(1+t_1)t_2\\
&\mathrm{Flip}(\mathsf{u}_3) = \{ba\overset{\bullet 3}{b}\}, && U_3(z,t_1,t_2,t_3,t_4) = \pi_a\pi_b^2 z^3(1+t_1)t_3\\
&\mathrm{Flip}(\mathsf{u}_4) = \{ba\overset{\circ 1}{b}\overset{\bullet 2}{a}\overset{\bullet 3}{b}\overset{\bullet 4}{a}\}, && U_4(z,t_1,t_2,t_3,t_4) = \pi_a^2\pi_b^2 z^4(1+t_1)(1+t_2)(1+t_3)t_4
\end{aligned}$$

The *bicolored decorated right extension set* E is

$$\mathsf{E} = \begin{pmatrix}
\emptyset & \emptyset & \{a\overset{\circ 1}{b}\overset{\bullet 3}{b}\} & \{a\overset{\circ 1}{b}\overset{\circ 2}{a}\overset{\bullet 3}{b}\overset{\bullet 4}{a}\}\\
\{\overset{\bullet 1}{b}\} & \{\overset{\circ 1}{b}\overset{\bullet 2}{a}\} & \{\overset{\circ 1}{b}\overset{\bullet 3}{b}\} & \{\overset{\circ 1}{b}\overset{\circ 2}{a}\overset{\bullet 3}{b}\overset{\bullet 4}{a}\}\\
\emptyset & \{\overset{\bullet 2}{a}\} & \{a\overset{\circ 1}{b}\overset{\bullet 3}{b}\} & \{a\overset{\circ 1}{b}\overset{\circ 2}{a}\overset{\bullet 3}{b}\overset{\bullet 4}{a}\}\\
\{\overset{\bullet 1}{b}\} & \{\overset{\circ 1}{b}\overset{\bullet 2}{a}\} & \{\overset{\circ 1}{b}\overset{\bullet 3}{b}\} & \{\overset{\circ 1}{b}\overset{\circ 2}{a}\overset{\bullet 3}{b}\overset{\bullet 4}{a}\}
\end{pmatrix}.$$

Therefore,

$$\mathbb{E}\,(z,t_1,t_2,t_3,t_4)$$
$$=\begin{pmatrix} 0 & 0 & \pi_a\pi_b z^2(1+t_1)t_3 & \pi_a^2\pi_b z^3(1+t_1)(1+t_2)(1+t_3)t_4 \\ \pi_b z t_1 & \pi_a\pi_b z^2(1+t_1)t_2 & \pi_b z(1+t_1)t_3 & \pi_a\pi_b z^2(1+t_1)(1+t_2)(1+t_3)t_4 \\ 0 & \pi_a z t_2 & \pi_a\pi_b z^2(1+t_1)t_3 & \pi_a^2\pi_b z^3(1+t_1)(1+t_2)(1+t_3)t_4 \\ \pi_b z t_1 & \pi_a\pi_b z^2(1+t_1)t_2 & \pi_b z(1+t_1)t_3 & \pi_a\pi_b z^2(1+t_1)(1+t_2)(1+t_3)t_4 \end{pmatrix}$$

Thus, the generating function of clusters is

$$\xi(z,t_1,t_2,t_3,t_4)=(U_1(z,\mathbf{t}),\ldots,U_4(z,\mathbf{t}))\cdot(\mathbb{I}-\mathbb{E}(z,\mathbf{t}))^{-1}\cdot\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}.$$

(2) For $\mathcal{U}=\{u_1,u_2\}=\{a^3,a^4\}$, we have

$$\mathsf{u}_1=aa\overset{\bullet 1}{a},\quad \mathsf{u}_2=aaa\overset{\bullet 2}{a}$$

and

$$\begin{aligned}\mathrm{Flip}(\mathsf{u}_1)&=\{aa\overset{\bullet 1}{a}\},\quad U_1(z,t_1,t_2)=\pi_a^3z^3t_1\\ \mathrm{Flip}(\mathsf{u}_2)&=\{aa\overset{\circ 1}{a}\overset{\bullet 2}{a}\},\quad U_2(z,t_1,t_2)=\pi_a^4z^4(1+t_1)^2t_2\end{aligned}$$

Then we have

$$\mathsf{E}=\begin{pmatrix}\{\overset{\bullet 1}{a},a\overset{\bullet 1}{a}\} & \{\overset{\circ 1}{a}\overset{\bullet 2}{a},a\overset{\circ 1}{a}\overset{\bullet 2}{a}\}\\ \{\overset{\bullet 1}{a},a\overset{\bullet 1}{a}\} & \{\overset{\bullet 2}{a},\overset{\circ 1}{a}\overset{\bullet 2}{a},a\overset{\circ 1}{a}\overset{\bullet 2}{a}\}\end{pmatrix}$$

and

$$\mathbb{E}\,(z,t_1,t_2)=\begin{pmatrix}\pi_a z t_1+\pi_a^2z^2t_1, & \pi_a^2z^2(1+t_1)^2t_2+\pi_a^3z^3(1+t_1)^2t_2\\ \pi_a z t_1+\pi_a^2z^2t_1, & \pi_a z(1+t_1)t_2+\pi_a^2z^2(1+t_1)^2t_2+\pi_a^3z^3(1+t_1)^2t_2\end{pmatrix}$$

and finally,

$$\xi(z,t_1,t_2) = (U_1(z,t_1,t_2), U_2(z,t_1,t_2)) \cdot (\mathbb{I} - \mathbb{E}(z,t_1,t_2))^{-1} \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{5.7}$$

6. INCLUSION-EXCLUSION METHOD FOR NON-REDUCED PATTERNS WITH A MARKOVIAN TEXT SOURCE

6.1 Generating function of clusters

We follow our notations from the previous chapter. The set of pattern words is $\mathcal{U} = \{u_1, \cdots, u_r\}$. The last letter of each pattern word is denoted by $\widehat{u_1}, \widehat{u_2}, \cdots, \widehat{u_r}$, respectively, where $\{\widehat{u_1}, \widehat{u_2}, \cdots, \widehat{u_r}\} \subseteq \mathcal{A}$.

Formula (5.6) provides the generating function of clusters in the Bernoulli case. When first order Markovian dependence is considered, and assuming the previous letter immediately before the cluster is a, we rewrite the generating function in the following form.

$$ {}^{a}\xi(z,\mathbf{t}) = ({}^{a}U_1(z,\mathbf{t}),\ldots,{}^{a}U_r(z,\mathbf{t}))\cdot(\mathbb{I}-{}^{\widehat{\mathcal{U}}}\mathbb{E}(z,\mathbf{t}))^{-1}\cdot\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = {}^{a}\eta(z,\mathbf{t})\cdot\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} \tag{6.1} $$

where

$$ {}^{a}\eta(z,\mathbf{t}) := ({}^{a}U_1(z,\mathbf{t}),\ldots,{}^{a}U_r(z,\mathbf{t}))\cdot(\mathbb{I}-{}^{\widehat{\mathcal{U}}}\mathbb{E}(z,\mathbf{t}))^{-1} \tag{6.2} $$

and

$$ {}^{\widehat{\mathcal{U}}}\mathbb{E}(z,\mathbf{t}) := \begin{pmatrix} {}^{\widehat{u_1}}E_{1,1}(z,\mathbf{t}) & {}^{\widehat{u_1}}E_{1,2}(z,\mathbf{t}) & \ldots & {}^{\widehat{u_1}}E_{1,r}(z,\mathbf{t}) \\ {}^{\widehat{u_2}}E_{2,1}(z,\mathbf{t}) & {}^{\widehat{u_2}}E_{2,2}(z,\mathbf{t}) & \ldots & {}^{\widehat{u_2}}E_{2,r}(z,\mathbf{t}) \\ \vdots & \vdots & \ddots & \vdots \\ {}^{\widehat{u_r}}E_{r,1}(z,\mathbf{t}) & {}^{\widehat{u_r}}E_{r,2}(z,\mathbf{t}) & \ldots & {}^{\widehat{u_r}}E_{r,r}(z,\mathbf{t}) \end{pmatrix}. \tag{6.3} $$

6.2 Generating function of a decorated text T

With the alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_\ell\}$ and the pattern set $\mathcal{U} = \{u_1, u_2, \ldots, u_r\}$, a decorated text right after a letter a_i could start with the following $1 + \ell + r$ possible cases:
– there is no text, or

– a letter a_j which does not belong to a cluster ($j \in \{1, 2, \cdots, \ell\}$), or, lastly,

– a cluster ending in $\widehat{\mathrm{Flip}(\mathsf{u}_s)}$, where $s \in \{1, 2, \cdots, r\}$.

The combinatorial structure of ${}^{a_i}\mathsf{T}$ is

$$
{}^{a_i}\mathsf{T} = [a_i] \begin{cases} \epsilon \\ a_1 \cdot {}^{a_1}\mathsf{T} \\ \vdots \\ a_\ell \cdot {}^{a_\ell}\mathsf{T} \\ (\text{a cluster ending in } \widehat{\mathrm{Flip}(\mathsf{u}_1)}) \cdot {}^{\widehat{u_1}}\mathsf{T} \\ \vdots \\ (\text{a cluster ending in } \widehat{\mathrm{Flip}(\mathsf{u}_r)}) \cdot {}^{\widehat{u_r}}\mathsf{T} \end{cases}
$$

We have

$$
{}^{a_i}T(z, \mathbf{t}) = 1 + \sum_{j=1}^{\ell} p_{a_i, a_j} \cdot z \cdot {}^{a_j}T(z, \mathbf{t}) + {}^{a_i}\eta(z, \mathbf{t}) \cdot \begin{pmatrix} {}^{\widehat{u_1}}T(z, \mathbf{t}) \\ \vdots \\ {}^{\widehat{u_r}}T(z, \mathbf{t}) \end{pmatrix} \tag{6.4}
$$

Therefore, we conclude the following theorem.

Theorem 6.2.1 Consider a random text that is generated by a first-order Markovian source with the alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_\ell\}$. Suppose that the set of pattern words is $\mathcal{U} = \{u_1, u_2, \ldots, u_r\}$. These patterns words could be non-reduced, i.e., some patterns may entirely cover others.

We can obtain ${}^{a_i}T(z, t_1, t_2, \ldots, t_r)$, the generating function of decorated text following the letter a_i ($i = 1, 2, \ldots, \ell$), by solving the following linear equations, for all ${}^{a_i}T(z, t_1, t_2, \ldots, t_r)$

$$
\begin{pmatrix} {}^{a_1}T(z, \mathbf{t}) \\ {}^{a_2}T(z, \mathbf{t}) \\ \vdots \\ {}^{a_\ell}T(z, \mathbf{t}) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + P_{(\ell \times \ell)} \cdot z \cdot \begin{pmatrix} {}^{a_1}T(z, \mathbf{t}) \\ {}^{a_2}T(z, \mathbf{t}) \\ \vdots \\ {}^{a_\ell}T(z, \mathbf{t}) \end{pmatrix} + \begin{pmatrix} {}^{a_1}\eta(z, \mathbf{t}) \\ {}^{a_2}\eta(z, \mathbf{t}) \\ \vdots \\ {}^{a_\ell}\eta(z, \mathbf{t}) \end{pmatrix} \cdot \begin{pmatrix} {}^{\widehat{u_1}}T(z, \mathbf{t}) \\ {}^{\widehat{u_2}}T(z, \mathbf{t}) \\ \vdots \\ {}^{\widehat{u_r}}T(z, \mathbf{t}) \end{pmatrix} \tag{6.5}
$$

where the notation ${}^{a_i}\eta(z, \mathbf{t})$ is formally defined by Equation (6.2).

We reemphasize here that each ${}^{a_i}\eta(z, \mathbf{t})$ (where $i \in \{1, 2, \ldots, \ell\}$) is a $1 \times r$ row vector, in the form of

$$\left((\cdots)\widehat{\mathrm{Flip}(\mathsf{u}_1)},\ (\cdots)\widehat{\mathrm{Flip}(\mathsf{u}_2)},\ \ldots,\ (\cdots)\widehat{\mathrm{Flip}(\mathsf{u}_r)} \right).$$

Then we have the generating functions ${}^{a_i}F(z, x_1, x_2, \ldots, x_\ell)$ that give the probability of occurrences for each pattern word:

$${}^{a_i}F(z, x_1, x_2, \ldots, x_r) = {}^{a_i}T(z, x_1 - 1, x_2 - 1, \ldots, x_r - 1)$$

Theorem 6.2.1 is the generalized version of Theorem 4.3.1. It is not surprising that Formula (6.5) seems identical to Formula (4.23). In fact, the difference between them is the process to obtain ${}^{a_i}\eta(z, \mathbf{t})$. In a case of reduced patterns, all bicolored decorated words in Formula (6.5) simply become monocolored. Thus, Formula (4.23) should apply.

6.3 An example

The pattern set in the following example is used in the previous chapter. We already discussed its right extension matrix and the generating functions of clusters in Examples 5.4.1 and 5.9.1.

Example 6.3.1 A binary text is generated by a first-order Markovian source with the transition matrix

$$P = \begin{pmatrix} p_{aa} & p_{ab} \\ p_{ba} & p_{bb} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 3/5 & 2/5 \end{pmatrix} \tag{6.6}$$

Consider a pattern set

$$\mathcal{U} = \{u_1, u_2, u_3, u_4\} = \{ab, aba, bab, baba\}$$

The last letter of each pattern word is

$$\widehat{u_1} = b, \quad \widehat{u_2} = a, \quad \widehat{u_3} = b, \quad \widehat{u_4} = a$$

Now we write the generating functions of the Flips for each pattern word. Since the text source is Markovian, we must consider the letter appearing before these bicolored decorated texts.

$\mathrm{Flip}(\mathsf{u}_1) = \{a\overset{\bullet 1}{b}\}$. The generating functions for $[a]\mathrm{Flip}(\mathsf{u}_1)$ and $[b]\mathrm{Flip}(\mathsf{u}_1)$ are respectively

$$
\begin{aligned}
{}^{a}U_1(z, t_1, t_2, t_3, t_4) &= p_{aa}p_{ab}z^2t_1 \\
{}^{b}U_1(z, t_1, t_2, t_3, t_4) &= p_{ba}p_{ab}z^2t_1
\end{aligned}
$$

$\mathrm{Flip}(\mathsf{u}_2) = \{a\overset{\circ 1}{b}\overset{\bullet 2}{a}\}$. For $[a]\mathrm{Flip}(\mathsf{u}_2)$ and $[b]\mathrm{Flip}(\mathsf{u}_2)$, we have

$$
\begin{aligned}
{}^{a}U_2(z, t_1, t_2, t_3, t_4) &= p_{aa}p_{ab}p_{ba}z^3(1+t_1)t_2 \\
{}^{b}U_2(z, t_1, t_2, t_3, t_4) &= p_{ba}^2p_{ab}z^3(1+t_1)t_2
\end{aligned}
$$

$\mathrm{Flip}(\mathsf{u}_3) = \{ba\overset{\circ 1}{\underset{\bullet 3}{b}}\}$. For $[a]\mathrm{Flip}(\mathsf{u}_3)$ and $[b]\mathrm{Flip}(\mathsf{u}_3)$, we have

$$
\begin{aligned}
{}^{a}U_3(z, t_1, t_2, t_3, t_4) &= p_{ab}^2p_{ba}z^3(1+t_1)t_3 \\
{}^{b}U_3(z, t_1, t_2, t_3, t_4) &= p_{bb}p_{ba}p_{ab}z^3(1+t_1)t_3
\end{aligned}
$$

$\mathrm{Flip}(\mathsf{u}_4) = \{ba\overset{\circ 1}{\underset{\bullet 3}{b}}\overset{\circ 2}{\underset{\bullet 4}{a}}\}$. For $[a]\mathrm{Flip}(\mathsf{u}_4)$ and $[b]\mathrm{Flip}(\mathsf{u}_4)$, we have

$$
\begin{aligned}
{}^{a}U_4(z, t_1, t_2, t_3, t_4) &= p_{ab}^2p_{ba}^2z^4(1+t_1)(1+t_2)(1+t_3)t_4 \\
{}^{b}U_4(z, t_1, t_2, t_3, t_4) &= p_{bb}p_{ba}^2p_{ab}z^4(1+t_1)(1+t_2)(1+t_3)t_4
\end{aligned}
$$

Next, we need the right extension matrix. With a Bernoulli text source, as discussed in Example 5.9.1(1), we obtained the bicolored right extension set E and the corresponding matrix of generating functions, as follows.

$$\mathsf{E}=\begin{pmatrix} \emptyset & \emptyset & \{\overset{\textcircled{1}}{\underset{\textcircled{3}}{ab}}\} & \{\overset{\textcircled{1}}{\underset{\textcircled{3}}{ab}}\overset{\textcircled{2}}{\underset{\textcircled{4}}{a}}\} \\ \{\underset{\textcircled{1}}{b}\} & \{\underset{\textcircled{1}}{b}\overset{\textcircled{2}}{\underset{\textcircled{2}}{a}}\} & \{\overset{\textcircled{1}}{\underset{\textcircled{3}}{b}}\} & \{\overset{\textcircled{1}}{\underset{\textcircled{3}}{b}}\overset{\textcircled{2}}{\underset{\textcircled{4}}{a}}\} \\ \emptyset & \{\underset{\textcircled{2}}{a}\} & \{\overset{\textcircled{1}}{\underset{\textcircled{3}}{ab}}\} & \{\overset{\textcircled{1}}{\underset{\textcircled{3}}{ab}}\overset{\textcircled{2}}{\underset{\textcircled{4}}{a}}\} \\ \{\underset{\textcircled{1}}{b}\} & \{\underset{\textcircled{1}}{b}\overset{\textcircled{2}}{\underset{\textcircled{2}}{a}}\} & \{\overset{\textcircled{1}}{\underset{\textcircled{3}}{b}}\} & \{\overset{\textcircled{1}}{\underset{\textcircled{3}}{b}}\overset{\textcircled{2}}{\underset{\textcircled{4}}{a}}\} \end{pmatrix}.$$

and

$$\begin{aligned}&\mathbb{E}\left(z,t_1,t_2,t_3,t_4\right)\\ &=\begin{pmatrix} 0 & 0 & \pi_a\pi_b z^2(1+t_1)t_3 & \pi_a^2\pi_b z^3(1+t_1)(1+t_2)(1+t_3)t_4 \\ \pi_b z t_1 & \pi_a\pi_b z^2(1+t_1)t_2 & \pi_b z(1+t_1)t_3 & \pi_a\pi_b z^2(1+t_1)(1+t_2)(1+t_3)t_4 \\ 0 & \pi_a z t_2 & \pi_a\pi_b z^2(1+t_1)t_3 & \pi_a^2\pi_b z^3(1+t_1)(1+t_2)(1+t_3)t_4 \\ \pi_b z t_1 & \pi_a\pi_b z^2(1+t_1)t_2 & \pi_b z(1+t_1)t_3 & \pi_a\pi_b z^2(1+t_1)(1+t_2)(1+t_3)t_4 \end{pmatrix}\end{aligned}$$

Now, we should revise the $\mathbb{E}\left(z,t_1,t_2,t_3,t_4\right)$ above to ${}^{\widehat{u}}\mathbb{E}(z,t_1,t_2,t_3,t_4)$ based on Formula (6.3). The result is the following.

In general, we have

$$\begin{aligned}{}^{\widehat{u}}\mathbb{E}(z,t_1,t_2,t_3,t_4)&=\begin{pmatrix} {}^{\widehat{u_1}}E_{1,1}(z,\mathbf{t}), & {}^{\widehat{u_1}}E_{1,2}(z,\mathbf{t}), & {}^{\widehat{u_1}}E_{1,3}(z,\mathbf{t}), & {}^{\widehat{u_1}}E_{1,4}(z,\mathbf{t}) \\ {}^{\widehat{u_2}}E_{2,1}(z,\mathbf{t}), & {}^{\widehat{u_2}}E_{2,2}(z,\mathbf{t}), & {}^{\widehat{u_2}}E_{2,3}(z,\mathbf{t}), & {}^{\widehat{u_2}}E_{2,4}(z,\mathbf{t}) \\ {}^{\widehat{u_3}}E_{3,1}(z,\mathbf{t}), & {}^{\widehat{u_3}}E_{3,2}(z,\mathbf{t}), & {}^{\widehat{u_3}}E_{3,3}(z,\mathbf{t}), & {}^{\widehat{u_3}}E_{3,4}(z,\mathbf{t}) \\ {}^{\widehat{u_4}}E_{4,1}(z,\mathbf{t}), & {}^{\widehat{u_4}}E_{4,2}(z,\mathbf{t}), & {}^{\widehat{u_4}}E_{4,3}(z,\mathbf{t}), & {}^{\widehat{u_4}}E_{4,4}(z,\mathbf{t}) \end{pmatrix}\\ &=\begin{pmatrix} {}^{b}E_{1,1}(z,\mathbf{t}), & {}^{b}E_{1,2}(z,\mathbf{t}), & {}^{b}E_{1,3}(z,\mathbf{t}), & {}^{b}E_{1,4}(z,\mathbf{t}) \\ {}^{a}E_{2,1}(z,\mathbf{t}), & {}^{a}E_{2,2}(z,\mathbf{t}), & {}^{a}E_{2,3}(z,\mathbf{t}), & {}^{a}E_{2,4}(z,\mathbf{t}) \\ {}^{b}E_{3,1}(z,\mathbf{t}), & {}^{b}E_{3,2}(z,\mathbf{t}), & {}^{b}E_{3,3}(z,\mathbf{t}), & {}^{b}E_{3,4}(z,\mathbf{t}) \\ {}^{a}E_{4,1}(z,\mathbf{t}), & {}^{a}E_{4,2}(z,\mathbf{t}), & {}^{a}E_{4,3}(z,\mathbf{t}), & {}^{a}E_{4,4}(z,\mathbf{t}) \end{pmatrix}\end{aligned}$$

In this case, this becomes

$$\begin{aligned}
&\widehat{\mathcal{U}}\mathbb{E}(z,t_1,t_2,t_3,t_4)\\
&=\begin{pmatrix}
0 & 0 & p_{ba}p_{ab}z^2(1+t_1)t_3 & p_{ba}^2p_{ab}z^3(1+t_1)(1+t_2)(1+t_3)t_4\\
p_{ab}zt_1 & p_{ab}p_{ba}z^2(1+t_1)t_2 & p_{ab}z(1+t_1)t_3 & p_{ab}p_{ba}z^2(1+t_1)(1+t_2)(1+t_3)t_4\\
0 & p_{ba}zt_2 & p_{ba}p_{ab}z^2(1+t_1)t_3 & p_{ba}^2p_{ab}z^3(1+t_1)(1+t_2)(1+t_3)t_4\\
p_{ab}zt_1 & p_{ab}p_{ba}z^2(1+t_1)t_2 & p_{ab}z(1+t_1)t_3 & p_{ab}p_{ba}z^2(1+t_1)(1+t_2)(1+t_3)t_4
\end{pmatrix}
\end{aligned} \tag{6.7}$$

Thus, using Formula (6.1), we obtain the generating functions of clusters.

$$\begin{aligned}
{}^a\xi(z,t_1,t_2,t_3,t_4) &= ({}^aU_1(z,\mathbf{t}),\dots,{}^aU_4(z,\mathbf{t}))\cdot(\mathbb{I}-\widehat{\mathcal{U}}\mathbb{E}(z,\mathbf{t}))^{-1}\cdot\begin{pmatrix}1\\1\\1\\1\end{pmatrix}\\
&= {}^a\eta(z,\mathbf{t})\cdot\begin{pmatrix}1\\1\\1\\1\end{pmatrix}
\end{aligned} \tag{6.8}$$

and

$$\begin{aligned}
{}^b\xi(z,t_1,t_2,t_3,t_4) &= \left({}^bU_1(z,\mathbf{t}),\dots,{}^bU_4(z,\mathbf{t})\right)\cdot(\mathbb{I}-\widehat{\mathcal{U}}\mathbb{E}(z,\mathbf{t}))^{-1}\cdot\begin{pmatrix}1\\1\\1\\1\end{pmatrix}\\
&= {}^b\eta(z,\mathbf{t})\cdot\begin{pmatrix}1\\1\\1\\1\end{pmatrix}
\end{aligned} \tag{6.9}$$

Hence, we can retrieve the values of ${}^a\eta(z,\mathbf{t})$ and ${}^b\eta(z,\mathbf{t})$ from (6.8) and (6.9). The full expressions of ${}^a\eta(z,\mathbf{t})$ and ${}^b\eta(z,\mathbf{t})$ are too long to display here. Thus, we omit their

expressions here and continue the procedures. The full expressions of ${}^a\eta(z,\mathbf{t})$ and ${}^b\eta(z,\mathbf{t})$ are provided in Appendix A.

In the next step, we will follow Formula (6.5) and compute the generating functions of decorated texts by solving the following two linear equations, for ${}^aT(z,\mathbf{t})$ and ${}^bT(z,\mathbf{t})$.

$$\begin{pmatrix} {}^aT(z,\mathbf{t}) \\ {}^bT(z,\mathbf{t}) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + P\cdot z\cdot \begin{pmatrix} {}^aT(z,\mathbf{t}) \\ {}^bT(z,\mathbf{t}) \end{pmatrix} + \begin{pmatrix} {}^a\eta(z,\mathbf{t}) \\ {}^b\eta(z,\mathbf{t}) \end{pmatrix} \cdot \begin{pmatrix} {}^bT(z,\mathbf{t}) \\ {}^aT(z,\mathbf{t}) \\ {}^bT(z,\mathbf{t}) \\ {}^aT(z,\mathbf{t}) \end{pmatrix} \tag{6.10}$$

Not surprisingly, the results of ${}^aT(z,\mathbf{t})$ and ${}^bT(z,\mathbf{t})$ are very long. Hence, we again skip the full expressions here. The full expressions of ${}^aT(z,\mathbf{t})$ and ${}^bT(z,\mathbf{t})$, along with the full expressions of ${}^aF(z,\mathbf{x})$ and ${}^bF(z,\mathbf{x})$ that are obtained by (6.11) and (6.12), are all provided in Appendix A.

The generating functions ${}^aF(z,\mathbf{x})$ and ${}^bF(z,\mathbf{x})$ are given by

$${}^aF(z,x_1,x_2,x_3,x_4) = {}^aT(z,x_1-1,x_2-1,x_3-1,x_4-1) \tag{6.11}$$

and

$${}^bF(z,x_1,x_2,x_3,x_4) = {}^bT(z,x_1-1,x_2-1,x_3-1,x_4-1) \tag{6.12}$$

The last step is computing the Taylor expansion series for ${}^aF(z,x_1,x_2,x_3,x_4)$ and for ${}^bF(z,x_1,x_2,x_3,x_4)$, at $z=0$. The first few terms are as follows.

$$\begin{aligned} {}^aF(z,x_1,x_2,x_3,x_4) =& 1+z+\left(\frac{x_1}{4}+\frac{3}{4}\right)z^2 \\ &+\left(\frac{19}{40}+\frac{(6x_2+6x_3+9)x_1}{40}\right)z^3 \\ &+\left(\frac{111}{400}+\frac{3x_1^2x_2x_3}{40}\right. \\ &\quad \left.+\frac{(36x_2x_3x_4+60x_2+48x_3+115)x_1}{400}\right)z^4+\mathrm{O}\left(z^5\right) \end{aligned} \tag{6.13}$$

and

$$
\begin{aligned}
{}^{b}F(z,x_1,x_2,x_3,x_4) = 1 + z + \left(\frac{3x_1}{10} + \frac{7}{10}\right) z^2 \\
+ \left(\frac{43}{100} + \frac{(18x_2 + 12x_3 + 27)x_1}{100}\right) z^3 \\
+ \left(\frac{247}{1000} + \frac{9x_1^2 x_2 x_3}{100}\right. \\
\left. + \frac{(72x_2x_3x_4 + 180x_2 + 96x_3 + 315)x_1}{1000}\right) z^4 + \mathrm{O}\left(z^5\right)
\end{aligned}
\tag{6.14}
$$

We can look into one of the terms in (6.13) and (6.14), and verify the probabilities they provide. For instance:

$$
\begin{aligned}
{}^{a}F(z,x_1,x_2,x_3,x_4) = \ldots + \left(\frac{111}{400} + \frac{3x_1^2 x_2 x_3}{40}\right. \\
\left. + \frac{(36x_2x_3x_4 + 60x_2 + 48x_3 + 115)x_1}{400}\right) z^4 + \mathrm{O}\left(z^5\right)
\end{aligned}
\tag{6.15}
$$

The interpretation of the z^4 term is that:

Assume a binary text is generated by a Markovian source of order 1, with the transition matrix (6.6). We are interested in counting four pattern words $\{u_1, u_2, u_3, u_4\} = \{ab, aba, bab, baba\}$. In a text of length 4 following a letter a, the probability that:

– none of the four patterns occurs, is 111/400.

– pattern $u_1 = ab$ occurs exactly once and no other patterns occur, is 115/400;

– pattern $u_1 = ab$ occurs exactly twice, $u_2 = aba$ occurs exactly once, $u_3 = bab$ occurs exactly once, and $u_4 = baba$ does not occur, is 3/40;

– each of the four patterns occurs exactly once, is 36/400;

– pattern $u_1 = ab$ occurs exactly once, $u_2 = aba$ occurs exactly once and no other patterns occur, is 60/400;

– pattern $u_1 = ab$ occurs exactly once, $u_3 = bab$ occurs exactly once and no other patterns occur, is 48/400;

and all other situations have 0 probability.

The probability values above sum up to 1.

$$\frac{111}{400} + \frac{115}{400} + \frac{3}{40} + \frac{36}{400} + \frac{60}{400} + \frac{48}{400} = 1.$$

7. MOMENTS OF OCCURRENCES FOR PATTERNS IN A BERNOULLI TEXT

In the next two chapters, we discuss the moments of pattern occurrences for two pattern sets, $\mathcal{U} = \{u_1, u_2, ..., u_r\}$, and $\mathcal{V} = \{v_1, v_2, ..., v_s\}$. The two sets may have non-empty intersection. Therefore, we denote $\mathcal{W} := \mathcal{U} \cap \mathcal{V}$.

In our context, both pattern sets $\mathcal{U}$ and $\mathcal{V}$ may be non-reduced. Namely, it is possible that a pattern word is entirely a part of another one.

To count the occurrences of patterns in $\mathcal{U}$ and $\mathcal{V}$, we define two random variables X_n and Y_n—the number of total pattern occurrences of $\mathcal{U}$ and $\mathcal{V}$, respectively, in a text of length n.

In this chapter, we analyze the approach to achieving the covariance of X_n and Y_n in Bernoulli models. The results were first obtained in the non-Markovian case by Bassino et al. [5]. Then in next chapter, we expand the results to Markovian models, and discuss the limit.

7.1 The first moment for one pattern set

When considering the random variable X_n, we do not differentiate each single pattern in the pattern set $\mathcal{U} = \{u_1, u_2, ..., u_r\}$. Instead, we count the total number of occurrences for all patterns in $\mathcal{U}$. The probability measure for the set $\mathcal{U}$ is the sum of the probabilities overall the elements of $\mathcal{U}$:

$$\pi(\mathcal{U}) = \sum_{u \in \mathcal{U}} \pi(u) \tag{7.1}$$

The first moment of X_n, i.e., $\mathbf{E}(X_n)$, can be obtained easily.

Theorem 7.1.1 The expected value of X_n is given by

$$\mathbf{E}(X_n) = \sum_{u \in \mathcal{U}} \pi(u)(n - |u| + 1) \tag{7.2}$$

where $|u|$ is the size of the pattern u.

Proof 3 Let $X_n(u)$ be the random variable representing the number of occurrences of pattern u in a text of length n. Since the pattern set is $\mathcal{U} = \{u_1, u_2, ..., u_r\}$, then, by the definition of X_n, we have

$$X_n = \sum_{u \in \mathcal{U}} X_n(u)$$

Therefore,

$$\begin{aligned} \mathbf{E}(X_n) &= \mathbf{E}\left(\sum_{u \in \mathcal{U}} X_n(u)\right) \\ &= \sum_{u \in \mathcal{U}} \mathbf{E}(X_n(u)) \\ &= \sum_{u \in \mathcal{U}} \pi(u)(n - |u| + 1) \end{aligned} \tag{7.3}$$

The last step in (7.3) considers substrings of length $|u|$ in a text of length n, for any specific pattern u. There are $n - |u| + 1$ such substrings. As the text source is Bernoulli, all these substrings have equal probability, $\pi(u)$, to be the pattern u.

■

Applying Theorem 7.1.1 to Y_n, we have $\mathbf{E}(Y_n)$.

$$\mathbf{E}(Y_n) = \sum_{v \in \mathcal{V}} \pi(v)(n - |v| + 1) \tag{7.4}$$

where $|v|$ is the size of the pattern v.

The computation of $\mathbf{Cov}(X_n, Y_n)$ is much more complex than the first moment. Next, we use several sections to organize the procedure.

7.2 Generating function for one pattern set

For the pattern set $\mathcal{U}$, the generating function of occurrences, $F(z, x)$, by definition (see Equation (3.1)), can be written in the following form.

$$F(z, x) = \sum_n \sum_k \Pr(X_n = k) \cdot z^n x^k \tag{7.5}$$

To reveal the relationship between the probability generating functions and moments of the discrete random variable X_n with values in $\mathbb{Z}_{\geq 0}$, we introduce Theorem 7.2.1. This

result has a long history. We refer to Flajolet and Sedgewick [13, Appendix A.3] for more discussion.

Theorem 7.2.1 The generating functions of the first two moments of X_n are

$$\sum_{n\geq 0} \mathbf{E}\left(X_n\right) z^n = \frac{\partial}{\partial x} F(z,x)\bigg|_{x=1} \tag{7.6}$$

and

$$\sum_{n\geq 0} \mathbf{E}\left(X_n^2\right) z^n = \frac{\partial^2}{\partial x^2} F(z,x)\bigg|_{x=1} + \frac{\partial}{\partial x} F(z,x)\bigg|_{x=1} \tag{7.7}$$

Proof 4 From (7.5), we have

$$\begin{aligned}
\frac{\partial}{\partial x} F(z,x)\bigg|_{x=1} &= \sum_n z^n \cdot \left(\sum_{k\geq 1} k \cdot \Pr(X_n = k) \right) \\
&= \sum_{n\geq 0} z^n \cdot \mathbf{E}\left(X_n\right)
\end{aligned}$$

This proves Equation (7.6). In addition, we have

$$\frac{\partial^2}{\partial x^2} F(z,x)\bigg|_{x=1} = \sum_n z^n \cdot \left(\sum_{k\geq 2} k(k-1) \cdot \Pr(X_n = k) \right)$$

Therefore, we obtain

$$\begin{aligned}
\frac{\partial^2}{\partial x^2} F(z,x)\bigg|_{x=1} + \frac{\partial}{\partial x} F(z,x)\bigg|_{x=1} &= \sum_n z^n \cdot \left(\sum_{k\geq 1} k^2 \cdot \Pr(X_n = k) \right) \\
&= \sum_{n\geq 0} z^n \cdot \mathbf{E}\left(X_n^2\right)
\end{aligned}$$

■

We emphasize that Theorem 7.2.1 applies to both Bernoulli models and Markovian models.

From the next section through the end of this chapter, we calculate the first and second moments of X_n and Y_n, including $\mathbf{Cov}(X_n, Y_n)$. We refer to Bassino et al. [5, Section 6] for the origin of the method that we use in this chapter.

7.3 Generating function for two pattern sets

When both pattern sets $\mathcal{U}$ and $\mathcal{V}$ are considered, the generating function of occurrences becomes

$$F(z,x,y) := \sum_n \sum_i \sum_j \Pr(X_n = i, Y_n = j) \cdot z^n x^i y^j \tag{7.8}$$

To obtain the generating function of clusters, we should be aware that the intersection $\mathcal{W} = \mathcal{U} \cap \mathcal{V}$ could be non-empty. Therefore, simply tracking the occurrences from $\mathcal{U}$ and $\mathcal{V}$ is not enough. We must regroup all patterns from $\mathcal{U} \cup \mathcal{V}$ into three disjoint and complete pattern sets: $\mathcal{U}\backslash\mathcal{V}$, $\mathcal{V}\backslash\mathcal{U}$, and $\mathcal{U} \cap \mathcal{V}$. In the following generating function for the decorated clusters, we explicitly count the occurrences for each of the three new sets, as shown in Fig. 7.1.

$$\begin{aligned} \Upsilon(z,t_1,t_2,t_3) &= \xi(z,\mathbf{t})\big|_{t_u=t_1 \text{ for } u\in\mathcal{U}\backslash\mathcal{V};\, t_u=t_2 \text{ for } u\in\mathcal{V}\backslash\mathcal{U};\, t_u=t_3 \text{ for } u\in\mathcal{U}\cap\mathcal{V}} \\ &= \sum_{\mathsf{c}\in\mathsf{C}} \pi(\mathsf{c}) \cdot z^{|\mathsf{c}|} \cdot t_1^{(\#\text{ distinguished occurrences of patterns in } \mathcal{U}\backslash\mathcal{V})} \\ &\qquad \cdot t_2^{(\#\text{ distinguished occurrences of patterns in } \mathcal{V}\backslash\mathcal{U})} \\ &\qquad \cdot t_3^{(\#\text{ distinguished occurrences of patterns in } \mathcal{U}\cap\mathcal{V})} \end{aligned} \tag{7.9}$$

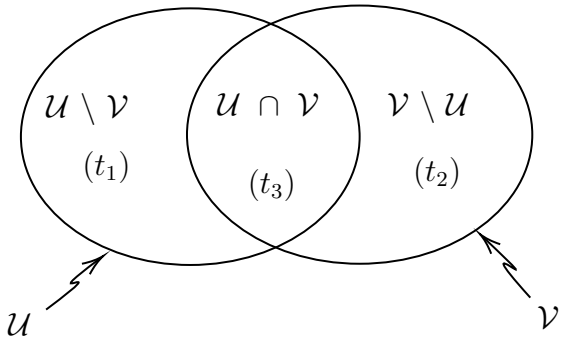


Figure 7.1. $\mathcal{U} \cup \mathcal{V} = (\mathcal{U}\backslash\mathcal{V}) \oplus (\mathcal{V}\backslash\mathcal{U}) \oplus (\mathcal{U} \cap \mathcal{V})$

In (7.9), t_1, t_2, and t_3 are used to respectively represent the distinguished occurrences from the three disjoint pattern sets: $\mathcal{U}\backslash\mathcal{V}$, $\mathcal{V}\backslash\mathcal{U}$, and $\mathcal{U} \cap \mathcal{V}$.

Apply Corollary 3.1.1 and Equation (4.2), we have

$$F(z,x,y)=\frac{1}{1-z-\Upsilon(z,x-1,y-1,xy-1)} \tag{7.10}$$

The generating function of cluster $\Upsilon(z,x-1,y-1,xy-1)$ in (7.10) comes from the following replacements to (7.9).

$$\begin{aligned} t_1 &\to x-1 \\ t_2 &\to y-1 \\ t_3 &\to xy-1 \end{aligned} \tag{7.11}$$

The reason for the mapping $t_3 \to xy-1$ is that every distinguished occurrence in $\mathcal{U}\cap\mathcal{V}$ should contribute to both the $\mathcal{U}$ set and the $\mathcal{V}$ set. If we only need to count the occurrences in one pattern set, say $\mathcal{U}$, we can simply perform the transformation

$$\{t_1 \to x-1; \quad t_2 \to 0; \quad t_3 \to x-1\}.$$

From (7.8) and (7.9), we have $F(z,1,1)=1/(1-z)$ and $\Upsilon(z,0,0,0)=0$.

7.4 Moments for two pattern sets

Similar to Theorem 7.2.1, we can obtain the following results from Equation (7.8).

$$\sum_{n\geq 0}\mathbf{E}\left(X_n\right)z^n=\left.\frac{\partial}{\partial x}F(z,x,y)\right|_{x=y=1} \tag{7.12}$$

$$\sum_{n\geq 0}\mathbf{E}\left(Y_n\right)z^n=\left.\frac{\partial}{\partial y}F(z,x,y)\right|_{x=y=1} \tag{7.13}$$

$$\sum_{n\geq 0}\mathbf{E}\left(X_nY_n\right)z^n=\left.\frac{\partial^2}{\partial x\partial y}F(z,x,y)\right|_{x=y=1} \tag{7.14}$$

The three equations will help us to obtain three expected values $\mathbf{E}\left(X_n\right)$, $\mathbf{E}\left(Y_n\right)$, and $\mathbf{E}\left(X_nY_n\right)$. Thus, we can compute $\mathbf{Cov}(X_n,Y_n)$.

In order to simplify the notations, we define

$$\Upsilon_i(z) = \frac{\partial}{\partial t_i}\Upsilon\left(z, t_1, t_2, t_3\right)\Big|_{t_1=t_2=t_3=0} \quad \text{for } i \in \{1,2,3\} \tag{7.15}$$

$$\Upsilon_{ij}(z) = \frac{\partial^2}{\partial t_i \partial t_j}\Upsilon\left(z, t_1, t_2, t_3\right)\Big|_{t_1=t_2=t_3=0} \quad \text{for } i,j \in \{1,2,3\} \tag{7.16}$$

Next, we use the transformation (7.11) to map $\Upsilon(z, t_1, t_2, t_3)$ to

$$\Upsilon(z, x-1, y-1, xy-1),$$

and we compute the following derivatives.

$$\begin{aligned}
&\frac{\partial \Upsilon(z, x-1, y-1, xy-1)}{\partial x}\Big|_{x=y=1} \\
&= \left(\frac{\partial \Upsilon(z, t_1, t_2, t_3)}{\partial z}\cdot\frac{\partial z}{\partial x} + \frac{\partial \Upsilon(z, t_1, t_2, t_3)}{\partial t_1}\cdot\frac{\partial t_1}{\partial x}\right. \\
&\qquad \left. + \frac{\partial \Upsilon(z, t_1, t_2, t_3)}{\partial t_2}\cdot\frac{\partial t_2}{\partial x} + \frac{\partial \Upsilon(z, t_1, t_2, t_3)}{\partial t_3}\cdot\frac{\partial t_3}{\partial x}\right)\Big|_{t_1=t_2=t_3=0} \\
&= \left(0 + \frac{\partial \Upsilon(z, t_1, t_2, t_3)}{\partial t_1} + 0 + \frac{\partial \Upsilon(z, t_1, t_2, t_3)}{\partial t_3}\cdot y\right)\Big|_{t_1=t_2=t_3=0,\, y=1} \\
&= \Upsilon_1(z) + \Upsilon_3(z)
\end{aligned} \tag{7.17}$$

Similarly,

$$\frac{\partial \Upsilon(z, x-1, y-1, xy-1)}{\partial y}\Big|_{x=y=1} = \Upsilon_2(z) + \Upsilon_3(z) \tag{7.18}$$

In addition,

$$
\begin{aligned}
&\left.\frac{\partial^2 \Upsilon(z, x-1, y-1, xy-1)}{\partial x \partial y}\right|_{x=y=1} \\
&= \left.\left(\frac{\partial}{\partial y}\left(\frac{\partial \Upsilon(z, t_1, t_2, t_3)}{\partial t_1} + \frac{\partial \Upsilon(z, t_1, t_2, t_3)}{\partial t_3} \cdot y\right)\right)\right|_{t_1=t_2=t_3=0,\, x=y=1} \\
&= \left(\frac{\partial}{\partial z}\left(\frac{\partial \Upsilon}{\partial t_1}\right)\frac{\partial z}{\partial y} + \frac{\partial}{\partial t_1}\left(\frac{\partial \Upsilon}{\partial t_1}\right)\frac{\partial t_1}{\partial y} + \frac{\partial}{\partial t_2}\left(\frac{\partial \Upsilon}{\partial t_1}\right)\frac{\partial t_2}{\partial y} + \frac{\partial}{\partial t_3}\left(\frac{\partial \Upsilon}{\partial t_1}\right)\frac{\partial t_3}{\partial y}\right. \\
&\quad + y\left(\frac{\partial}{\partial z}\left(\frac{\partial \Upsilon}{\partial t_3}\right)\frac{\partial z}{\partial y} + \frac{\partial}{\partial t_1}\left(\frac{\partial \Upsilon}{\partial t_3}\right)\frac{\partial t_1}{\partial y} + \frac{\partial}{\partial t_2}\left(\frac{\partial \Upsilon}{\partial t_3}\right)\frac{\partial t_2}{\partial y} + \frac{\partial}{\partial t_3}\left(\frac{\partial \Upsilon}{\partial t_3}\right)\frac{\partial t_3}{\partial y}\right) \\
&\quad \left.+ \frac{\partial \Upsilon}{\partial t_3}\right)\Bigg|_{t_1=t_2=t_3=0,\, x=y=1} \\
&= \left(0 + 0 + \frac{\partial^2 \Upsilon}{\partial t_1 \partial t_2} + x \cdot \frac{\partial^2 \Upsilon}{\partial t_1 \partial t_3}\right. \\
&\quad \left. + 0 + 0 + y \cdot \frac{\partial^2 \Upsilon}{\partial t_2 \partial t_3} + xy \cdot \frac{\partial^2 \Upsilon}{\partial t_3^2} + \frac{\partial \Upsilon}{\partial t_3}\right)\Bigg|_{t_1=t_2=t_3=0,\, x=y=1} \\
&= \Upsilon_{12}(z) + \Upsilon_{13}(z) + \Upsilon_{23}(z) + \Upsilon_{33}(z) + \Upsilon_3(z)
\end{aligned} \tag{7.19}
$$

7.4.1 Computing derivatives

We focus on a few partial derivatives of $\Upsilon(z, t_1, t_2, t_3)$ in this subsection. The results will be handy to use in later computations.

Combining (7.9) and (7.15), we have

$$
\Upsilon_i(z) = \left.\frac{\partial}{\partial t_i}\Upsilon(z, t_1, t_2, t_3)\right|_{t_1=t_2=t_3=0} = \sum_{\mathsf{c}' \in \mathsf{C}_i} \pi(\mathsf{c}') \cdot z^{|\mathsf{c}'|} \tag{7.20}
$$

where C_i is the set of clusters with exactly 1 distinguished occurrence of a pattern word in $\mathcal{U} \backslash \mathcal{V}$ (for $i = 1$), $\mathcal{V} \backslash \mathcal{U}$ (for $i = 2$), or $\mathcal{W} = \mathcal{U} \cap \mathcal{V}$ (for $i = 3$). A cluster with exactly 1 distinguished occurrence must be a pattern word. Therefore,

$$
\Upsilon_3(z) = \sum_{w \in \mathcal{W}} \pi(w) \cdot z^{|w|} \tag{7.21}
$$

$$\Upsilon_1(z) + \Upsilon_3(z) = \sum_{u \in \mathcal{U}} \pi(u) \cdot z^{|u|} \tag{7.22}$$

$$\Upsilon_2(z) + \Upsilon_3(z) = \sum_{v \in \mathcal{V}} \pi(v) \cdot z^{|v|} \tag{7.23}$$

A Taylor expansion at $z = 1$ gives

$$\Upsilon_i(z) = \Upsilon_i(1) - (1-z)\Upsilon_i'(1) + o(1-z) \qquad \text{for } i = 1, 2, 3 \tag{7.24}$$

where $\Upsilon_i'(1)$ is obtained from (7.21), (7.22), (7.23):

$$\Upsilon_3'(1) = \frac{\partial}{\partial z}\Upsilon_3(z)\bigg|_{z=1} = \sum_{w \in \mathcal{W}} \pi(w) \cdot |w| \tag{7.25}$$

$$\Upsilon_1'(1) + \Upsilon_3'(1) = \sum_{u \in \mathcal{U}} \pi(u) \cdot |u| \tag{7.26}$$

$$\Upsilon_2'(1) + \Upsilon_3'(1) = \sum_{v \in \mathcal{V}} \pi(v) \cdot |v| \tag{7.27}$$

Similarly, for $\Upsilon_{ij}(z)$, we combine (7.9) and (7.16). For $i \neq j$, we have:

$$\begin{aligned} \Upsilon_{ij}(z) &= \frac{\partial^2}{\partial t_i \partial t_j}\Upsilon(z, t_1, t_2, t_3)\bigg|_{t_1=t_2=t_3=0} \\ &= \sum_{\mathsf{c}' \in \mathsf{C}_{ij}} \pi(\mathsf{c}') \cdot z^{|\mathsf{c}'|} \end{aligned} \tag{7.28}$$

where C_{ij} is the set of clusters with exactly 1 distinguished occurrence of i and exactly 1 distinguished occurrence of j.

Hence, we obtain

$$\Upsilon_{13}(1) = \sum_{u \in \mathcal{U} \setminus \mathcal{V}} \sum_{w \in \mathcal{U} \cap \mathcal{V}} \Big(\pi(u) \cdot |u|_w + \pi(w) \cdot |w|_u + \pi(u) \cdot \pi(\mathcal{E}_{u,w}) + \pi(w) \cdot \pi(\mathcal{E}_{w,u}) \Big) \tag{7.29}$$

$$\Upsilon_{23}(1) = \sum_{v\in\mathcal{V}\setminus\mathcal{U}} \sum_{w\in\mathcal{U}\cap\mathcal{V}} \Big(\pi(v)\cdot|v|_w + \pi(w)\cdot|w|_v + \pi(v)\cdot\pi\left(\mathcal{E}_{v,w}\right) + \pi(w)\cdot\pi\left(\mathcal{E}_{w,v}\right)\Big) \tag{7.30}$$

$$\Upsilon_{12}(1) = \sum_{u\in\mathcal{U}\setminus\mathcal{V}} \sum_{v\in\mathcal{V}\setminus\mathcal{U}} \Big(\pi(u)\cdot|u|_v + \pi(v)\cdot|v|_u + \pi(u)\cdot\pi\left(\mathcal{E}_{u,v}\right) + \pi(v)\cdot\pi\left(\mathcal{E}_{v,u}\right)\Big) \tag{7.31}$$

To clarify the notations, $|u|$ represents the length of u, whereas $|u|_v$ $(u \neq v)$ stands for the number of v occurring in u. For instance, $|aabbaabb| = 8$, $|aabbaabb|_{aab} = 2$, and $|aabbaabb|_{baa} = 1$.

Akin to (7.24), we have

$$\Upsilon_{ij}(z) = \Upsilon_{ij}(1) - (1-z)\Upsilon'_{ij}(1) + o(1-z) \qquad \text{for } i = 1,2,3 \tag{7.32}$$

Applying (7.28), the derivative gives, for $i \neq j$,

$$\begin{aligned}\Upsilon'_{ij}(1) &= \left(\frac{\partial\Upsilon_{ij}(z)}{\partial z}\right)\Big|_{z=1} \\ &= \left(\sum_{\mathsf{c}'\in\mathsf{C}_{ij}} \pi\left(\mathsf{c}'\right)\cdot|\mathsf{c}'|\cdot z^{|\mathsf{c}'|-1}\right)\Big|_{z=1} \\ &= \sum_{\mathsf{c}'\in\mathsf{C}_{ij}} \pi\left(\mathsf{c}'\right)\cdot|\mathsf{c}'|\end{aligned} \tag{7.33}$$

Therefore, we obtain

$$\begin{aligned}\Upsilon'_{13}(1) = \sum_{u\in\mathcal{U}\setminus\mathcal{V}} \sum_{w\in\mathcal{U}\cap\mathcal{V}} \Bigg(&\pi(u)\cdot|u|\cdot|u|_w + \pi(w)\cdot|w|\cdot|w|_u \\ &+ \sum_{\mathsf{c}'\in u\mathcal{E}_{u,w}} \pi\left(\mathsf{c}'\right)\cdot|\mathsf{c}'| + \sum_{\mathsf{c}'\in w\mathcal{E}_{w,u}} \pi\left(\mathsf{c}'\right)\cdot|\mathsf{c}'|\Bigg)\end{aligned} \tag{7.34}$$

$$\Upsilon'_{23}(1) = \sum_{v\in\mathcal{V}\setminus\mathcal{U}} \sum_{w\in\mathcal{U}\cap\mathcal{V}} \Bigg(\pi(v)\cdot|v|\cdot|v|_w + \pi(w)\cdot|w|\cdot|w|_v + \sum_{\mathsf{c}'\in v\mathcal{E}_{v,w}} \pi\left(\mathsf{c}'\right)\cdot|\mathsf{c}'| + \sum_{\mathsf{c}'\in w\mathcal{E}_{w,v}} \pi\left(\mathsf{c}'\right)\cdot|\mathsf{c}'| \Bigg) \tag{7.35}$$

$$\Upsilon'_{12}(1) = \sum_{u\in\mathcal{U}\setminus\mathcal{V}} \sum_{v\in\mathcal{V}\setminus\mathcal{U}} \Bigg(\pi(u)\cdot|u|\cdot|u|_v + \pi(v)\cdot|v|\cdot|v|_u + \sum_{\mathsf{c}'\in u\mathcal{E}_{u,v}} \pi\left(\mathsf{c}'\right)\cdot|\mathsf{c}'| + \sum_{\mathsf{c}'\in v\mathcal{E}_{v,u}} \pi\left(\mathsf{c}'\right)\cdot|\mathsf{c}'| \Bigg) \tag{7.36}$$

We emphasize that (7.28) and (7.33) are only valid for $i \neq j$. When $i = j = 3$, we have

$$\Upsilon_{33}(z) = \left(\frac{\partial^2\Upsilon\left(z,t_1,t_2,t_3\right)}{\partial t_3^2}\right)\Bigg|_{t_1=t_2=t_3=0} = \sum_{\mathsf{c}'\in\mathsf{C}_{33}} 2\cdot\pi\left(\mathsf{c}'\right)\cdot z^{|\mathsf{c}'|} \tag{7.37}$$

where C_{33} is the set of clusters with exactly 2 occurrences of patterns in $\mathcal{W} = \mathcal{U}\cap\mathcal{V}$.

Replace z with 1 in (7.37), we have

$$\begin{aligned}\Upsilon_{33}(1) &= \sum_{\mathsf{c}'\in\mathsf{C}_{33}} 2\cdot\pi\left(\mathsf{c}'\right) \\ &= \sum_{w_1\in\mathcal{W}} \sum_{w_2\in\mathcal{W}} 2\cdot\left(\pi\left(w_1\right)\cdot|w_1|_{w_2}\cdot[\![w_1\neq w_2]\!] + \pi\left(w_1\right)\cdot\pi\left(\mathcal{E}_{w_1,w_2}\right)\right)\end{aligned} \tag{7.38}$$

where $[\![w_1 \neq w_2]\!]$ is an Iverson indicator notation, namely

$$[\![w_1\neq w_2]\!] = \begin{cases} 0 & \text{if } w_1 = w_2 \\ 1 & \text{if } w_1\neq w_2 \end{cases}$$

To interpret (7.38), we recall that a cluster $\mathsf{c}' \in \mathsf{C}_{33}$ includes exactly 2 occurrences of patterns in $\mathcal{W}$ — 1 occurrence of pattern $w_1 \in \mathcal{W}$ and 1 occurrence of pattern $w_2 \in \mathcal{W}$.

In the case that $w_1 \neq w_2$, one pattern word may

(1) entirely cover the other one (corresponding to $\pi\left(w_1\right)\cdot|w_1|_{w_2}$); or

(2) the second occurrence follows the end of the first occurrence with a right extension (corresponding to $\pi\left(w_1\right)\cdot\pi\left(\mathcal{E}_{w_1,w_2}\right)$).

On the contrary, when $w_1 = w_2$, only situation (2) is allowed. Because in situation (1), although a pattern word entirely covers itself, we would only consider w_1 and w_2 as 1 occurrence, instead of 2. Therefore, the indicator $[\![w_1 \neq w_2]\!]$ is to ensure that we would not count situation (1) when $w_1 = w_2$. (Recall that $|w_1|_{w_2} = 1$ if $w_1 = w_2$.)

We proceed from (7.38), and have

$$\begin{aligned}
\Upsilon'_{33}(1) &= \left(\frac{\partial \Upsilon_{33}(z)}{\partial z}\right)\Bigg|_{z=1} \\
&= \left(\sum_{\mathsf{c}' \in \mathsf{C}_{33}} 2 \cdot \pi\left(\mathsf{c}'\right) \cdot |\mathsf{c}'| \cdot z^{|\mathsf{c}'|-1}\right)\Bigg|_{z=1} \\
&= \sum_{\mathsf{c}' \in \mathsf{C}_{33}} 2 \cdot \pi\left(\mathsf{c}'\right) \cdot |\mathsf{c}'| \\
&= \sum_{w_1 \in \mathcal{W}} \sum_{w_2 \in \mathcal{W}} 2 \cdot \left(\pi\left(w_1\right) \cdot |w_1| \cdot |w_1|_{w_2} \cdot [\![w_1 \neq w_2]\!] + \sum_{\mathsf{c}' \in w_1 \mathcal{E}_{w_1,w_2}} \pi\left(\mathsf{c}'\right) \cdot |\mathsf{c}'|\right)
\end{aligned} \tag{7.39}$$

Now we have prepared all needed derivatives of Υ for later use. Apart from them, the following three expansion formulas will also be used.

$$\begin{aligned}
\frac{1}{1-z} &= \sum_{n=0}^{\infty} z^n \\
\frac{1}{(1-z)^2} &= \sum_{n=0}^{\infty} (n+1) \cdot z^n \\
\frac{1}{(1-z)^3} &= \sum_{n=0}^{\infty} \frac{(n+1)(n+2)}{2} \cdot z^n
\end{aligned} \tag{7.40}$$

7.4.2 First moment for each pattern set

While Theorem 7.1.1 already gives $\mathbf{E}\left(X_n\right)$, we emphasize here that $\mathbf{E}\left(X_n\right)$ can be also obtained from the partial derivative $\frac{\partial F(z,x,y)}{\partial x}\Big|_{x=y=1}$, according to Theorem 7.2.1. It does not matter which approach we use when the text source is Bernoulli. However, when the text source is Markovian, we have to calculate the partial derivatives of the generating functions to obtain the moments of pattern occurrences. (We will discuss the Markovian case in Chapter 8.)

We provide the calculation details of Equation (7.41) in Appendix B, and give the result here.

$$\begin{aligned}\sum_{n=0}^{\infty} \mathbf{E}\left(X_n\right) \cdot z^n &= \left.\frac{\partial F(z,x,y)}{\partial x}\right|_{x=y=1} \\ &= \sum_{n=0}^{\infty} z^n \cdot \left(\sum_{u\in\mathcal{U}} \pi(u)\cdot(n+1-|u|)\right)\end{aligned} \tag{7.41}$$

Hence, we obtain

$$\mathbf{E}(X_n) = \sum_{u\in\mathcal{U}} \pi(u)\cdot(n+1-|u|) \tag{7.42}$$

Similarly, for Y_n, we have

$$\mathbf{E}(Y_n) = \sum_{v\in\mathcal{V}} \pi(v)\cdot(n+1-|v|) \tag{7.43}$$

As expected, (7.42) and (7.43) are respectively identical to (7.2) and (7.4). The approach of (7.41) will be used again in the second-order derivative of $F(z,x,y)$.

7.4.3 Second-order derivative

We rely on a few results that we calculated in Section 7.4.1 to obtain $\left.\frac{\partial^2 F(z,x,y)}{\partial x \partial y}\right|_{x=y=1}$. Here we skip the long details of the calculation and provide the result in Equation (7.44). One can refer to Appendix B for the details.

$$
\begin{aligned}
&\sum_{n=0}^{\infty} \mathbf{E}\left(X_n Y_n\right) \cdot z^n \\
&= \left.\frac{\partial^2 F(z,x,y)}{\partial x \partial y}\right|_{x=y=1} \\
&= \sum_{n=0}^{\infty} z^n \cdot (n+1)(n+2)\Big(\sum_{u\in\mathcal{U}} \pi(u)\Big)\Big(\sum_{v\in\mathcal{V}} \pi(v)\Big) \\
&\quad - \sum_{n=0}^{\infty} z^n \cdot 2 \cdot (n+1)\Bigg(\Big(\sum_{u\in\mathcal{U}} \pi(u)\Big)\Big(\sum_{v\in\mathcal{V}} \pi(v)\cdot|v|\Big) + \Big(\sum_{v\in\mathcal{V}} \pi(v)\Big)\Big(\sum_{u\in\mathcal{U}} \pi(u)\cdot|u|\Big)\Bigg) \\
&\quad + \sum_{n=0}^{\infty} z^n \cdot 2 \cdot \Big(\sum_{u\in\mathcal{U}} \pi(u)\cdot|u|\Big)\Big(\sum_{v\in\mathcal{V}} \pi(v)\cdot|v|\Big) \\
&\quad + \sum_{n=0}^{\infty} z^n \cdot (n+1) \cdot \Big(\Upsilon_{12}(1) + \Upsilon_{13}(1) + \Upsilon_{23}(1) + \Upsilon_{33}(1) + \Upsilon_{3}(1)\Big) \\
&\quad - \sum_{n=0}^{\infty} z^n \cdot \Big(\Upsilon'_{12}(1) + \Upsilon'_{13}(1) + \Upsilon'_{23}(1) + \Upsilon'_{33}(1) + \Upsilon'_{3}(1)\Big)
\end{aligned} \tag{7.44}
$$

Therefore, we obtain $\mathbf{E}\left(X_n Y_n\right)$.

$$
\begin{aligned}
&\mathbf{E}\left(X_n Y_n\right) \\
&= (n+1)(n+2) \cdot \Big(\sum_{u\in\mathcal{U}} \pi(u)\Big) \cdot \Big(\sum_{v\in\mathcal{V}} \pi(v)\Big) \\
&\quad - 2 \cdot (n+1)\left(\Big(\sum_{u\in\mathcal{U}} \pi(u)\Big)\Big(\sum_{v\in\mathcal{V}} \pi(v)\cdot|v|\Big) + \Big(\sum_{v\in\mathcal{V}} \pi(v)\Big)\Big(\sum_{u\in\mathcal{U}} \pi(u)\cdot|u|\Big)\right) \\
&\quad + 2 \cdot \Big(\sum_{u\in\mathcal{U}} \pi(u)\cdot|u|\Big)\Big(\sum_{v\in\mathcal{V}} \pi(v)\cdot|v|\Big) \\
&\quad + (n+1) \cdot \Big(\Upsilon_{12}(1) + \Upsilon_{13}(1) + \Upsilon_{23}(1) + \Upsilon_{33}(1) + \Upsilon_{3}(1)\Big) \\
&\quad - \Big(\Upsilon'_{12}(1) + \Upsilon'_{13}(1) + \Upsilon'_{23}(1) + \Upsilon'_{33}(1) + \Upsilon'_{3}(1)\Big)
\end{aligned} \tag{7.45}
$$

7.4.4 Covariance

Here we are eventually able to compute $\mathbf{Cov}(X_n, Y_n)$. According to (7.42), (7.43) and (7.45), the covariance of X_n and Y_n gives

$$\mathbf{Cov}(X_n, Y_n)$$

$$
\begin{aligned}
&= \mathbf{E}\left(X_n Y_n\right) - \mathbf{E}\left(X_n\right) \cdot \mathbf{E}\left(Y_n\right) \\
&= (n+1)(n+2) \cdot \left(\sum_{u \in \mathcal{U}} \pi(u)\right) \left(\sum_{v \in \mathcal{V}} \pi(v)\right) \\
&\quad - 2 \cdot (n+1) \left(\left(\sum_{u \in \mathcal{U}} \pi(u)\right) \left(\sum_{v \in \mathcal{V}} \pi(v) \cdot |v|\right) + \left(\sum_{v \in \mathcal{V}} \pi(v)\right) \left(\sum_{u \in \mathcal{U}} \pi(u) \cdot |u|\right)\right) \\
&\quad + 2 \cdot \left(\sum_{u \in \mathcal{U}} \pi(u) \cdot |u|\right) \left(\sum_{v \in \mathcal{V}} \pi(v) \cdot |v|\right) \\
&\quad + (n+1) \cdot \Big(\Upsilon_{12}(1) + \Upsilon_{13}(1) + \Upsilon_{23}(1) + \Upsilon_{33}(1) + \Upsilon_{3}(1)\Big) \\
&\quad - \Big(\Upsilon'_{12}(1) + \Upsilon'_{13}(1) + \Upsilon'_{23}(1) + \Upsilon'_{33}(1) + \Upsilon'_{3}(1)\Big) \\
&\quad - \left(\sum_{u \in \mathcal{U}} \pi(u) \cdot (n+1-|u|)\right) \cdot \left(\sum_{v \in \mathcal{V}} \pi(v) \cdot (n+1-|v|)\right) \\
&= n \cdot \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} \pi(u) \cdot \pi(v) \cdot (1 + |u| + |v|) \\
&\quad + (n+1) \cdot \Big(\Upsilon_{12}(1) + \Upsilon_{13}(1) + \Upsilon_{23}(1) + \Upsilon_{33}(1) + \Upsilon_{3}(1)\Big) \\
&\quad - 2 \cdot (n+1) \cdot \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} \pi(u) \cdot \pi(v) \cdot (|u| + |v|) + o(n) \\
&= n \cdot \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} \pi(u) \cdot \pi(v) \cdot (1 - |u| - |v|) \\
&\quad + n \cdot \Big(\Upsilon_{12}(1) + \Upsilon_{13}(1) + \Upsilon_{23}(1) + \Upsilon_{33}(1) + \Upsilon_{3}(1)\Big) + o(n)
\end{aligned}
$$

Therefore, we have the conclusion.

Theorem 7.4.1 Consider a random Bernoulli text of size n and two pattern sets $\mathcal{U}$ and $\mathcal{V}$. We allow for the case in which the intersection of the two pattern sets, $\mathcal{W} = \mathcal{U} \cap \mathcal{V}$, may not be empty. Two random variables X_n and Y_n are defined as follows: X_n is the number of pattern occurrences from $\mathcal{U}$ in the text of size n;
Y_n is the number of pattern occurrences from $\mathcal{V}$ in the text of size n.

We have

$$
\begin{aligned}
&\lim_{n \to \infty} \frac{1}{n} \mathbf{Cov}\left(X_n, Y_n\right) \\
&= \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} \pi(u) \cdot \pi(v) \cdot (1 - |u| - |v|)
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{u\in\mathcal{U}\setminus\mathcal{V}} \sum_{v\in\mathcal{V}\setminus\mathcal{U}} \Big(\pi(u)\cdot|u|_v + \pi(v)\cdot|v|_u + \pi(u)\cdot\pi\left(\mathcal{E}_{u,v}\right) + \pi(v)\cdot\pi\left(\mathcal{E}_{v,u}\right) \Big) \\
&+ \sum_{u\in\mathcal{U}\setminus\mathcal{V}} \sum_{w\in\mathcal{U}\cap\mathcal{V}} \Big(\pi(u)\cdot|u|_w + \pi(w)\cdot|w|_u + \pi(u)\cdot\pi\left(\mathcal{E}_{u,w}\right) + \pi(w)\cdot\pi\left(\mathcal{E}_{w,u}\right) \Big) \\
&+ \sum_{v\in\mathcal{V}\setminus\mathcal{U}} \sum_{w\in\mathcal{U}\cap\mathcal{V}} \Big(\pi(v)\cdot|v|_w + \pi(w)\cdot|w|_v + \pi(v)\cdot\pi\left(\mathcal{E}_{v,w}\right) + \pi(w)\cdot\pi\left(\mathcal{E}_{w,v}\right) \Big) \\
&+ \sum_{w_1\in\mathcal{W}} \sum_{w_2\in\mathcal{W}} 2\cdot\Big(\pi\left(w_1\right)\cdot|w_1|_{w_2}\cdot[\![w_1\neq w_2]\!] + \pi\left(w_1\right)\cdot\pi\left(\mathcal{E}_{w_1,w_2}\right)\Big) \\
&+ \sum_{w\in\mathcal{W}} \pi\left(w\right) \\
&+ o(1)
\end{aligned}
\tag{7.46}
$$

Proof 5 We already have the result of $\mathbf{Cov}(X_n, Y_n)$, which gives

$$
\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathbf{Cov}\left(X_n, Y_n\right) = &\sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} \pi(u)\cdot\pi(v)\cdot(1-|u|-|v|) \\
&+ \Big(\Upsilon_{12}(1)+\Upsilon_{13}(1)+\Upsilon_{23}(1)+\Upsilon_{33}(1)+\Upsilon_{3}(1)\Big) + o(1)
\end{aligned}
$$

The term $\Big(\Upsilon_{12}(1)+\Upsilon_{13}(1)+\Upsilon_{23}(1)+\Upsilon_{33}(1)+\Upsilon_{3}(1)\Big)$ can be computed by results from (7.21), (7.29)–(7.31) and (7.38), in Section 7.4.1. This term becomes

$$
\begin{aligned}
&\sum_{u\in\mathcal{U}\setminus\mathcal{V}} \sum_{v\in\mathcal{V}\setminus\mathcal{U}} \Big(\pi(u)\cdot|u|_v + \pi(v)\cdot|v|_u + \pi(u)\cdot\pi\left(\mathcal{E}_{u,v}\right) + \pi(v)\cdot\pi\left(\mathcal{E}_{v,u}\right) \Big) \\
&+ \sum_{u\in\mathcal{U}\setminus\mathcal{V}} \sum_{w\in\mathcal{U}\cap\mathcal{V}} \Big(\pi(u)\cdot|u|_w + \pi(w)\cdot|w|_u + \pi(u)\cdot\pi\left(\mathcal{E}_{u,w}\right) + \pi(w)\cdot\pi\left(\mathcal{E}_{w,u}\right) \Big) \\
&+ \sum_{v\in\mathcal{V}\setminus\mathcal{U}} \sum_{w\in\mathcal{U}\cap\mathcal{V}} \Big(\pi(v)\cdot|v|_w + \pi(w)\cdot|w|_v + \pi(v)\cdot\pi\left(\mathcal{E}_{v,w}\right) + \pi(w)\cdot\pi\left(\mathcal{E}_{w,v}\right) \Big) \\
&+ \sum_{w_1\in\mathcal{W}} \sum_{w_2\in\mathcal{W}} 2\cdot\Big(\pi\left(w_1\right)\cdot|w_1|_{w_2}\cdot[\![w_1\neq w_2]\!] + \pi\left(w_1\right)\cdot\pi\left(\mathcal{E}_{w_1,w_2}\right)\Big) \\
&+ \sum_{w\in\mathcal{W}} \pi\left(w\right)
\end{aligned}
$$

Thus, Formula (7.46) is obtained. ■

We also have the following corollary.

Corollary 7.4.1 Let variable X_n count the number of pattern occurrences from pattern set $\mathcal{U}$ in a random Bernoulli text of size n. We have

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left(X_n\right)=\sum_{u_1\in\mathcal{U}}\sum_{u_2\in\mathcal{U}}\pi(u_1)\cdot\pi(u_2)\cdot(1-|u_1|-|u_2|)$$
$$+\sum_{u_1\in\mathcal{U}}\sum_{u_2\in\mathcal{U}}2\cdot\left(\pi\left(u_1\right)\cdot|u_1|_{u_2}\cdot[\![u_1\neq u_2]\!]+\pi\left(u_1\right)\cdot\pi\left(\mathcal{E}_{u_1,u_2}\right)\right) \tag{7.47}$$
$$+\sum_{u\in\mathcal{U}}\pi\left(u\right)+o(1)$$

Proof 6 Since $\mathbf{Var}\left(X_n\right)=\mathbf{Cov}\left(X_n,X_n\right)$, we respectively replace $\mathcal{V}$ with $\mathcal{U}$, and Y_n with X_n, in Theorem 7.4.1. Equation (7.46) is then simplified because $\mathcal{U}\backslash\mathcal{U}=\emptyset$, and $\mathcal{U}\cap\mathcal{U}=\mathcal{U}$. Therefore, we obtain (7.47). ■

7.5 An example

We make an example in this section to demonstrate the power of Theorem 7.4.1.

Example 7.5.1 Consider two pattern sets $\mathcal{U}=\{a^2,a^3\}$ and $\mathcal{V}=\{a^3,a^4\}$. Let X_n,Y_n respectively count the number of pattern occurrences from $\mathcal{U}$ and $\mathcal{V}$ in a random Bernoulli text of size n, where $p=\pi(a)$ in this Bernoulli model. Our aim is the variance-covariance matrix of X_n and Y_n.

We start with the computation of $\mathbf{Cov}\left(X_n,Y_n\right)$ using Theorem 7.4.1. First, we calculate

$$\begin{aligned}
&\sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}}\pi(u)\cdot\pi(v)\cdot(1-|u|-|v|)\\
&=\pi(a^2)\cdot\pi(a^3)\cdot(1-|a^2|-|a^3|)+\pi(a^2)\cdot\pi(a^4)\cdot(1-|a^2|-|a^4|)\\
&\quad+\pi(a^3)\cdot\pi(a^3)\cdot(1-|a^3|-|a^3|)+\pi(a^3)\cdot\pi(a^4)\cdot(1-|a^3|-|a^4|)\\
&=p^2\cdot p^3\cdot(1-2-3)+p^2\cdot p^4\cdot(1-2-4)\\
&\quad+p^3\cdot p^3\cdot(1-3-3)+p^3\cdot p^4\cdot(1-3-4)\\
&=-4p^5-10p^6-6p^7
\end{aligned}$$

As $\mathcal{U}$ and $\mathcal{V}$ have one pattern in common, we list three disjoint sets:

$$\mathcal{U}\backslash\mathcal{V}=\{a^2\},\quad \mathcal{V}\backslash\mathcal{U}=\{a^4\},\quad \mathcal{W}=\mathcal{U}\cap\mathcal{V}=\{a^3\}$$

Therefore,

$$
\begin{aligned}
&\sum_{u\in\mathcal{U}\setminus\mathcal{V}}\sum_{v\in\mathcal{V}\setminus\mathcal{U}}\Big(\pi(u)\cdot|u|_v+\pi(v)\cdot|v|_u+\pi(u)\cdot\pi\left(\mathcal{E}_{u,v}\right)+\pi(v)\cdot\pi\left(\mathcal{E}_{v,u}\right)\Big)\\
&+\sum_{u\in\mathcal{U}\setminus\mathcal{V}}\sum_{w\in\mathcal{U}\cap\mathcal{V}}\Big(\pi(u)\cdot|u|_w+\pi(w)\cdot|w|_u+\pi(u)\cdot\pi\left(\mathcal{E}_{u,w}\right)+\pi(w)\cdot\pi\left(\mathcal{E}_{w,u}\right)\Big)\\
&+\sum_{v\in\mathcal{V}\setminus\mathcal{U}}\sum_{w\in\mathcal{U}\cap\mathcal{V}}\Big(\pi(v)\cdot|v|_w+\pi(w)\cdot|w|_v+\pi(v)\cdot\pi\left(\mathcal{E}_{v,w}\right)+\pi(w)\cdot\pi\left(\mathcal{E}_{w,v}\right)\Big)\\
&+\sum_{w_1\in\mathcal{W}}\sum_{w_2\in\mathcal{W}}2\cdot\Big(\pi\left(w_1\right)\cdot|w_1|_{w_2}+\pi\left(w_1\right)\cdot\pi\left(\mathcal{E}_{w_1,w_2}\right)\Big)\\
&+\sum_{w\in\mathcal{W}}\pi\left(w\right)\\
=\;&\pi(a^2)\cdot|a^2|_{a^4}+\pi(a^4)\cdot|a^4|_{a^2}+\pi(a^2)\cdot\pi\left(\mathcal{E}_{a^2,a^4}\right)+\pi(a^4)\cdot\pi\left(\mathcal{E}_{a^4,a^2}\right)\\
&+\pi(a^2)\cdot|a^2|_{a^3}+\pi(a^3)\cdot|a^3|_{a^2}+\pi(a^2)\cdot\pi\left(\mathcal{E}_{a^2,a^3}\right)+\pi(a^3)\cdot\pi\left(\mathcal{E}_{a^3,a^2}\right)\\
&+\pi(a^4)\cdot|a^4|_{a^3}+\pi(a^3)\cdot|a^3|_{a^4}+\pi(a^4)\cdot\pi\left(\mathcal{E}_{a^4,a^3}\right)+\pi(a^3)\cdot\pi\left(\mathcal{E}_{a^3,a^4}\right)\\
&+2\cdot\Big(\pi\left(a^3\right)\cdot\left|a^3\right|_{a^3}+\pi\left(a^3\right)\cdot\pi\left(\mathcal{E}_{a^3,a^3}\right)\Big)\\
&+\pi\left(a^3\right)\\
=\;&0+3p^4+p^2\cdot p^3+p^4\cdot p\\
&+0+2p^3+p^2\cdot p^2+p^3\cdot p\\
&+2p^4+0+p^4\cdot(p+p^2)+p^3\cdot(p^2+p^3)\\
&+2\cdot\Big(0+p^3\cdot(p+p^2)\Big)\\
&+p^3\\
=\;&3p^3+9p^4+6p^5+2p^6
\end{aligned}
$$

According to Theorem 7.4.1, we have the covariance:

$$
\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathbf{Cov}\left(X_n,Y_n\right)&=(-4p^5-10p^6-6p^7)+(3p^3+9p^4+6p^5+2p^6)+o(1)\\
&=p^3\cdot\left(3+9p+2p^2-8p^3-6p^4\right)+o(1)
\end{aligned}
$$

Next, we compute $\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left(X_n\right)$ and $\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left(Y_n\right)$ by applying Corollary 7.4.1:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left(X_n\right) = &\sum_{u_1\in\mathcal{U}}\sum_{u_2\in\mathcal{U}}\pi(u_1)\cdot\pi(u_2)\cdot(1-|u_1|-|u_2|)\\ &+\sum_{u_1\in\mathcal{U}}\sum_{u_2\in\mathcal{U}}2\cdot\left(\pi\left(u_1\right)\cdot|u_1|_{u_2}+\pi\left(u_1\right)\cdot\pi\left(\mathcal{E}_{u_1,u_2}\right)\right)\\ &+\sum_{u\in\mathcal{U}}\pi\left(u\right)+o(1)\end{aligned}$$

In this case, we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left(X_n\right) = &\pi(a^2)\cdot\pi(a^2)\cdot(1-|a^2|-|a^2|)+\pi(a^2)\cdot\pi(a^3)\cdot(1-|a^2|-|a^3|)\\ &+\pi(a^3)\cdot\pi(a^2)\cdot(1-|a^3|-|a^2|)+\pi(a^3)\cdot\pi(a^3)\cdot(1-|a^3|-|a^3|)\\ &+2\cdot\left(\pi\left(a^2\right)\cdot\pi\left(\mathcal{E}_{a^2,a^2}\right)+\pi\left(a^2\right)\cdot\pi\left(\mathcal{E}_{a^2,a^3}\right)\right.\\ &\left.+\pi\left(a^3\right)\cdot\left|a^3\right|_{a^2}+\pi\left(a^3\right)\cdot\pi\left(\mathcal{E}_{a^3,a^2}\right)+\pi\left(a^3\right)\cdot\pi\left(\mathcal{E}_{a^3,a^3}\right)\right)\\ &+\pi(a^2)+\pi(a^3)+o(1)\\ = &\ p^2\cdot p^2\cdot(1-2-2)+p^2\cdot p^3\cdot(1-2-3)\\ &+p^3\cdot p^2\cdot(1-3-2)+p^3\cdot p^3\cdot(1-3-3)\\ &+2\cdot\left(p^2\cdot p+p^2\cdot p^2+2p^3+p^3\cdot p+p^3\cdot(p+p^2)\right)\\ &+p^2+p^3+o(1)\\ = &\ p^2\cdot\left(1+7p+3p^2-6p^3-5p^4\right)+o(1)\end{aligned}$$

The variance of Y_n can be obtained similarly. We omit the details of computation and give the result.

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left(Y_n\right)=p^3\cdot\left(1+7p+8p^2+p^3-10p^4-7p^5\right)+o(1)$$

7.6 Higher moments

If needed, the tools that we use to compute the first and second moment of pattern occurrence counting can be further expand to obtain higher moments. The mth order partial derivative with respect to x on both sides of Equation (7.5) gives

$$\sum_{n\geq 0} \mathbf{E}\Big(X_n(X_n-1)\cdots(X_n-m+1)\Big)z^n = \frac{\partial^m}{\partial x^m}F(z,x)\bigg|_{x=1}$$

This is the generalized version of Theorem 7.2.1, and is known as the mth factorial moment. (See the text by Flajolet and Sedgewick [13, Appendix A.3].)

Therefore, by mathematical induction, we should be able to obtain the moment in any higher order, though the computation is expected to be more complex.

8. MOMENTS OF OCCURRENCES FOR PATTERNS IN A MARKOVIAN TEXT

Suppose that our context is the same as that in a Bernoulli text. Namely, we have an alphabet $\mathcal{A} = \{a_1, a_2, \dots\ a_\ell\}$ and two sets of pattern words, $\mathcal{U} = \{u_1, u_2, ..., u_r\}$, and $\mathcal{V} = \{v_1, v_2, ..., v_s\}$. The two sets may have non-empty intersection $\mathcal{W} = \mathcal{U} \cap \mathcal{V}$.

In Chapter 7, we use two random variables X_n and Y_n to respectively count the occurrences of patterns from each pattern set. However, in a first-order Markovian context, we need to specify a letter $a_i \in \mathcal{A}$ that the text follows. Hence, the number of random variables immediately jumps from 2 (i.e., X_n and Y_n) to $2|\mathcal{A}|$, i.e., ${}^{a_1}X_n, {}^{a_1}Y_n, {}^{a_2}X_n, {}^{a_2}Y_n, \dots, {}^{a_\ell}X_n, {}^{a_\ell}Y_n$.

To obtain the first moment for any of these random variables, the corresponding generating function ${}^{a_j}F(z, x, y)$ is needed. Thus, we should follow the procedures specified in Theorem 6.2.1, which requires solving linear equations

$$\begin{pmatrix} {}^{a_1}T(z,\mathbf{t}) \\ {}^{a_2}T(z,\mathbf{t}) \\ \vdots \\ {}^{a_\ell}T(z,\mathbf{t}) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + P_{(\ell\times\ell)} \cdot z \cdot \begin{pmatrix} {}^{a_1}T(z,\mathbf{t}) \\ {}^{a_2}T(z,\mathbf{t}) \\ \vdots \\ {}^{a_\ell}T(z,\mathbf{t}) \end{pmatrix} + \begin{pmatrix} {}^{a_1}\eta(z,\mathbf{t}) \\ {}^{a_2}\eta(z,\mathbf{t}) \\ \vdots \\ {}^{a_\ell}\eta(z,\mathbf{t}) \end{pmatrix} \cdot \begin{pmatrix} \widehat{u_1}T(z,\mathbf{t}) \\ \widehat{u_2}T(z,\mathbf{t}) \\ \vdots \\ \widehat{u_r}T(z,\mathbf{t}) \end{pmatrix} \tag{8.1}$$

Recall that each ${}^{a_j}\eta(z,\mathbf{t})$ (where $j \in \{1, 2, \dots, \ell\}$) is a $1 \times r$ row vector, in the form of

$$\left((\cdots)\widehat{\mathrm{Flip}(\mathsf{u}_1)},\ (\cdots)\widehat{\mathrm{Flip}(\mathsf{u}_2)},\ \dots,\ (\cdots)\widehat{\mathrm{Flip}(\mathsf{u}_r)} \right).$$

Without knowing the specific pattern words in $\mathcal{U}$ and $\mathcal{V}$, it is impractical for us to go any further than the general form of linear equations. Moreover, as the pattern words in one set may have different ending letters, counting them together prevents us from systematically determining the ending letter of a cluster.

Therefore, unlike texts with Bernoulli sources, we do not have an elegant result as Theorem 7.4.1 when a Markovian source is involved.

However, by utilizing the results in Chapter 6, we can still calculate the moments of occurrences of pattern words (not pattern sets) in a Markovian text. In the next section, we provide an example of Markovian case in which the first and second moment of pattern occurrences are calculated.

8.1 Moments in a Markovian text

The generating functions of decorated texts can be obtained by Theorem 6.2.1. In Example 6.3.1, we obtained two generating functions, namely, ${}^{a}F(z, x_1, x_2, x_3, x_4)$ from (6.11) and ${}^{b}F(z, x_1, x_2, x_3, x_4)$ from (6.12).

The following example is designed so that we can use the result of those generating functions and then focus on calculating the moments.

Example 8.1.1 A binary text is generated by a first-order Markovian source with the transition matrix

$$P = \begin{pmatrix} p_{aa} & p_{ab} \\ p_{ba} & p_{bb} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 3/5 & 2/5 \end{pmatrix} \tag{8.2}$$

Consider two patterns

$$u_1 = ab,\ u_2 = aba$$

and define four random variables ${}^{a}X_n, {}^{b}X_n, {}^{a}Y_n$ and ${}^{b}Y_n$ as follows:

${}^{a}X_n$ is the number of occurrences of $u_1 = ab$ in a text of length n following a letter a;

${}^{b}X_n$ is the number of occurrences of $u_1 = ab$ in a text of length n following a letter b;

${}^{a}Y_n$ is the number of occurrences of $u_2 = aba$ in a text of length n following a letter a;

${}^{b}Y_n$ is the number of occurrences of $u_2 = aba$ in a text of length n following a letter b.

In order to calculate the moments of ${}^{a}X_n, {}^{b}X_n, {}^{a}Y_n$ and ${}^{b}Y_n$, we need to obtain the generating functions of the pattern occurrences, i.e., ${}^{a}F(z, x, y)$ and ${}^{b}F(z, x, y)$, where x, y is the dummy variable to count the occurrence of u_1 and u_2, respectively.

Observe that the two patterns u_1 and u_2 in this example are identical to u_1 and u_2 in Example 6.3.1. Therefore, we can obtain ${}^aF(z,x,y)$ and ${}^bF(z,x,y)$ directly by making the following replacements to ${}^aF(z,x_1,x_2,x_3,x_4)$ and ${}^bF(z,x_1,x_2,x_3,x_4)$ of Example 6.3.1.

$$x_1 \to x, \quad x_2 \to y, \quad x_3 \to 1, \quad x_4 \to 1.$$

Thus, we have

$${}^aF(z,x,y) = \frac{100 - 3x(y-1)z^3 + (5 + (25 - 30y)x)z^2 + 10z}{100 + 12x(y-1)z^3 + (20 - 30xy)z^2 - 90z} \tag{8.3}$$

and

$${}^bF(z,x,y) = \frac{50 - 15(y-1)xz^2 + 5z}{50 + 6x(y-1)z^3 + (10 - 15xy)z^2 - 45z} \tag{8.4}$$

Now we calculate the first moments of ${}^aX_n$ by Theorem 7.2.1. The generating functions of the first moment of ${}^aX_n$, ${}^bX_n$, ${}^aY_n$ and ${}^bY_n$ can be obtained by Theorem 7.2.1 as follows.

$$\sum_{n\geq 0} \mathbf{E}\left({}^aX_n\right) z^n = \frac{\partial}{\partial x} {}^aF(z,x,y)\Big|_{x=y=1} = \frac{z^2(z+5)}{2(z-1)^2(z+10)} \tag{8.5}$$

The Taylor expansion on the RHS of (8.5) gives

$$\frac{z^2(z+5)}{2(z-1)^2(z+10)} = \sum_{n\geq 0} \left(\frac{3n}{11} - \frac{71}{242} - \frac{25}{121}\left(-\frac{1}{10}\right)^n\right) \cdot z^n$$

Hence, we have

$$\mathbf{E}\left({}^aX_n\right) = \frac{3n}{11} - \frac{71}{242} - \frac{25}{121} \cdot \left(-\frac{1}{10}\right)^n \tag{8.6}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \mathbf{E}\left({}^aX_n\right) = \frac{3}{11} \tag{8.7}$$

The first moment of ${}^bX_n, {}^aY_n$ and ${}^bY_n$ can be obtained by performing similar procedures. We skip the calculations and provide the results here.

For ${}^bX_n$, we have

$$\mathbf{E}\left({}^bX_n\right) = \frac{3}{121}\left(11n - 10 + 10 \cdot \left(-\frac{1}{10}\right)^n\right) \tag{8.8}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{E}\left({}^{b}X_n\right)=\frac{3}{11} \tag{8.9}$$

For ${}^{a}Y_n$, we have

$$\mathbf{E}\left({}^{a}Y_n\right)=\frac{3}{1210}\left(66n-137+500\cdot\left(-\frac{1}{10}\right)^{n}\right) \tag{8.10}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{E}\left({}^{a}Y_n\right)=\frac{9}{55} \tag{8.11}$$

For ${}^{b}Y_n$, we have

$$\mathbf{E}\left({}^{b}Y_n\right)=\frac{9}{605}\left(11n-21-100\left(-\frac{1}{10}\right)^{n}\right) \tag{8.12}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{E}\left({}^{b}Y_n\right)=\frac{9}{55} \tag{8.13}$$

Next, we move forward to the variance of ${}^{a}X_n$, ${}^{b}X_n$, ${}^{a}Y_n$ and ${}^{b}Y_n$. We perform the calculation of $\mathbf{Var}({}^{a}X_n)$, and then give the results of $\mathbf{Var}({}^{b}X_n)$, $\mathbf{Var}({}^{a}Y_n)$, and $\mathbf{Var}({}^{b}Y_n)$.

According to Theorem 7.2.1, we have

$$\begin{aligned}\sum_{n\geq 0}\mathbf{E}\left({}^{a}X_n^2\right)z^n&=\frac{\partial^2}{\partial x^2}{}^{a}F(z,x,y)\Big|_{x=y=1}+\frac{\partial}{\partial x}{}^{a}F(z,x,y)\Big|_{x=y=1}\\&=\frac{z^2(z+5)\left(5z^2-9z+10\right)}{2(z-1)^3(z+10)^2}\end{aligned} \tag{8.14}$$

The Taylor expansion on the RHS of (8.14) gives

$$\begin{aligned}\frac{z^2(z+5)\left(5z^2-9z+10\right)}{2(z-1)^3(z+10)^2}=\sum_{n\geq 0}z^n\cdot\Bigg(&\frac{9n^2}{121}-\frac{129n}{1331}+\frac{4255}{29282}\\&-\frac{1500}{1331}n\left(-\frac{1}{10}\right)^{n}+\frac{34475}{14641}\cdot\left(-\frac{1}{10}\right)^{n}\Bigg)\end{aligned}$$

Hence, we have

$$\mathbf{E}\left({}^{a}X_n^2\right)=\frac{9n^2}{121}-\frac{129n}{1331}+\frac{4255}{29282}-\frac{1500}{1331}n\left(-\frac{1}{10}\right)^{n}+\frac{34475}{14641}\cdot\left(-\frac{1}{10}\right)^{n} \tag{8.15}$$

and

$$\lim_{n\to\infty}\frac{1}{n^2}\mathbf{E}\left({}^aX_n^2\right) = \frac{9}{121} \tag{8.16}$$

The variance of ${}^aX_n$ is given by

$$\begin{aligned}\mathbf{Var}\left({}^aX_n\right) &= \mathbf{E}\left({}^aX_n^2\right) - \left(\mathbf{E}\left({}^aX_n\right)\right)^2 \\ &= \frac{84n}{1331} + \frac{3469}{58564} - \frac{1350}{1331}n\left(-\frac{1}{10}\right)^n + \frac{32700}{14641}\left(-\frac{1}{10}\right)^n - \frac{625}{14641}\left(-\frac{1}{10}\right)^{2n}\end{aligned} \tag{8.17}$$

and therefore,

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left({}^aX_n\right) = \frac{84}{1331} \tag{8.18}$$

The other variance values can be obtain similarly. The results are as follows.

For ${}^bX_n$, we have

$$\mathbf{Var}\left({}^bX_n\right) = \frac{84n}{1331} + \frac{870}{14641} + \frac{1620}{1331}n\left(-\frac{1}{10}\right)^n + \frac{30}{14641}\left(-\frac{1}{10}\right)^n - \left(\frac{30}{121}\right)^2\left(-\frac{1}{10}\right)^{2n} \tag{8.19}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left({}^bX_n\right) = \frac{84}{1331} \tag{8.20}$$

For ${}^aY_n$, we have

$$\begin{aligned}\mathbf{Var}\left({}^aY_n\right) = &\frac{4122n}{33275} - \frac{409263}{1464100} + \frac{44460}{1331}n\left(-\frac{1}{10}\right)^n \\ &- \frac{1985520}{14641}\left(-\frac{1}{10}\right)^n - \left(\frac{150}{121}\right)^2\left(-\frac{1}{10}\right)^{2n}\end{aligned} \tag{8.21}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left({}^aY_n\right) = \frac{4122}{33275} \tag{8.22}$$

For ${}^bY_n$, we have

$$\begin{aligned}\mathbf{Var}\left({}^bY_n\right) = &\frac{4122n}{33275} - \frac{19224}{73205} - \frac{53352}{1331}n\left(-\frac{1}{10}\right)^n \\ &+ \frac{1195812}{14641}\left(-\frac{1}{10}\right)^n - \left(\frac{180}{121}\right)^2\left(-\frac{1}{10}\right)^{2n}\end{aligned} \tag{8.23}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left({}^{b}Y_{n}\right)=\frac{4122}{33275} \tag{8.24}$$

Lastly, we calculate the covariance value $\mathbf{Cov}\left({}^{a}X_{n},{}^{a}Y_{n}\right)$, and also give the result of the other covariance $\mathbf{Cov}\left({}^{b}X_{n},{}^{b}Y_{n}\right)$.

Note that $\mathbf{Cov}\left({}^{a}X_{n},{}^{b}Y_{n}\right)$ and $\mathbf{Cov}\left({}^{b}X_{n},{}^{a}Y_{n}\right)$ do not exist, as the counting of u_1 and u_2 must be performed on the same text, which either follows a letter a or a letter b.

$$\begin{aligned}\sum_{n\geq 0}\mathbf{E}\left({}^{a}X_{n}{}^{a}Y_{n}\right)z^{n}&=\left.\frac{\partial^{2}}{\partial x\partial y}{}^{a}F(z,x,y)\right|_{x=y=1}\\&=\frac{3z^{3}\left(2z^{2}+5z-25\right)}{5(z-1)^{3}(z+10)^{2}}\end{aligned} \tag{8.25}$$

The Taylor expansion of the RHS of (8.25) gives

$$\begin{aligned}\frac{3z^{3}\left(2z^{2}+5z-25\right)}{5(z-1)^{3}(z+10)^{2}}=\sum_{n\geq 0}z^{n}\cdot\Bigg(&\frac{27n^{2}}{605}-\frac{486n}{6655}-\frac{2346}{73205}\\&+\frac{750}{1331}n\left(-\frac{1}{10}\right)^{n}-\frac{17100}{14641}\left(-\frac{1}{10}\right)^{n}\Bigg)\end{aligned}$$

Hence, we have

$$\mathbf{E}\left({}^{a}X_{n}{}^{a}Y_{n}\right)=\frac{27n^{2}}{605}-\frac{486n}{6655}-\frac{2346}{73205}+\frac{750}{1331}n\left(-\frac{1}{10}\right)^{n}-\frac{17100}{14641}\left(-\frac{1}{10}\right)^{n} \tag{8.26}$$

The covariance $\mathbf{Cov}\left({}^{a}X_{n},{}^{a}Y_{n}\right)$ is given by

$$\begin{aligned}\mathbf{Cov}\left({}^{a}X_{n},{}^{a}Y_{n}\right)&=\mathbf{E}\left({}^{a}X_{n}{}^{a}Y_{n}\right)-\mathbf{E}\left({}^{a}X_{n}\right)\mathbf{E}\left({}^{a}Y_{n}\right)\\&=\frac{90n}{1331}-\frac{7713}{58564}+\frac{345}{1331}n\left(-\frac{1}{10}\right)^{n}\\&\quad-\frac{25605}{29282}\left(-\frac{1}{10}\right)^{n}+\frac{3750}{14641}\left(-\frac{1}{10}\right)^{2n}\end{aligned} \tag{8.27}$$

and therefore,

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Cov}\left({}^{a}X_{n},{}^{a}Y_{n}\right)=\frac{90}{1331} \tag{8.28}$$

A similar procedure gives the covariance $\mathbf{Cov}\left({}^{b}X_n, {}^{b}Y_n\right)$ as follows

$$\begin{aligned}\mathbf{Cov}\left({}^{b}X_n, {}^{b}Y_n\right) = \frac{90n}{1331} - \frac{1854}{14641} - \frac{414}{1331}n\left(-\frac{1}{10}\right)^n \\ - \frac{3546}{14641}\left(-\frac{1}{10}\right)^n + \frac{5400}{14641}\left(-\frac{1}{10}\right)^{2n}\end{aligned} \tag{8.29}$$

and therefore,

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Cov}\left({}^{b}X_n, {}^{b}Y_n\right) = \frac{90}{1331} \tag{8.30}$$

8.2 Higher order moments

Similar to the first and second moments for Markovian text, a closed form for higher moments does not exist. However, when the transition matrix and pattern set are given, one can certainly follow the procedure in Section 8.1, and compute any desired moment of interest.

It is worth emphasizing that a drawback to the approach (comparing to that with a Bernoulli source) is that we are not able to compute the moment of occurrences for an entire pattern set with a closed form. The moments of the number of occurrences (of any order) in a Markovian order 1 text, can be obtained for individual pattern words.

8.3 Remarks

One may have observed that in Example 8.1.1, we have the following results.

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{E}\left({}^{a}X_n\right) = \lim_{n\to\infty}\frac{1}{n}\mathbf{E}\left({}^{b}X_n\right) = \frac{3}{11}$$

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{E}\left({}^{a}Y_n\right) = \lim_{n\to\infty}\frac{1}{n}\mathbf{E}\left({}^{b}Y_n\right) = \frac{9}{55}$$

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left({}^{a}X_n\right) = \lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left({}^{b}X_n\right) = \frac{84}{1331}$$

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left({}^{a}Y_n\right)=\lim_{n\to\infty}\frac{1}{n}\mathbf{Var}\left({}^{b}Y_n\right)=\frac{4122}{33275}$$

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Cov}\left({}^{a}X_n,{}^{a}Y_n\right)=\lim_{n\to\infty}\frac{1}{n}\mathbf{Cov}\left({}^{b}X_n,{}^{b}Y_n\right)=\frac{90}{1331}$$

These results indicate that in a long run ($n\to\infty$), the moments of a pattern occurrence in a Markovian text may be independent of the letter that the text follows. Namely, the moments of ${}^{j}X_n$ or ${}^{j}Y_n$ are asymptotically independent of the letter j.

8.3.1 First moment in Markovian texts converges to Bernoulli

Without proof, we expect the statement is true with a time-homogeneous Markov chain that is aperiodic and irreducible. Let the stationary distribution for the alphabet $\mathcal{A}=\{a_1,a_2,\cdots,a_\ell\}$ be

$$\begin{pmatrix}\pi_1 & \pi_2 & \cdots & \pi_\ell\end{pmatrix}$$

Suppose a pattern $u=\alpha_1\alpha_2\alpha_3\cdots\alpha_{|u|}$, where each α_i is a single letter from the alphabet $\mathcal{A}$. Then we have, in a long run,

$${}^{a_j}\pi(u)=\pi\left(\alpha_1\right)\cdot P_{\alpha_1,\alpha_2}\cdot P_{\alpha_2,\alpha_3}\cdots P_{\alpha_{(|u|-1)},\alpha_{|u|}}\quad\text{for }\forall j\in\{1,2,\cdots,\ell\}$$

It is independent of a_j, the letter that the entire text follows. i.e.,

$${}^{a_j}\pi(u)=\pi(u)\quad\text{for }\forall j\in\{1,2,\cdots,\ell\}$$

Therefore, in a Markovian context, we have

$$\lim_{n\to\infty}\mathbf{E}\left({}^{a_j}X_n\right)=n\cdot\pi(u)+o(n)$$

or equivalently,

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{E}\left({}^{a_j}X_n\right)=\pi(u)+o(1)\quad\text{for }\forall j\in\{1,2,\cdots,\ell\}\tag{8.31}$$

Comparing to the first moment in Bernoulli texts (7.3) or (7.42):

$$\mathbf{E}\left(X_n\right) = \sum_{u\in\mathcal{U}} \pi(u)(n - |u| + 1)$$

We have, in Bernoulli texts,

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left(X_n\right) = \sum_{u\in\mathcal{U}} \pi(u) + o(1) \tag{8.32}$$

The limit of the Markovian moment (8.31) is consistent with that of the Bernoulli moment (8.32), with the constraint that only one pattern is allowed in $\mathcal{U}$.

The discussion above is easily verified with our results in Example 8.1.1. Our transition matrix

$$P = \begin{pmatrix} p_{aa} & p_{ab} \\ p_{ba} & p_{bb} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 3/5 & 2/5 \end{pmatrix}$$

has a stationary probabilities $\pi(a)$ and $\pi(b)$ where

$$\pi(a) = \frac{6}{11}, \quad \pi(b) = \frac{5}{11}$$

For large n, the probabilities of the two patterns, $\pi(ab)$ and $\pi(aba)$, are respectively

$$\pi(ab) = \pi(a) \cdot p_{ab} = \frac{3}{11}$$

and

$$\pi(aba) = \pi(a) \cdot p_{ab} \cdot p_{ba} = \frac{9}{55}$$

Therefore,

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left({}^aX_n\right) = \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left({}^bX_n\right) = \pi(ab)$$

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left({}^aY_n\right) = \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left({}^bY_n\right) = \pi(aba)$$

These results are consistent with the non-Markovian case.

9. SUMMARY

This thesis is a methodology and combinatorial structural analysis of counting the occurrences of patterns in a Markovian text.

We first introduced the fundamentals of generating functions and the inclusion-exclusion method. We demonstrated how Bassino et al.'s novel approach [5] counts the pattern occurrences on a sequence from a Bernoulli source. Then we build the combinatorial structure from a recursive point of view, in order to adapt the Markovian dependence. For Bernoulli texts, our model leads to the same results of Bassino et al. [5].

The difficulty of counting pattern occurrences in a Markovian text is the fact that each letter is dependent on what have appeared directly before itself. Therefore, no matter for a letter, a cluster, or a text, the construction of the generating function must consider all possible letters that comes before it. In addition, since a text is considered a sequence of letters and clusters, we also need to keep track of the ending letter of every block (e.g., a letter of a cluster) as it is the pre-letter of the next block.

It is expected to cause a large amount of computation when the alphabet is huge. This situation could happen when we need to consider higher order Markovian dependence, even if for a small alphabet set. As for higher order Markovian, we can always expand the alphabet and the corresponding transition matrix in order to boil down to the Markovian of order 1.

Due to the Markovian dependence, our results are given in the form of a linear matrix equation. The generating functions that depend on all possible letters can be obtained. We computed the first and second moments of pattern occurrences, and compared the results in the Bernoulli case and our first-order Markovian case. We found that for a time-homogeneous Markov chain that is aperiodic and irreducible, the first moment of pattern occurrences in Markovian texts will converge to the Bernoulli case in a long run. The second moment in Markovian texts does converge, but not to the corresponding Bernoulli case. The behavior of the convergence could be a future direction of this topic.

REFERENCES

[1] A. V. Aho and M. J. Corasick. Efficient string matching: An aid to bibliographic search. *Commun. ACM*, 18(6):333–340, June 1975.

[2] A. Apostolico. The myriad virtues of suffix trees. *Combinatorial Algorithms on Words, vol 12 of NATO Advance Science Institutes, Series F: Computer and Systems Sciences*, pages 85–96, 1985.

[3] F. Bassino, J. Clément, J. Fayolle, and P. Nicodème. Counting occurrences for a finite set of words: an inclusion-exclusion approach. *DMTCS Proceedings, 2007 Conference on Analysis of Algorithms, AofA 07*, pages 29–44, June 2007.

[4] F. Bassino, J. Clément, J. Fayolle, and P. Nicodème. Constructions for clumps statistics. *2008 Discrete Mathematics and Theoretical Computer Science (DMTCS)*, pages 179–194, April 2008.

[5] F. Bassino, J. Clément, and P. Nicodème. Counting occurrences for a finite set of words: Combinatorial methods. *ACM Trans. Algorithms*, 8(3):31:1–31:28, July 2012.

[6] G. Blom and D. Thorburn. How many random digits are required until given sequences are obtained? *Journal of Applied Probability*, 19(3):518–531, 1982.

[7] S. Breen, M. S. Waterman, and N. Zhang. Renewal theory for several patterns. *Journal of Applied Probability*, 22(1):228–234, 1985.

[8] O. Chryssaphinou and S. Papastavridis. The occurrence of sequence patterns in repeated dependent experiments. *Theory of Probability & Its Applications*, 35:145–152, January 1991.

[9] M. Crochemore, C. Hancart, and T. Lecroq. *Algorithms on Strings*. Cambridge University Press, 2007.

[10] M. Crochemore and W. Rytter. *Jewels of Stringology*. World Scientific, 2002.

[11] J. Fayolle and M. D. Ward. Analysis of the average depth in a suffix tree under a Markov model. *2005 International Conference on Analysis of Algorithms, DMTCS Proceedings*, AD:95–104, 2005.

[12] P. Flajolet, X. Gourdon, and C. Martínez. Patterns in random binary search trees. *Random Struct. Algorithms*, 11:223–244, October 1997.

[13] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[14] H. U. Gerber and S.-Y. R. Li. The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. *Stochastic Processes and their Applications*, 11(1):101–108, 1981.

[15] I. Gheorghiciuc and M. D. Ward. On correlation polynomials and subword complexity. *Discrete Mathematics & Theoretical Computer Science*, DMTCS Proceedings vol. AH, 2007 Conference on Analysis of Algorithms (AofA 07), January 2007.

[16] I. Goulden and D. Jackson. *Combinatorial Enumeration*. John Wiley, New York, 1983.

[17] I. P. Goulden and D. M. Jackson. An inversion theorem for cluster decompositions of sequences with distinguished subsequences. *Journal of the London Mathematical Society*, s2-20(3):567–576, 1979.

[18] L. Guibas and A. Odlyzko. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory, Series A*, 30(2):183–208, 1981.

[19] L. J. Guibas and A. M. Odlyzko. Periods in strings. *Journal of Combinatorial Theory, Series A*, 30(1):19–42, 1981.

[20] P. Jacquet and W. Szpankowski. Analysis of digital tries with Markovian dependency. *IEEE Transactions on Information Theory*, 37(5):1470–1475, 1991.

[21] P. Jacquet and W. Szpankowski. Autocorrelation on words and its applications: Analysis of suffix trees by string-ruler approach. *Journal of Combinatorial Theory, Series A*, 66(2):237–269, 1994.

[22] D. Knuth, J. H. Morris, and V. Pratt. Fast pattern matching in strings. *SIAM Journal on Computing*, 6(2):323–350, 1977.

[23] M. Lothaire. *Applied Combinatorics on Words*. Encyclopedia of Mathematics and its Applications, vol. 105. Cambridge University Press, 2005.

[24] P. Nicodème. Regexpcount, a symbolic package for counting problems on regular expressions and words. *Fundamenta Informaticae*, 56:71–88, July 2003.

[25] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 287(2):593–617, 2002.

[26] J. Noonan and D. Zeilberger. The Goulden-Jackson cluster method: extensions, applications and implementations. *Journal of Difference Equations and Applications*, 5(4-5):355–377, 1999.

[27] G. Park, H.-K. Hwang, P. Nicodème, and W. Szpankowski. Profiles of tries. *SIAM Journal on Computing*, 38(5):1821–1880, 2009.

[28] B. Prum, F. Rodolphe, and E. de Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):205–220, 1995.

[29] M. Régnier and A. Denise. Rare events and conditional events on random strings. *Discret. Math. Theor. Comput. Sci.*, 6:191–214, 2004.

[30] M. Régnier and W. Szpankowski. On the approximate pattern occurrences in a text. *Proceedings. Compression and Complexity of SEQUENCES 1997*, pages 253–264, June 1997.

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

[32] B. Vallée. Dynamical sources in information theory: Fundamental intervals and word prefixes. *Algorithmica*, 29:262–306, February 2001.

[33] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, IT-23(3):337–343, May 1977.

A. FULL EXPRESSIONS

Expressions from Example 6.3.1 are provided in this appendix.

Let ${}^a\eta(z,\mathbf{t}) = \Big({}^a\eta_1(z,\mathbf{t}),\ \ {}^a\eta_2(z,\mathbf{t}),\ \ {}^a\eta_3(z,\mathbf{t}),\ \ {}^a\eta_4(z,\mathbf{t}) \Big)$. Then we have:

$$
\begin{aligned}
{}^a\eta_1(z,\mathbf{t}) = {} & \Big[t_1 z^2\Big(-50 + 3(1+t_1)z^2(5(1+t_2)(t_3+t_4+t_3t_4) \\
& \quad - 3(t_2t_3 + (1+t_2)(1+t_3)t_4)z)\Big)\Big] \Big/ \\
& \Big[2\Big(-100 + 30(1+t_1)(t_3+t_4+t_3t_4+t_2(1+t_3)(1+t_4))z^2 \\
& \quad + 9(1+t_1)(-t_2t_3 + t_1(1+t_2)(1+t_3)t_4)z^4\Big)\Big] \\
{}^a\eta_2(z,\mathbf{t}) = {} & -\Big[3(1+t_1)t_2z^3\Big(50 + 3z\Big(10t_3 - 5(t_3 + (1+t_1)(1+t_2)(1+t_3)t_4)z \\
& \quad + 3(1+t_1)(1+t_2)(1+t_3)t_4z^2\Big)\Big)\Big] \Big/ \\
& \Big[10\Big(-100 + 30(1+t_1)(t_3+t_4+t_3t_4+t_2(1+t_3)(1+t_4))z^2 \\
& \quad + 9(1+t_1)(-t_2t_3 + t_1(1+t_2)(1+t_3)t_4)z^4\Big)\Big] \\
{}^a\eta_3(z,\mathbf{t}) = {} & \Big[3(1+t_1)t_3z^3\Big(-10 - 5(t_1+t_2+t_1t_2)z + 3(1+t_1)t_2z^2\Big)\Big] \Big/ \\
& \Big[-200 + 6(1+t_1)z^2\Big(10(t_3+t_4+t_3t_4+t_2(1+t_3)(1+t_4)) \\
& \quad + 3(-t_2t_3 + t_1(1+t_2)(1+t_3)t_4)z^2\Big)\Big] \\
{}^a\eta_4(z,\mathbf{t}) = {} & \Big[9(1+t_1)(1+t_2)(1+t_3)t_4z^4\Big(-10 - 5(t_1+t_2+t_1t_2)z + 3(1+t_1)t_2z^2\Big)\Big] \Big/ \\
& \Big[-1000 + 30(1+t_1)z^2\Big(10(t_3+t_4+t_3t_4+t_2(1+t_3)(1+t_4)) \\
& \quad + 3(-t_2t_3 + t_1(1+t_2)(1+t_3)t_4)z^2\Big)\Big]
\end{aligned}
$$

Let ${}^b\eta(z,\mathbf{t}) = \Big(\ {}^b\eta_1(z,\mathbf{t}),\ \ {}^b\eta_2(z,\mathbf{t}),\ \ {}^b\eta_3(z,\mathbf{t}),\ \ {}^b\eta_4(z,\mathbf{t})\ \Big)$. Then we have:

$$
\begin{aligned}
{}^b\eta_1(z,\mathbf{t}) = -\Big[&3t_1z^2\Big(50+3(1+t_1)z^2(-5(1+t_2)(t_3+t_4+t_3t_4)\\
&\quad+2(t_2t_3+(1+t_2)(1+t_3)t_4)z)\Big)\Big]\Big/\\
&\Big[5\Big(-100+30(1+t_1)(t_3+t_4+t_3t_4+t_2(1+t_3)(1+t_4))z^2\\
&\quad+9(1+t_1)(-t_2t_3+t_1(1+t_2)(1+t_3)t_4)z^4\Big)\Big]\\
{}^b\eta_2(z,\mathbf{t}) = -\Big[&9(1+t_1)t_2z^3\Big(50+z\Big(20t_3-15(t_3+(1+t_1)(1+t_2)(1+t_3)t_4)z\\
&\quad+6(1+t_1)(1+t_2)(1+t_3)t_4z^2\Big)\Big)\Big]\Big/\\
&\Big[25\Big(-100+30(1+t_1)(t_3+t_4+t_3t_4+t_2(1+t_3)(1+t_4))z^2\\
&\quad+9(1+t_1)(-t_2t_3+t_1(1+t_2)(1+t_3)t_4)z^4\Big)\Big]\\
{}^b\eta_3(z,\mathbf{t}) = \Big[&3(1+t_1)t_3z^3\Big(-20-15(t_1+t_2+t_1t_2)z+6(1+t_1)t_2z^2\Big)\Big]\Big/\\
&\Big[-500+15(1+t_1)z^2\Big(10(t_3+t_4+t_3t_4+t_2(1+t_3)(1+t_4))\\
&\quad+3(-t_2t_3+t_1(1+t_2)(1+t_3)t_4)z^2\Big)\Big]\\
{}^b\eta_4(z,\mathbf{t}) = \Big[&9(1+t_1)(1+t_2)(1+t_3)t_4z^4\Big(-20-15(t_1+t_2+t_1t_2)z+6(1+t_1)t_2z^2\Big)\Big]\Big/\\
&\Big[25\Big(-100+30(1+t_1)(t_3+t_4+t_3t_4+t_2(1+t_3)(1+t_4))z^2\\
&\quad+9(1+t_1)(-t_2t_3+t_1(1+t_2)(1+t_3)t_4)z^4\Big)\Big]
\end{aligned}
$$

The full expressions of ${}^aT(z,\mathbf{t})$ and ${}^bT(z,\mathbf{t})$ in Example 6.3.1 are respectively listed as follows.

$$
\begin{aligned}
{}^aT(z,\mathbf{t}) = \Big[& -200 + z\Big(-20 + 10(t_1 + 6t_2 + 6t_1t_2 + 6t_3 + 6t_1t_3 + 6t_2t_3 + 6t_1t_2t_3 \\
& + 6(1+t_1)(1+t_2)(1+t_3)t_4)z + 6(1+t_1)(1+t_3)(t_2 + t_4 + t_2t_4)z^2 \\
& + 15(1+t_1)(-t_2t_3 + t_1(1+t_2)(1+t_3)t_4)z^3\Big)\Big] \Big/ \\
& \Big[2\Big(-100 + z\Big(90 + 10(1 + 3t_3 + 3(1+t_3)t_4 + 3t_2(1+t_3)(1+t_4) \\
& + 3t_1(1+t_2)(1+t_3)(1+t_4))z \\
& - 3(1+t_1)(4t_2 + 5t_3 + 9t_2t_3 + 9(1+t_2)(1+t_3)t_4)z^2 \\
& + 6(1+t_1)(t_2t_3 + (1+t_2)(1+t_3)t_4)z^3\Big)\Big)\Big] \\
{}^bT(z,\mathbf{t}) = \Big[& -500 + z\Big(-50 + 3(1+t_1)z\Big(50(t_3 + t_4 + t_3t_4 + t_2(1+t_3)(1+t_4)) \\
& + 5(1+t_2)(t_3 + t_4 + t_3t_4)z \\
& + 3(-4t_2t_3 + (1+5t_1)(1+t_2)(1+t_3)t_4)z^2\Big)\Big)\Big] \Big/ \\
& \Big[5\Big(-100 + z\Big(90 + 10(1 + 3t_3 + 3(1+t_3)t_4 \\
& + 3t_2(1+t_3)(1+t_4) + 3t_1(1+t_2)(1+t_3)(1+t_4))z \\
& - 3(1+t_1)(4t_2 + 5t_3 + 9t_2t_3 + 9(1+t_2)(1+t_3)t_4)z^2 \\
& + 6(1+t_1)(t_2t_3 + (1+t_2)(1+t_3)t_4)z^3\Big)\Big)\Big]
\end{aligned}
$$

The full expressions of ${}^{a}F(z,\mathbf{x})$ and ${}^{b}F(z,\mathbf{x})$ in Example 6.3.1 are respectively listed as follows.

$$
\begin{aligned}
{}^{a}F(z,\mathbf{x}) = &\Big[-200+z\Big(-20+z\Big(-10+x_1\Big(-50+60x_2x_3x_4+6x_3(-1+x_2x_4)z\\
&\quad+15(-1+x_2+x_3-x_1x_2x_3+(-1+x_1)x_2x_3x_4)z^2\Big)\Big)\Big)\Big]\Big/\\
&\Big[-200+2z(90+z(-20+6x_1z(-x_3(-2+z)+z)\\
&\quad+3x_1x_2(x_3x_4(-2+z)-z)(-5+2z)))\Big]\\
{}^{b}F(z,\mathbf{x}) = &\Big[-500+z\Big(-50+3x_1z\Big(-50+50x_2x_3x_4+5x_2(-1+x_3x_4)z\\
&\quad+3(4(-1+x_3)+x_2(4+5x_1x_3(-1+x_4)-4x_3x_4))z^2\Big)\Big)\Big]\Big/\\
&\Big[-500+5z(90+z(-20+6x_1z(-x_3(-2+z)+z)\\
&\quad+3x_1x_2(x_3x_4(-2+z)-z)(-5+2z)))\Big]
\end{aligned}
$$

B. CALCULATIONS

Here is a detailed calculation of Equation (7.41) in Section 7.4.2.

The following computation used the results from (7.10), (7.12), (7.17), (7.22), (7.24), and (7.26).

$$\begin{aligned}
\sum_{n=0}^{\infty} \mathbf{E}\left(X_n\right) \cdot z^n &= \left.\frac{\partial F(z,x,y)}{\partial x}\right|_{x=y=1} \\
&= \left.\frac{1}{\left(1-z-\Upsilon(z,0,0,0)\right)^2} \cdot \frac{\partial \Upsilon(z,x-1,y-1,xy-1)}{\partial x}\right|_{x=y=1} \\
&= \frac{1}{(1-z)^2} \cdot \left(\Upsilon_1(z)+\Upsilon_3(z)\right) \\
&= \frac{1}{(1-z)^2} \cdot \Big(\Upsilon_1(1)+\Upsilon_3(1)-(1-z)\cdot\left(\Upsilon_1'(1)+\Upsilon_3'(1)\right)+o(1-z)\Big) \qquad \text{(B.1)} \\
&= \frac{\Upsilon_1(1)+\Upsilon_3(1)}{(1-z)^2} - \frac{\Upsilon_1'(1)+\Upsilon_3'(1)}{(1-z)} \\
&= \sum_{n=0}^{\infty}(n+1)\cdot z^n \cdot \sum_{u\in\mathcal{U}} \pi\left(u\right) - \sum_{n=0}^{\infty} z^n \sum_{u\in\mathcal{U}} \pi\left(u\right)\cdot|u| \\
&= \sum_{n=0}^{\infty} z^n \cdot \left(\sum_{u\in\mathcal{U}} \pi(u)\cdot(n+1-|u|)\right)
\end{aligned}$$

Here is a detailed calculation of Equation (7.44) in Section 7.4.3.

First we provide the detailed calculation of Equation (7.44). The following equations in Section 7.4.1 are used: (7.10), (7.14), (7.17)–(7.19), (7.29)–(7.31), (7.32), (7.34)–(7.36), (7.38), and (7.39). We know that $\sum_{n=0}^{\infty} \mathbf{E}\left(X_n Y_n\right)\cdot z^n = \left.\frac{\partial^2 F(z,x,y)}{\partial x \partial y}\right|_{x=y=1}$. It follows that

$$\begin{aligned}
\sum_{n=0}^{\infty} \mathbf{E}\left(X_n Y_n\right)\cdot z^n = 2\cdot &\left.\frac{1}{\left(1-z-\Upsilon(z,0,0,0)\right)^3}\cdot\frac{\partial\Upsilon}{\partial y}\cdot\frac{\partial\Upsilon}{\partial x}\right|_{x=y=1} \\
&+ \left.\frac{1}{\left(1-z-\Upsilon(z,0,0,0)\right)^2}\cdot\frac{\partial^2\Upsilon}{\partial x\partial y}\right|_{x=y=1}
\end{aligned}$$

and now we calculate

$$\begin{aligned}
\sum_{n=0}^{\infty} \mathbf{E}\left(X_n Y_n\right) \cdot z^n &= 2 \cdot \frac{1}{(1-z)^3} \cdot \frac{\partial \Upsilon}{\partial y} \cdot \frac{\partial \Upsilon}{\partial x}\bigg|_{x=y=1} + \frac{1}{(1-z)^2} \cdot \frac{\partial^2 \Upsilon}{\partial x \partial y}\bigg|_{x=y=1} \\
&= \frac{2 \cdot (\Upsilon_1(z) + \Upsilon_3(z)) \cdot (\Upsilon_2(z) + \Upsilon_3(z))}{(1-z)^3} \\
&\quad + \frac{\Upsilon_{12}(z) + \Upsilon_{13}(z) + \Upsilon_{23}(z) + \Upsilon_{33}(z) + \Upsilon_3(z)}{(1-z)^2} \\
&= \frac{2}{(1-z)^3} \cdot \Big(\Upsilon_1(1) + \Upsilon_3(1) - (1-z) \cdot (\Upsilon_1'(1) + \Upsilon_3'(1)) + o(1-z)\Big) \\
&\qquad\qquad \cdot \Big(\Upsilon_2(1) + \Upsilon_3(1) - (1-z) \cdot (\Upsilon_2'(1) + \Upsilon_3'(1)) + o(1-z)\Big) \\
&\quad + \frac{1}{(1-z)^2} \cdot \Big(\Big(\Upsilon_{12}(1) + \Upsilon_{13}(1) + \Upsilon_{23}(1) + \Upsilon_{33}(1) + \Upsilon_3(1)\Big) \\
&\qquad - (1-z) \cdot \Big(\Upsilon_{12}'(1) + \Upsilon_{13}'(1) + \Upsilon_{23}'(1) + \Upsilon_{33}'(1) + \Upsilon_3'(1)\Big) + o(1-z)\Big) \\
&= \frac{2}{(1-z)^3} \cdot \Big(\Upsilon_1(1) + \Upsilon_3(1)\Big) \cdot \Big(\Upsilon_2(1) + \Upsilon_3(1)\Big) \\
&\quad - \frac{2}{(1-z)^2} \cdot \Big((\Upsilon_1(1) + \Upsilon_3(1))(\Upsilon_2'(1) + \Upsilon_3'(1)) \\
&\qquad + (\Upsilon_2(1) + \Upsilon_3(1))(\Upsilon_1'(1) + \Upsilon_3'(1))\Big) \\
&\quad + \frac{2}{(1-z)} \cdot (\Upsilon_1'(1) + \Upsilon_3'(1)) \cdot (\Upsilon_2'(1) + \Upsilon_3'(1)) \\
&\quad + \frac{1}{(1-z)^2} \cdot \Big(\Upsilon_{12}(1) + \Upsilon_{13}(1) + \Upsilon_{23}(1) + \Upsilon_{33}(1) + \Upsilon_3(1)\Big) \\
&\quad - \frac{1}{(1-z)} \cdot \Big(\Upsilon_{12}'(1) + \Upsilon_{13}'(1) + \Upsilon_{23}'(1) + \Upsilon_{33}'(1) + \Upsilon_3'(1)\Big) \\
&= \sum_{n=0}^{\infty} z^n \cdot (n+1)(n+2)\Big(\sum_{u \in \mathcal{U}} \pi(u)\Big)\Big(\sum_{v \in \mathcal{V}} \pi(v)\Big) \\
&\quad - \sum_{n=0}^{\infty} z^n \cdot 2 \cdot (n+1)\Big(\Big(\sum_{u \in \mathcal{U}} \pi(u)\Big)\Big(\sum_{v \in \mathcal{V}} \pi(v) \cdot |v|\Big) \\
&\qquad + \Big(\sum_{v \in \mathcal{V}} \pi(v)\Big)\Big(\sum_{u \in \mathcal{U}} \pi(u) \cdot |u|\Big)\Big) \\
&\quad + \sum_{n=0}^{\infty} z^n \cdot 2 \cdot \Big(\sum_{u \in \mathcal{U}} \pi(u) \cdot |u|\Big)\Big(\sum_{v \in \mathcal{V}} \pi(v) \cdot |v|\Big) \\
&\quad + \sum_{n=0}^{\infty} z^n \cdot (n+1) \cdot \Big(\Upsilon_{12}(1) + \Upsilon_{13}(1) + \Upsilon_{23}(1) + \Upsilon_{33}(1) + \Upsilon_3(1)\Big) \\
&\quad - \sum_{n=0}^{\infty} z^n \cdot \Big(\Upsilon_{12}'(1) + \Upsilon_{13}'(1) + \Upsilon_{23}'(1) + \Upsilon_{33}'(1) + \Upsilon_3'(1)\Big)
\end{aligned}$$

VITA

Yucong Zhang was born in Changsha, China. He is the only child of Gang Zhang and Xiaogui Zhang. After graduating from First High School of Changsha, He enrolled at Fudan University in Shanghai and earned a Bachelor of Science degree in Physics in 2011. He joined Purdue University in 2013 in the Ph.D. program of Physics. Apart from the coursework in Physics, he also took courses in Statistics and started to read analytic combinatorics. He was greatly motivated and decided to do his Ph.D. research in Statistics with Professor Mark Daniel Ward. In 2015, he graduated with a Master of Science degree in Physics and started his Ph.D. in Statistics at Purdue. He was a lecturer and teaching assistant for several courses, including elementary statistical methods (STAT 301) for two years, applied regression analysis (STAT 512), methods of theoretical physics (PHYS 600), thermal and statistical physics (honors) (PHYS 416), and general physics (PHYS 220). He has worked as a research assistant with the Center for Science of information. This is an interdisciplinary initiative from NSF (grant CCF-0939370). As a student in the Center, he established and maintained online courses of Data Science and Probability, which already served thousands of students worldwide. He will finish his Ph.D. in December 2020. Throughout his Ph.D. years he enjoys working with Dr. Ward very much, and appreciates all the support from his mentor and friend.