# ASYMPTOTIC ANALYSIS OF THE $k$th SUBWORD COMPLEXITY

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Lida Ahmadi

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

# THE PURDUE UNIVERSITY GRADUATE SCHOOL
# STATEMENT OF DISSERTATION APPROVAL

Dr. Mark Daniel Ward, Chair

Department of Statistics

Dr. Steve Bell

Department of Mathematics

Dr. Wojciech Szpankowski

Department of Computer Science

Dr. Nung Kwan Aaron Yip

Department of Mathematics

**Approved by:**

Dr. David Goldberg

Associate Head for Graduate Studies

**To my mom, Soheila:**

I love you and I appreciate all you have done for me throughout my life. You are a symbol of what a strong woman should be.

**To my sister, Aida:**

You have been my inspiration for wanting to learn and grow. I love you, and thank you for always being there for me.

# ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my PhD supervisor, Dr. Mark Daniel Ward, who has guided me through my graduate studies and has been exceptionally supportive of my work and endeavors. Your understanding and encouragement kept me motivated and excited about my graduate research.

My appreciation extends to my PhD committee, Dr. Steve Bell, Dr. Wojciech Szpankowski, and Dr. Aaron Yip: Thank you Dr. Bell for guiding me throughout the first two years of my Ph.D., and for always being willing to help me out. Thank you Dr. Szpankowski for the insightful comments during our discussions about my thesis. Thank you Dr. Yip for teaching one of the most enjoyable classes I have taken at Purdue.

Lastly, a heartfelt thank you to my husband, Vlad, for his love and support. I appreciate your patience and assistance through the stressful times. With your encouraging words and reminders to stay positive you made this path easier for me.

## TABLE OF CONTENTS

Page

# ABSTRACT

Ahmadi Lida Ph.D., Purdue University, August 2019. Asymptotic Analysis of the $k$th Subword Complexity. Major Professor: Dr. Mark Daniel Ward.

The Subword Complexity of a character string refers to the number of distinct substrings of any length that occur as contiguous patterns in the string. The $k$th Subword Complexity in particular, refers to the number of distinct substrings of length $k$ in a string of length $n$. In this work, we evaluate the expected value and the second factorial moment of the $k$th Subword Complexity for the binary strings over memoryless sources. We first take a combinatorial approach to derive a probability generating function for the number of occurrences of patterns in strings of finite length. This enables us to have an exact expression for the two moments in terms of patterns' auto-correlation and correlation polynomials. We then investigate the asymptotic behavior for values of $k = \Theta(\log n)$. In the proof, we compare the distribution of the $k$th Subword Complexity of binary strings to the distribution of distinct prefixes of independent strings stored in a trie. The methodology that we use involves complex analysis, analytical poissonization and depoissonization, the Mellin transform, and saddle point analysis.

# 1. INTRODUCTION

Analyzing and understanding occurrences of patterns in a character string is helpful for extracting useful information regarding the nature of a string. We classify strings to low complexity and high complexity, according to their level of randomness. For instance, we take the binary string $X = 10101010...$, which is constructed by repetitions of the pattern $w = 10$. This string is periodic, and therefore has low randomness. Such periodic strings are classified as low-complexity strings, whereas strings that do not show periodicity are considered to have high complexity. An effective way of measuring a string's randomness is to count all distinct patterns that appear as contiguous subwords in the string. This value is called the Subword Complexity. The name is given by Ehrenfeucht, Lee, and Rozenberg [1], and initially was introduced by Morse and Hedlund in 1938 [2]. The higher the Subword Complexity, the more complex the string is considered to be.

Assessing information about the distribution of the Subword Complexity enables us to better characterize strings, and determine atypically random or periodic strings that have complexities far from the average complexity [3]. This type of string classification has applications in fields such as data compression [4], genome analysis (see [5], [6], [7], [8], and [9]), and plagiarism detection [10]. For example, in data compression, a data set is considered compressible if it has low complexity, since it consists of repeated subwords. In computational genomics, Subword Complexity (known as the number of $k$-mers) is used in detection of repeated sequences and DNA barcoding [11], [12].

There are two variations for the definition of the Subword Complexity: The one that counts all distinct subwords of a given string (also known as Complexity Index, and Sequence Complexity [13]), and the one that only counts the subwords of the same length, say $k$, that appear in the string. In our work, we analyze the latter, and we call it the $k$th Subword Complexity to avoid any confusion.

## 1.1 Thesis Statement

Throughout this thesis, we consider the $k$th Subword Complexity of a random binary string of length $n$ over a memory-less source, and we denote it by $X_{n,k}$. We analyze the first and second factorial moments of $X_{n,k}$ for the range $k = \Theta(\log n)$, as $n \to \infty$. More precisely, we are interested in the range $k = a \log n$, for the following intervals of $a$

*i.* $\dfrac{1}{\log q^{-1}} < a < \dfrac{2}{\log q^{-1} + \log p^{-1}}$,

*ii.* $\dfrac{2}{\log q^{-1} + \log p^{-1}} < a < \dfrac{1}{q \log q^{-1} + p \log p^{-1}}$, and

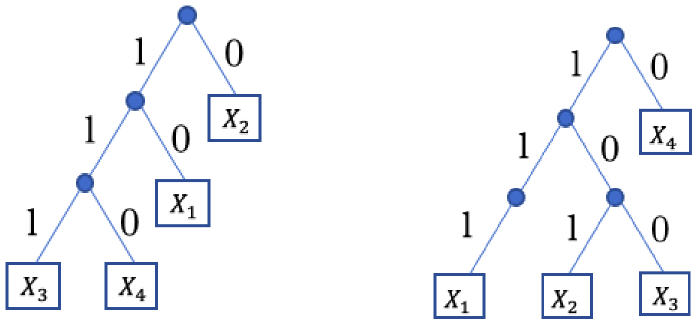
*iii.* $\dfrac{1}{q \log q^{-1} + p \log p^{-1}} < a < \dfrac{1}{\log p^{-1}}$.

## 1.2 Analysis Outline

Our approach involves two major steps. At first we choose a suitable model for the asymptotic analysis, and afterwards we provide proofs for the derivation of the asymptotic expansion of the first two factorial moments.

### 1.2.1 Part I

This part of the analysis is inspired by the earlier work of Jacquet and Szpankowski [14] on the analysis of suffix trees by comparing them to independent tries. A trie, first introduced by René de la Briandais in 1959 (see [15]), is a search tree that stores $n$ strings, according to their prefixes. A suffix tree, introduced by Weiner in 1973 (see [16]), is a trie where the strings are suffixes of a given string. An example of these data structures are given in Figure 1.1.



(a) Suffix Tree

(b) Trie

Figure 1.1. The suffix tree in (a) is built over the first four suffixes of string $X = 101110...$, and the trie in (b) is build over strings $X_1 = 111...$, $X_2 = 101...$, $X_3 = 100$, and $X_4 = 010...$.

A direct asymptotic analysis of the moments is a difficult task, since patterns in a string are not independent from each other. However, we note that each pattern in a string can be regarded as a prefix of a suffix of the string. Therefore, the number of distinct patterns of length $k$ in a string is actually the number of nodes of the suffix tree at level $k$ and lower. It is shown by I. Gheorghiciuc and M. D. Ward [17] that the expected value of the $k$-th Subword Complexity of a Bernoulli string of length $n$ is asymptotically comparable to the expected value of the number of nodes at level $k$ of a trie built over $n$ independent strings generated by a memory-less source.

We extend this analysis to the desired range for $k$, and we prove that the result holds for when $k$ grows logarithmically with $n$. Additionally, we show that asymptotically, the second factorial moment of the $k$-th Subword Complexity can also be estimated by admitting the same independent model generated by a memory-less source. The proof of this theorem heavily relies on the characterization of the overlaps of the patterns with themselves and with one another. Auto-correlation and correlation polynomials explicitly describe these overlaps. The analytic properties of these polynomials are key to understanding repetitions of patterns in large Bernoulli strings. This, in conjunction with Cauchy's integral formula (used to compare the generating functions in the two models) and the residue theorem, provides solid verification that the second factorial moment in the Subword Complexity behaves the same as in the independent model.

To make this comparison, we derive the generating functions of the first two factorial moments in both settings. In a paper published by F. Bassino, J. Clément, and P. Nicodème in 2012 [18], the authors provide a multivariate probability generating function $f(z, x)$ for the number of occurrences of patterns in a finite Bernoulli string. That is, given a pattern $w$, the coefficient of the term $z^n x^m$ in $f(z, x)$ is the probability in the Bernoulli model that a random string of size $n$ has exactly $m$ occurrences of the pattern $w$. Following their technique, we derive the exact expression for the generating functions of the first two factorial moments of the $k$th Subword Complexity. In the independent model, the generating functions are obtained by basic probability concepts.

### 1.2.2 Part II

This part of the proof is analogous to the analysis of profile of tries [19]. To capture the asymptotic behavior, the expressions for the first two factorial moments in the independent trie are further improved by means of a Poisson process. The poissonized version yields generating functions in the form of harmonic sums for each of the moments. The Mellin transform and the inverse Mellin transforms of these harmonic sums establish a connection between the asymptotic expansion and singularities of the transformed function. This methodology is sufficient for when the length $k$ of the patterns are fixed. However, allowing $k$ to grow with $n$, makes the analysis more challenging. This is because for large $k$, the dominant term of the poissonized generating function may come from the term involving $k$, and singularities may not be significant compared to the growth of $k$. This issue is treated by combining the singularity analysis with a saddle point method [20]. The outcome of the analysis is a precise first-order asymptotics of the moments in the poissonized model. Depoissonization theorems are then applied to obtain the desired result in the Bernoulli model. References [21], [22], [13], [23], [24], [25], [26], [27], [28], [29] have been very useful in studying the methodology utilized in this work.

# 2. MAIN RESULTS

We let $X$ be a binary string over a Bernoulli Model; That is, $X = X_1 X_2 ... X_n$, where for $i \in 1, ..., n$, $X_i$'s are independent and identically distributed random variables over the alphabet $\mathcal{A} = \{0, 1\}$. We assume that $\mathbf{P}(X_i = 1) = p$, $\mathbf{P}(X_i = 0) = q = 1 - p$, and $p > q$. We let $X_{n,k}$ denote the the number of distinct patterns of length $k$ that appear as substrings of $X$, called the $k$th Subword Complexity. In this work, we intend to find the average and the second factorial moment of $X_{n,k}$, namely $\mathbf{E}[X_{n,k}]$ and $\mathbf{E}[(X_{n,k})_2]$, asymptotically. We perform the analysis for large $n$, where $k$ grows as a function of $n$. The desired range in our analysis is $k = \Theta(\log n)$. More precisely, $k = a \log n$, in the following ranges for $a$

*i.* $\dfrac{1}{\log q^{-1}} < a < \dfrac{2}{\log q^{-1} + \log p^{-1}}$,

*ii.* $\dfrac{2}{\log q^{-1} + \log p^{-1}} < a < \dfrac{1}{q \log q^{-1} + p \log p^{-1}}$, and

*iii.* $\dfrac{1}{q \log q^{-1} + p \log p^{-1}} < a < \dfrac{1}{\log p^{-1}}$.

The first challenge in our analysis is that patterns in a string are not independent from each other and they overlap one another. This makes the direct analysis of the $k$-th Subword Complexity quite difficult. For this reason, we compare the $k$th Subword Complexity to an independent model constructed in the following way: We store a set of $n$ independently generated strings (by a memory-less source) in a trie. We denote the number of distinct prefixes of length $k$ in the trie by $\hat{X}_{n,k}$, and we call it *the $k$th prefix complexity.* We then show that the average and the second moment of $X_{n,k}$ is asymptotically comparable to $\hat{X}_{n,k}$, when $k = \Theta(\log n)$.

In Chapter 3, we provide a summary of the methodology that we use to derive the generating functions for the average and the second factorial moment of the $k$th

Subword Complexity. The approach is borrowed from the paper by Bassino, Clément, and Nicodème on counting occurrences for a finite set of words (see [18]). In Chapter 4, we utilize the techniques presented in Chapter 3, and we derive a closed form for the desired generating functions in Theorems 4.1.1, and 4.2.1.

In order to proceed with our analysis, we need more information on the analytic properties of the generating functions given in Theorem 4.1.1. We show that for large enough $k$, the polynomials $D_w(z)$ and $D_{w,w'}(z)$ have exactly one root in a disk of radius $\rho$, where $\rho > 1$. The proof regarding $D_w(z)$ is shown by Jacquet and Szpankowski in [30]. In Chapter 5, we provide some lemmas and the proof for the existence of a unique root of $D_{w,w'}(z)$ in a disk of radius $\rho$.

In Chapter 6, in Theorem 6.2.1 we prove that for $k = \Theta(\log n)$, $\mathbf{E}[X_{n,k}]$ and $\mathbf{E}[\hat{X}_{n,k}]$ have the same first order asymptotic growth. In other words, we show that

$$\mathbf{E}[X_{n,k}] - \mathbf{E}[\hat{X}_{n,k}] = O(n^{-M}),$$

where $M$ is a positive real value. The proof involves considering the two generating functions $H_k(z)$ and $\hat{H}_k(z)$ (whose coefficients are $\mathbf{E}[X_{n,k}]$ and $\mathbf{E}[\hat{X}_{n,k}]$, respectively). We apply the Cauchy's coefficient formula to express the coefficient $[z^n](H(z)-\hat{H}(z))$ in precise terms. By a residue analysis, and the application of the Mellin transform, we prove that $[z^n](H(z)-\hat{H}(z))$ tends to zero for large values of $n$ even when $k$ grows logarithmically with $n$. We apply the same methodology for comparing $\mathbf{E}[(X_{n,k})_2]$ and $\mathbf{E}[(\hat{X}_{n,k})_2]$ asymptotically. In Theorem 6.3.1, we prove the following

$$\mathbf{E}[(X_{n,k})_2] - \mathbf{E}[(\hat{X}_{n,k})_2] = O(n^{-\epsilon}),$$

for $\epsilon$ a positive real value.

The final chapter of this dissertation is devoted to the analysis of the first order asymptotics of the average and the second factorial moment of the $k$th Prefix Complexity. The results hold true for the average and the second factorial moment of the

$k$th Subword Complexity as we proved in Chapter 6. The methodology presented in this chapter is analogues to the asymptotic analysis of profile of tries [19]. To achieve complete independence throughout the trie, we embed the Bernoulli model into a Poisson process, where the fixed value $n$ is replaced by a Poisson random variable $N_z$ with mean equal to $z$. We perform all of our analysis in the poissonized model, and we derive the corresponding results in the Bernoulli model by a depoissonization process. The poissonized generating functions for the average and the second factorial moment are respectively

$$\tilde{E}_k(z) = \sum_{w \in \mathcal{A}^k} \left(1 - e^{-z\mathbf{P}(w)}\right),$$

and

$$\tilde{G}_k(z) = \left(\tilde{E}_k(z)\right)^2 - \sum_{w \in \mathcal{A}^k} \left(1 - 2e^{-\mathbf{P}(w)z} + e^{-2\mathbf{P}(w)z}\right).$$

Since we are dealing with the harmonic sums, the Mellin transform is a quite useful technique for deriving the asymptotic expansions for $\tilde{E}_k(z)$ and $\tilde{G}_k(z)$. When estimating the inverse Mellin transform, we obtain integrals that involve $k$ which is a large parameter. For this reason, we compute the integrals through a combination of singularity analysis and the saddle point method. The fundamental strip of the Mellin integral corresponds to

$$\frac{2}{\log p^{-1} + \log q^{-1}} < a < \frac{1}{p \log p^{-1} + q \log q^{-1}},$$

In this range, there is no coalescence between the saddle points and the singularities of the integrals, and therefore we proceed by a saddle point method. For the range

$$\frac{1}{p \log p^{-1} + q \log q^{-1}} < a < \frac{1}{\log p^{-1}},$$

we take into account the dominant pole of the integrand at $s = 0$, as well as a saddle point analysis. And finally, for the range

$$\frac{1}{\log q^{-1}} < a < \frac{2}{\log p^{-1} + \log q^{-1}},$$

The pole at $s = -1$ has the dominant contribution to the asymptotic growth.

We prove the following theorems.

**Theorem 7.3.1** *The average of the kth Prefix Complexity has the following asymptotic expansion*

*i. For a as in* (7.13),

$$\mathbf{E}[\hat{X}_{n,k}] = 2^k - \Phi_1((1+\log p)\log_{p/q} n)\frac{n^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{2.1}$$

*where* $\nu = -r_0 + a\log(p^{-r_0} + q^{-r_0})$, *and*

$$\Phi_1(x) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2\pi}\log p/q}\sum_{j\in\mathbb{Z}}\Gamma(r_0 + it_j)e^{-2\pi ijx}$$

*is a bounded periodic function.*

*ii. For a as in* (7.14),

$$\mathbf{E}[\hat{X}_{n,k}] = \Phi_1((1+\log p)\log_{p/q} n)\frac{n^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right).$$

*iii. For a as in* (7.15)

$$\mathbf{E}[\hat{X}_{n,k}] = n + O(n^{\nu_0}),$$

*for some* $\nu_0 < 1$.

**Theorem 7.4.1** *The second factorial moment of the kth Prefix Complexity has the following asymptotic expansion.*

*i. For a as in* (7.13),

$$\mathbf{E}[(\hat{X}_{n,k})_2] = \left(2^k - \Phi_1(\log_{p/q} n(1+\log p))\frac{n^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right)\right)^2.$$

*ii. For a as in* (7.14),

$$\mathbf{E}[(\hat{X}_{n,k})_2] = \Phi_1^2(\log_{p/q} n(1+\log p))\frac{n^{2\nu}}{\log n}\left(1 + O\left(\frac{1}{\log n}\right)\right).$$

*ii. For $a$ as in* (7.15),

$$\mathbf{E}[(\hat{X}_{n,k})_2] = n^2 + O(n^{2\nu_0}).$$

The periodic function $\Phi_1(x)$ in Theorems 7.3.1, and 7.4.1 is shown in Figures 9.5 and 9.6. For a fixed $p$, the amplitude increases as $r_0$ increases. For a fixed $r_0$, the amplitude tends to 0 as $p \to 1/2^+$.

# 3. PRELIMINARIES

In this chapter, we provide a summary of some of the results that are discussed in [18]. We recall that $X = X_1X_2...X_n$ denotes a binary string, where the characters $\{X_i \,|\, i = 1...n\}$ are independent and identically distributed Bernoulli random variables over the alphabet $\mathcal{A} = \{0, 1\}$. We assume that for each $i$, $\mathbf{P}(X_i = 1) = p$, $\mathbf{P}(X_i = 0) = q = 1-p$, and $p > q$. For a random binary string, we are able to obtain a multivariate generating function that gives the probability that $r$ distinct patterns $u_1, u_2, ..., u_r$ occur exactly $n_1, n_2, .., n_r$ times, respectively, in the string. The analysis for derivation of such generating functions dates back to Régnier and Szpankowski [31], and it relies on understanding the overlaps of patterns with themselves and with each other. Here, we will present the analysis for occurrences of both a single pattern and a pair of patterns $\{w, w'\}$ of the same length. The approach for the case of $r > 2$ patterns can be found in [18].

## 3.1 Word Overlaps: Auto-correlation and Correlation Polynomials

We begin by introducing a few terminologies that describe the self-overlaps in a single pattern $w$ and the overlaps between a pair $\{w, w'\}$. The notations we use in this work are borrowed from Jacquet and Szpankowski [30].

**Definition 3.1.1** *Let $w = w_1...w_k$ be a binary word of length $k$. The auto-correlation set $\mathcal{S}_w$ of the word $w$ is defined in the following way*

$$\mathcal{S}_w = \{w_{i+1}..w_k \,|\, w_1...w_i = w_{k-i+1}..w_k\}. \tag{3.1}$$

*We also define the auto-correlation index set to be*

$$\mathcal{P}(w) = \{i \,|\, w_1...w_i = w_{k-i+1}..w_k\}. \tag{3.2}$$

*Finally, we define the auto-correlation polynomial as*

$$S_w(z) = \sum_{i \in \mathcal{P}(w)} \mathbf{P}(w_{i+1}...w_k) z^{k-i}. \tag{3.3}$$

**Definition 3.1.2** *Let $w = w_1...w_k$ and $w' = w'_1...w'_k$ be two distinct binary words of length $k$. The correlation set $\mathcal{S}_{w,w'}$ of the words $w$ and $w'$ is*

$$\mathcal{S}_{w,w'} = \{w'_{i+1}...w'_k \,|\, w'_1...w'_i = w_{k-i+1}...w_k\}. \tag{3.4}$$

*The correlation index set is defined as*

$$\mathcal{P}(w, w') = \{i \,|\, w'_1...w'_i = w_{k-i+1}..w_k\}. \tag{3.5}$$

*And the correlation polynomial is given as*

$$S_{w,w'}(z) = \sum_{i \in \mathcal{P}(w,w')} \mathbf{P}(w'_{i+1}...w'_k) z^{k-i}. \tag{3.6}$$

## 3.2 The Occurrence PGF for a Single Pattern

We explain the analysis by providing a simple example. Consider the binary string $X = 01101000111101$ and the pattern $w = 11$. We can distinguish occurrences of $w$ in the given string by marking the ending position of each occurrence with a bullet notation. Below are a few ways of distinguishing occurrences of $w$ in the string $X$.

$$X_1 = 01\overset{\bullet}{1}010001\overset{\bullet}{1}\overset{\bullet}{1}\overset{\bullet}{1}01 \qquad X_2 = 011010001\overset{\bullet}{1}\overset{\bullet}{1}\overset{\bullet}{1}01 \qquad X_3 = 01\overset{\bullet}{1}010001\overset{\bullet}{1}1\overset{\bullet}{1}01$$

Paper [18] refers to the above strings as *decorated strings*. We let $\mathcal{X}$ denote the class of all decorated strings. We can easily observe that $\mathcal{X}$ can be written as a sequence of arbitrary letters and decorated words (possibly followed by a sequence of their nontrivial decorated overlaps). In other words, $\mathcal{X} = \mathrm{SEQ}\left(\mathcal{A} + \overset{\bullet}{w} \cdot \mathrm{SEQ}(\mathcal{S}_{\overset{\bullet}{w}} - \epsilon)\right)$.

Note that the bullet in $\overset{\bullet}{w}$ emphasizes that $w$ is a decorated word. This yields the probability generating function

$$F_w(z,t) = \frac{1}{1 - A(z) - \dfrac{t\mathbf{P}(w)z^k}{1 - t(S_w(z) - 1)}}, \tag{3.7}$$

where $z$ marks the length of the string $X$, $t$ marks the number of distinguished occurrences of $w$, and $A(z)$ is the probability generating function for the alphabet. We then apply the substitution $t \to x - 1$ to indicate whether an occurrence is distinguished or not. This way, overcounting of occurrences is prevented and we obtain the probability generating function

$$F_w(z,x-1) = \frac{1}{1 - A(z) - \dfrac{(x-1)\mathbf{P}(w)z^k}{1 - (x-1)(S_w(z) - 1)}}, \tag{3.8}$$

where the coefficient $[z^n x^m]F_w(z,x-1)$ is the probability that a random binary string of length $n$ has $m$ occurrences of the pattern $w$.

## 3.3 The Occurrence PGF for Two Distinct Patterns

We again consider the example provided above. This time, we let $w = 11$ and $w' = 01$. We use the asterisk notation to distinguish occurrences of $w'$, and we recall that the bullet notation is used to distinguish occurrences of $w$. Below are a few examples of distinguishing both patterns in $X$.

$$X_1 = 0\overset{*}{1}\overset{\bullet}{1}0100011\overset{\bullet}{1}\overset{\bullet}{1}\overset{\bullet}{1}01 \qquad X_2 = 0\overset{*}{1}10100011\overset{\bullet}{1}\overset{\bullet}{1}\overset{\bullet}{1}01 \qquad X_3 = 0\overset{*}{1}101000\overset{*}{1}\overset{\bullet}{1}1\overset{\bullet}{1}01$$

In this case, $\mathcal{X}$ can be written as a sequence of arbitrary letters and decorated $w$ and $w'$ (possibly followed by a sequence of their nontrivial decorated self-overlaps or the decorated overlaps between $w$ and $w'$). We can describe the decorated part of the string like the following

$$\mathcal{M} = \begin{pmatrix} \overset{\bullet}{w} & \overset{*}{w}' \end{pmatrix} \mathrm{SEQ}\left( \begin{pmatrix} \mathcal{S}_{\overset{\bullet}{w}-\varepsilon} & \mathcal{S}_{\overset{\bullet}{w},\overset{*}{w}'} \\ \mathcal{S}_{\overset{*}{w}',\overset{\bullet}{w}} & \mathcal{S}_{\overset{*}{w}'} - \varepsilon \end{pmatrix} \right) \begin{pmatrix} \varepsilon \\ \varepsilon \end{pmatrix}. \tag{3.9}$$

Then we have $\mathcal{X} = \mathrm{SEQ}\,(\mathcal{A} + \mathcal{M})$, which results in the generating function below

$$F_{w,w'}(z, t_1, t_2) = \frac{1}{1 - A(z) - M(z, t_1, t_2)}, \tag{3.10}$$

where

$$M(z, t_1, t_2) = \begin{pmatrix} \mathbf{P}(w) z^k t_1 & \mathbf{P}(w') z^k t_2 \end{pmatrix} \left( \mathbb{I} - \begin{pmatrix} (S_w(z) - 1)t_1 & S_{w,w'}(z) t_2 \\ S_{w',w}(z) t_1 & (S_{w'}(z) - 1)t_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

is the generating function for $\mathcal{M}$, $z$ marks the length of the string, $t_1$ marks the number of distinguished occurrences of $w$, and $t_2$ marks the number of distinguished occurrences of $w'$. Like for the singular pattern, we apply the substitutions $t_1 \to x_1 - 1$ and $t_2 \to x_2 - 1$ to avoid overcounting the occurrences of $w$ and $w'$. Then the coefficient $[z^n x_1^{m_1} x_2^{m_2}] F_{w,w'}(z, x_1 - 1, x_2 - 1)$ is the probability that there are $m_1$ occurrences of $w$ and $m_2$ occurrences of $w'$ in a random string of length $n$.

We will utilize the above results to find mathematical expressions for the first two factorial moments of the $k$th Subword Complexity in the next chapter.

# 4. SUBWORD COMPLEXITY VS. PREFIX COMPLEXITY

We recall that in chapter one we defined the $k$th Subword Complexity $X_{n,k}$ of a string $X$ (of length $n$) to be the number of distinct patterns of length $k$ that appear in $X$. Due to overlaps between the subwords of the string $X$, direct asymptotic evaluation of the first and second moments of $X_{n,k}$ is quite complicated. To alleviate this, we note that each subword in a string can be regarded as a prefix of a suffix. Clearly, the suffixes of a string are highly dependent on each other. The analysis will be simplified by comparing to a model with strings that were generated independently. We can show that for same ranges of the parameters, the factorial moments in the original problem are asymptotically comparable to the this new version, where the strings are independent from one another. In the independent model, we first construct a set of $n$ independently generated strings by a memory-less source. We denote the number of distinct prefixes of length $k$ of the strings by $\hat{X}_{n,k}$, and we call it the *$k$th Prefix Complexity*. In chapter six, we will show that for $k = \Theta(\ln n)$, both $\mathbf{E}[X_{n,k}] - \mathbf{E}[\hat{X}_{n,k}]$ and $\mathbf{E}[(X_{n,k})_2] - \mathbf{E}[(\hat{X}_{n,k})_2]$ are asymptotically negligible. A reasonable approach for comparing the moments of the $k$th Subword Complexity to those of the $k$th Prefix Complexity, is to derive the generating function of the first two moments in both settings. Therefore, the focus of this chapter is on the analysis which leads to obtaining a closed form for each of the desired generating functions.

## 4.1 On the $k$th Subword Complexity

The methodology for obtaining the generating functions for $\mathbf{E}[X_{n,k}]$ and $\mathbf{E}[(X_{n,k})_2]$ is centered on the results shown in the previous chapter. We will present these generating functions in the following theorem.

**Theorem 4.1.1** *Let $H_k(z) = \sum_{n\geq 0} \mathbf{E}[X_{n,k}]z^n$ and $G_k(z) = \sum_{n\geq 0} \mathbf{E}[(X_{n,k})_2]z^n$ denote the generating functions for the first and second moments of $X_{n,k}$, $\mathbf{E}[X_{n,k}]$ and $\mathbf{E}[(X_{n,k})_2]$ respectively. We have*

*i.*

$$H_k(z) = \sum_{w\in\mathcal{A}^k} \left( \frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} \right), \tag{4.1}$$

*where $D_w(z) = \mathbf{P}(w)z^k + (1-z)S_w(z)$, and*

*ii.*

$$G_k(z) = \sum_{\substack{w,w^{'}\in\mathcal{A}^k \\ w\neq w'}} \left( \frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} - \frac{S_{w'}(z)}{D_{w'}(z)} + \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)} \right), \tag{4.2}$$

*where*

$$\begin{aligned} D_{w,w'}(z) &= (1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)) \\ &\quad + z^k \left( \mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z)) \right). \end{aligned} \tag{4.3}$$

**Proof** *i.* Let $F_w(z,x)$ denote the occurrence probability generating function for $w$ as in (3.8). Recall that $[z^n x^m]F_w(z,x)$ is the probability that there are exactly $m$ occurrences of $w$ in a randomly generated binary string of length $n$. To compute $\mathbf{E}[X_{n,k}]$, We consider all binary patterns $w$ of length $k$ and define

$$X_{n,k}^{(w)} = \begin{cases} 1 & \text{if } w \text{ appears at least once in string } X \\ 0 & \text{otherwise.} \end{cases}$$

It follows that $X_{n,k} = \sum_{w\in\mathcal{A}^k} X_{n,k}^{(w)}$, and by linearity of expectation, we have

$$\mathbf{E}[X_{n,k}] = \sum_{w\in\mathcal{A}^k} \mathbf{E}[X_{n,k}^{(w)}], \tag{4.4}$$

Using the properties of indicator variables, we have

$$\begin{aligned}\mathbf{E}[X_{n,k}^{(w)}] &= \mathbf{P}(X_{n,k}^{(w)} = 1)\\ &= 1 - P(X_{n,k}^{(w)} = 0)\\ &= 1 - [z^n x^0]F_w(z,x). \qquad (4.5)\end{aligned}$$

Note that $[z^n x^0]F_w(z,x) = [z^n]F_w(z,0)$. We define $f_w(z) = F_w(z,0)$. By (3.8), we obtain

$$f_w(z) = \frac{S_w(z)}{\mathbf{P}(w)z^k + (1-z)S_w(z)}. \qquad (4.6)$$

Therefore, the generating function $H_k(z)$ for the average $k$th Subword Complexity is the following

$$\begin{aligned}H(z) &= \sum_{n\geq 0} \mathbf{E}[X_{n,k}]z^n\\ &= \sum_{n\geq 0}\sum_{w\in\mathcal{A}^k} (1 - [z^n]f_w(z))z^n\\ &= \sum_{w\in\mathcal{A}^k}\left(\frac{1}{1-z} - f_w(z)\right)\\ &= \sum_{w\in\mathcal{A}^k}\left(\frac{1}{1-z} - \frac{S_w(z)}{D_w(z)}\right). \qquad (4.7)\end{aligned}$$

*ii*. To find the generating function for $\mathbf{E}[(X_{n,k})_2]$, we observe that by linearity of expectation

$$\begin{aligned}\mathbf{E}[(X_{n,k})_2] &= \mathbf{E}[X_{n,k}^2] - \mathbf{E}[X_{n,k}]\\ &= \mathbf{E}\left[(X_{n,k}^{(w)} + ... + X_{n,k}^{(w^{(r)})})^2\right] - \mathbf{E}\left[X_{n,k}^{(w)} + ... + X_{n,k}^{(w^{(r)})}\right]\\ &= \sum_{w\in\mathcal{A}^k}\mathbf{E}\left[(X_{n,k}^{(w)})^2\right] + \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\mathbf{E}\left[X_{n,k}^{(w)}X_{n,k}^{(w')}\right] - \sum_{w\in\mathcal{A}^k}\mathbf{E}\left[X_{n,k}^{(w)}\right]\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\mathbf{E}\left[X_{n,k}^{(w)}X_{n,k}^{(w')}\right]. \qquad (4.8)\end{aligned}$$

Since each $X_{n,k}^{(w)}$ is an indicator random variable, we have $\mathbf{E}[(X_{n,k}^{(w)})^2] = \mathbf{E}[X_{n,k}^{(w)}]$, and we get

$$\mathbf{E}[(X_{n,k})_2] = \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}\left[X_{n,k}^{(w)} X_{n,k}^{(w')}\right]. \tag{4.9}$$

To compute $\mathbf{E}[X_{n,k}^{(w)} X_{n,k}^{(w')}]$, we first note that

$$X_{n,k}^{(w)} X_{n,k}^{(w')} = \begin{cases} 1 & \text{if } X_{n,k}^{(w)} = X_{n,k}^{(w')} = 1 \\ 0 & \text{otherwise,} \end{cases}$$

which yields

$$\begin{aligned} \mathbf{E}[X_{n,k}^{(w)} X_{n,k}^{(w')}] &= \mathbf{P}\left(X_{n,k}^{(w)} = 1, X_{n,k}^{(w')} = 1\right) \\ &= 1 - \mathbf{P}\left(X_{n,k}^{(w)} = 0 \cup X_{n,k}^{(w')} = 0\right) \\ &= 1 - \mathbf{P}\left(X_{n,k}^{(w)} = 0\right) - \mathbf{P}\left(X_{n,k}^{(w')} = 0\right) + \mathbf{P}\left(X_{n,k}^{(w)} = 0, X_{n,k}^{(w')} = 0\right). \end{aligned}$$

The above expression gives the following

$$\mathbf{E}[(X_{n,k})_2] = \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left(1 - [z^n]\, f_w(z) - [z^n] f_{w'}(z) + [z^n] f_{ww'}(z)\right), \tag{4.10}$$

where $f_{w,w'}(z) = F_{w,w'}(z, 0, 0)$ and $[z^n]F_{w,w'}(z, 0, 0) = [z^n x_1^0 x_2^0]F_{w,w'}(z, x_1, x_2)$. Following (3.9), and (3.10), we arrive at

$$f_{w,w'}(z) = \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}. \tag{4.11}$$

Finally, we obtain

$$\begin{aligned}
G_k(z) &= \sum_{n\geq 0}\mathbf{E}[(X_{n,k})_2]z^n \\
&= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\sum_{n\geq 0}\left(1-[z^n]f_w(z)-[z^n]f_{w'}(z)+[z^n]f_{w,w'}(z)\right)z^n \\
&= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{1}{1-z}-f_w(z)-f_{w'}(z)+f_{w,w'}(z)\right) \\
&= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{1}{1-z}-\frac{S_w(z)}{D_w(z)}-\frac{S_{w'}(z)}{D_{w'}(z)}+\frac{S_w(z)S_{w'}(z)-S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}\right).
\end{aligned} \tag{4.12}$$

■

## 4.2 On the $k$th Prefix Complexity

We recall that, in the independent model, we construct a set $P$ of $n$ independent strings generated by a memory-less source. Here, we present the generating functions of the first two factorial moments for the $k$th Prefix Complexity. We will see in the proof of the following theorem that the analysis is much simpler than the one given in Theorem 4.1.1. This is due to assuming independence between the strings.

**Theorem 4.2.1** *Let $\hat{H}_k(z) = \sum_{n\geq 0}\mathbf{E}[\hat{X}_{n,k}]z^n$ and $\hat{G}_k(z) = \sum_{n\geq 0}\mathbf{E}[(\hat{X}_{n,k})_2]z^n$ denote the generating functions for $\mathbf{E}[\hat{X}_{n,k}]$ and $\mathbf{E}[(\hat{X}_{n,k})_2]$ respectively. We have*

*i.*

$$\hat{H}_k(z) = \sum_{w\in\mathcal{A}^k}\left(\frac{1}{1-z}-\frac{1}{1-(1-\mathbf{P}(w))z}\right). \tag{4.13}$$

*ii.*

$$\hat{G}_k(z) = \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \left( \frac{1}{1-z} - \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{1}{1-(1-\mathbf{P}(w'))z} \right) + \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z}. \tag{4.14}$$

**Proof** *i*. We begin by defining the following indicator variable.

$$\hat{X}_{n,k}^{(w)} = \begin{cases} 1 & \text{if } w \text{ is a prefix of at least one string in } P \\ 0 & \text{otherwise.} \end{cases}$$

For each $\hat{X}_{n,k}^{(w)}$, we compute that

$$\begin{aligned} \mathbf{E}[\hat{X}_{n,k}^{(w)}] &= \mathbf{P}(\hat{X}_{n,k}^{(w)} = 1) \\ &= 1 - P(\hat{X}_{n,k}^{(w)} = 0) \\ &= 1 - (1-\mathbf{P}(w))^n . \end{aligned} \tag{4.15}$$

We then sum over all words $w$ of length k, and obtain the generating function below

$$\begin{aligned} \hat{H}(z) &= \sum_{n \geq 0} \mathbf{E}[\hat{X}_{n,k}] z^n \\ &= \sum_{w \in \mathcal{A}^k} \left( \frac{1}{1-z} - \frac{1}{1-(1-\mathbf{P}(w))z} \right) . \end{aligned} \tag{4.16}$$

*ii*. Similar to what we did in (4.8) and (4.10), we see that

$$\begin{aligned}\mathbf{E}[(\hat{X}_{n,k})_2] &= \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}[\hat{X}_{n,k}^{(w)} \hat{X}_{n,k}^{(w')}] \\ &= \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \left(1 - (1 - \mathbf{P}(w))^n - (1 - \mathbf{P}(w'))^n + (1 - \mathbf{P}(w) - \mathbf{P}(w'))^n\right),\end{aligned} \tag{4.17}$$

And this yields the following

$$\begin{aligned}\hat{G}(z) &= \sum_{n \geq 0} \mathbf{E}[(\hat{X}_{n,k})_2] z^n \\ &= \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \sum_{n \geq 0} \left(1 - (1 - \mathbf{P}(w))^n - (1 - \mathbf{P}(w'))^n + (1 - \mathbf{P}(w) - \mathbf{P}(w'))^n\right) z^n \\ &= \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \left( \frac{1}{1-z} - \frac{1}{1 - (1 - \mathbf{P}(w))z} - \frac{1}{1 - (1 - \mathbf{P}(w'))z} \right) \\ &\quad + \sum_{\substack{w,w^{'} \in \mathcal{A}^k \\ w \neq w'}} \frac{1}{1 - (1 - \mathbf{P}(w) - \mathbf{P}(w'))z}.\end{aligned} \tag{4.18}$$

■

In Chapter 6, we will compare the coefficients of the generating functions in the two models. We will show that the coefficients are asymptotically equivalent in the specified range for $k$.

# 5. SOME ANALYTIC PROPERTIES

In Theorem 4.1.1, we derived the generating functions for the first two factorial moments of the $k$th Subword Complexity. In this chapter, we will provide some results regarding the roots of $D_w(z)$ and $D_{w,w'}(z)$ seen in Theorem 4.1.1. The polynomial $D_w(z)$ has a term involving the auto-correlation polynomial and $D_{w,w'}(z)$ has terms involving both the auto-correlation and correlation polynomials. For this reason, we begin by stating some lemmas on the analytic properties of $S_w(z)$ (cf. [14]) and $S_{w,w'}(z)$. As stated in chapter two, our work concerns binary strings in the Bernoulli model with $\mathbf{P}(1) = p$, $\mathbf{P}(0) = q$, and $p > q$.

**Lemma 5.0.1 (Jacquet and Szpankowski, 1994)** *For most words $w$, the auto-correlation polynomial $S_w(z)$ is very close to 1, with high probably. More precisely, if $w$ is a binary word of length $k$ and $\delta = \sqrt{p}$, there exists $\rho > 1$, such that $\rho\delta < 1$ and*

$$\sum_{w \in \mathcal{A}^k} [\![ |S_w(\rho) - 1| \leq (\rho\delta)^k \theta ]\!] \mathbf{P}(w) \geq 1 - \theta\delta^k, \tag{5.1}$$

*where $\theta = (1-p)^{-1}$. We use Iverson notation*

$$[\![A]\!] = \begin{cases} 1 & \textit{if } A \textit{ holds} \\ 0 & \textit{otherwise} \end{cases}$$

**Proof** We follow the method of proof given in [32]. If the minimal degree of $S_w(z) - 1$ is greater than $\lfloor k/2 \rfloor$, we have

$$|S_w(\rho) - 1| \leq \sum_{i > \lfloor k/2 \rfloor} (\rho p)^i \leq \rho^k \frac{p^{k/2}}{1-p}. \tag{5.2}$$

we define $\delta = \sqrt{p}$ , $\theta = (1-p)^{-1}$, then $[\![|S_w(\rho)-1| \leq (\rho\delta)^k\theta\,]\!] = 1$ for those $S_w(z)-1$ with minimal degree greater than $\lfloor k/2 \rfloor$. Thus, we simplify the problem to finding a lower bound for

$$\sum_{w\in\mathcal{A}^k} [\![S_w(z)-1 \text{ has minimal degree} > \lfloor k/2\rfloor]\!]\mathbf{P}(w).$$

First we note that, for $w = w_1...w_iw_{i+1}...w_k$, we have

$$\begin{aligned}
&\sum_{w\in\mathcal{A}^k} [\![S_w(z)-1 \text{ has minimal degree} \leq \lfloor k/2\rfloor\,]\!]\mathbf{P}(w)\\
&\qquad = \sum_{i=1}^{\lfloor k/2\rfloor}\sum_{w\in\mathcal{A}^k} [\![S_w(z)-1\text{has minimal degree} = i\,]\!]\mathbf{P}(w)\\
&\qquad = \sum_{i=1}^{\lfloor k/2\rfloor}\sum_{w_1...w_i\in\mathcal{A}^i} \mathbf{P}(w_1...w_i)\\
&\qquad\qquad \sum_{w_{i+1}...w_k\in\mathcal{A}^{k-i}} [\![S_w(z)-1 \text{ has minimal degree} = i\,]\!]\mathbf{P}(w_{i+1}...w_k)\\
&\qquad \leq \sum_{i=1}^{\lfloor k/2\rfloor}\sum_{w_1..w_i\in\mathcal{A}^i} \mathbf{P}(w_1...w_i)p^{k-i}\\
&\qquad = \sum_{i=1}^{\lfloor k/2\rfloor} p^{k-i}\sum_{w_1...w_i\in\mathcal{A}^i} \mathbf{P}(w_1...w_i)\\
&\qquad = \sum_{i=1}^{\lfloor k/2\rfloor} p^{k-i}\\
&\leq \frac{p^{k-\lfloor k/2\rfloor}}{1-p}.
\end{aligned} \tag{5.3}$$

And this yields the following

$$\begin{aligned}
&\sum_{w\in\mathcal{A}^k} [\![S_w(z)-1 \text{ has minimal degree} > \lfloor k/2\rfloor]\!]\mathbf{P}(w)\\
&\qquad = 1 - \sum_{w\in\mathcal{A}^k} [\![S_w(z)-1 \text{ has minimal degree} \leq \lfloor k/2\rfloor]\!]\mathbf{P}(w)\\
&\qquad \geq 1 - \frac{p^{\lceil k/2\rceil}}{1-p}\\
&\qquad \geq 1-\theta\delta^k.
\end{aligned} \tag{5.4}$$

■

**Lemma 5.0.2 (Jacquet and Szpankowski, 1994)** *There exist $K > 0$ and $\rho > 1$, such that $p\rho < 1$, and for every binary word $w$ with length $k \geq K$ and $|z| \leq \rho$, we have*

$$|S_w(z)| > 0. \tag{5.5}$$

*In other words, $S_w(z)$ does not have any roots in $|z| \leq \rho$.*

**Proof** We follow the method of proof and notations from [30]. We let $i$ denote the minimal degree of $S_w(z) - 1$, and consider the following two cases.

Case $i$. If $i > \lfloor k/2 \rfloor$, then there exists $K_1 > 0$, such that, for $w$ of length $k \geq K_1$ and $|z| \leq \rho$, we have

$$|S_w(z)| \geq 1 - \left| \sum_{j=i}^{k-1} \mathbf{P}(w_{j+1}...w_k) z^{k-j} \right| \geq 1 - \frac{(p\rho)^i}{1 - p\rho} \geq 1 - \frac{(p\rho)^{k/2}}{1 - p\rho} > 0. \tag{5.6}$$

Case $ii$. If $i \leq \lfloor k/2 \rfloor$, we define $q = \lfloor k/i \rfloor$. Then $w = u^q v$ where $u$ is a prefix of length $i$ of word $w$. Thus

$$\begin{aligned} S_w(z) &= 1 + \mathbf{P}(u)z^i + \mathbf{P}(u^2)(z^i)^2 + ... + \mathbf{P}(u^{q-1})(z^i)^{q-1} S_{uv}(z) \\ &= 1 + \mathbf{P}(u)z^i + \big(\mathbf{P}(u)z^i\big)^2 + ... + \big(\mathbf{P}(u)z^i\big)^{q-1} S_{uv}(z), \end{aligned} \tag{5.7}$$

where the second equality follows by the independence assumption of the probability metric $\mathbf{P}$. Therefore, there exists $K_2 > 0$, such that for $w$ of length $k$, with $k > q(i-1) \geq K_2$ and $|z| \leq \rho$, we have

$$\begin{aligned} |S_w(z)| &\geq \left| 1 + \mathbf{P}(u)z^i + \big(\mathbf{P}(u)z^i\big)^2 + ... + \big(\mathbf{P}(u)z^i\big)^{q-2} \right| - \left| \big(\mathbf{P}(u)z^i\big)^{q-1} S_{uv}(z) \right| \\ &\geq \frac{1 - (p\rho)^{(q-1)i}}{1 + (p\rho)^i} - (p\rho)^{(q-1)i} \cdot \frac{1}{1 - p\rho} > 0. \end{aligned} \tag{5.8}$$

We complete the proof by setting $K = \max\{K_1, K_2\}$, so that the assumptions of cases $i$ and $ii$ are both satisfied. ■

In a similar manner, we present Lemmas 5.0.3 and 5.0.4 below.

**Lemma 5.0.3** *With high probability, for most distinct pairs $\{w, w'\}$, the correlation polynomial $S_{w,w'}(z)$ is very close to 0. More precisely, if $w$ and $w'$ are two distinct binary words of length $k$ and $\delta = \sqrt{p}$, there exists $\rho > 1$, such that $\rho\delta < 1$ and*

$$\sum_{w \in \mathcal{A}^k} [\![ |S_{w,w'}(\rho)| \leq (\rho\delta)^k \theta ]\!] \mathbf{P}(w) \geq 1 - \theta\delta^k \tag{5.9}$$

**Proof** The proof is similar to Lemma 5.0.1. If the minimal degree of $S_{w,w'}(z)$ is greater than $> \lfloor k/2 \rfloor$, then

$$|S_{w,w'}(\rho)| \leq (\rho\delta)^k \theta. \tag{5.10}$$

for $\theta = (1-p)^{-1}$. For a fixed $w'$, we have

$$\begin{aligned}
&\sum_{w \in \mathcal{A}^k} [\![ S_{w,w'}(z) \text{ has minimal degree} \leq \lfloor k/2 \rfloor ]\!] \, \mathbf{P}(w) \\
&= \sum_{i=1}^{\lfloor k/2 \rfloor} \sum_{w \in \mathcal{A}^k} [\![ S_{w,w'}(z) \text{ has minimal degree } = i ]\!] \, \mathbf{P}(w) \\
&= \sum_{i=1}^{\lfloor k/2 \rfloor} \sum_{w_1 \ldots w_i \in \mathcal{A}^i} \mathbf{P}(w_1 \ldots w_i) \\
&\qquad \sum_{w_{i+1} \ldots w_k \in \mathcal{A}^{k-i}} [\![ S_{w,w'}(z) \text{ has minimal degree } = i ]\!] \mathbf{P}(w_{i+1} \ldots w_k) \\
&\leq \sum_{i=1}^{\lfloor k/2 \rfloor} \sum_{w_1 .. w_i \in \mathcal{A}^i} \mathbf{P}(w_{i+1} \ldots w_k) p^{k-i} \\
&= \sum_{i=1}^{\lfloor k/2 \rfloor} p^{k-i} \sum_{w_1 .. w_i \in \mathcal{A}^i} \mathbf{P}(w_1 \ldots w_i) \\
&= \sum_{i=1}^{\lfloor k/2 \rfloor} p^{k-i} \leq \frac{p^{k - \lfloor k/2 \rfloor}}{1-p}.
\end{aligned} \tag{5.11}$$

And this leads to the following

$$\begin{aligned}
&\sum_{w \in \mathcal{A}^k} [\![ \text{ every term of } S_{w,w'}(z) \text{ is of degree} > \lfloor k/2 \rfloor ]\!] \mathbf{P}(w) \\
&= 1 - \sum_{w \in \mathcal{A}^k} [\![ S_{w,w'}(z) \text{ has a term of degree} \leq \lfloor k/2 \rfloor ]\!] \mathbf{P}(w) \\
&\geq 1 - \frac{p^{\lceil k/2 \rceil}}{1-p} \geq 1 - \theta\delta^k.
\end{aligned} \tag{5.12}$$

■

**Lemma 5.0.4** *There exist $K' > 0$, and $\rho > 1$ such that $p\rho < 1$, and such that, for every pair of distinct words $w$, and $w'$ of length $k \geq K'$, and for $|z| \leq \rho$, we have*

$$|S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)| > 0. \tag{5.13}$$

*In other words, $S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)$ does not have any roots in $|z| \leq \rho$.*

**Proof** We consider the three following cases.
Case *i*. When either $S_w(z) = 1$ or $S_{w'}(z) = 1$, then every term of $S_{w,w'}(z)S_{w',w}(z)$ has degree $k$ or larger, and therefore

$$|S_{w,w'}(z)S_{w',w}(z)| \leq k\frac{(p\rho)^k}{1 - p\rho}. \tag{5.14}$$

There exists $K_1 > 0$, such that for $k > K_1$, we have $\lim_{k\to\infty} k\frac{(p\rho)^k}{1-p\rho} = 0$. This yields

$$\begin{aligned}|S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)| &\geq |S_w(z)S_{w'}(z)| - |S_{w,w'}(z)S_{w',w}(z)| \\ &\geq 1 - k\frac{(p\rho)^k}{1 - p\rho} > 0. \end{aligned} \tag{5.15}$$

Case *ii*. If the minimal degree for $S_w(z) - 1$ or $S_{w'}(z) - 1$ is greater than $\lfloor k/2 \rfloor$, then every term of $S_{w,w'}(z)S_{w',w}(z)$ has degree at least $k/2$. We also note that, by Lemma 5.0.2, $|S_w(z)S_{w'}(z)| > 0$. Therefore, there exists $K_2 > 0$, such that

$$\begin{aligned}|S_w(z)S_{w'}(z) - S_{w',w}(z)S_{w,w'}(z)| &\geq |S_w(z)S_{w'}(z)| - |S_{w',w}(z)S_{w,w'}(z)| \\ &> 0 \quad \text{for } k > K_2. \end{aligned} \tag{5.16}$$

Case *iii*. The only remaining case is where the minimal degree for $S_w(z) - 1$ and $S_{w'}(z) - 1$ are both less than or equal to $\lfloor k/2 \rfloor$. If $w = w_1...w_k$, then $w' = uw_1...w_{k-m}$, where $u$ is a word of length $m \geq 1$. Then we have

$$S_{w',w}(z) = \mathbf{P}(w_{k-m+1}...w_k)z^m \left(S_w(z) - O\left((pz)^{k-m}\right)\right). \tag{5.17}$$

There exists $K_3 > 0$, such that

$$\begin{aligned} |S_{w',w}(z)| &\leq (p\rho)^m \left(|S_w(z)| + O\left((p\rho)^{k-m}\right)\right. \\ &= (p\rho)^m |S_w(z)| + O\left((p\rho)^k\right) \\ &< |S_w(z)| \quad \text{for } k > K_3 \ . \end{aligned} \tag{5.18}$$

Similarly, we can show that there exists $K_3'$, such that $|S_{w,w'}(z)| < |S_{w'}(z)|$. Therefore, for $k > K_3'$ we have

$$\begin{aligned} |S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)| &\geq |S_w(z)||S_{w'}(z)| - |S_{w,w'}(z)||S_{w',w}(z)| \\ &> |S_w(z)||S_{w'}(z)| - |S_w(z)||S_{w'}(z)| = 0. \end{aligned} \tag{5.19}$$

We complete the proof by setting $K' = \max\{K_1, K_2, K_3, K_3'\}$. ■

We now use the above lemmas to show that both $D_w(z)$ and $D_{w,w'}(z)$ have exactly one root in the disk $|z| \leq \rho$. For Theorem 5.0.5 and Remark 5.0.6, we use the notations and methodology presented in [30].

**Theorem 5.0.5 (Jacquet and Szpankowski, 1994)** *There exist $K_w > 0$ and $\rho > 1$ such that, $p\rho < 1$, and for every word $w$ of length $k \geq K_w$, the polynomial $D_w(z) = (1-z)S_w(z) + \mathbf{P}(w)z^k$ has exactly one root in the disk $|z| \leq \rho$.*

**Proof** We first show that $|\mathbf{P}(w)z^k|$ is bounded above by $|(1-z)S_w(z)|$, for large enough $k$. Note that $|\mathbf{P}(w)z^k| \leq (p\rho)^k$, and by lemma 5.0.2, for $|z| \leq \rho$, and $w$ of length $k > K$, There exists $\alpha > 0$ such that $|S_w(z)| \geq \alpha$. Therefore, if we set $\bar{K} > 0$ to be such that $(p\rho)^{\bar{K}} < \alpha(\rho - 1)$, for $k \geq K_w = \max\{K, \bar{K}\}$, we will have

$$\begin{aligned} |\mathbf{P}(w)z^k| &\leq (p\rho)^k \\ &\leq \alpha(\rho - 1) \\ &< |(1-z)S_w(z)|. \end{aligned} \tag{5.20}$$

Therefore, by Rouché's theorem (cf. [33]), the polynomial $D_w(z)$ must have the same number of roots as $(1-z)S_w(z)$ in the disk $|z| \leq \rho$. But $(1-z)S_w(z)$ has only one root in $|z| \leq \rho$, and this completes the proof. ■

**Remark 5.0.6** *Following the notations in [30], we denote the root within the disk $|z| \leq \rho$ of $D_w(z)$ by $A_w$, and by bootstrapping we obtain*

$$A_w = 1 + \frac{1}{S_w(1)}\mathbf{P}(w) + O\left(\mathbf{P}(w)^2\right). \tag{5.21}$$

*We also denote the derivative of $D_w(z)$ at the root $A_w$, by $B_w$, and we obtain*

$$B_w = -S_w(1) + \left(k - \frac{2S'_w(1)}{S_w(1)}\mathbf{P}(w)\right) + O\left(\mathbf{P}(w)^2\right). \tag{5.22}$$

**Theorem 5.0.7** *There exist $K_{w,w'} > 0$ and $\rho > 1$ such that $p\rho < 1$, and for every word $w$ and $w'$ of length $k \geq K_{w,w'}$, the polynomial*

$$\begin{aligned} D_{w,w'}(z) = (1-z)&(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)) \\ &+ z^k\left(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z))\right), \end{aligned} \tag{5.23}$$

*has exactly one root in the disk $|z| \leq \rho$.*

**Proof** First note that

$$\begin{aligned} |S_w(z) - S_{w',w}(z)| &\leq |S_w(z)| + |S_{w',w}(z)| \\ &\leq \frac{1}{1-p\rho} + \frac{p\rho}{1-p\rho} = \frac{1+p\rho}{1-p\rho}. \end{aligned} \tag{5.24}$$

This yields

$$\begin{aligned} &\left|z^k\left(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z))\right)\right| \\ &\quad\leq (p\rho)^k\left(|S_w(z) - S_{w',w}(z)| + |S_{w'}(z) - S_{w,w'}(z)|\right) \\ &\quad\leq (p\rho)^k\left(\frac{2(1+p\rho)}{1-p\rho}\right). \end{aligned} \tag{5.25}$$

There exist $K'$, $K''$ large enough, such that, for $k > K'$, we have

$$|(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z))| \geq \beta > 0,$$

and for $k > K''$,

$$(p\rho)^k\left(\frac{2(1+p\rho)}{1-p\rho}\right) < (\rho-1)\beta.$$

If we define $K_{w,w'} = \max\{K', K''\}$, then we have, for $k \geq K_{w,w'}$ ,

$$\begin{aligned}(p\rho)^k \left(\frac{2(1+p\rho)}{1-p\rho}\right) &< (\rho-1)\beta \\ &< |(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z))|. \qquad (5.26)\end{aligned}$$

by Rouché's theorem, since $(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z))$ has only one root in $|z| \leq \rho$, then also $D_{w,w'}(z)$ has exactly one root in $|z| \leq \rho$. ■

We present this root in the following Remark.

**Remark 5.0.8** *We denote the root within the disk $|z| \leq \rho$ of $D_{w,w'}(z)$ by $\alpha_{w,w'}$, and by bootstrapping we obtain*

$$\begin{aligned}\alpha_{w,w'} = 1 &+ \frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) \\ &+ \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w') + O(p^{2k}). \qquad (5.27)\end{aligned}$$

*We also denote the derivative of $D_{w,w'}(z)$ at the root $\alpha_{w,w'}$, by $\beta_{w,w'}$, and we obtain*

$$\beta_{w,w'} = S_{w,w'}(1)S_{w',w}(1) - S_w(1)S_{w'}(1) + O(kp^k). \qquad (5.28)$$

We will refer to Remarks 5.0.6, and 5.0.8 in the residue analysis that we present in the next chapter.

# 6. ASYMPTOTIC DIFFERENCE

The results discussed in the previous chapters enable us to show that the first two factorial moments of the $k$th Subword Complexity asymptotically behave the same way as the first two factorial moments of the $k$th prefix complexity. The proof for this analysis includes Cauchy's coefficient formula, the residue theorem, and the Mellin transform. Before conducting this analysis, we provide a brief introduction to the Mellin transform and its asymptotic properties. A more detailed discussion on the Mellin transforms can be found in [34], [35] and chapter nine of [36]. We will adopt the notations used in [34, 36] in the section below.

## 6.1 Review of The Mellin Transform

**Definition 6.1.1** *Let $f(z)$ be a complex valued function defined over positive real values. The Mellin transform $f^*(s)$ of $f(z)$ is defined as*

$$f^*(s) = \int_0^\infty f(z) z^{s-1} dz, \tag{6.1}$$

*where $s \in \mathbb{C}$.*

To ensure the existence of the Mellin transform for a continuous function $f(z)$, we use $\alpha$ and $\beta$ in characterizing the asymptotics of $f(z)$, as follows:

$$f(z) = \begin{cases} O(z^\alpha) & z \to 0, \\ O(z^\beta) & z \to \infty. \end{cases}$$

We then have

$$\begin{aligned} \left| \int_0^\infty f(z) z^{s-1} dz \right| &\leq \int_0^1 |f(z)| z^{\Re(s)-1} dz + \int_1^\infty |f(z)| z^{\Re(s)-1} dz \\ &\leq c_1 \int_0^1 z^{\Re(s)+\alpha-1} dz + c_2 \int_1^\infty z^{\Re(s)+\beta-1} dz, \end{aligned} \tag{6.2}$$

where $c_1$ and $c_2$ are constants. The above integrals exist only when $\Re(s) > -\alpha$ for the first integral and $\Re(s) < -\beta$ for the second one. The strip $-\alpha < \Re(s) < -\beta$ is called the *fundamental strip* of the Mellin transform and is denoted by $\langle -\alpha, -\beta \rangle$.

The inverse of the Mellin transform of the function $f^*(s)$ with the fundamental strip $\langle -\alpha, -\beta \rangle$ can be represented as the following integral

$$f(z) = \frac{1}{2\pi i} \int_{m-i\infty}^{m+i\infty} f^*(s) z^{-s} ds, \tag{6.3}$$

where $m \in \langle -\alpha, -\beta \rangle$.

What draws our attention to the Mellin transform in our analysis is its asymptotic properties discussed in [36]. There exists a direct mapping between the asymptotic expansion of a function $f(z)$ near infinity and the set of singularities of $f^*(s)$ in $\mathbb{C}$. If $f^*(s)$ is a meromorphic function that can be analytically continued to $\langle -\alpha, M \rangle$ for some $M > -\beta$, then

$$f(z) = -\sum_{\lambda_i \in \Lambda} \mathrm{Res}[f^*(s) z^{-s}, s = \lambda_i] + O(z^{-M}) \quad \text{as } z \to \infty, \tag{6.4}$$

where $\Lambda$ is the set of singularities and $M$ is as large as desired. This is easy to see, for to solve the inverse Mellin integral $f(z)$, we can consider a large rectangular contour $\gamma$ as in Figure 6.1 with its left edge at $m$, its right edge at $M$, while the top and bottom edges approach $\pm\infty$. Therefore,

$$\begin{aligned} \int_\gamma f^*(s) z^{-s} ds = \int_{m-i\infty}^{m+i\infty} f^*(s) z^{-s} ds + \int_{m+i\infty}^{M+i\infty} f^*(s) z^{-s} ds \\ + \int_{M+i\infty}^{M-i\infty} f^*(s) z^{-s} ds + \int_{M-i\infty}^{m-i\infty} f^*(s) z^{-s} ds. \end{aligned} \tag{6.5}$$

The integrals over the top and bottom edges are relatively insignificant, since for an $r$-times differentiable function $f(z)$, we have $f^*(r + it) = o(|t|^{-r})$ as $|t| \to \pm\infty$. Also, for the integral over the line $\Re(s) = M$, we have

$$\left| \int_{M+i\infty}^{M-i\infty} f^*(s) z^{-s} ds \right| \leq |z^{-M}| \int_{\infty}^{-\infty} |f^*(r + it)||z^{-it}| dt = O(z^{-M}). \tag{6.6}$$

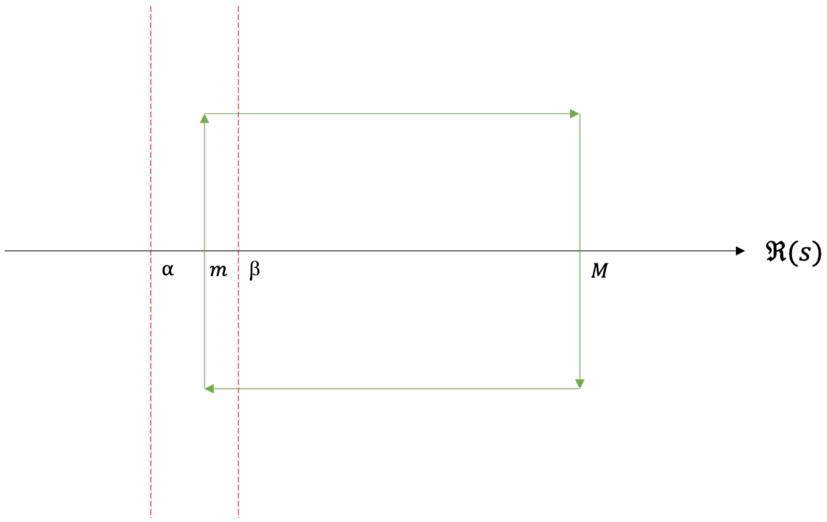


Figure 6.1. The integral contour $\gamma$ for estimating the inverse of the Mellin transform.

Therefore, the result in (6.4) follows. We will use the above property in the proofs of Theorems 6.2.1 and 6.3.1.

## 6.2 Comparison of The Expected Values

Here, we prove that for large $n$, and for $k = \Theta(\log n)$, the expected values $\mathbf{E}[X_{n,k}]$ and $\mathbf{E}[\hat{X}_{n,k}]$ have similar asymptotic growth.

**Theorem 6.2.1** *For large values of $n$, and for $k = \Theta(\log n)$, there exists $M > 0$ such that $\mathbf{E}[X_{n,k}] - \mathbf{E}[\hat{X}_{n,k}] = O(n^{-M})$.*

**Proof** Recall that in Theorems 4.1.1 and 4.2.1, we stated that the generating functions representing $\mathbf{E}[X_{n,k}]$ and $\mathbf{E}[\hat{X}_{n,k}]$ are respectively

$$H(z) = \sum_{w \in \mathcal{A}^k} \left( \frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} \right),$$

and

$$\hat{H}(z) = \sum_{w \in \mathcal{A}^k} \left( \frac{1}{1-z} - \frac{1}{1-(1-\mathbf{P}(w))z} \right).$$

We have

$$H(z) - \hat{H}(z) = \sum_{w \in \mathcal{A}^k} \left( \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)} \right). \tag{6.7}$$

We define

$$\Delta_w(z) = \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)}. \tag{6.8}$$

Therefore, by Cauchy integral formula (see [20], page 236), we have

$$[z^n]\Delta_w(z) = \frac{1}{2\pi i} \oint \Delta_w(z) \frac{dz}{z^{n+1}} = \mathrm{Res}_{z=0}\, \Delta_w(z) \frac{dz}{z^{n+1}}, \tag{6.9}$$

where the path of integration is a circle about zero with counterclockwise orientation. We note that the above integrand has poles at $z = 0$, $z = \dfrac{1}{1-\mathbf{P}(w)}$, and $z = A_w$ (refer to Remark 5.0.6). Therefore, we define

$$I^w(\rho) := \frac{1}{2\pi i} \int_{|z|=\rho} \Delta_w(z) \frac{dz}{z^{n+1}}, \tag{6.10}$$

where the circle of radius $\rho$ contains all of the above poles. By the residue theorem, we have

$$\begin{aligned} I^w(\rho) &= \mathrm{Res}_{z=0} \frac{\Delta_w(z)}{z^{n+1}} + \mathrm{Res}_{z=A_w} \frac{\Delta_w(z)}{z^{n+1}} + \mathrm{Res}_{z=1/1-\mathbf{P}(w)} \frac{\Delta_w(z)}{z^{n+1}} \\ &= [z^n]\Delta_w(z) - \mathrm{Res}_{z=A_w} \frac{H_w(z)}{z^{n+1}} + \mathrm{Res}_{z=1/1-\mathbf{P}(w)} \frac{\hat{H}_w(z)}{z^{n+1}} \end{aligned} \tag{6.11}$$

We observe that

$$\begin{aligned} &\mathrm{Res}_{z=A_w} \frac{\Delta_w(z)}{z^{n+1}} = \frac{S_w(A_w)}{B_w\, A_w^{n+1}}, \quad \text{where } B_w \text{ is as in Remark 5.0.6} \\ &\mathrm{Res}_{z=1/1-\mathbf{P}(w)} \frac{\hat{H}_w(z)}{z^{n+1}} = -(1-\mathbf{P}(w))^{n+1}. \end{aligned}$$

Then we obtain

$$[z^n]\Delta_w = I^w(\rho) - \frac{S_w(A_w)}{B_w\, A_w^{n+1}} - (1-\mathbf{P}(w))^{n+1}, \tag{6.12}$$

and finally, we have

$$\begin{aligned} [z^n](H(z) - \hat{H}(z)) &= \sum_{w \in \mathcal{A}^k} [z^n]\Delta_w \\ &= \sum_{w \in \mathcal{A}^k} I_n^w(\rho) - \sum_{w \in \mathcal{A}^k} \left( \frac{S_w(A_w)}{B_w\, A_w^{n+1}} + (1-\mathbf{P}(w))^{n+1} \right). \end{aligned} \tag{6.13}$$

First we show that, for sufficiently large $n$, the sum $\sum_{w\in\mathcal{A}^k}\left(\frac{S_w(A_w)}{B_w\,A_w^{n+1}}+(1-\mathbf{P}(w))^{n+1}\right)$ approaches zero.

**Lemma 6.2.2** *For large enough $n$, and for $k=\Theta(\log n)$, there exists $M>0$ such that*

$$\sum_{w\in\mathcal{A}^k}\left(\frac{S_w(A_w)}{B_w\,A_w^{n+1}}+(1-\mathbf{P}(w))^{n+1}\right)=O(n^{-M}). \tag{6.14}$$

**Proof** We let

$$r_w(z)=(1-\mathbf{P}(w))^z+\frac{S_w(A_w)}{B_wA_w^z}. \tag{6.15}$$

The Mellin transform of the above function is

$$r_w^*(s)=\Gamma(s)\log^{-s}\left(\frac{1}{1-\mathbf{P}(w)}\right)-\frac{S_w(A_w)}{B_w}\Gamma(s)\log^{-s}(A_w). \tag{6.16}$$

We define

$$C_w=\frac{S_w(A_w)}{B_w}=\frac{S_w(A_w)}{-S_w(1)+O(k\mathbf{P}(w))}, \tag{6.17}$$

which is negative and uniformly bounded for all $w$. Also, for a fixed $s$, we have

$$\begin{aligned}\ln^{-s}\left(\frac{1}{1-\mathbf{P}(w)}\right)&=\ln^{-s}\left(1+\mathbf{P}(w)+O\left(\mathbf{P}(w)^2\right)\right)\\&=\left(\mathbf{P}(w)+O\left(\mathbf{P}(w)^2\right)\right)^{-s}\\&=\mathbf{P}(w)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right)^{-s}\\&=\mathbf{P}(w)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right),\end{aligned} \tag{6.18}$$

$$\begin{aligned}\ln^{-s}(A_w)&=\ln^{-s}\left(1-\left(-\frac{\mathbf{P}(w)}{S_w(1)}+O\left(\mathbf{P}(w)^2\right)\right)\right)\\&=\left(\frac{\mathbf{P}(w)}{S_w(1)}+O\left(\mathbf{P}(w)^2\right)\right)^{-s}\\&=\left(\frac{\mathbf{P}(w)}{S_w(1)}\right)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right)^{-s}\\&=\left(\frac{\mathbf{P}(w)}{S_w(1)}\right)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right),\end{aligned} \tag{6.19}$$

and therefore, we obtain

$$r_w^*(s) = \Gamma(s)\mathbf{P}(w)^{-s}\left(1 - \frac{1}{S_w(1)^{-s}}\right)O(1). \tag{6.20}$$

From this expression, and noticing that the function has a removable singularity at $s = 0$, we can see that the Mellin transform $r_w^*(s)$ exists on the strip where $\Re(s) > -1$. We still need to investigate the Mellin strip for the sum $\sum_{w\in\mathcal{A}^k} r_w^*(s)$. In other words, we need to examine whether summing $r_w^*(s)$ over all words of length $k$ (where $k$ grows with $n$) has any effect on the analyticity of the function. We observe that

$$\begin{aligned}\sum_{w\in\mathcal{A}^k} |r_w^*(s)| &= \sum_{w\in\mathcal{A}^k}\left|\Gamma(s)\mathbf{P}(w)^{-s}\left(1 - \frac{1}{S_w(1)^{-s}}\right)O(1)\right| \\ &\leq |\Gamma(s)| \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)^{-\Re(s)}\left(1 - \frac{1}{S_w(1)^{-\Re(s)}}\right)O(1) \\ &= (q^k)^{-\Re(s)-1}|\Gamma(s)| \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)(1 - S_w(1)^{\Re(s)})O(1).\end{aligned}$$

Lemma 5.0.1 allow us to split the above sum between the words for which $S_w(1) \leq 1 + O(\delta^k)$ and words that have $S_w(1) > 1 + O(\delta^k)$. Such a split yields the following

$$\sum_{w\in\mathcal{A}^k} |r_w^*(s)| = (q^k)^{-\Re(s)-1}|\Gamma(s)|O(\delta^k). \tag{6.21}$$

This shows that $\sum_{w\in\mathcal{A}^k} r_w^*(s)$ is bounded above for $\Re(s) > -1$ and therefore, it is analytic. This argument holds for $k = \Theta(\log n)$ as well, since $(q^k)^{-\Re(s)-1}$ would still be bounded above by a constant $M_{s,k}$ that depends on $s$ and $k$.

We would like to approximate $\sum_{w\in\mathcal{A}^k} r_w^*(s)$ when $z \to \infty$. By the inverse Mellin transform, we have

$$\sum_{w\in\mathcal{A}^k} r_w(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\Big(\sum_{w\in\mathcal{A}^k} r_w^*(s)\Big)z^{-s}ds. \tag{6.22}$$

We choose $c \in (-1, M)$ for a fixed $M > 0$. Then by (6.4), we obtain

$$\sum_{w\in\mathcal{A}^k} r_w(z) = O(z^{-M}). \tag{6.23}$$

and subsequently, we get

$$\sum_{w\in\mathcal{A}^k}\left(\frac{S_w(A_w)}{B_w\,A_w^{n+1}}+(1-\mathbf{P}(w))^{n+1}\right)=O(n^{-M}). \tag{6.24}$$

■

We next prove the asymptotic smallness of $I_n^w(\rho)$ in (6.10).

**Lemma 6.2.3** *Let*

$$I_n^w(\rho)=\frac{1}{2\pi i}\int_{|z|=\rho}\left(\frac{1}{1-(1-\mathbf{P}(w))z}-\frac{S_w(z)}{D_w(z)}\right)\frac{dz}{z^{n+1}}. \tag{6.25}$$

*For large $n$ and $k=\Theta(\log n)$, we have*

$$\sum_{w\in\mathcal{A}^k} I_n^w(\rho)=O\left(\rho^{-n}(\rho\delta)^k\right). \tag{6.26}$$

**Proof** We observe that

$$|I_n^w(\rho)|\leq\frac{1}{2\pi}\int_{|z|=\rho}\left|\frac{\mathbf{P}(w)z\left(z^{k-1}-S_w(z)\right)}{D_w(z)(1-(1-\mathbf{P}(w))z)}\frac{1}{z^{n+1}}\right|dz. \tag{6.27}$$

For $|z|=\rho$, we show that the denominator in (6.27) is bounded away from zero.

$$\begin{aligned}|D_w(z)|&=|(1-z)S_w(z)+\mathbf{P}(w)z^k|\\&\geq|1-z||S_w(z)|-\mathbf{P}(w)|z^k|\\&\geq(\rho-1)\alpha-(p\rho)^k,\quad\text{where }\alpha\text{ is as in the proof of Theorem 5.0.5.}\\&>0,\qquad\text{since }(p\rho)^k<\alpha(\rho-1)\text{ by the assumption in Theorem 5.0.5.}\end{aligned} \tag{6.28}$$

To find a lower bound for $|1-(1-\mathbf{P}(w))z|$, we can choose $K_w$ in Theorem 5.0.5 large enough such that

$$\begin{aligned}|1-(1-\mathbf{P}(w))z|&\geq|1-(1-\mathbf{P}(w))|z||\\&\geq|1-\rho(1-p^{K_w})|\\&>0.\end{aligned} \tag{6.29}$$

We now move on to finding an upper bound for the numerator in (6.27), for $|z| = \rho$.

$$\begin{aligned} |z^{k-1} - S_w(z)| &\leq |S_w(z) - 1| + |1 - z^{k-1}| \\ &\leq (S_w(\rho) - 1) + (1 + \rho^{k-1}) \\ &= (S_w(\rho) - 1) + O(\rho^k). \end{aligned} \tag{6.30}$$

Therefore, there exists a constant $\mu > 0$ such that

$$\begin{aligned} |I_n^w| &\leq \mu\rho\mathbf{P}(w)\left((S_w(\rho) - 1) + O(\rho^k)\right)\frac{1}{\rho^{n+1}} \\ &= O(\rho^{-n})\left(\mathbf{P}(w)(S_w(\rho) - 1) + \mathbf{P}(w)O(\rho^k)\right). \end{aligned} \tag{6.31}$$

Summing over all patterns $w$, and applying Lemma 5.0.1, we obtain

$$\begin{aligned} \sum_{w\in\mathcal{A}^k} |I_n^w(\rho)| &= O(\rho^{-n})\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)(S_w(\rho) - 1) + O(\rho^{-n+k})\sum_{w\in\mathcal{A}^k}\mathbf{P}(w) \\ &= O(\rho^{-n})\left(\theta(\rho\delta)^k + \frac{p\rho}{1 - p\rho}\theta\delta^k\right) + O(\rho^{-n+k}) \\ &= O(\rho^{-n}(\rho\delta)^k), \end{aligned} \tag{6.32}$$

which approaches zero as $n \to \infty$ and $k = \Theta(\log n)$. This completes the proof of of Theorem 6.2.1. ■

### 6.3 Comparison of The Second Factorial Moments

Similar to the previous section, we provide a proof to show that the second factorial moments of the $k$th Subword Complexity and the $k$th Prefix Complexity, have the same first order asymptotic behavior.

**Theorem 6.3.1** *For large values of $n$, and for $k = \Theta(\log n)$, there exists $\epsilon > 0$ such that $\mathbf{E}[(X_{n,k})_2] - \mathbf{E}[(\hat{X}_{n,k})_2] = O(n^{-\epsilon})$.*

**Proof** As discussed in Theorems 4.1.1 and 4.2.1, the generating functions representing $\mathbf{E}[(X_{n,k})_2]$ and $\mathbf{E}[(\hat{X}_{n,k})_2]$ respectively, are

$$G(z) = \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}}\left(\frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} - \frac{S_{w'}(z)}{D_{w'}(z)} + \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}\right),$$

And

$$\hat{G}(z) = \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left( \frac{1}{1-z} - \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{1}{1-(1-\mathbf{P}(w'))z} \right) + \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z}.$$

Note that

$$G(z) - \hat{G}(z) = \sum_{\substack{w' \in \mathcal{A}^k \\ w \neq w'}} \sum_{w \in \mathcal{A}^k} \left( \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)} \right) \tag{6.33}$$

$$+ \sum_{\substack{w \in \mathcal{A}^k \\ w \neq w'}} \sum_{w' \in \mathcal{A}^k} \left( \frac{1}{1-(1-\mathbf{P}(w'))z} - \frac{S_{w'}(z)}{D_{w'}(z)} \right) \tag{6.34}$$

$$+ \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left( \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z} - \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)} \right) \tag{6.35}$$

In Theorem 6.2.1, we proved that for every $M > 0$ (which does not depend on $n$ or $k$), we have

$$H(z) - \hat{H}(z) = \sum_{w \in \mathcal{A}^k} \left( \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)} \right) = O(n^{-M}).$$

Therefore, both (6.33) and (6.34) are of order $(2^k - 1)O(n^{-M}) = O(n^{-M+a\log 2})$ for $k = a\log n$. Thus, in order to show the asymptotic smallness, it is enough to choose $M = a\log 2 + \epsilon$, where $\epsilon$ is a small positive value. Now, it only remains to show (6.35) is asymptotically negligible as well. We define

$$\Delta_{w,w'}(z) = \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z} - \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}. \tag{6.36}$$

Next, we extract the coefficient of $z^n$

$$[z^n]\Delta_{w,w'}(z) = \frac{1}{2\pi i} \oint \Delta_{w,w'}(z) \frac{dz}{z^{n+1}}, \tag{6.37}$$

where the path of integration is a circle about the origin with counterclockwise orientation. We define

$$I_n^{w,w'}(\rho) = \frac{1}{2\pi i}\int_{|z|=\rho} \Delta_{w,w'}(z)\frac{dz}{z^{n+1}}, \tag{6.38}$$

The above integrand has poles at $z = 0$, $z = \alpha_{w,w'}$ (as in Remark 5.0.7), and $z = \frac{1}{1-\mathbf{P}(w)-\mathbf{P}(w')}$. We have chosen $\rho$ such that the poles are all inside the circle $|z| = \rho$. It follows that

$$I_n^{w,w'}(\rho) = \mathrm{Res}_{z=0}\frac{\Delta_{w,w'}(z)}{z^{n+1}} + \mathrm{Res}_{z=\alpha_{w,w'}}\frac{\Delta_{w,w'}(z)}{z^{n+1}} + \mathrm{Res}_{z=\frac{1}{1-\mathbf{P}(w)-\mathbf{P}(w')}}\frac{\Delta_w(z)}{z^{n+1}}, \tag{6.39}$$

and the residues give us the following.

$$\mathrm{Res}_{z=\frac{1}{1-\mathbf{P}(w)-\mathbf{P}(w')}}\frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z)z^{n+1}} = -(1-\mathbf{P}(w)-\mathbf{P}(w'))^{n+1},$$

and

$$\mathrm{Res}_{z=\alpha_{w,w'}}\frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)} = \frac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}\alpha_{w,w'}^{n+1}},$$

where $\beta_{w,w'}$ is as in Remark 5.0.7. Therefore, we get

$$\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}[z^n]\Delta_{w,w'}(z) = \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} I_n^{w,w'}(\rho) - \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}\alpha_{w,w'}^{n+1}} + (1-\mathbf{P}(w)-\mathbf{P}(w'))^{n+1}\right). \tag{6.40}$$

We now show that the above two terms are asymptotically small.

**Lemma 6.3.2** *There exists $\epsilon > 0$ where the sum*

$$\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}\alpha_{w,w'}^{n+1}} + (1-\mathbf{P}(w)-\mathbf{P}(w'))^{n+1}\right)$$

*is of order $O(n^{-\epsilon})$.*

**Proof** We define

$$r_{w,w'}(z) = \frac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}\alpha^z_{w,w'}} + (1 - \mathbf{P}(w) - \mathbf{P}(w'))^z.$$

The Mellin transform of the above function is

$$r^*_{w,w'}(s) = \Gamma(s)\log^{-s}\left(\frac{1}{1 - \mathbf{P}(w) - \mathbf{p}(w')}\right) + C_{w,w'}\Gamma(s)\log^{-s}(\alpha_{w,w'}), \qquad (6.41)$$

where $C_{w,w'} = \dfrac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}}$. We note that $C_{w,w'}$ is negative and uniformly bounded from above for all $w, w' \in \mathcal{A}^k$.For a fixes $s$, we also have,

$$\begin{aligned}
\ln^{-s}\left(\frac{1}{1 - \mathbf{P}(w) - \mathbf{P}(w')}\right) &= \ln^{-s}\left(1 + \mathbf{P}(w) + \mathbf{P}(w') + O\left(p^{2k}\right)\right) \\
&= \left(\mathbf{P}(w) + \mathbf{P}(w') + O\left(p^{2k}\right)\right)^{-s} \\
&= (\mathbf{P}(w) + \mathbf{P}(w'))^{-s}\left(1 + O\left(p^k\right)\right)^{-s} \\
&= (\mathbf{P}(w) + \mathbf{P}(w'))^{-s}\left(1 + O\left(p^k\right)\right), \qquad (6.42)
\end{aligned}$$

and

$$\begin{aligned}
\ln^{-s}(\alpha_{w,w'}) = \Big(&\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) \\
&+ \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w') + O(p^{2k})\Big)^{-s} \\
= \Big(&\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) \\
&+ \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\Big)^{-s}\left(1 + O(p^k)\right).
\end{aligned} \qquad (6.43)$$

Therefore, we have

$$\begin{aligned}
r^*_{w,w'}(s) = \Gamma(s)&\left(\mathbf{P}(w) + \mathbf{P}(w')\right)^{-s}\left(1 + O(p^k)\right) \\
&- \Gamma(s)\Bigg(\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) \\
&\qquad + \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\Bigg)^{-s}\left(1 + O(p^k)\right)O(1).
\end{aligned} \qquad (6.44)$$

To find the Mellin strip for the sum $\sum_{w\in\mathcal{A}^k} r^*_{w,w'}(s)$, we first note that

$$(x+y)^a \leq x^a + y^a, \quad \text{for any real } x, y > 0 \text{ and } a \leq 1.$$

Since $-\Re(s) < 1$, we have

$$(\mathbf{P}(w) + \mathbf{P}(w'))^{-\Re(s)} \leq \mathbf{P}(w)^{-\Re(s)} + \mathbf{P}(w')^{-\Re(s)}, \tag{6.45}$$

and

$$\begin{aligned}\left(\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) + \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\right)^{-\Re(s)} \\ \leq \left(\frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w)\right)^{-\Re(s)} \\ + \left(\frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\right)^{-\Re(s)}.\end{aligned} \tag{6.46}$$

Therefore, we get

$$\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} |r^*_{w,w'}(s)| \le |\Gamma(s)|O(1)$$

$$\left(\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \mathbf{P}(w)^{-\Re(s)}\left(1-\left(\frac{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}{S_{w'}(1)-S_{w,w'}(1)}\right)^{\Re(s)}\right)\right.$$

$$\left.+\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \mathbf{P}(w')^{-\Re(s)}\left(1-\left(\frac{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}{S_w(1)-S_{w',w}(1)}\right)^{\Re(s)}\right)\right)$$

$$\le (q^k)^{-\Re(s)-1}|\Gamma(s)|O(1)$$

$$\left(\sum_{\substack{w'\in\mathcal{A}^k\\ w\neq w'}}\sum_{w\in\mathcal{A}^k} \mathbf{P}(w)\left(1-(S_w(1))^{\Re(s)}\left(1-\frac{S_{w,w'}(1)}{S_{w'}(1)}\right)^{-\Re(s)}\right)\right. \tag{6.47}$$

$$+\sum_{\substack{w'\in\mathcal{A}^k\\ w\neq w'}}\sum_{w\in\mathcal{A}^k} \mathbf{P}(w)S_{w,w'}(1)^{\Re(s)}\left(\frac{S_{w'}(1)-S_{w,w'}(1)}{S_{w',w}(1)}\right)^{-\Re(s)} \tag{6.48}$$

$$+\sum_{\substack{w\in\mathcal{A}^k\\ w\neq w'}}\sum_{w'\in\mathcal{A}^k} \mathbf{P}(w')\left(1-(S_{w'}(1))^{\Re(s)}\left(1-\frac{S_{w',w}(1)}{S_w(1)}\right)^{-\Re(s)}\right) \tag{6.49}$$

$$\left.+\sum_{\substack{w\in\mathcal{A}^k\\ w\neq w'}}\sum_{w'\in\mathcal{A}^k} \mathbf{P}(w')S_{w',w}(1)^{\Re(s)}\left(\frac{S_w(1)-S_{w',w}(1)}{S_{w,w'}(1)}\right)^{-\Re(s)}\right). \tag{6.50}$$

By Lemma 5.0.3, with high probability, a randomly selected $w$ has the property $S_{w,w'}(1)=O(\delta^k)$ , and thus

$$\left(1-\frac{S_{w,w'}(1)}{S_{w'}(1)}\right)^{-\Re(s)} = 1+O(\delta^k).$$

With that and by Lemma 5.0.1, for most words $w$,

$$1-S_w(1)^{\Re(s)}(1+O(\delta^k)) = O(\delta^k).$$

Therefore, both sums (6.47) and (6.49) are of the form $(2^k - 1)O(\delta^k)$. The sums (6.48) and (6.50) are also of order $(2^k - 1)O(\delta^k)$ by Lemma 5.0.3. Combining all these terms we will obtain

$$\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} |r^*_{w,w'}(s)| \leq (2^k - 1)(q^k)^{-\Re(s)-1} |\Gamma(s)| O(\delta^k) O(1). \tag{6.51}$$

By the inverse Mellin transform, for $k = a \log n$, $M = a \log 2 + \epsilon$ and $c \in (-1, M)$, we have

$$\begin{aligned} \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} r_{w,w'}(z) &= \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \Big( \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} r^*_{w,w'}(s) \Big) z^{-s} ds = O(z^{-M}) O(2^k) \\ &= O(z^{-\epsilon}). \end{aligned} \tag{6.52}$$

■

In the following lemma we show that the first term in (6.41) is asymptotically small.

**Lemma 6.3.3** *Recall that*

$$I_n^{w,w'}(\rho) = \frac{1}{2\pi i} \int_{|z|=\rho} \Delta_{w,w'}(z) \frac{dz}{z^{n+1}}.$$

*We have*

$$\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} I_n^{w,w'}(\rho) = O\left(\rho^{-n+2k} \delta^k\right). \tag{6.53}$$

**Proof** First note that

$$\begin{aligned} \Delta_{w,w'}(z) &= \frac{1}{1 - (1 - \mathbf{P}(w) - \mathbf{P}(w'))z} - \frac{S_w(z) S_{w'}(z) - S_{w,w'}(z) S_{w',w}(z)}{D_{w,w'}(z)} \\ &= \frac{z\mathbf{P}(w) \left( S_{w,w'}(z) S_{w',w}(z) - S_w(z) S_{w'}(z) + z^{k-1} S_{w'}(z) - z^{k-1} S_{w,w'}(z) \right)}{(1 - (1 - \mathbf{P}(w) - \mathbf{P}(w'))z) \, D_{w,w'}(z)} \\ &\quad + \frac{z\mathbf{P}(w') \left( S_{w',w}(z) S_{w,w'}(z) - S_{w'}(z) S_w(z) + z^{k-1} S_w(z) - z^{k-1} S_{w',w}(z) \right)}{(1 - (1 - \mathbf{P}(w) - \mathbf{P}(w'))z) \, D_{w,w'}(z)}. \end{aligned} \tag{6.54}$$

We saw in (6.29) that $|1-(1-\mathbf{P}(w'))z| \geq c_2$, and therefore, it follows that

$$|1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z| \geq c_1 \tag{6.55}$$

For $z=\rho$, $|D_{w,w'}(z)|$ is also bounded below as the following

$$\begin{aligned}
|D_{w,w'}(z)| &= \big|(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)) \\
&\qquad + z^k\left(\mathbf{P}(w)(S_{w'}(z)-S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z)-S_{w',w}(z))\right)\big| \\
&\geq |(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z))| \\
&\qquad - \left|z^k\right| |(\mathbf{P}(w)(S_{w'}(z)-S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z)-S_{w',w}(z)))| \\
&\geq (\rho-1)\beta - (p\rho)^k\left(\frac{2(1+p\rho)}{1-p\rho}\right),
\end{aligned} \tag{6.56}$$

which is bounded away from zero by the assumption of Theorem 5.0.7. Additionally, we show that the numerator in (6.54) is bounded above, as follows

$$\begin{aligned}
&|S_{w,w'}(z)S_{w',w}(z) - S_w(z)S_{w'}(z) + z^{k-1}S_{w'}(z) - z^{k-1}S_{w,w'}(z)| \leq \\
&\quad |S_{w'}(z)(z^{k-1} - S_w(z))| + |S_{w,w'}(z)(S_{w',w}(z) - z^{k-1})| \\
&\quad \leq S_{w'}(\rho)\left((S_w(\rho)-1) + O(\rho^k)\right) + S_{w,w'}(\rho)\left(S_{w',w}(\rho) + O(\rho^k)\right).
\end{aligned} \tag{6.57}$$

This yields

$$\begin{aligned}
\sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} |I_n^{w,w'}| &\leq O(\rho^{-n}) \sum_{\substack{w'\in\mathcal{A}^k \\ w\neq w'}} S_{w'}(\rho) \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)\left((S_w(\rho)-1) + O(\rho^k)\right) \\
&\quad + O(\rho^{-n}) \sum_{\substack{w'\in\mathcal{A}^k \\ w\neq w'}} \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)S_{w,w'}(\rho)\left(S_{w',w}(\rho) + O(\rho^k)\right).
\end{aligned} \tag{6.58}$$

By (6.31), the first term above is of order $(2^k-1)O(\rho^{-n+k})$ and by Lemma 5.0.3 and an analysis similar to (6.31), the second term yields $(2^k-1)O(\rho^{-n+k})$ as well. Finally, we have

$$\sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} |I_n^{w,w'}| \leq O(\rho^{-n+2k}\delta^k).$$

Which goes to zero asymptotically, for $k=\Theta(\log n)$. ∎

This lemma completes our proof of Theorem 6.3.1.

# 7. ASYMPTOTIC ANALYSIS OF THE $k$th PREFIX COMPLEXITY

In this final chapter, we proceed by analyzing the asymptotic moments of the $k$th Prefix Complexity. The results obtained hold true for the moments of the $k$th Subword Complexity, as we proved in chapter 6. Our methodology here is analogous to the analysis of profile of tries (see [19], [3]), and it involves poissonization, saddle point analysis (the complex version of Laplace's method [37]), and depoissonization. Below, we will give a brief introduction to poissonization (cf. [36]) and the saddle point method (cf. [20, 36]). Afterwards, we conclude with the asymptotic analysis of the average and the second factorial moment of the $k$th Subword Complexity.

## 7.1 Review of Poissonization and Depoissonization

Poissonization is a probabilistic approach in which a Bernoulli model is replaced by a Poisson process. A Bernoulli process still suffers from some sort of dependence. As for instance, for the case of $n$ randomly and independently generated strings, the number of prefixes starting with a 0, influences the number of prefixes that start with a 1. In other words, the sum of the number of prefixes that start with a 0 and the number of prefixes that start with a 1 is always equal to $n$. A remedy for this is to replace the fixed value $n$ with a Poisson random variable $N$ with mean equal to $n$. This way, the number of prefixes that start with 0, and the number of those that start with 1 are two a Poisson random variables that are independent from each other. Below is the formal definition of the Poisson transform.

**Definition 7.1.1** *Let $g_n$ be a sequence of size n over a Bernoulli model. The Poisson transform of $g_n$ is defines as*

$$\tilde{G}(z) = \mathbf{E}[g_N] = \sum_{n \geq 0} g_n \frac{z^n}{n!} e^{-z}, \tag{7.1}$$

*where $z$ is often viewed as the mean of the Poisson random variable $N$.*

Once we solve the poissonized problem, we need to convert it back to its original Bernoulli form. This process is called depoissonization. The following depoissonization theorem (see [36] for proof) gives an expression for the Bernoulli sequence in terms of its poissonized generating function.

**Theorem 7.1.1** *(Jacquet and Szpankowski, 1998) Let $\tilde{G}(z)$ be the Poisson transform of a sequence $g_n$. If $\tilde{G}(z)$ is analytic in a linear cone $S_\theta$ with $\theta < \pi/2$, and if the following two conditions hold:*

*(I) For $z \in S_\theta$ and real values $B$, $r > 0$, $\nu$*

$$|z| > r \rightarrow |\tilde{G}(z)| \leq B|z^\nu|\Psi(|z|), \tag{7.2}$$

*where $\Psi(x)$ is such that, for fixed $t$, $\lim_{x\to\infty} \dfrac{\Psi(tx)}{\Psi(x)} = 1$;*

*(O) For $z \notin S_\theta$ and $A, \alpha < 1$*

$$|z| > r \rightarrow |\tilde{G}(z)e^z| \leq Ae^{\alpha|z|}. \tag{7.3}$$

*Then, for every nonnegative integer n, we have*

$$g_n = \tilde{G}(n) + O(n^{\nu-1}\Psi(n)).$$

## 7.2 Review of Laplace and Saddle Point Methods

One approach to solving a complex integral is to adopt a contour that crosses one or multiple saddle points of the integrand. This is especially useful for when the integrand involves a large parameter $n$ such as the following

$$I(n) = \int_C f(z)e^{-nh(z)}dz. \tag{7.4}$$

This method is known as the Laplace method, in which the path $C$ is a real interval and $z$ is a real value. For this case, we have the following theorem.

**Theorem 7.2.1** *(Laplace's Method) Let $h(t)$ and $f(t)$ be twice differentiable functions on the interval [a,b]. Assume that $h(t)$ has only one minimum in $(a, b)$ occurring at a point $t_0$ (i.e. $h'(t_0) = 0$ and $h''(t_0) > 0$). Then for $n \to \infty$, we have*

$$I(n) = \int_a^b f(t)e^{-nh(t)}dt = f(t_0)\sqrt{\frac{2\pi}{n\,h''(t_0)}}e^{-nh(t_0)}(1 + O(n^{-1/2})). \tag{7.5}$$

The integral that shows up in our work in (7.10) will be of the form

$$I(n) = \int_{c-i\infty}^{c+i\infty} f(z)e^{-nh(z)}dz, \tag{7.6}$$

which is over a complex line. By a change of variables, we transform the path into a real interval, and we obtain

$$I(n) = \int_{-\infty}^{\infty} f(c+it)e^{-nh(c+it)}dt. \tag{7.7}$$

Like in the Laplace case, we are interested in finding a point $z_0$ that minimizes the surface $|h(z)|$, i.e. $h'(z_0) = 0$. The point $z_0$ is called the saddle point. It is expected that the main contribution for estimation of integrals like (7.7) will come from a small neighborhood around the saddle point(s) of $h(z)$. This analysis is known as the saddle point method. Below we provide a road-map for the saddle point analysis to approximate the integral given in (7.6).

*i*. We choose an integration line that crosses the saddle point(s).

*ii*. We split the integration line $l$ into $l_0 \cup l_1$, and the range $l_0$ is chosen such that $h''(t)\delta \to \infty$, and $h^{(3)}\delta \to 0$. (The goal is to get a good approximation from the quadratic expansion of $h(t)$.)

*iii*. Tails pruning: We show that the integral over $l_1$ doesn't contribute much to the integral estimation.

*iv.* Central approximation: Theorem 7.2.1 holds for the range $l_0$.

A detailed discussion on the saddle point method can be found in [20, 36].

### 7.3 On the Expected Value

To transform the sequence of interest, $(\mathbf{E}[\hat{X}_{n,k}])_{n\geq 0}$, into a Poisson model, we recall that in (4.15) we found

$$\mathbf{E}[\hat{X}_{n,k}] = \sum_{w\in\mathcal{A}^k} \left(1-(1-\mathbf{P}(w))^n\right).$$

Thus, the Poisson transform is

$$\begin{aligned}\tilde{E}_k(z) &= \sum_{n=0}^{\infty} \mathbf{E}[\hat{X}_{n,k}]\frac{z^n}{n!}e^{-z} \\ &= \sum_{n=0}^{\infty}\sum_{w\in\mathcal{A}^k} \left(1-(1-\mathbf{P}(w))^n\right)\frac{z^n}{n!}e^{-z} \\ &= \sum_{w\in\mathcal{A}^k}\left(1-e^{-z\mathbf{P}(w)}\right). \qquad (7.8)\end{aligned}$$

To asymptotically evaluate this harmonic sum, we turn our attention to the Mellin Transform once more. The Mellin transform of $\tilde{E}_k(z)$ is

$$\begin{aligned}\tilde{E}_k^*(s) &= -\Gamma(s)\sum_{w\in\mathcal{A}^k} P(w)^{-s} \\ &= -\Gamma(s)(p^{-s}+q^{-s})^k, \qquad (7.9)\end{aligned}$$

which has the fundamental strip $s \in \langle -1, 0\rangle$. For $c \in (-1, 0)$, the inverse Mellin integral is the following

$$\begin{aligned}\tilde{E}_k(z) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \tilde{E}_k^*(s)\cdot z^{-s}ds \\ &= \frac{-1}{2\pi i}\int_{c-i\infty}^{c+i\infty} z^{-s}\Gamma(s)(p^{-s}+q^{-s})^k ds \\ &= \frac{-1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \Gamma(s)e^{-k(s\frac{\log z}{k}-\log(p^{-s}+q^{-s}))}ds \\ &= \frac{-1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \Gamma(s)e^{-kh(s)}ds, \qquad (7.10)\end{aligned}$$

where we define $h(s) = \frac{s}{a} - \log(p^{-s} + q^{-s})$ for $k = a \log z$. We emphasize that the above integral involves $k$, and $k$ grows with $n$. We evaluate the integral through the saddle point analysis. Therefore, we choose the line of integration to cross the saddle point $r_0$. To find the saddle point $r_0$, we let $h'(r_0) = 0$, and we obtain

$$(p/q)^{-r_0} = \frac{a \log p^{-1} - 1}{1 - a \log q^{-1}}, \tag{7.11}$$

and therefore,

$$r_0 = \frac{-1}{\log p/q} \log \left( \frac{a \log q^{-1} - 1}{1 - a \log p^{-1}} \right), \tag{7.12}$$

where $\frac{1}{\log q^{-1}} < a < \frac{1}{\log p^{-1}}$.

By (7.11) and the fact that $(p/q)^{it_j} = 1$ for $t_j = \frac{2\pi j}{\log p/q}$ and $j \in \mathbb{Z}$, we can see that there are actually infinitely many saddle points $z_j$ of the form $r_0 + it_j$ on the line of integration.

We remark that the location of $r_0$ depends on the value of $a$. We have $r_0 \to \infty$ as $a \to \frac{1}{\log q^{-1}}$, and $r_0 \to -\infty$ as $a \to \frac{1}{\log p^{-1}}$. We divide the analysis into three parts, for the three ranges $r_0 \in (0, \infty)$, $r_0 \in (-1, 0)$, and $r_0 \in (-\infty, -1)$.

In the first range, which corresponds to

$$\frac{1}{\log q^{-1}} < a < \frac{2}{\log q^{-1} + \log p^{-1}}, \tag{7.13}$$

we perform a residue analysis, taking into account the dominant pole at $s = -1$. In the second range, we have

$$\frac{2}{\log q^{-1} + \log p^{-1}} < a < \frac{1}{q \log q^{-1} + p \log p^{-1}}, \tag{7.14}$$

and we get the asymptotic result through the saddle point method. The last range corresponds to

$$\frac{1}{q \log q^{-1} + p \log p^{-1}} < a < \frac{1}{\log p^{-1}}, \tag{7.15}$$

and we approach it with a combination of residue analysis at $s = 0$, and the saddle point method.

We prove the following theorem.

**Theorem 7.3.1** *The average of kth Prefix Complexity has the following asymptotic expansion*

*i. For $a$ as in* (7.13),

$$\mathbf{E}[\hat{X}_{n,k}] = 2^k - \Phi_1((1+\log p)\log_{p/q} n)\frac{n^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \quad (7.16)$$

*where $\nu = -r_0 + a\log(p^{-r_0} + q^{-r_0})$, and*

$$\Phi_1(x) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2\pi}\log p/q}\sum_{j\in\mathbb{Z}}\Gamma(r_0 + it_j)e^{-2\pi ijx} \quad (7.17)$$

*is a bounded periodic function.*

*ii. For $a$ as in* (7.14),

$$\mathbf{E}[\hat{X}_{n,k}] = \Phi_1(\log_{p/q} n(1+\log p))\frac{n^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right). \quad (7.18)$$

*iii. For $a$ as in* (7.15)

$$\mathbf{E}[\hat{X}_{n,k}] = n + O(n^{\nu_0}), \quad (7.19)$$

*for some $\nu_0 < 1$.*

**Proof** We begin with proving part *ii* which requires a saddle point analysis. We rewrite the inverse Mellin transform with integration line at $\Re(s) = r_0$ as

$$\begin{aligned}\tilde{E}_k(z) &= \frac{-1}{2\pi}\int_{-\infty}^{\infty} z^{-(r_0+it)}\Gamma(r_0+it)(p^{-(r_0+it)} + q^{-(r_0+it)})^k dt \\ &= \frac{-1}{2\pi}\int_{-\infty}^{\infty}\Gamma(r_0+it)e^{-k((r_0+it)\frac{\log z}{k} - \log(p^{-(r_0+it)}+q^{-(r_0+it)}))}dt. \quad (7.20)\end{aligned}$$

**Step one: Saddle points' contribute to the integral estimation**

First, we are able to show those saddle points with $|t_j| > \sqrt{\log n}$ do not have a significant asymptotic contribution to the integral. To show this, we let

$$T_k(z) = \int_{|t|>\sqrt{\log n}} z^{-r_0-it}\Gamma(r_0+it)(p^{-r_0-it} + q^{-r_0-it})^k dt. \quad (7.21)$$

Since $|\Gamma(r_0+it)| = O(|t|^{r_0-\frac{1}{2}}e^{\frac{-\pi|t|}{2}})$ as $|t| \to \pm\infty$, we observe that

$$\begin{aligned}
T_k(z) &= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k \int_{\sqrt{\log n}}^{\infty} t^{r_0/2-1/2}e^{-\pi t/2}dt\right) \\
&= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k (\log n)^{r_0/4-1/4} \int_{\sqrt{\log n}}^{\infty} e^{-\pi t/2}dt\right) \\
&= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k (\log n)^{r_0/4-1/4} e^{-\pi\sqrt{\log n}/2}\right) \\
&= O\left((\log n)^{r_0/4-1/4} e^{-\pi\sqrt{\log n}/2}\right), \qquad (7.22)
\end{aligned}$$

which is very small for large $n$. Note that for $t \in (\sqrt{\log n}, \infty)$, $t^{r_0/2-1/2}$ is decreasing, and bounded above by $(\log n)^{r_0/4-1/4}$.

**Step two: Partitioning the integral**

There are now only finitely many saddle points to work with. We split the integral range into sub-intervals, each of which contains exactly one saddle point. This way, each integral has a contour traversing a single saddle-point, and we will be able to estimate the dominant contribution in each integral from a small neighborhood around the saddle point. Assuming that $j^*$ is the largest $j$ for which $\frac{2\pi j}{\log p/q} \leq \sqrt{\log n}$, we split the integral $\tilde{E}_k(z)$ as following

$$\begin{aligned}
\tilde{E}_k(z) = &-\frac{1}{2\pi}\left(\sum_{|j|<j^*} \int_{|t-t_j|\leq \frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt\right) \\
&- \frac{1}{2\pi}\int_{\frac{\pi}{\log p/q}\leq |t_j^*|<\sqrt{\log n}} \Gamma(r+it)z^{-r_0+it}(p^{-r_0-it}+q^{-r_0-it})^k dt. \qquad (7.23)
\end{aligned}$$

By the same argument as in (7.22), the second term in (7.23) is also asymptotically negligible. Therefore, we are only left with

$$\tilde{E}_k(z) = \sum_{|j|<j^*} S_j(z), \qquad (7.24)$$

where $S_j(z) = -\frac{1}{2\pi}\int_{|t-t_j|\leq \frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt.$

**Step three: Splitting the saddle contour**

For each integral $S_j$, we write the expansion of $h(t)$ about $t_j$, as follows

$$h(t) = h(t_j) + \frac{1}{2}h^{''}(t_j)(t-t_j)^2 + O((t-t_j)^3). \tag{7.25}$$

The main contribution for the integral estimate should come from an small integration path that reduces $kh(t)$ to its quadratic expansion about $t_j$. In other words, we want the integration path to be such that

$$k(t-t_j)^2 \to \infty, \qquad \text{and} \qquad k(t-t_j)^3 \to 0. \tag{7.26}$$

The above conditions are true when $|t-t_j| \gg k^{-1/2}$ and $|t-tj| \ll k^{-1/3}$. Thus, we choose the integration path to be $|t-t_j| \leq k^{-2/5}$. Therefore, we have

$$\begin{aligned} S_j(z) = &-\frac{1}{2\pi}\int_{|t-t_j|\leq k^{-2/5}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt \\ &\qquad -\frac{1}{2\pi}\int_{k^{-2/5}<|t-t_j|<\frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt. \end{aligned} \tag{7.27}$$

**Saddle Tails Pruning.**

We show that the integral is small for $k^{-2/5} < |t-t_j| < \dfrac{\pi}{\log p/q}$. We define

$$S_j^{(1)}(z) = -\frac{1}{2\pi}\int_{k^{-2/5}<|t-t_j|<\frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt. \tag{7.28}$$

Note that for $|t-t_j| \leq \dfrac{\pi}{\log p/q}$, we have

$$\begin{aligned} |p^{-r_0-it}+q^{-r_0-it}| &= (p^{-r_0}+q^{-r_0})\sqrt{1-\frac{2p^{-r_0}q^{-r_0}}{(p^{-r_0}+q^{-r_0})^2}(1-\cos(t\log p/q))} \\ &\leq (p^{-r_0}+q^{-r_0})\left(1-\frac{p^{-r_0}q^{-r_0}}{(p^{-r_0}+q^{-r_0})^2}(1-\cos(t-t_j)\log p/q)\right) \\ &\qquad\qquad \text{since}\sqrt{1-x} \leq 1-\frac{x}{2} \text{ for } x\in[0,1] \\ &\leq (p^{-r_0}+q^{-r_0})\left(1-\frac{2p^{-r_0}q^{-r_0}}{\pi^2(p^{-r_0}+q^{-r_0})^2}((t-t_j)\log p/q)^2\right) \\ &\qquad\qquad \text{since } 1-\cos x \geq \frac{2x^2}{\pi^2} \text{ for } |x|\leq\pi \\ &\leq (p^{-r_0}+q^{-r_0})e^{-(t-t_j)^2}, \end{aligned} \tag{7.29}$$

where $\gamma = \dfrac{2p^{-r_0}q^{-r_0}\log^2 p/q}{\pi^2(p^{-r_0}+q^{-r_0})^2}$. Thus,

$$\begin{aligned}
S_j^{(1)}(z) &= O\left(z^{-r_0}|\Gamma(r_0+it)|\int_{k^{-2/5}<|t-t_j|<\frac{\pi}{\log p/q}} |p^{-r_0-it}+q^{-r_0-it}|dt\right)\\
&= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k\int_{k^{-2/5}}^{\infty} e^{-\gamma k u^2}du\right)\\
&= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k k^{-3/5}e^{-\gamma k^{1/5}}\right), \text{since } \mathrm{erf}(x) = O\left(e^{-x^2}/x\right). \qquad (7.30)
\end{aligned}$$

**Central Approximation.**

Over the main path, the integrals are of the form

$$\begin{aligned}
S_j^{(0)}(z) &= -\frac{1}{2\pi}\int_{|t-t_j|\le k^{-2/5}} \Gamma(r_0+it)z^{-r_0+it}(p^{-r_0-it}+q^{-r_0-it})^k dt\\
&= -\frac{1}{2\pi}\int_{|t-t_j|\le k^{-2/5}} \Gamma(r_0+it)e^{-kh(t)}dt.
\end{aligned}$$

We have

$$h''(t_j) = \frac{\log^2 p/q}{((p/q)^{-r_0/2}+(p/q)^{r_0/2})^2}, \qquad (7.31)$$

and

$$p^{-r_0-it_j}+q^{-r_0-it_j} = p^{-it_j}(p^{-r_0}+q^{-r_0}). \qquad (7.32)$$

Therefore, by Theorem 7.2.1, we obtain

$$\begin{aligned}
S_j^{(0)}(z) &= \frac{1}{\sqrt{2\pi k h''(t_j)}}\Gamma(r_0+it_j)e^{-kh(t_j)}(1+O(k^{-1/2}))\\
&= \frac{(p/q)^{-r_0/2}+(p/q)^{r_0/2}}{\sqrt{2\pi}\log p/q}\\
&\qquad \times z^{-r_0}(p^{-r_0}+q^{-r_0})^k\Gamma(r_0+it_j)z^{-it_j}p^{-ikt_j}k^{-1/2}\left(1+O\left(\frac{1}{\sqrt{k}}\right)\right).
\end{aligned} \qquad (7.33)$$

We finally sum over all $j$ ($|j|<j^*$), and we get

$$\tilde{E}_k(z) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2\pi}\log p/q} \\ \times \sum_{|j|<j^*} z^{-r_0}(p^{-r_0} + q^{-r_0})^k \Gamma(r_0 + it_j) z^{-it_j} p^{-ikt_j} k^{-1/2} \left(1 + O\left(\frac{1}{\sqrt{k}}\right)\right). \tag{7.34}$$

We can rewrite $\tilde{E}_k(z)$ as

$$\tilde{E}_k(z) = \Phi_1((1 + a\log p)\log_{p/q} n)\frac{z^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{7.35}$$

where $\nu = -r_0 + a\log(p^{-r_0} + q^{-r_0})$, and

$$\Phi_1(x) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2a\pi}\log p/q} \sum_{|j|<j^*} \Gamma(r_0 + it_j) e^{-2\pi ijx}. \tag{7.36}$$

For part *ii*, we move the line of integration to $r_0 \in (0, \infty)$. Note that in this range, we must consider the contribution of the pole at $s = 0$. We have

$$\tilde{E}_k(z) = \mathrm{Res}_{s=0}\tilde{E}_k^*(s)z^{-s} + \int_{r_0-i\infty}^{r_0+i\infty} \tilde{E}_k^*(z)z^{-s}ds. \tag{7.37}$$

Computing the residue at $s = 0$, and following the same analysis as in part *i* for the above integral, we arrive at

$$\tilde{E}_k(z) = 2^k - \Phi_1((1 + a\log p)\log_{p/q} n)\frac{z^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right). \tag{7.38}$$

For part *iii*. of Theorem 7.3.1, we shift the line of integration to $c_0 \in (-2, -1)$, then we have

$$\begin{aligned} \tilde{E}_k(z) &= \mathrm{Res}_{s=-1}\tilde{E}_k^*(s)z^{-s} + \int_{c-i\infty}^{c+i\infty} \tilde{E}_k^*(z)z^{-s}ds \\ &= z + O\left(z^{-c_0}(p^{-c_0} + q^{-c_0})^k\right) \\ &= z^{a\log 2} + O(z^{\nu_0}), \end{aligned} \tag{7.39}$$

where $\nu_0 = -c_0 + a\log(p^{-c_0} + q^{-c_0}) < 1$.

**Step four: Asymptotic depoissonization**

To show that both conditions in (7.1.1) hold for $\tilde{E}_k(z)$, we extend the real values $z$ to complex values $z = ne^{i\theta}$, where $|\theta| < \pi/2$. To prove (7.2), we note that

$$|e^{-i\theta(r_0+it)}\Gamma(r_0+it)| = O(|t|^{r_0-1/2}e^{t\theta-\pi|t|/2}), \tag{7.40}$$

and therefore

$$\tilde{E}_k(ne^{i\theta}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\theta(r_0+it)}n^{-r_0-it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt \tag{7.41}$$

is absolutely convergent for $|\theta| < \pi/2$. The same saddle point analysis applies here and we obtain

$$|\tilde{E}_k(z)| \leq B\frac{|z^\nu|}{\sqrt{\log n}}, \tag{7.42}$$

where $B = |\Phi_1((1+a\log p)\log_{p/q} n)|$, and $\nu$ is as in 7.35. Condition (7.2) is therefore satisfied. To prove condition (7.3) We see that for a fixed $k$,

$$\begin{aligned} |\tilde{E}_k(z)e^z| &\leq \sum_{w\in\mathcal{A}^k} |e^z - e^{z(1-\mathbf{P}(w))}| \\ &\leq 2^{k+1}e^{|z|\cos(\theta)}. \end{aligned} \tag{7.43}$$

Therefore, we have

$$\mathbf{E}[\hat{X}_{n,k}] = \tilde{E}(n) + O\left(\frac{n^{\nu-1}}{\sqrt{\log n}}\right). \tag{7.44}$$

This completes the proof of Theorem 7.3.1. ■

## 7.4 On the Second Factorial Moment

We poissonize the sequence $(\mathbf{E}[(\hat{X}_{n,k})_2])_{n\geq 0}$ as well. By the analysis in (4.17),

$$\mathbf{E}[(\hat{X}_{n,k})_2] = \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left(1-(1-\mathbf{P}(w))^n-(1-\mathbf{P}(w'))^n+(1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right),$$

which gives the following poissonized form

$$\begin{aligned}
\tilde{G}(z) &= \sum_{n\geq 0} \mathbf{E}[(\hat{X}_{n,k})_2]\frac{z^n}{n!}e^{-z} \\
&= \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} 1 - e^{-\mathbf{P}(w)z} - e^{-\mathbf{P}(w')z} + e^{-(\mathbf{P}(w)+\mathbf{P}(w'))z} \\
&= \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left(1 - e^{-\mathbf{P}(w')z}\right)\left(1 - e^{-\mathbf{P}(w)z}\right) \\
&= \left(\sum_{w\in\mathcal{A}^k}\left(1 - e^{-\mathbf{P}(w)z}\right)\right)^2 - \sum_{w\in\mathcal{A}^k}\left(1 - e^{-\mathbf{P}(w)z}\right)^2 \\
&= (\tilde{E}_k(z))^2 - \sum_{w\in\mathcal{A}^k}\left(1 - e^{-\mathbf{P}(w)z}\right)^2 \\
&= (\tilde{E}_k(z))^2 - \sum_{w\in\mathcal{A}^k}\left(1 - 2e^{-\mathbf{P}(w)z} + e^{-2\mathbf{P}(w)z}\right). \qquad (7.45)
\end{aligned}$$

We show that in all ranges of $a$ the leftover sum in (7.45) has a lower order contribution to $\tilde{G}_k(z)$ compared to $(\tilde{E}_k(z))^2$. We define

$$\tilde{L}_k(z) = \sum_{w\in\mathcal{A}^k}\left(1 - 2e^{-\mathbf{P}(w)z} + e^{-2\mathbf{P}(w)z}\right). \qquad (7.46)$$

In the first range for $k$, we take the Mellin transform of $\tilde{L}_k(z)$, which is

$$\begin{aligned}
\tilde{L}_k^*(s) &= -2\Gamma(s)\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)^{-s} + \Gamma(s)\sum_{w\in\mathcal{A}^k}(2\mathbf{P}(w))^{-s} \\
&= -2\Gamma(s)(p^{-s}+q^{-s})^k + \Gamma(s)2^{-s}(p^{-s}+q^{-s})^k \\
&= \Gamma(s)(p^{-s}+q^{-s})^k(2^{-s-1}-1), \qquad (7.47)
\end{aligned}$$

and we note that the fundamental strip for this Mellin transform of is $\langle -2, 0\rangle$ as well. The inverse Mellin transform for $c \in (-2, 0)$ is

$$\begin{aligned}
\tilde{L}_k(z) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\tilde{L}_k^*(s)z^{-s}ds \\
&= \frac{1}{\pi i}\int_{c-i\infty}^{c+i\infty}\Gamma(s)(p^{-s}+q^{-s})^k(2^{-s-1}-1)z^{-s}ds \qquad (7.48)
\end{aligned}$$

We note that this range of $r_0$ corresponds to

$$\frac{2}{\log q^{-1} + \log p^{-1}} < a < \frac{p^2 + q^2}{q^2 \log q^{-1} + p^2 \log p^{-1}}. \tag{7.49}$$

The integrand in (7.48) is quite similar to the one seen in (7.10). The only difference is the extra term $2^{-s-1} - 1$. However, we notice that $2^{-s-1} - 1$ is analytic and bounded. Thus, we obtain the same saddle points with the real part as in (7.12) and the same imaginary parts in the form of $\frac{2\pi ij}{\log p/q}$, $j \in \mathbb{Z}$. Thus, the same saddle point analysis for the integral in (7.10) applies to $\tilde{L}_k(z)$ as well. We avoid repeating the similar steps, and we skip to the central approximation, where by theorem 7.2.1, we get

$$\begin{aligned}\tilde{L}_k(z) = &\frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2\pi} \log p/q} \\ &\times \sum_{|j|<j^*} z^{-r_0} (p^{-r_0} + q^{-r_0})^k (2^{-r_0-1-it_j} - 1) \\ &\times \Gamma(r_0 + it_j) z^{-it_j} p^{-ikt_j} k^{-1/2} \left(1 + O\left(\frac{1}{\sqrt{k}}\right)\right),\end{aligned} \tag{7.50}$$

which can be represented as

$$\tilde{L}_k(z) = \Phi_2((1 + a \log p) \log_{p/q} n) \frac{z^\nu}{\sqrt{\log n}} \left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{7.51}$$

where

$$\Phi_2(x) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2a\pi} \log p/q} \sum_{|j|<j^*} (2^{-r_0-1-it_j} - 1) \Gamma(r_0 + it_j) e^{-2\pi ijx}. \tag{7.52}$$

This shows that $\tilde{L}_k(z) = O\left(\frac{z^\nu}{\sqrt{\log n}}\right)$, when

$$\frac{2}{\log q^{-1} + \log p^{-1}} < a < \frac{p^2 + q^2}{q^2 \log q^{-1} + p^2 \log p^{-1}}.$$

Subsequently, for $\frac{1}{\log q^{-1}} < a < \frac{2}{\log q^{-1} + \log p^{-1}}$, we get

$$\tilde{L}_k(z) = 2^k - \Phi_2((1 + a \log p) \log_{p/q} n) \frac{z^\nu}{\sqrt{\log n}} \left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{7.53}$$

and for $\dfrac{p^2+q^2}{q^2\log q^{-1}+p^2\log p^{-1}} < a < \dfrac{1}{\log p^{-1}}$, we get

$$\tilde{L}_k(z) = O(n^2). \tag{7.54}$$

It is not difficult to see that for each range of $a$ as stated above, $\tilde{L}_k(z)$ has a lower order contribution to the asymptotic expansion of $\tilde{G}_k(z)$, compared to $(\tilde{E}_k(z))^2$. Therefore, this leads us to the following theorem.

**Theorem 7.4.1** *The second factorial moment of the kth Prefix Complexity has the following asymptotic expansion.*

*i. For $a$ as in* (7.13),

$$\mathbf{E}[(\hat{X}_{n,k})_2] = \left(2^k - \Phi_1(\log_{p/q} n(1+\log p))\frac{n^\nu}{\sqrt{\log n}}\left(1+O\left(\frac{1}{\sqrt{\log n}}\right)\right)\right)^2.$$

*ii. For $a$ as in* (7.14),

$$\mathbf{E}[(\hat{X}_{n,k})_2] = \Phi_1^2(\log_{p/q} n(1+\log p))\frac{n^{2\nu}}{\log n}\left(1+O\left(\frac{1}{\log n}\right)\right). \tag{7.55}$$

*ii. For $a$ as in* (7.15),

$$\mathbf{E}[(\hat{X}_{n,k})_2] = n^2 + O(n^{2\nu_0}). \tag{7.56}$$

**Proof** It is only left to show that the two depoissonization conditions hold: For condition (7.2) in Theorem 7.1.1, from (7.42) we have

$$|\tilde{G}_k(z)| \le B^2\frac{|z^{2\nu}|}{\log n}, \tag{7.57}$$

and for condition (7.3), we have, for fixed $k$,

$$\begin{aligned} |\tilde{G}_k(z)e^z| &\le \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \left| e^z - e^{(1-\mathbf{P}(w))z} - e^{(1-\mathbf{P}(w'))z} + e^{(1-(\mathbf{P}(w)+\mathbf{P}(w')))z}\right| \\ &\le 4^k e^{|z|\cos\theta}. \end{aligned} \tag{7.58}$$

Therefore both depoissonization conditions are satisfied and the desired result follows.

■

### 7.5 A Remark on the Second Moment and the Variance

For the second moment we have

$$\begin{aligned}\mathbf{E}\left[(\hat{X}_{n,k})^2\right] &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \mathbf{E}\left[\hat{X}_{n,k}^{(w)}\hat{X}_{n,k}^{(w')}\right] + \sum_{w\in\mathcal{A}^k} \mathbf{E}[\hat{X}_{n,k}^{(w)}]\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \left(1-(1-\mathbf{P}(w))^n-(1-\mathbf{P}(w'))^n+(1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right)\\ &\qquad + \sum_{w\in\mathcal{A}^k}\left(1-(1-\mathbf{P}(w))^n\right). \end{aligned} \tag{7.59}$$

Therefore, by (7.8) and (7.45) the Poisson transform of the second moment, which we denote by $\tilde{G}_k^{(2)}(z)$ is

$$\tilde{G}_k^{(2)}(z) = (\tilde{E}_k(z))^2 + \tilde{E}_k(z) - \sum_{w\in\mathcal{A}^k}\left(1-2e^{-\mathbf{P}(w)z}+e^{-2\mathbf{P}(w)z}\right), \tag{7.60}$$

which results in the same first order asymptotic as the second factorial moment. Also, it is not difficult to extend the proof in Chapter 6 to show that the second moments of the two models are asymptotically the same. For the variance we have

$$\begin{aligned}\mathrm{Var}[\hat{X}_{n,k}] &= \mathbf{E}\left[(\hat{X}_{n,k})^2\right] - \left(\mathbf{E}\left[\hat{X}_{n,k}\right]\right)^2\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \left(1-(1-\mathbf{P}(w))^n-(1-\mathbf{P}(w'))^n+(1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right)\\ &\qquad + \sum_{w\in\mathcal{A}^k}\left(1-(1-\mathbf{P}(w))^n\right)\\ &- \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \left(1-(1-\mathbf{P}(w))^n-(1-\mathbf{P}(w'))^n+(1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right)\\ &\qquad - \sum_{w\in\mathcal{A}^k}\left(1-(1-\mathbf{P}(w))^n-(1-\mathbf{P}(w))^n+(1-\mathbf{P}(w))^{2n}\right)\\ &= \sum_{w\in\mathcal{A}^k}\left((1-\mathbf{P}(w))^n-(1-\mathbf{P}(w))^{2n}\right). \end{aligned} \tag{7.61}$$

Therefore the Poisson transform, which we denote by $\tilde{G}_k^{\mathrm{var}}(z)$ is

$$\tilde{G}_k^{\mathrm{var}}(z) = \sum_{w \in \mathcal{A}^k} \left( e^{-\mathbf{P}(w)z} - e^{-(2\mathbf{P}(w)+(\mathbf{P}(w))^2)z} \right). \tag{7.62}$$

The Mellin transform of the above function has the following form

$$\tilde{G^*}_k^{\mathrm{var}}(z) = \Gamma(s)(p^{-s} + q^{-s})^k(-1 + O(\mathbf{P}(w))). \tag{7.63}$$

This is quite similar to what we saw in (7.9), which indicates that the variance has the same asymptotic growth as the expected value. But the variance of the two models do not behave in the same way (cf. Figure 9.4).

# 8. SUMMARY

We studied the first order asymptotic growth of the first two (factorial) moments of the $k$th Subwoed Complexity. We recall that the $k$th Subword Complexity of a string of length $n$ is denoted by $X_{n,k}$, and is defined as the number of distinct subwords of length $k$, that appear in the string. We are interested in the asymptotic analysis for when $k$ grows as a function of the string's length. More specifically, we conduct the analysis for $k = \Theta(\log n)$, and as $n \to \infty$.

The analysis is inspired by the earlier work of Jacquet and Szpankowski on the analysis of suffix trees, where they are compared to independent tries (cf. [14]). In our work, we compare the first two moments of the $k$th Subword Complexity to the $k$th Prefix Complexity over a random trie built over $n$ independently generated binary strings. We recall that we define the $k$th Prefix Complexity as the number of distinct prefixes that appear in the trie at level $k$ and lower.

We obtain the generating functions representing the expected value and the second factorial moments as their coefficients, in both settings. We prove that the first two moments have the same asymptotic growth in both models. For deriving the asymptotic behavior, we split the range for $k$ into three intervals. We analyze each range using the saddle point method, in combination with residue analysis. We close our work with some remarks regarding the comparison of the second moment and the variance to the $k$th Prefix Complexity.

# 9. FIGURES

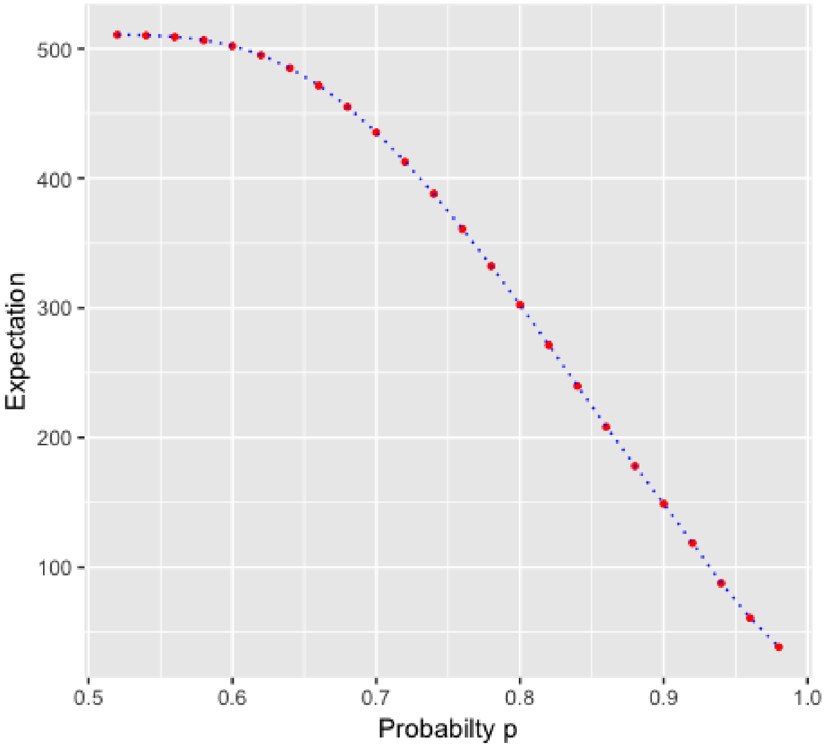


Figure 9.1. Approximated expectations of the $k$th Subword Complexity (red), and the $k$th Prefix Complexity (blue), for n=4000, at different probability levels, averaged over 10,000 iterations.

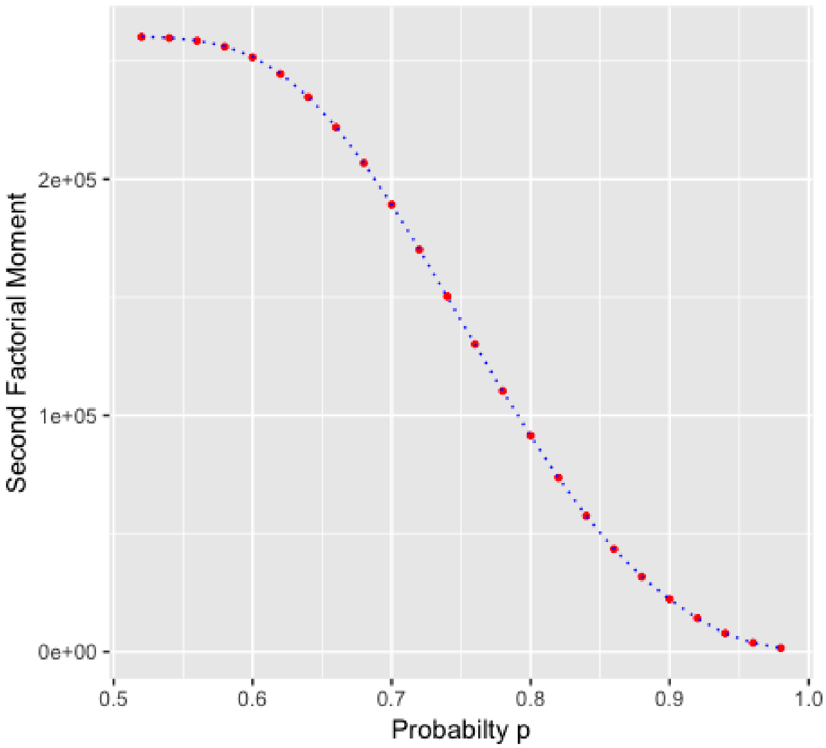


Figure 9.2. Approximated second factorial moments of the $k$th Subword Complexity (red), and the $k$th Prefix Complexity (blue), for n=4000, at different probability levels, averaged over 10,000 iterations.

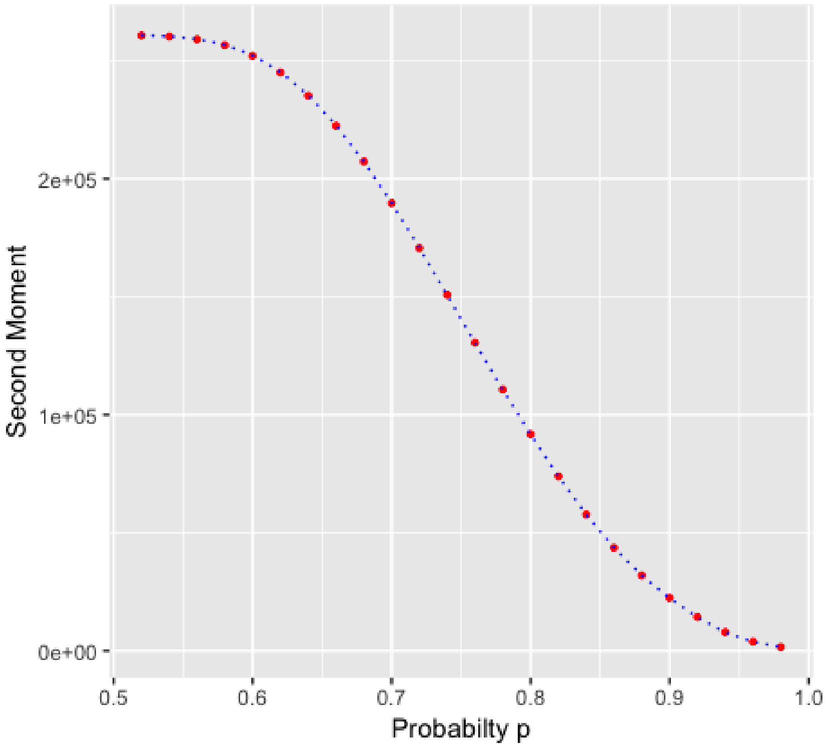


Figure 9.3. Approximated second moments of the $k$th Subword Complexity (red), and the $k$th Prefix Complexity (blue), for n=4000, at different probability levels, averaged over 10,000 iterations.

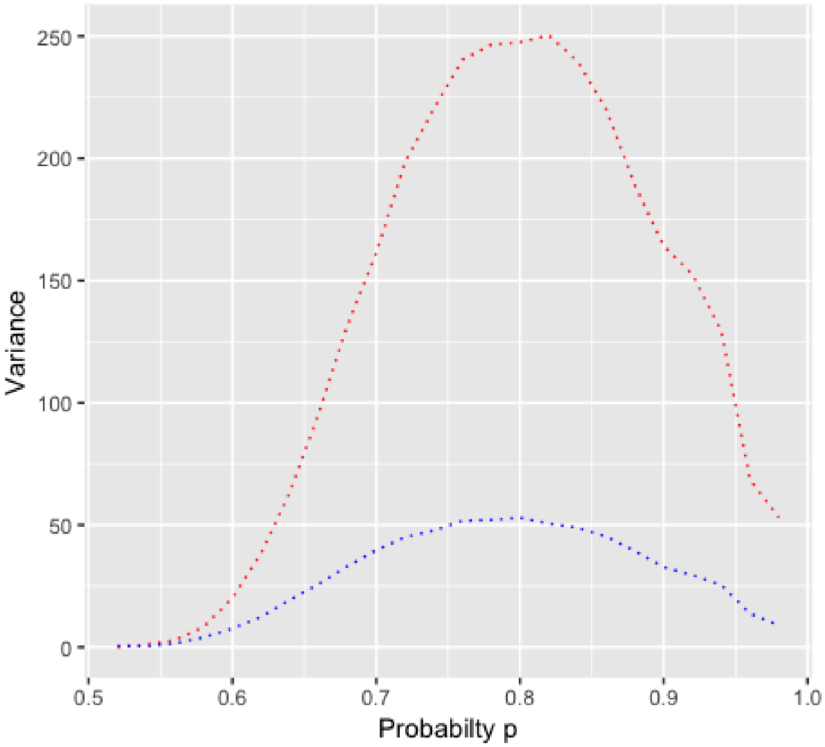


Figure 9.4. Approximated variances of the *k*th Subword Complexity (red), and the *k*th Prefix Complexity (blue), for n=4000, at different probability levels, averaged over 10,000 iterations.

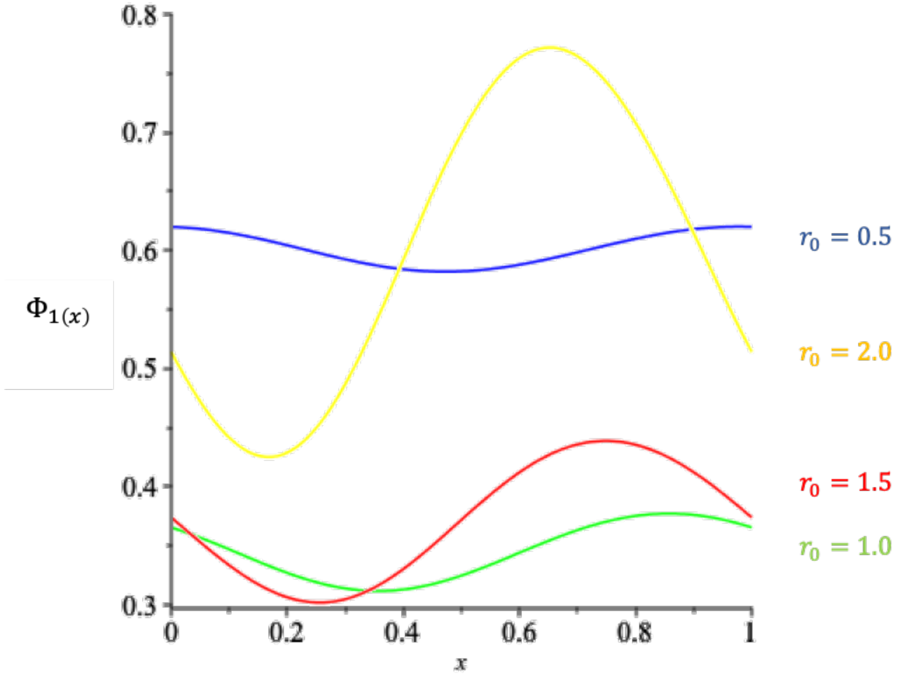


Figure 9.5. $\Phi_1(x)$ at $p = 0.90$, and various levels of $r_0$. The amplitude increases as $r_0$ increases.

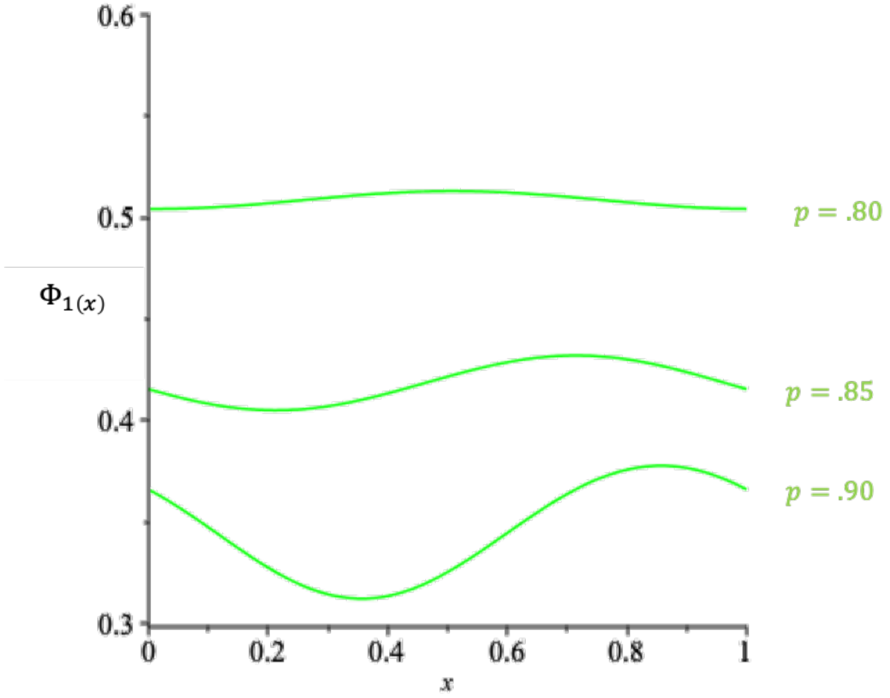


Figure 9.6. $\Phi_1(x)$ at $r_0 = 1$, and various levels of $p$. The amplitude tends to zero as $p \to 1/2^+$.

# REFERENCES

[1] A. Ehrenfeucht, K. Lee, and G. Rozenberg, "Subword complexities of various classes of deterministic developmental languages without interactions," *Theoretical Computer Science*, vol. 1, no. 1, pp. 59–75, 1975.

[2] M. Morse and G. A. Hedlund, "Symbolic Dynamics," *American Journal of Mathematics*, vol. 60, no. 4, pp. 815–866, 1938.

[3] P. Jacquet and W. Szpankowski, *Analytic Pattern Matching: From DNA to Twitter*. Cambridge University Press, 2015.

[4] T. C. Bell, J. G. Cleary, and I. H. Witten, *Text Compression*. Prentice-Hall, Inc., 1990.

[5] C. Burge, A. M. Campbell, and S. Karlin, "Over-and under-representation of short oligonucleotides in dna sequences," *Proceedings of the National Academy of Sciences*, vol. 89, no. 4, pp. 1358–1362, 1992.

[6] J. W. Fickett, D. C. Torney, and D. R. Wolf, "Base compositional structure of genomes," *Genomics*, vol. 13, no. 4, pp. 1056–1064, 1992.

[7] S. Karlin, C. Burge, and A. M. Campbell, "Statistical analyses of counts and distributions of restriction sites in DNA sequences," *Nucleic Acids Research*, vol. 20, no. 6, pp. 1363–1370, 1992.

[8] S. Karlin, J. Mrázek, and A. M. Campbell, "Frequent Oligonucleotides and Peptides of the Haemophilus Influenzae Genome," *Nucleic Acids Research*, vol. 24, no. 21, pp. 4263–4272, 1996.

[9] P. A. Pevzner, M. Y. Borodovsky, and A. A. Mironov, "Linguistics of Nucleotide Sequences ii: Stationary Words in Genetic Texts and the Zonal Structure of DNA," *Journal of Biomolecular Structure and Dynamics*, vol. 6, no. 5, pp. 1027–1038, 1989.

[10] X. Chen, B. Francia, M. Li, B. Mckinnon, and A. Seker, "Shared information and program plagiarism detection," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1545–1551, 2004.

[11] B. Chor, D. Horn, N. Goldman, Y. Levy, and T. Massingham, "Genomic DNA k-mer spectra: models and modalities," *Genome Biology*, vol. 10, no. 10, p. R108, 2009.

[12] A. L. Price, N. C. Jones, and P. A. Pevzner, "De novo identification of repeat families in large genomes," *Bioinformatics*, vol. 21, no. suppl_1, pp. i351–i358, 2005.

[13] S. Janson, S. Lonardi, and W. Szpankowski, "On the Average Sequence Complexity," in *Annual Symposium on Combinatorial Pattern Matching*. Springer, 2004, pp. 74–88.

[14] 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*, vol. 66, no. 2, pp. 237–269, 1994.

[15] F. M. Liang, "Word hy-phen-a-tion by com-put-er," Calif. Univ. Stanford. Comput. Sci. Dept., Tech. Rep., 1983.

[16] P. Weiner, "Linear pattern matching algorithms," in *14th Annual Symposium on Switching and Automata Theory (swat 1973).* IEEE, 1973, pp. 1–11.

[17] I. Gheorghiciuc and M. D. Ward, "On correlation Polynomials and Subword Complexity," in *Discrete Mathematics and Theoretical Computer Science.* Discrete Mathematics and Theoretical Computer Science, 2007, pp. 1–18.

[18] F. Bassino, J. Clément, and P. Nicodème, "Counting occurrences for a finite set of words: Combinatorial methods," *ACM Transactions on Algorithms (TALG)*, vol. 8, no. 3, p. 31, 2012.

[19] G. Park, H.-K. Hwang, P. Nicodème, and W. Szpankowski, "Profile of Tries," in *Latin American Symposium on Theoretical Informatics.* Springer, 2008, pp. 1–11.

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

[21] P. Flajolet, X. Gourdon, and C. Martínez, "Patterns in random binary search trees," *Random Structures & Algorithms*, vol. 11, no. 3, pp. 223–244, 1997.

[22] A. Bertoni, C. Choffrut, M. Goldwurm, and V. Lonati, "On the number of occurrences of a symbol in words of regular languages," *Theoretical Computer Science*, vol. 302, no. 1-3, pp. 431–456, 2003.

[23] P. Flajolet, M. Régnier, and R. Sedgewick, "Some uses of the Mellin integral transform in the analysis of algorithms," in *Combinatorial Algorithms on Words.* Springer, 1985, pp. 241–254.

[24] M. Lothaire, *Combinatorics on words.* Cambridge University Press, 1997, vol. 17.

[25] A. M. Odlyzko, "Asymptotic Enumeration Methods," *Handbook of Combinatorics*, vol. 2, no. 1063, p. 1229, 1995.

[26] P. Flajolet and A. Odlyzko, "Singularity Analysis of Generating Functions," *SIAM Journal on Discrete Mathematics*, vol. 3, no. 2, pp. 216–240, 1990.

[27] P. Jacquet and W. Szpankowski, "Asymptotic behavior of the Lempel-Ziv parsing scheme and digital search trees," *Theoretical Computer Science*, vol. 144, no. 1-2, pp. 161–197, 1995.

[28] B. Rais, P. Jacquet, and W. Szpankowski, "Limiting Distribution for the Depth in PATRICIA Tries," *SIAM Journal on Discrete Mathematics*, vol. 6, no. 2, pp. 197–213, 1993.

[29] B. L. van der Waerden, "On the method of saddle points," *Applied Scientific Research, Section B*, vol. 2, no. 1, pp. 33–45, 1952.

[30] M. Lothaire, *Applied Combinatorics on Words.* Cambridge University Press, 2005, vol. 105.

[31] M. Régnier and W. Szpankowski, *On the approximate pattern occurrences in a text.* IEEE, 1997.

[32] J. Fayolle and M. D. Ward, "Analysis of the average depth in a suffix tree under a Markov model," in *International Conference on Analysis of Algorithms DMTCS proc. AD*, vol. 95, 2005, p. 104.

[33] L. V. Ahlfors, "Complex Analysis. 1979," 1973.

[34] P. Flajolet, X. Gourdon, and P. Dumas, "Mellin transforms and asymptotics: Harmonic sums," *Theoretical Computer Science*, vol. 144, no. 1-2, pp. 3–58, 1995.

[35] J. Bertrand, P. Bertrand, and J.-P. Ovarlez, "The Mellin transform," *ONERA, TP no. 1994-98*, 1994.

[36] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences.* John Wiley & Sons, 2011, vol. 50.

[37] D. V. Widder, *The Laplace Transform (PMS-6).* Princeton University Press, 2015.

[38] W. Szpankowski, "On the height of digital trees and related problems," *Algorithmica*, vol. 6, no. 1-6, pp. 256–277, 1991.

[39] L. Devroye, W. Szpankowski, and B. Rais, "A Note on the Height of Suffix Trees," *SIAM Journal on Computing*, vol. 21, no. 1, pp. 48–53, 1992.

[40] H. Mahmoud and W. Szpankowski, "An Analytic Approach for the Asymptotic Distribution of the Height of an Incomplete Digital Tree," 1994.

[41] M. D. Ward and W. Szpankowski, "Analysis of Randomized Selection Algorithm Motivated by the LZ'77 Scheme," in *ALENEX/ANALCO*, 2004, pp. 153–160.

[42] M. D. Ward, "The average Profile of Suffix Trees," in *2007 Proceedings of the Fourth Workshop on Analytic Algorithmics and Combinatorics (ANALCO).* SIAM, 2007, pp. 183–193.

[43] R. De La Briandais, "File searching using variable length keys," in *Papers presented at the the March 3-5, 1959, Western Joint Computer Conference.* ACM, 1959, pp. 295–298.

[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons, 2012.

# VITA

Lida Ahmadi was born on May 22, 1989 in Tehran, Iran. Throughout her school years, she always enjoyed math classes. After finishing high school, she was very excited about being admitted to the mathematics program at University of Tehran. There, Lida's interest to advance her studies in mathematics grew and she decided to obtain her Ph.D. degree abroad. She received her undergraduate degree in January 2012, and enrolled for the graduate program at Purdue University in Fall 2012.

In the third year of her Ph.D., Lida was introduced to her current research area, Analytic Combinatorics and Probability, through a talk held by her advisor Dr. Mark Daniel Ward. She then decided to attended Dr. Ward's seminar classes and she has been enjoying working in this filed of research since. Along with her research, Lida has been teaching a variety of mathematics and statistics classes at Purdue University. She has also obtained a graduate certificate in Applied Statistics from Purdue University to supplement her teaching and research.

Lida will finish her Ph.D. in August 2019, and will join California State University, San Bernardino in September 2019, as an Assistant Professor of Statistics.