entropy

MDPI

Article

Asymptotic Analysis of the kth Subword Complexity

Lida Ahmadi [1,†,*] and Mark Daniel Ward [2]

[1] Department of Mathematics, Purdue University, West Lafayette 47907, IN, USA
[2] Department of Statistics, Purdue University, West Lafayette 47907, IN, USA; mdw@purdue.edu
* Correspondence: lida.ahmadi@csusb.edu
† Current address: 5500 University Parkway, San Bernardino 92407, CA, USA

Received: 25 December 2019; Accepted: 4 February 2020; Published: 12 February 2020

Abstract: Patterns within strings enable us to extract vital information regarding a string's randomness. Understanding whether a string is random (Showing no to little repetition in patterns) or periodic (showing repetitions in patterns) are described by a value that is called the kth Subword Complexity of the character string. By definition, the kth Subword Complexity is the number of distinct substrings of length k that appear in a given string. In this paper, we evaluate the expected value and the second factorial moment (followed by a corollary on the second moment) of the kth Subword Complexity for the binary strings over memory-less 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 kth 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..

Keywords: subword complexity; asymptotics; generating functions; saddle point method; probability; the Mellin transform; moments.

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 [5–9]), and plagiarism detection [10]. For example, in data compression, a data set is considered compressible if it has low complexity, as consists of repeated subwords. In computational genomics, Subword Complexity (known as k-mers) is used in detection of repeated sequences and DNA barcoding [11,12]. k-mers are composed of A, T, G, and C nucleotides. For instance, 7-mers for a DNA sequence GTAGAGCTGT is four, meaning that there are 4-hour distinct

substrings of length 7 in the given DNA sequence. Counting k-mers becomes challenging for longer DNA sequences. Our results can be easily extended to the alphabet $\{A, T, G, C\}$ and directly applied in theoretical analysis of the genomic k-mer distributions under the Bernoulli probabilistic model, particularly when the length n of the sequence approaches infinity.

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 kth Subword Complexity to avoid any confusion.

Throughout this work, we consider the kth 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 $X_{n,k}$ (1) for the range $k = \Theta(\log n)$, as $n \to \infty$. More precisely, will divide the analysis into three ranges as follows.

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

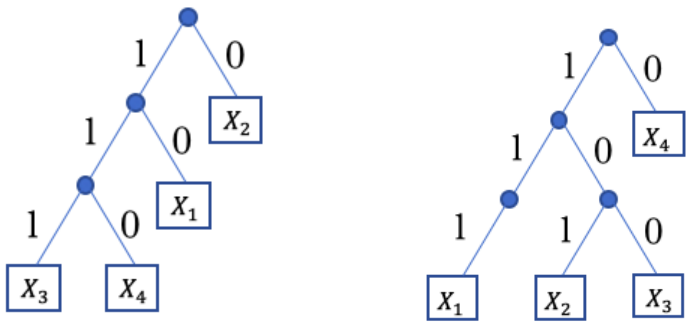
ii. $\dfrac{2}{\log q^{-1} + \log p^{-1}} \log n < k < \dfrac{1}{q \log q^{-1} + p \log p^{-1}} \log n$, and

iii. $\dfrac{1}{q \log q^{-1} + p \log p^{-1}} \log n < k < \dfrac{1}{\log p^{-1}} \log n$,

Our approach involves two major steps. 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.

1.1. 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.



(**a**)Suffix Tree (**b**)Trie

Figure 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, as 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. Autocorrelation 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 kth Subword Complexity. In the independent model, the generating functions are obtained by basic probability concepts.

1.2. 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.

2. Results

For a binary string $X = X_1X_2...X_n$, where X_i's ($i = 1, ..., n$) are independent and identically distributed random variables , we assume that $\mathbf{P}(X_i = 1) = p$, $\mathbf{P}(X_i = 0) = q = 1 - p$, and $p > q$. We define the kth Subword Complexity, $X_{n,k}$, to be the number of distinct substrings of length k that appear in a random string X with the above assumptions. In this work, we obtain the first order asymptotics for the average and the second factorial moment of $X_{n,k}$. The analysis is done in the range $k = \Theta(\log n)$. We rewrite this range as $k = a \log n$, and by performing a saddle point analysis, we will show that

$$1/\log q^{-1} < a < 1/\log p^{-1} \tag{1}$$

In the first step, we compare the kth 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. This means that each string is a sequence of independent and identically distributed Bernoulli random variables from the binary alphabet $\mathcal{A} = \{0, 1\}$, with $\mathbf{P}(1) = p$, $\mathbf{P}(0) = q = 1 - p$. We denote the number of distinct prefixes of length k in the trie by $\hat{X}_{n,k}$, and we call it *the kth prefix complexity*.

Before proceeding any further, we remind that factorial moments of a random variable are defined as following.

Definition 1. *The jth factorial moment of a random variable X is defined as*

$$\mathbf{E}[(X)_j] = \mathbf{E}[(X)(X-1)(X-2)...(X-j+1)], \tag{2}$$

where j=1,2,... .

will show that the first and second factorial moments of $X_{n,k}$ are asymptotically comparable to those of $\hat{X}_{n,k}$, when $k = \Theta(\log n)$. We have the following theorems.

Theorem 1. *For large values of n, and for $k = \Theta(\log n)$, there exists $M > 0$ such that*

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

We also prove a similar result for the second factorial moments of the kth Subword Complexity and the kth Prefix Complexity:

Theorem 2. *For large values of n, and for $k = \Theta(\log n)$, there exists $\epsilon > 0$ such that*

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

In the second part of our analysis, we derive the first order asymptotics of the kth Prefix Complexity. The methodology used here is analogous to the analysis of profile of tries [19]. The rate of the asymptotic growth depends on the location of the value a as seen in (1) . For instance, for the average kth Subword Complexity ,$\mathbf{E}[X_{n,k}]$, we have the following observations.

i. For the range $I_1 : \dfrac{1}{\log q^{-1}} < a < \dfrac{2}{\log q^{-1} + \log p^{-1}}$, the growth rate is of order $O(2^k)$,

ii. in the range $I_2 : \dfrac{2}{\log q^{-1} + \log p^{-1}} < a < \dfrac{1}{q \log q^{-1} + p \log p^{-1}}$, we observe some oscillations with n, and

iii. in the range $I_3 : \dfrac{1}{q \log q^{-1} + p \log p^{-1}} < a < \dfrac{1}{\log p^{-1}}$, the average has a linear growth $O(n)$.

The above observations will be discussed in depth in the proofs of the following theorems.

Theorem 3. *The average of the kth Prefix Complexity has the following asymptotic expansion*

i. For $a \in I_1$,

$$\mathbf{E}[\hat{X}_{n,k}] = 2^k - \Phi_1((1 + \log p) \log_{p/q} n) \frac{n^\nu}{\sqrt{\log n}} \left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{3}$$

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

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

is a bounded periodic function.

ii. For $a \in I_2$,

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

iii. For $a \in I_3$

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

for some $\nu_0 < 1$.

Theorem 4. *The second factorial moment of the kth Prefix Complexity has the following asymptotic expansion.*

i. For $a \in I_1$,

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

ii. For $a \in I_2$,

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

ii. For $a \in I_3$,

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

The periodic function $\Phi_1(x)$ in Theorems 3, and 4 is shown in Figure 2.

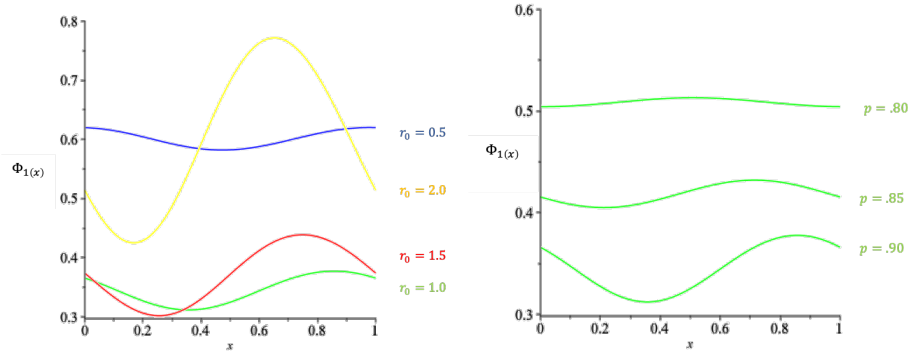


Figure 2. Left: $\Phi_1(x)$ at $p = 0.90$, and various levels of r_0. The amplitude increases as r_0 increases. Right: $\Phi_1(x)$ at $r_0 = 1$, and various levels of p. The amplitude tends to zero as $p \to 1/2^+$.

The results in Theorem 4 will follow for the second moment of the kth Subword Complexity as the analysis can be easily extended from the second factorial moment to the second moment. The variance however, as seen in Figure 3, does not show the same asymptotic behavior as the variance of kth Subword Complexity.

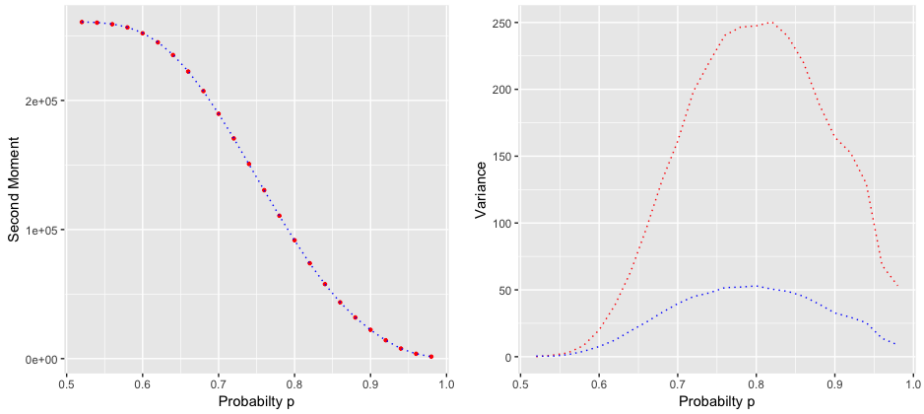


Figure 3. Approximated second moments (**left**), and variances (**right**) of the kth Subword Complexity (**red**), and the kth Prefix Complexity (**blue**), for n=4000, at different probability levels, averaged over 10,000 iterations.

3. Proofs and Methods

3.1. Groundwork

We first introduce a few terminologies and lemmas regarding overlaps of patterns and their number of occurrences in texts. Some of the notations we use in this work are borrowed from [18] and [21].

Definition 2. *For a binary word $w = w_1...w_k$ of length k, The autocorrelation set $\mathcal{S}_w$ of the word w is defined in the following way.*

$$\mathcal{S}_w = \{w_{i+1}..w_k \,|\, w_1...w_i = w_{k-i+1}..w_k\}. \tag{4}$$

The autocorrelation index set is

$$\mathcal{P}(w) = \{i \,|\, w_1...w_i = w_{k-i+1}..w_k\}, \tag{5}$$

And the autocorrelation polynomial is

$$S_w(z) = \sum_{i \in \mathcal{P}(w)} \mathbf{P}(w_{i+1}...w_k) z^{k-i}. \tag{6}$$

Definition 3. *For the distinct binary words $w = w_1...w_k$ and $w' = w'_1...w'_k$, the correlation set $\mathcal{S}_{w,w'}$ of the words w and w' is*

$$\mathcal{S}_{w,w'} = \{w'_{i+1}...w'_k \,|\, w'_1...w'_i = w_{k-i+1}...w_k\}. \tag{7}$$

The correlation index set is

$$\mathcal{P}(w, w') = \{i \,|\, w'_1...w'_i = w_{k-i+1}..w_k\}, \tag{8}$$

The correlation polynomial is

$$S_{w,w'}(z) = \sum_{i \in \mathcal{P}(w,w')} \mathbf{P}(w'_{i+1}...w'_k) z^{k-i}. \tag{9}$$

The following two lemmas present the probability generating functions for the number of occurrences of a single pattern and a pair of distinct pattern, respectively, in a random text of length n. For a detailed dissection on obtaining such generating functions, refer to [18].

Lemma 1. *The Occurrence probability generating function for a single pattern w in a binary text over a memoryless source is given by $F_w(z, x-1)$, where*

$$F_w(z,t) = \frac{1}{1 - A(z) - \dfrac{t\mathbf{P}(w)z^k}{1 - t(S_w(z) - 1)}}, \tag{10}$$

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.

Lemma 2. *The Occurrence PGF for two distinct Patterns of length k in a Bernoulli random text is given by $F_{w,w'}(z, x_1 - 1, x_2 - 1)$ where,*

$$F_{w,w'}(z,t_1,t_2) = \frac{1}{1 - A(z) - M(z,t_1,t_2)}, \tag{11}$$

and

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

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.

The above results will be used to find the generating functions for the first two factorial moments of the kth Subword Complexity in the following section.

3.2. Derivation of Generating Functions

Lemma 3. *For generating functions $H_k(z) = \sum_{n\geq 0} \mathbf{E}[X_{n,k}]z^n$ and $G_k(z) = \sum_{n\geq 0} \mathbf{E}[(X_{n,k})_2]z^n$, we have*

i.

$$H_k(z) = \sum_{w\in\mathcal{A}^k} \left(\frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} \right), \tag{12}$$

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

ii.

$$G_k(z) = \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left(\frac{1}{1-z} - \frac{S_w(z)}{D_w(z)} - \frac{S_{w'}(z)}{D_{w'}(z)} + \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)} \right), \tag{13}$$

where

$$\begin{aligned} D_{w,w'}(z) = {} & (1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)) \\ & + z^k \left(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z)) \right). \end{aligned} \tag{14}$$

Proof. *i*. We define

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

This yields

$$\begin{aligned} \mathbf{E}[X_{n,k}^{(w)}] &= \mathbf{P}(X_{n,k}^{(w)} = 1) \\ &= 1 - P(X_{n,k}^{(w)} = 0) \\ &= 1 - [z^n x^0] F_w(z, x). \end{aligned} \tag{15}$$

We observe that $[z^n x^0] F_w(z, x) = [z^n] F_w(z, 0)$. By defining $f_w(z) = F_w(z, 0)$ and from (10), we obtain

$$f_w(z) = \frac{S_w(z)}{\mathbf{P}(w) z^k + (1 - z) S_w(z)}. \tag{16}$$

Having the above function, we derive the following result.

$$\begin{aligned} H(z) &= \sum_{n \geq 0} \mathbf{E}[X_{n,k}] z^n \\ &= \sum_{n \geq 0} \sum_{w \in \mathcal{A}^k} (1 - [z^n] f_w(z)) z^n \\ &= \sum_{w \in \mathcal{A}^k} \left(\frac{1}{1 - z} - f_w(z) \right) \\ &= \sum_{w \in \mathcal{A}^k} \left(\frac{1}{1 - z} - \frac{S_w(z)}{D_w(z)} \right). \end{aligned} \tag{17}$$

ii. For this part, we first note that

$$\begin{aligned} \mathbf{E}[(X_{n,k})_2] &= \mathbf{E}[X_{n,k}^2] - \mathbf{E}[X_{n,k}] \\ &= \mathbf{E}\left[(X_{n,k}^{(w)} + ... + X_{n,k}^{(w^{(r)})})^2 \right] - \mathbf{E}\left[X_{n,k}^{(w)} + ... + X_{n,k}^{(w^{(r)})} \right] \\ &= \sum_{w \in \mathcal{A}^k} \mathbf{E}\left[(X_{n,k}^{(w)})^2 \right] + \sum_{\substack{w, w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}\left[X_{n,k}^{(w)} X_{n,k}^{(w')} \right] - \sum_{w \in \mathcal{A}^k} \mathbf{E}\left[X_{n,k}^{(w)} \right] \\ &= \sum_{\substack{w, w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}\left[X_{n,k}^{(w)} X_{n,k}^{(w')} \right]. \end{aligned} \tag{18}$$

Due to properties of indicator random variables, we observe that the expected value of the second factorial moment has only one term:

$$\mathbf{E}[(X_{n,k})_2] = \sum_{\substack{w, w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}\left[X_{n,k}^{(w)} X_{n,k}^{(w')} \right]. \tag{19}$$

We proceed by defining a second indicator variable as following.

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

This gives

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

Finally, we are able to express $\mathbf{E}[(X_{n,k})_2]$ in the following

$$\mathbf{E}[(X_{n,k})_2] = \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} \left(1-[z^n]\,f_w(z)-[z^n]f_{w'}(z)+[z^n]f_{ww'}(z)\right), \tag{20}$$

where $f_{w,w'}(z)=F_{w,w'}(z,0,0)$ and $[z^n]F_{w,w'}(z,0,0)=[z^n x_1^0 x_2^0]F_{w,w'}(z,x_1,x_2)$. By (11) we have

$$f_{w,w'}(z) = \frac{S_w(z)S_{w'}(z)-S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)} \tag{21}$$

Having the above expression, we finally obtain

$$\begin{aligned}G_k(z) &= \sum_{n\geq 0}\mathbf{E}[(X_{n,k})_2]z^n\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\sum_{n\geq 0}\left(1-[z^n]f_w(z)-[z^n]f_{w'}(z)+[z^n]f_{w,w'}(z)\right)z^n\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{1}{1-z}-f_w(z)-f_{w'}(z)+f_{w,w'}(z)\right)\\ &= \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{1}{1-z}-\frac{S_w(z)}{D_w(z)}-\frac{S_{w'}(z)}{D_{w'}(z)}+\frac{S_w(z)S_{w'}(z)-S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}\right).\end{aligned} \tag{22}$$

□

In the following lemma, we present the generating functions for the first two factorial moments for the kth Prefix Complexity in the independent model.

Lemma 4. *For $\hat{H}_k(z)=\sum_{n\geq 0}\mathbf{E}[\hat{X}_{n,k}]z^n$ and $\hat{G}_k(z)=\sum_{n\geq 0}\mathbf{E}[(\hat{X}_{n,k})_2]z^n$, which are the generating functions for $\mathbf{E}[\hat{X}_{n,k}]$ and $\mathbf{E}[(\hat{X}_{n,k})_2]$ respectively, we have*

i.

$$\hat{H}_k(z) = \sum_{w\in\mathcal{A}^k}\left(\frac{1}{1-z}-\frac{1}{1-(1-\mathbf{P}(w))z}\right). \tag{23}$$

ii.

$$\begin{aligned}\hat{G}_k(z) = &\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{1}{1-z}-\frac{1}{1-(1-\mathbf{P}(w))z}-\frac{1}{1-(1-\mathbf{P}(w'))z}\right)\\ &+\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z}.\end{aligned} \tag{24}$$

Proof. *i*. We define the indicator variable $\hat{X}_{n,k}^{(w)}$ as follows.

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

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

$$\begin{aligned} \mathbf{E}[\hat{X}_{n,k}^{(w)}] &= \mathbf{P}(\hat{X}_{n,k}^{(w)} = 1) \\ &= 1 - P(\hat{X}_{n,k}^{(w)} = 0) \\ &= 1 - (1 - \mathbf{P}(w))^n . \end{aligned} \tag{25}$$

Summing over all words w of length k, determines the generating function $\hat{H}(z)$:

$$\begin{aligned} \hat{H}(z) &= \sum_{n \geq 0} \mathbf{E}[\hat{X}_{n,k}] z^n \\ &= \sum_{w \in \mathcal{A}^k} \left(\frac{1}{1-z} - \frac{1}{1 - (1 - \mathbf{P}(w))z} \right). \end{aligned} \tag{26}$$

ii. Similar to in (18) and (20), we obtain

$$\begin{aligned} \mathbf{E}[(\hat{X}_{n,k})_2] &= \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \mathbf{E}[\hat{X}_{n,k}^{(w)} \hat{X}_{n,k}^{(w')}] \\ &= \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left(1 - (1 - \mathbf{P}(w))^n - (1 - \mathbf{P}(w'))^n + (1 - \mathbf{P}(w) - \mathbf{P}(w'))^n \right). \end{aligned} \tag{27}$$

Subsequently, we obtain the generating function below.

$$\begin{aligned} \hat{G}(z) &= \sum_{n \geq 0} \mathbf{E}[(\hat{X}_{n,k})_2] z^n \\ &= \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \sum_{n \geq 0} \left(1 - (1 - \mathbf{P}(w))^n - (1 - \mathbf{P}(w'))^n + (1 - \mathbf{P}(w) - \mathbf{P}(w'))^n \right) z^n \\ &= \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left(\frac{1}{1-z} - \frac{1}{1 - (1 - \mathbf{P}(w))z} - \frac{1}{1 - (1 - \mathbf{P}(w'))z} \right) \\ &\quad + \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \frac{1}{1 - (1 - \mathbf{P}(w) - \mathbf{P}(w'))z}. \end{aligned} \tag{28}$$

□

Our first goal is to compare the coefficients of the generating functions in the two models. The coefficients are expected to be asymptotically equivalent in the desired range for k. To compare the coefficients, we need more information on the analytic properties of these generating functions. This will be discussed in Section 3.3.

3.3. Analytic Properties of the Generating Functions

Here, we turn our attention to the smallest singularities of the two generating functions given in Lemma 3. It has been shown by Jacquet and Szpankowski [21] that $D_w(z)$ has exactly one root in the

disk $|z| \leq \rho$. Following the notations in [21], we denote the root within the disk $|z| \leq \rho$ of $D_w(z)$ by A_w, and by bootstrapping we obtain

$$A_w = 1 + \frac{1}{S_w(1)}\mathbf{P}(w) + O\left(\mathbf{P}(w)^2\right). \tag{29}$$

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

$$B_w = -S_w(1) + \left(k - \frac{2S'_w(1)}{S_w(1)}\mathbf{P}(w)\right) + O\left(\mathbf{P}(w)^2\right). \tag{30}$$

In this paper, we will prove a similar result for the polynomial $D_{w,w'}(z)$ through the following work.

Lemma 5. *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*

$$\sum_{w \in \mathcal{A}^k} [\![|S_{w,w'}(\rho)| \leq (\rho\delta)^k \theta]\!] \mathbf{P}(w) \geq 1 - \theta\delta^k. \tag{31}$$

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

$$|S_{w,w'}(\rho)| \leq (\rho\delta)^k \theta. \tag{32}$$

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

$$\begin{aligned}
&\sum_{w \in \mathcal{A}^k} [\![S_{w,w'}(z) \text{ has minimal degree} \leq \lfloor k/2 \rfloor]\!] \, \mathbf{P}(w) \\
&\quad = \sum_{i=1}^{\lfloor k/2 \rfloor} \sum_{w \in \mathcal{A}^k} [\![S_{w,w'}(z) \text{ has minimal degree } = i]\!] \, \mathbf{P}(w) \\
&\quad = \sum_{i=1}^{\lfloor k/2 \rfloor} \sum_{w_1 \ldots w_i \in \mathcal{A}^i} \mathbf{P}(w_1 \ldots w_i) \\
&\qquad \sum_{w_{i+1} \ldots w_k \in \mathcal{A}^{k-i}} [\![S_{w,w'}(z) \text{ has minimal degree } = i]\!] \mathbf{P}(w_{i+1} \ldots w_k) \\
&\quad \leq \sum_{i=1}^{\lfloor k/2 \rfloor} \sum_{w_1 .. w_i \in \mathcal{A}^i} \mathbf{P}(w_{i+1} \ldots w_k) p^{k-i} \\
&\quad = \sum_{i=1}^{\lfloor k/2 \rfloor} p^{k-i} \sum_{w_1 .. w_i \in \mathcal{A}^i} \mathbf{P}(w_1 \ldots w_i) \\
&\quad = \sum_{i=1}^{\lfloor k/2 \rfloor} p^{k-i} \leq \frac{p^{k-\lfloor k/2 \rfloor}}{1-p}.
\end{aligned} \tag{33}$$

This leads to the following

$$\begin{aligned}
&\sum_{w \in \mathcal{A}^k} [\![\text{ every term of } S_{w,w'}(z) \text{ is of degree} > \lfloor k/2 \rfloor]\!] \mathbf{P}(w) \\
&\quad = 1 - \sum_{w \in \mathcal{A}^k} [\![S_{w,w'}(z) \text{ has a term of degree} \leq \lfloor k/2 \rfloor]\!] \mathbf{P}(w) \\
&\quad \geq 1 - \frac{p^{\lceil k/2 \rceil}}{1-p} \geq 1 - \theta\delta^k.
\end{aligned} \tag{34}$$

□

Lemma 6. *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*

$$|S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)| > 0. \tag{35}$$

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. There are three cases to consider :
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

$$|S_{w,w'}(z)S_{w',w}(z)| \leq k\frac{(p\rho)^k}{1-p\rho}. \tag{36}$$

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

$$\begin{aligned} |S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)| &\geq |S_w(z)S_{w'}(z)| - |S_{w,w'}(z)S_{w',w}(z)| \\ &\geq 1 - k\frac{(p\rho)^k}{1-p\rho} > 0. \end{aligned} \tag{37}$$

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 9, $|S_w(z)S_{w'}(z)| > 0$. Therefore, there exists $K_2 > 0$, such that

$$\begin{aligned} |S_w(z)S_{w'}(z) - S_{w',w}(z)S_{w,w'}(z)| &\geq |S_w(z)S_{w'}(z)| - |S_{w',w}(z)S_{w,w'}(z)| \\ &> 0 \quad \text{for } k > K_2. \end{aligned} \tag{38}$$

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

$$S_{w',w}(z) = \mathbf{P}(w_{k-m+1}...w_k)z^m \left(S_w(z) - O\left((pz)^{k-m}\right)\right). \tag{39}$$

There exists $K_3 > 0$, such that

$$\begin{aligned} |S_{w',w}(z)| &\leq (p\rho)^m \left(|S_w(z)| + O\left((p\rho)^{k-m}\right)\right. \\ &= (p\rho)^m |S_w(z)| + O\left((p\rho)^k\right) \\ &< |S_w(z)| \quad \text{for } k > K_3\,. \end{aligned} \tag{40}$$

$$\tag{41}$$

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

$$\begin{aligned} |S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)| &\geq |S_w(z)||S_{w'}(z)| - |S_{w,w'}(z)||S_{w',w}(z)| \\ &> |S_w(z)||S_{w'}(z)| - |S_w(z)||S_{w'}(z)| = 0. \end{aligned} \tag{42}$$

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

Lemma 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*

$$D_{w,w'}(z) = (1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)) + z^k \left(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z))\right), \quad (43)$$

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

Proof. First note that

$$|S_w(z) - S_{w',w}(z)| \leq |S_w(z)| + |S_{w',w}(z)| \leq \frac{1}{1-p\rho} + \frac{p\rho}{1-p\rho} = \frac{1+p\rho}{1-p\rho}. \quad (44)$$

This yields

$$\begin{aligned} &\left|z^k \left(\mathbf{P}(w)(S_{w'}(z) - S_{w,w'}(z)) + \mathbf{P}(w')(S_w(z) - S_{w',w}(z))\right)\right| \\ &\quad \leq (p\rho)^k \left(|S_w(z) - S_{w',w}(z)| + |S_{w'}(z) - S_{w,w'}(z)|\right) \\ &\quad \leq (p\rho)^k \left(\frac{2(1+p\rho)}{1-p\rho}\right). \end{aligned} \quad (45)$$

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 $k \geq K_{w,w'}$,

$$\begin{aligned} (p\rho)^k \left(\frac{2(1+p\rho)}{1-p\rho}\right) &< (\rho-1)\beta \\ &< |(1-z)(S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z))|. \end{aligned} \quad (46)$$

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

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

$$\begin{aligned} \alpha_{w,w'} = 1 &+ \frac{S_{w'}(1) - S_{w,w'}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w) \\ &+ \frac{S_w(1) - S_{w',w}(1)}{S_w(1)S_{w'}(1) - S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w') + O(p^{2k}). \end{aligned} \quad (47)$$

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

$$\beta_{w,w'} = S_{w,w'}(1)S_{w',w}(1) - S_w(1)S_{w'}(1) + O(kp^k). \quad (48)$$

We will refer to these expressions in the residue analysis that we present in the next section.

3.4. Asymptotic Difference

We begin this section by the following lemmas on the autocorrelation polynomials.

Lemma 8 (Jacquet and Szpankowski, 1994). *For most words w, the autocorrelation 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*

$$\sum_{w \in \mathcal{A}^k} [\![|S_w(\rho) - 1| \leq (\rho\delta)^k \theta]\!] \mathbf{P}(w) \geq 1 - \theta\delta^k, \tag{49}$$

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

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

Lemma 9 (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*

$$|S_w(z)| > 0. \tag{50}$$

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

Lemma 10. *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*

$$\sum_{w \in \mathcal{A}^k} [\![|S_{w,w'}(\rho)| \leq (\rho\delta)^k \theta]\!] \mathbf{P}(w) \geq 1 - \theta\delta^k \tag{51}$$

We will use the above results to prove that the expected values in the Bernoulli model and the model built over a trie are asymptotically equivalent. We now prove Theorem 1 below.

Proof (of Theorem 1). From Lemmas 3 and 4, we have

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

and

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

subtracting the two generating functions, we obtain

$$H(z) - \hat{H}(z) = \sum_{w \in \mathcal{A}^k} \left(\frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)} \right). \tag{52}$$

We define

$$\Delta_w(z) = \frac{1}{1-(1-\mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)}. \tag{53}$$

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

$$[z^n]\Delta_w(z) = \frac{1}{2\pi i} \oint \Delta_w(z) \frac{dz}{z^{n+1}} = \mathrm{Res}_{z=0}\, \Delta_w(z) \frac{dz}{z^{n+1}}, \tag{54}$$

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 expression (29)). Therefore, we define

$$I^w(\rho) := \frac{1}{2\pi i}\int_{|z|=\rho} \Delta_w(z)\frac{dz}{z^{n+1}}, \tag{55}$$

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

$$\begin{aligned} I^w(\rho) &= \mathrm{Res}_{z=0}\frac{\Delta_w(z)}{z^{n+1}} + \mathrm{Res}_{z=A_w}\frac{\Delta_w(z)}{z^{n+1}} + \mathrm{Res}_{z=1/1-\mathbf{P}(w)}\frac{\Delta_w(z)}{z^{n+1}} \\ &= [z^n]\Delta_w(z) - \mathrm{Res}_{z=A_w}\frac{H_w(z)}{z^{n+1}} + \mathrm{Res}_{z=1/1-\mathbf{P}(w)}\frac{\hat{H}_w(z)}{z^{n+1}} \end{aligned} \tag{56}$$

We observe that

$$\begin{aligned} &\mathrm{Res}_{z=A_w}\frac{\Delta_w(z)}{z^{n+1}} = \frac{S_w(A_w)}{B_w\, A_w^{n+1}}, \quad \text{where } B_w \text{ is as in (30)} \\ &\mathrm{Res}_{z=1/1-\mathbf{P}(w)}\frac{\hat{H}_w(z)}{z^{n+1}} = -(1-\mathbf{P}(w))^{n+1}. \end{aligned}$$

Then we obtain

$$[z^n]\Delta_w = I^w(\rho) - \frac{S_w(A_w)}{B_w\, A_w^{n+1}} - (1-\mathbf{P}(w))^{n+1}, \tag{57}$$

and finally, we have

$$\begin{aligned} [z^n](H(z)-\hat{H}(z)) &= \sum_{w\in\mathcal{A}^k}[z^n]\Delta_w \\ &= \sum_{w\in\mathcal{A}^k} I_n^w(\rho) - \sum_{w\in\mathcal{A}^k}\left(\frac{S_w(A_w)}{B_w\, A_w^{n+1}} + (1-\mathbf{P}(w))^{n+1}\right). \end{aligned} \tag{58}$$

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

Lemma 11. *For large enough n, and for $k = \Theta(\log n)$, there exists $M > 0$ such that*

$$\sum_{w\in\mathcal{A}^k}\left(\frac{S_w(A_w)}{B_w\, A_w^{n+1}} + (1-\mathbf{P}(w))^{n+1}\right) = O(n^{-M}). \tag{59}$$

Proof. We let

$$r_w(z) = (1-\mathbf{P}(w))^z + \frac{S_w(A_w)}{B_w A_w^z}. \tag{60}$$

The Mellin transform of the above function is

$$r_w^*(s) = \Gamma(s)\log^{-s}\left(\frac{1}{1-\mathbf{P}(w)}\right) - \frac{S_w(A_w)}{B_w}\Gamma(s)\log^{-s}(A_w). \tag{61}$$

We define

$$C_w = \frac{S_w(A_w)}{B_w} = \frac{S_w(A_w)}{-S_w(1) + O(k\mathbf{P}(w))}, \tag{62}$$

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

$$\begin{aligned}\ln^{-s}\left(\frac{1}{1-\mathbf{P}(w)}\right) &= \ln^{-s}\left(1+\mathbf{P}(w)+O\left(\mathbf{P}(w)^2\right)\right)\\ &= \left(\mathbf{P}(w)+O\left(\mathbf{P}(w)^2\right)\right)^{-s}\\ &= \mathbf{P}(w)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right)^{-s}\\ &= \mathbf{P}(w)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right),\end{aligned} \tag{63}$$

$$\begin{aligned}\ln^{-s}(A_w) &= \ln^{-s}\left(1-\left(-\frac{\mathbf{P}(w)}{S_w(1)}+O\left(\mathbf{P}(w)^2\right)\right)\right)\\ &= \left(\frac{\mathbf{P}(w)}{S_w(1)}+O\left(\mathbf{P}(w)^2\right)\right)^{-s}\\ &= \left(\frac{\mathbf{P}(w)}{S_w(1)}\right)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right)^{-s}\\ &= \left(\frac{\mathbf{P}(w)}{S_w(1)}\right)^{-s}\left(1+O\left(\mathbf{P}(w)\right)\right),\end{aligned} \tag{64}$$

and therefore, we obtain

$$r_w^*(s) = \Gamma(s)\mathbf{P}(w)^{-s}\left(1-\frac{1}{S_w(1)^{-s}}\right)O(1). \tag{65}$$

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

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

Lemma 8 allows us to split the above sum between the words for which $S_w(1) \le 1+O(\delta^k)$ and words that have $S_w(1) > 1+O(\delta^k)$.

Such a split yields the following

$$\sum_{w\in\mathcal{A}^k}|r_w^*(s)| = (q^k)^{-\Re(s)-1}|\Gamma(s)|O(\delta^k). \tag{66}$$

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, as $(q^k)^{-\Re(s)-1}$ 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

$$\sum_{w\in\mathcal{A}^k} r_w(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\Big(\sum_{w\in\mathcal{A}^k} r_w^*(s)\Big)z^{-s}ds. \tag{67}$$

We choose $c \in (-1, M)$ for a fixed $M > 0$. Then by the direct mapping theorem [22], we obtain

$$\sum_{w \in \mathcal{A}^k} r_w(z) = O(z^{-M}). \tag{68}$$

and subsequently, we get

$$\sum_{w \in \mathcal{A}^k} \left(\frac{S_w(A_w)}{B_w \, A_w^{n+1}} + (1 - \mathbf{P}(w))^{n+1} \right) = O(n^{-M}). \tag{69}$$

□

We next prove the asymptotic smallness of $I_n^w(\rho)$ in (55).

Lemma 12. *Let*

$$I_n^w(\rho) = \frac{1}{2\pi i} \int_{|z|=\rho} \left(\frac{1}{1 - (1 - \mathbf{P}(w))z} - \frac{S_w(z)}{D_w(z)} \right) \frac{dz}{z^{n+1}}. \tag{70}$$

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

$$\sum_{w \in \mathcal{A}^k} I_n^w(\rho) = O\left(\rho^{-n} (\rho\delta)^k \right). \tag{71}$$

Proof. We observe that

$$|I_n^w(\rho)| \leq \frac{1}{2\pi} \int_{|z|=\rho} \left| \frac{\mathbf{P}(w) z \left(z^{k-1} - S_w(z) \right)}{D_w(z)(1 - (1 - \mathbf{P}(w))z)} \frac{1}{z^{n+1}} \right| dz. \tag{72}$$

For $|z| = \rho$, we show that the denominator in (72) is bounded away from zero.

$$\begin{aligned} |D_w(z)| &= |(1-z)S_w(z) + \mathbf{P}(w)z^k| \\ &\geq |1-z||S_w(z)| - \mathbf{P}(w)|z^k| \\ &\geq (\rho - 1)\alpha - (p\rho)^k, \quad \text{where } \alpha > 0 \text{ by Lemma 9 .} \\ &> 0, \qquad \text{we assume } k \text{ to be large enough such that } (p\rho)^k < \alpha(\rho - 1). \end{aligned} \tag{73}$$

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

$$\begin{aligned} |1 - (1 - \mathbf{P}(w))z| &\geq |1 - (1 - \mathbf{P}(w))|z|| \\ &\geq |1 - \rho(1 - p^{K_w})| \\ &> 0. \end{aligned} \tag{74}$$

We now move on to finding an upper bound for the numerator in (72), for $|z| = \rho$.

$$\begin{aligned} |z^{k-1} - S_w(z)| &\leq |S_w(z) - 1| + |1 - z^{k-1}| \\ &\leq (S_w(\rho) - 1) + (1 + \rho^{k-1}) \\ &= (S_w(\rho) - 1) + O(\rho^k). \end{aligned} \tag{75}$$

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

$$\begin{aligned} |I_n^w| &\leq \mu\rho\mathbf{P}(w)\left((S_w(\rho)-1)+O(\rho^k)\right)\frac{1}{\rho^{n+1}} \\ &= O(\rho^{-n})\left(\mathbf{P}(w)(S_w(\rho)-1)+\mathbf{P}(w)O(\rho^k)\right). \end{aligned} \tag{76}$$

Summing over all patterns w, and applying Lemma 8, we obtain

$$\begin{aligned} \sum_{w\in\mathcal{A}^k}|I_n^w(\rho)| &= O(\rho^{-n})\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)(S_w(\rho)-1)+O(\rho^{-n+k})\sum_{w\in\mathcal{A}^k}\mathbf{P}(w) \\ &= O(\rho^{-n})\left(\theta(\rho\delta)^k+\frac{p\rho}{1-p\rho}\theta\delta^k\right)+O(\rho^{-n+k}) \\ &= O(\rho^{-n}(\rho\delta)^k), \end{aligned} \tag{77}$$

which approaches zero as $n\to\infty$ and $k=\Theta(\log n)$. This completes the proof of of Theorem 1. □

Similar to Theorem 1, we provide a proof to show that the second factorial moments of the kth Subword Complexity and the kth Prefix Complexity, have the same first order asymptotic behavior. We are now ready to state the proof of Theorem 2.

Proof (of Theorem 2). As discussed in Lemmas 3 and 4, 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

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

Note that

$$G(z)-\hat{G}(z)=\sum_{\substack{w'\in\mathcal{A}^k\\ w\neq w'}}\sum_{w\in\mathcal{A}^k}\left(\frac{1}{1-(1-\mathbf{P}(w))z}-\frac{S_w(z)}{D_w(z)}\right) \tag{78}$$

$$+\sum_{\substack{w\in\mathcal{A}^k\\ w\neq w'}}\sum_{w'\in\mathcal{A}^k}\left(\frac{1}{1-(1-\mathbf{P}(w'))z}-\frac{S_{w'}(z)}{D_{w'}(z)}\right) \tag{79}$$

$$+\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z}-\frac{S_w(z)S_{w'}(z)-S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}\right) \tag{80}$$

In Theorem 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 (78) and (79) are of order $(2^k-1)O(n^{-M}) = O(n^{-M+a\log 2})$ for $k = a\log n$. Thus, to show the asymptotic smallness, it is enough to choose $M = a\log 2 + \epsilon$, where ϵ is a small positive value. Now, it only remains to show (80) is asymptotically negligible as well. We define

$$\Delta_{w,w'}(z) = \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z} - \frac{S_w(z)S_{w'}(z) - S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}. \tag{81}$$

Next, we extract the coefficient of z^n

$$[z^n]\Delta_{w,w'}(z) = \frac{1}{2\pi i}\oint \Delta_{w,w'}(z)\frac{dz}{z^{n+1}}, \tag{82}$$

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

$$I_n^{w,w'}(\rho) = \frac{1}{2\pi i}\int_{|z|=\rho} \Delta_{w,w'}(z)\frac{dz}{z^{n+1}}, \tag{83}$$

The above integrand has poles at $z=0$, $z=\alpha_{w,w'}$ (as in (47)), and $z=\frac{1}{1-\mathbf{P}(w)-\mathbf{P}(w')}$. We have chosen ρ such that the poles are all inside the circle $|z|=\rho$. It follows that

$$I_n^{w,w'}(\rho) = \mathrm{Res}_{z=0}\frac{\Delta_{w,w'}(z)}{z^{n+1}} + \mathrm{Res}_{z=\alpha_{w,w'}}\frac{\Delta_{w,w'}(z)}{z^{n+1}} + \mathrm{Res}_{z=\frac{1}{1-\mathbf{P}(w)-\mathbf{P}(w')}}\frac{\Delta_w(z)}{z^{n+1}}, \tag{84}$$

and the residues give us the following.

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

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 (48). Therefore, we get

$$\begin{aligned}\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}[z^n]\Delta_{w,w'}(z) = &\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} I_n^{w,w'}(\rho)\\ &- \sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\left(\frac{S_w(\alpha_{w,w'})S_{w'}(\alpha_{w,w'}) - S_{w,w'}(\alpha_{w,w'})S_{w',w}(\alpha_{w,w'})}{\beta_{w,w'}\alpha_{w,w'}^{n+1}}\right.\\ &\left. + (1-\mathbf{P}(w)-\mathbf{P}(w'))^{n+1}\right).\end{aligned} \tag{85}$$

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

Lemma 13. *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

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

The Mellin transform of the above function is

$$r^*_{w,w'}(s) = \Gamma(s)\log^{-s}\left(\frac{1}{1-\mathbf{P}(w)-\mathbf{p}(w')}\right) + C_{w,w'}\Gamma(s)\log^{-s}(\alpha_{w,w'}), \tag{86}$$

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,

$$\begin{aligned}
\ln^{-s}\left(\frac{1}{1-\mathbf{P}(w)-\mathbf{P}(w')}\right) &= \ln^{-s}\left(1+\mathbf{P}(w)+\mathbf{P}(w')+O\left(p^{2k}\right)\right)\\
&= \left(\mathbf{P}(w)+\mathbf{P}(w')+O\left(p^{2k}\right)\right)^{-s}\\
&= (\mathbf{P}(w)+\mathbf{P}(w'))^{-s}\left(1+O\left(p^{k}\right)\right)^{-s}\\
&= (\mathbf{P}(w)+\mathbf{P}(w'))^{-s}\left(1+O\left(p^{k}\right)\right),
\end{aligned} \tag{87}$$

and

$$\begin{aligned}
\ln^{-s}(\alpha_{w,w'}) &= \Big(\frac{S_{w'}(1)-S_{w,w'}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w)\\
&\qquad + \frac{S_w(1)-S_{w',w}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w') + O(p^{2k})\Big)^{-s}\\
&= \Big(\frac{S_{w'}(1)-S_{w,w'}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w)\\
&\qquad + \frac{S_w(1)-S_{w',w}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\Big)^{-s}\left(1+O(p^k)\right).
\end{aligned} \tag{88}$$

Therefore, we have

$$\begin{aligned}
r^*_{w,w'}(s) &= \Gamma(s)\left(\mathbf{P}(w)+\mathbf{P}(w')\right)^{-s}(1+O(p^k))\\
&\quad - \Gamma(s)\Bigg(\frac{S_{w'}(1)-S_{w,w'}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w)\\
&\qquad + \frac{S_w(1)-S_{w',w}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\Bigg)^{-s}\left(1+O(p^k)\right)O(1).
\end{aligned} \tag{89}$$

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

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

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

$$\left(\mathbf{P}(w)+\mathbf{P}(w')\right)^{-\Re(s)} \le \mathbf{P}(w)^{-\Re(s)} + \mathbf{P}(w')^{-\Re(s)}, \tag{90}$$

and

$$\left(\frac{S_{w'}(1)-S_{w,w'}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w)ßn+\frac{S_w(1)-S_{w',w}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\right)^{-\Re(s)}$$
$$\leq\left(\frac{S_{w'}(1)-S_{w,w'}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w)\right)^{-\Re(s)}$$
$$+\left(\frac{S_w(1)-S_{w',w}(1)}{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}\mathbf{P}(w')\right)^{-\Re(s)}. \tag{91}$$

Therefore, we get

$$\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}|r^*_{w,w'}(s)|\leq|\Gamma(s)|O(1)$$
$$\left(\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\mathbf{P}(w)^{-\Re(s)}\left(1-\left(\frac{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}{S_{w'}(1)-S_{w,w'}(1)}\right)^{\Re(s)}\right)\right.$$
$$\left.+\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}}\mathbf{P}(w')^{-\Re(s)}\left(1-\left(\frac{S_w(1)S_{w'}(1)-S_{w,w'}(1)S_{w',w}(1)}{S_w(1)-S_{w',w}(1)}\right)^{\Re(s)}\right)\right)$$
$$\leq(q^k)^{-\Re(s)-1}|\Gamma(s)|O(1)$$
$$\left(\sum_{\substack{w'\in\mathcal{A}^k\\ w\neq w'}}\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)\left(1-(S_w(1))^{\Re(s)}\left(1-\frac{S_{w,w'}(1)}{S_{w'}(1)}\right)^{-\Re(s)}\right)\right. \tag{92}$$
$$+\sum_{\substack{w'\in\mathcal{A}^k\\ w\neq w'}}\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)S_{w,w'}(1)^{\Re(s)}\left(\frac{S_{w'}(1)-S_{w,w'}(1)}{S_{w',w}(1)}\right)^{-\Re(s)} \tag{93}$$
$$+\sum_{\substack{w\in\mathcal{A}^k\\ w\neq w'}}\sum_{w'\in\mathcal{A}^k}\mathbf{P}(w')\left(1-(S_{w'}(1))^{\Re(s)}\left(1-\frac{S_{w',w}(1)}{S_w(1)}\right)^{-\Re(s)}\right) \tag{94}$$
$$\left.+\sum_{\substack{w\in\mathcal{A}^k\\ w\neq w'}}\sum_{w'\in\mathcal{A}^k}\mathbf{P}(w')S_{w',w}(1)^{\Re(s)}\left(\frac{S_w(1)-S_{w',w}(1)}{S_{w,w'}(1)}\right)^{-\Re(s)}\right). \tag{95}$$

By Lemma 10, 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 8, for most words w,

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

Therefore, both sums (92) and (94) are of the form $(2^k-1)O(\delta^k)$. The sums (93) and (95) are also of order $(2^k-1)O(\delta^k)$ by Lemma 10. Combining all these terms we will obtain

$$\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} |r^*_{w,w'}(s)| \leq (2^k-1)(q^k)^{-\Re(s)-1}|\Gamma(s)|O(\delta^k)O(1). \tag{96}$$

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

$$\begin{aligned}\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} r_{w,w'}(z) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\Big(\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} r^*_{w,w'}(s)\Big)z^{-s}ds = O(z^{-M})O(2^k)\\ &= O(z^{-\epsilon}).\end{aligned} \tag{97}$$

□

In the following lemma we show that the first term in (86) is asymptotically small.

Lemma 14. *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

$$\sum_{\substack{w,w'\in\mathcal{A}^k\\ w\neq w'}} I_n^{w,w'}(\rho) = O\left(\rho^{-n+2k}\delta^k\right). \tag{98}$$

Proof. First note that

$$\begin{aligned}\Delta_{w,w'}(z) &= \frac{1}{1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z} - \frac{S_w(z)S_{w'}(z)-S_{w,w'}(z)S_{w',w}(z)}{D_{w,w'}(z)}\\ &= \frac{z\mathbf{P}(w)\left(S_{w,w'}(z)S_{w',w}(z)-S_w(z)S_{w'}(z)+z^{k-1}S_{w'}(z)-z^{k-1}S_{w,w'}(z)\right)}{(1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z)\,D_{w,w'}(z)}\\ &\quad+\frac{z\mathbf{P}(w')\left(S_{w',w}(z)S_{w,w'}(z)-S_{w'}(z)S_w(z)+z^{k-1}S_w(z)-z^{k-1}S_{w',w}(z)\right)}{(1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z)\,D_{w,w'}(z)}.\end{aligned} \tag{99}$$

We saw in (74) that $|1-(1-\mathbf{P}(w'))z|\geq c_2$, and therefore, it follows that

$$|1-(1-\mathbf{P}(w)-\mathbf{P}(w'))z|\geq c_1 \tag{100}$$

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

$$\begin{aligned}|D_{w,w'}(z)| &= \big|(1-z)(S_w(z)S_{w'}(z)-S_{w,w'}(z)S_{w',w}(z))\\ &\qquad+z^k\left(\mathbf{P}(w)(S_{w'}(z)-S_{w,w'}(z))+\mathbf{P}(w')(S_w(z)-S_{w',w}(z))\right)\big|\\ &\geq |(1-z)(S_w(z)S_{w'}(z)-S_{w,w'}(z)S_{w',w}(z))|\\ &\qquad-\left|z^k\right|\left|\left(\mathbf{P}(w)(S_{w'}(z)-S_{w,w'}(z))+\mathbf{P}(w')(S_w(z)-S_{w',w}(z))\right)\right|\\ &\geq(\rho-1)\beta-(p\rho)^k\left(\frac{2(1+p\rho)}{1-p\rho}\right),\end{aligned} \tag{101}$$

which is bounded away from zero by the assumption of Lemma 7. Additionally, we show that the numerator in (99) is bounded above, as follows

$$\begin{aligned}|S_{w,w'}(z)S_{w',w}(z) - S_w(z)S_{w'}(z) + z^{k-1}S_{w'}(z) - z^{k-1}S_{w,w'}(z)| \leq \\ |S_{w'}(z)(z^{k-1} - S_w(z))| + |S_{w,w'}(z)(S_{w',w}(z) - z^{k-1})| \\ \leq S_{w'}(\rho)\left((S_w(\rho) - 1) + O(\rho^k)\right) + S_{w,w'}(\rho)\left(S_{w',w}(\rho) + O(\rho^k)\right). \end{aligned} \tag{102}$$

This yields

$$\begin{aligned}\sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} |I_n^{w,w'}| \leq O(\rho^{-n}) \sum_{\substack{w' \in \mathcal{A}^k \\ w \neq w'}} S_{w'}(\rho) \sum_{w \in \mathcal{A}^k} \mathbf{P}(w)\left((S_w(\rho) - 1) + O(\rho^k)\right) \\ + O(\rho^{-n}) \sum_{\substack{w' \in \mathcal{A}^k \\ w \neq w'}} \sum_{w \in \mathcal{A}^k} \mathbf{P}(w) S_{w,w'}(\rho)\left(S_{w',w}(\rho) + O(\rho^k)\right). \end{aligned} \tag{103}$$

By (76), the first term above is of order $(2^k - 1)O(\rho^{-n+k})$ and by Lemma 10 and an analysis similar to (76), 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 2.

3.5. Asymptotic Analysis of the kth Prefix Complexity

We finally proceed to analyzing the asymptotic moments of the kth Prefix Complexity. The results obtained hold true for the moments of the kth Subword Complexity. Our methodology involves poissonization, saddle point analysis (the complex version of Laplace's method [23]), and depoissonization.

Lemma 15. *(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_θ with $\theta < \pi/2$, and if the following two conditions hold:*
(I) For $z \in S_\theta$ and real values $B, r > 0, \nu$

$$|z| > r \to |\tilde{G}(z)| \leq B|z^\nu|\Psi(|z|), \tag{104}$$

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 $A, \alpha < 1$

$$|z| > r \to |\tilde{G}(z)e^z| \leq Ae^{\alpha|z|}. \tag{105}$$

Then, for every non-negative integer n, we have

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

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 (25) 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

$$\begin{aligned}\tilde{E}_k(z) &= \sum_{n=0}^{\infty} \mathbf{E}[\hat{X}_{n,k}]\frac{z^n}{n!}e^{-z}\\ &= \sum_{n=0}^{\infty}\sum_{w\in\mathcal{A}^k}\left(1-(1-\mathbf{P}(w))^n\right)\frac{z^n}{n!}e^{-z}\\ &= \sum_{w\in\mathcal{A}^k}\left(1-e^{-z\mathbf{P}(w)}\right).\end{aligned} \tag{106}$$

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

$$\begin{aligned}\tilde{E}_k^*(s) &= -\Gamma(s)\sum_{w\in\mathcal{A}^k} P(w)^{-s}\\ &= -\Gamma(s)(p^{-s}+q^{-s})^k,\end{aligned} \tag{107}$$

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

$$\begin{aligned}\tilde{E}_k(z) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\tilde{E}_k^*(s)\cdot z^{-s}ds\\ &= \frac{-1}{2\pi i}\int_{c-i\infty}^{c+i\infty} z^{-s}\Gamma(s)(p^{-s}+q^{-s})^k ds\\ &= \frac{-1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\Gamma(s)e^{-k(s\frac{\log z}{k}-\log(p^{-s}+q^{-s}))}ds\\ &= \frac{-1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\Gamma(s)e^{-kh(s)}ds,\end{aligned} \tag{108}$$

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

$$(p/q)^{-r_0} = \frac{a\log p^{-1}-1}{1-a\log q^{-1}}, \tag{109}$$

and therefore,

$$r_0 = \frac{-1}{\log p/q}\log\left(\frac{a\log q^{-1}-1}{1-a\log p^{-1}}\right), \tag{110}$$

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

By (109) 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

$$\frac{1}{\log q^{-1}} < a < \frac{2}{\log q^{-1} + \log p^{-1}}, \tag{111}$$

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

$$\frac{2}{\log q^{-1} + \log p^{-1}} < a < \frac{1}{q \log q^{-1} + p \log p^{-1}}, \tag{112}$$

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

$$\frac{1}{q \log q^{-1} + p \log p^{-1}} < a < \frac{1}{\log p^{-1}}, \tag{113}$$

and we approach it with a combination of residue analysis at $s = 0$, and the saddle point method. We now proceed by stating the proof of theorem 3.

Proof (of Theorem 3). 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

$$\begin{aligned} \tilde{E}_k(z) &= \frac{-1}{2\pi} \int_{-\infty}^{\infty} z^{-(r_0+it)} \Gamma(r_0 + it)(p^{-(r_0+it)} + q^{-(r_0+it)})^k dt \\ &= \frac{-1}{2\pi} \int_{-\infty}^{\infty} \Gamma(r_0 + it) e^{-k((r_0+it)\frac{\log z}{k} - \log(p^{-(r_0+it)} + q^{-(r_0+it)}))} dt. \end{aligned} \tag{114}$$

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

$$T_k(z) = \int_{|t| > \sqrt{\log n}} z^{-r_0 - it} \Gamma(r_0 + it)(p^{-r_0 - it} + q^{-r_0 - it})^k dt. \tag{115}$$

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

$$\begin{aligned} T_k(z) &= O\left(z^{-r_0} (p^{-r_0} + q^{-r_0})^k \int_{\sqrt{\log n}}^{\infty} t^{r_0/2 - 1/2} e^{-\pi t/2} dt \right) \\ &= O\left(z^{-r_0} (p^{-r_0} + q^{-r_0})^k (\log n)^{r_0/4 - 1/4} \int_{\sqrt{\log n}}^{\infty} e^{-\pi t/2} dt \right) \\ &= O\left(z^{-r_0} (p^{-r_0} + q^{-r_0})^k (\log n)^{r_0/4 - 1/4} e^{-\pi\sqrt{\log n}/2} \right) \\ &= O\left((\log n)^{r_0/4 - 1/4} e^{-\pi\sqrt{\log n}/2} \right), \end{aligned} \tag{116}$$

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

$$\tilde{E}_k(z) = -\frac{1}{2\pi}\left(\sum_{|j|<j^*}\int_{|t-t_j|\leq\frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt\right) - \frac{1}{2\pi}\int_{\frac{\pi}{\log p/q}\leq|t_j^*|<\sqrt{\log n}} \Gamma(r+it)z^{-r_0+it}(p^{-r_0-it}+q^{-r_0-it})^k dt. \quad (117)$$

By the same argument as in (116), the second term in (117) is also asymptotically negligible. Therefore, we are only left with

$$\tilde{E}_k(z) = \sum_{|j|<j^*} S_j(z), \quad (118)$$

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

$$h(t) = h(t_j) + \frac{1}{2}h''(t_j)(t-t_j)^2 + O((t-t_j)^3). \quad (119)$$

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

$$k(t-t_j)^2 \to \infty, \qquad \text{and} \qquad k(t-t_j)^3 \to 0. \quad (120)$$

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

$$S_j(z) = -\frac{1}{2\pi}\int_{|t-t_j|\leq k^{-2/5}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt - \frac{1}{2\pi}\int_{k^{-2/5}<|t-t_j|<\frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt. \quad (121)$$

Saddle Tails Pruning.

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

$$S_j^{(1)}(z) = -\frac{1}{2\pi}\int_{k^{-2/5}<|t-t_j|<\frac{\pi}{\log p/q}} z^{-r_0+it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt. \quad (122)$$

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

$$
\begin{aligned}
|p^{-r_0-it}+q^{-r_0-it}| &= (p^{-r_0}+q^{-r_0})\sqrt{1-\frac{2p^{-r_0}q^{-r_0}}{(p^{-r_0}+q^{-r_0})^2}(1-\cos(t\log p/q))} \\
&\le (p^{-r_0}+q^{-r_0})\left(1-\frac{p^{-r_0}q^{-r_0}}{(p^{-r_0}+q^{-r_0})^2}(1-\cos(t-t_j)\log p/q)\right) \\
&\qquad \text{since}\sqrt{1-x}\le 1-\frac{x}{2} \text{ for } x\in[0,1] \\
&\le (p^{-r_0}+q^{-r_0})\left(1-\frac{2p^{-r_0}q^{-r_0}}{\pi^2(p^{-r_0}+q^{-r_0})^2}((t-t_j)\log p/q)^2\right) \\
&\qquad \text{since } 1-\cos x\ge\frac{2x^2}{\pi^2} \text{ for } |x|\le\pi \\
&\le (p^{-r_0}+q^{-r_0})e^{-\gamma(t-t_j)^2}, \qquad (123)
\end{aligned}
$$

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

$$
\begin{aligned}
S_j^{(1)}(z) &= O\left(z^{-r_0}|\Gamma(r_0+it)|\int_{k^{-2/5}<|t-t_j|<\frac{\pi}{\log p/q}}|p^{-r_0-it}+q^{-r_0-it}|dt\right) \\
&= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k\int_{k^{-2/5}}^{\infty}e^{-\gamma ku^2}du\right) \\
&= O\left(z^{-r_0}(p^{-r_0}+q^{-r_0})^k k^{-3/5}e^{-\gamma k^{1/5}}\right), \text{since } \operatorname{erf}(x)=O\left(e^{-x^2}/x\right). \qquad (124)
\end{aligned}
$$

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

$$
h''(t_j)=\frac{\log^2 p/q}{((p/q)^{-r_0/2}+(p/q)^{r_0/2})^2}, \qquad (125)
$$

and

$$
p^{-r_0-it_j}+q^{-r_0-it_j}=p^{-it_j}(p^{-r_0}+q^{-r_0}). \qquad (126)
$$

Therefore, by Laplace's theorem (refer to [22]) we obtain

$$
\begin{aligned}
S_j^{(0)}(z) &= \frac{1}{\sqrt{2\pi k h''(t_j)}}\Gamma(r_0+it_j)e^{-kh(t_j)}(1+O(k^{-1/2})) \\
&= \frac{(p/q)^{-r_0/2}+(p/q)^{r_0/2}}{\sqrt{2\pi}\log p/q} \\
&\qquad \times z^{-r_0}(p^{-r_0}+q^{-r_0})^k\Gamma(r_0+it_j)z^{-it_j}p^{-ikt_j}k^{-1/2}\left(1+O\left(\frac{1}{\sqrt{k}}\right)\right). \qquad (127)
\end{aligned}
$$

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

$$\tilde{E}_k(z) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2\pi}\log p/q} \times \sum_{|j|<j^*} z^{-r_0}(p^{-r_0}+q^{-r_0})^k \Gamma(r_0+it_j) z^{-it_j} p^{-ikt_j} k^{-1/2} \left(1 + O\left(\frac{1}{\sqrt{k}}\right)\right). \tag{128}$$

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

$$\tilde{E}_k(z) = \Phi_1((1+a\log p)\log_{p/q} n)\frac{z^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{129}$$

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

$$\Phi_1(x) = \frac{(p/q)^{-r_0/2} + (p/q)^{r_0/2}}{\sqrt{2a\pi}\log p/q} \sum_{|j|<j^*} \Gamma(r_0+it_j) e^{-2\pi i j x}. \tag{130}$$

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

$$\tilde{E}_k(z) = \mathrm{Res}_{s=0}\tilde{E}_k^*(s)z^{-s} + \int_{r_0-i\infty}^{r_0+i\infty} \tilde{E}_k^*(z)z^{-s}ds. \tag{131}$$

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

$$\tilde{E}_k(z) = 2^k - \Phi_1((1+a\log p)\log_{p/q} n)\frac{z^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right). \tag{132}$$

For part *iii*. of Theorem 3, we shift the line of integration to $c_0 \in (-2,-1)$, then we have

$$\begin{aligned}\tilde{E}_k(z) &= \mathrm{Res}_{s=-1}\tilde{E}_k^*(s)z^{-s} + \int_{c-i\infty}^{c+i\infty} \tilde{E}_k^*(z)z^{-s}ds \\ &= z + O\left(z^{-c_0}(p^{-c_0}+q^{-c_0})^k\right) \\ &= z^{a\log 2} + O(z^{\nu_0}),\end{aligned} \tag{133}$$

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

Step four: Asymptotic depoissonization

To show that both conditions in (15) 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 (104), we note that

$$|e^{-i\theta(r_0+it)}\Gamma(r_0+it)| = O(|t|^{r_0-1/2}e^{t\theta-\pi|t|/2}), \tag{134}$$

and therefore

$$\tilde{E}_k(ne^{i\theta}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\theta(r_0+it)}n^{-r_0-it}\Gamma(r_0+it)(p^{-r_0-it}+q^{-r_0-it})^k dt \tag{135}$$

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

$$|\tilde{E}_k(z)| \leq B\frac{|z^\nu|}{\sqrt{\log n}}, \tag{136}$$

where $B = |\Phi_1((1 + a \log p) \log_{p/q} n)|$, and ν is as in 129. Condition (104) is therefore satisfied. To prove condition (105) We see that for a fixed k,

$$\begin{aligned} |\tilde{E}_k(z)e^z| &\leq \sum_{w\in\mathcal{A}^k} |e^z - e^{z(1-\mathbf{P}(w))}| \\ &\leq 2^{k+1}e^{|z|\cos(\theta)}. \end{aligned} \tag{137}$$

Therefore, we have

$$\mathbf{E}[\hat{X}_{n,k}] = \tilde{E}(n) + O\left(\frac{n^{\nu-1}}{\sqrt{\log n}}\right). \tag{138}$$

This completes the proof of Theorem 3. □

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 (27),

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

which gives the following poissonized form

$$\begin{aligned} \tilde{G}(z) &= \sum_{n\geq 0} \mathbf{E}[(\hat{X}_{n,k})_2]\frac{z^n}{n!}e^{-z} \\ &= \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} 1 - e^{-\mathbf{P}(w)z} - e^{-\mathbf{P}(w')z} + e^{-(\mathbf{P}(w)+\mathbf{P}(w'))z} \\ &= \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left(1 - e^{-\mathbf{P}(w')z}\right)\left(1 - e^{-\mathbf{P}(w)z}\right) \\ &= \left(\sum_{w\in\mathcal{A}^k} \left(1 - e^{-\mathbf{P}(w)z}\right)\right)^2 - \sum_{w\in\mathcal{A}^k} \left(1 - e^{-\mathbf{P}(w)z}\right)^2 \\ &= (\tilde{E}_k(z))^2 - \sum_{w\in\mathcal{A}^k} \left(1 - e^{-\mathbf{P}(w)z}\right)^2 \\ &= (\tilde{E}_k(z))^2 - \sum_{w\in\mathcal{A}^k} \left(1 - 2e^{-\mathbf{P}(w)z} + e^{-2\mathbf{P}(w)z}\right). \end{aligned} \tag{139}$$

We show that in all ranges of a the leftover sum in (139) has a lower order contribution to $\tilde{G}_k(z)$ compared to $(\tilde{E}_k(z))^2$. We define

$$\tilde{L}_k(z) = \sum_{w\in\mathcal{A}^k} \left(1 - 2e^{-\mathbf{P}(w)z} + e^{-2\mathbf{P}(w)z}\right). \tag{140}$$

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

$$\begin{aligned}\tilde{L}_k^*(s) &= -2\Gamma(s)\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)^{-s} + \Gamma(s)\sum_{w\in\mathcal{A}^k}(2\mathbf{P}(w))^{-s}\\ &= -2\Gamma(s)(p^{-s}+q^{-s})^k + \Gamma(s)2^{-s}(p^{-s}+q^{-s})^k\\ &= \Gamma(s)(p^{-s}+q^{-s})^k(2^{-s-1}-1),\end{aligned} \tag{141}$$

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

$$\begin{aligned}\tilde{L}_k(z) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\tilde{L}_k^*(s)z^{-s}ds\\ &= \frac{1}{\pi i}\int_{c-i\infty}^{c+i\infty}\Gamma(s)(p^{-s}+q^{-s})^k(2^{-s-1}-1)z^{-s}ds\end{aligned} \tag{142}$$

We note that this range of r_0 corresponds to

$$\frac{2}{\log q^{-1}+\log p^{-1}} < a < \frac{p^2+q^2}{q^2\log q^{-1}+p^2\log p^{-1}}. \tag{143}$$

The integrand in (142) is quite similar to the one seen in (108). 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 (110) 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 (108) applies to $\tilde{L}_k(z)$ as well. We avoid repeating the similar steps, and we skip to the central approximation, where by Laplace's theorem (ref. [22]), we get

$$\begin{aligned}\tilde{L}_k(z) = \frac{(p/q)^{-r_0/2}+(p/q)^{r_0/2}}{\sqrt{2\pi}\log p/q}&\\ \times\sum_{|j|<j^*} z^{-r_0}(p^{-r_0}+q^{-r_0})^k(2^{-r_0-1-it_j}-1)&\\ \times\Gamma(r_0+it_j)z^{-it_j}p^{-ikt_j}k^{-1/2}\left(1+O\left(\frac{1}{\sqrt{k}}\right)\right),&\end{aligned} \tag{144}$$

which can be represented as

$$\tilde{L}_k(z) = \Phi_2((1+a\log p)\log_{p/q}n)\frac{z^\nu}{\sqrt{\log n}}\left(1+O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{145}$$

where

$$\Phi_2(x) = \frac{(p/q)^{-r_0/2}+(p/q)^{r_0/2}}{\sqrt{2a\pi}\log p/q}\sum_{|j|<j^*}(2^{-r_0-1-it_j}-1)\Gamma(r_0+it_j)e^{-2\pi ijx}. \tag{146}$$

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

$$\tilde{L}_k(z) = 2^k - \Phi_2((1 + a\log p)\log_{p/q} n)\frac{z^\nu}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right), \tag{147}$$

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

$$\tilde{L}_k(z) = O(n^2). \tag{148}$$

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 Theorem 4, which will be proved bellow.

Proof (of Theorem 4). It is only left to show that the two depoissonization conditions hold: For condition (104) in Theorem 15, from (136) we have

$$|\tilde{G}_k(z)| \le B^2 \frac{|z^{2\nu}|}{\log n}, \tag{149}$$

and for condition (105), we have, for fixed k,

$$\begin{aligned} |\tilde{G}_k(z)e^z| &\le \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left| e^z - e^{(1-\mathbf{P}(w))z} - e^{(1-\mathbf{P}(w'))z} + e^{(1-(\mathbf{P}(w)+\mathbf{P}(w')))z} \right| \\ &\le 4^k e^{|z|\cos\theta}. \end{aligned} \tag{150}$$

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

Corollary. A Remark on the Second Moment and the Variance
For the second moment we have

$$\begin{aligned} \mathbf{E}\left[(\hat{X}_{n,k})^2\right] &= \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \mathbf{E}\left[\hat{X}^{(w)}_{n,k}\hat{X}^{(w')}_{n,k}\right] + \sum_{w\in\mathcal{A}^k} \mathbf{E}[\hat{X}^{(w)}_{n,k}] \\ &= \sum_{\substack{w,w'\in\mathcal{A}^k \\ w\neq w'}} \left(1 - (1-\mathbf{P}(w))^n - (1-\mathbf{P}(w'))^n + (1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right) \\ &\qquad + \sum_{w\in\mathcal{A}^k} \left(1 - (1-\mathbf{P}(w))^n\right). \end{aligned} \tag{151}$$

Therefore, by (106) and (139) the Poisson transform of the second moment, which we denote by $\tilde{G}^{(2)}_k(z)$ is

$$\tilde{G}^{(2)}_k(z) = (\tilde{E}_k(z))^2 + \tilde{E}_k(z) - \sum_{w\in\mathcal{A}^k}\left(1 - 2e^{-\mathbf{P}(w)z} + e^{-2\mathbf{P}(w)z}\right), \tag{152}$$

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

$$\begin{aligned}
\mathrm{Var}[\hat{X}_{n,k}] &= \mathbf{E}\left[(\hat{X}_{n,k})^2\right] - \left(\mathbf{E}\left[\hat{X}_{n,k}\right]\right)^2 \\
&= \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left(1 - (1-\mathbf{P}(w))^n - (1-\mathbf{P}(w'))^n + (1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right) \\
&\qquad + \sum_{w \in \mathcal{A}^k} \left(1 - (1-\mathbf{P}(w))^n\right) \\
&- \sum_{\substack{w,w' \in \mathcal{A}^k \\ w \neq w'}} \left(1 - (1-\mathbf{P}(w))^n - (1-\mathbf{P}(w'))^n + (1-\mathbf{P}(w)-\mathbf{P}(w'))^n\right) \\
&\qquad - \sum_{w \in \mathcal{A}^k} \left(1 - (1-\mathbf{P}(w))^n - (1-\mathbf{P}(w))^n + (1-\mathbf{P}(w))^{2n}\right) \\
&= \sum_{w \in \mathcal{A}^k} \left((1-\mathbf{P}(w))^n - (1-\mathbf{P}(w))^{2n}\right). \qquad (153)
\end{aligned}$$

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

$$\tilde{G}_k^{\mathrm{var}}(z) = \sum_{w \in \mathcal{A}^k} \left(e^{-\mathbf{P}(w)z} - e^{-(2\mathbf{P}(w)+(\mathbf{P}(w))^2)z}\right). \qquad (154)$$

The Mellin transform of the above function has the following form

$$\tilde{G}^{*\mathrm{var}}_k(z) = \Gamma(s)(p^{-s}+q^{-s})^k(-1+O(\mathbf{P}(w))). \qquad (155)$$

This is quite similar to what we saw in (107), 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 2).

4. Summary and Conclusions

We studied the first-order asymptotic growth of the first two (factorial) moments of the kth Subword Complexity. We recall that the kth 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 kth Subword Complexity to the kth Prefix Complexity over a random trie built over n independently generated binary strings. We recall that we define the kth 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 kth Prefix Complexity.

5. Future Challenges

The intervals' endpoints for a in Theorems 3 and 4 are not investigated in this work. The asymptotic analysis of the end points can be studied using van der Waerden saddle point method [24].

The analogous results are not (yet) known in the case where the underlying probability source has Markovian dependence or in the case of dynamical sources.

Author Contributions: This paper is based on a Ph.D. dissertation conducted by the first author under the supervision of the second author.

Funding: M.D. Ward's research is supported by FFAR Grant 534662, by the USDA NIFA Food and Agriculture Cyberinformatics and Tools (FACT) initiative, by NSF Grant DMS-1246818, by the NSF Science & Technology Center for Science of Information Grant CCF-0939370, and by the Society Of Actuaries.

Acknowledgments: The authors thank Wojciech Szpankowski and Mireille Régnier for insightful conversations on this topic.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PGF	Probabilty Generating Function
P	Probability
E	Expected value
Var	Variance
$\mathbf{E}[(X_{n,k})_2]$	The second factorial moment of $X_{n,k}$

References

1. Ehrenfeucht, A.; Lee, K.; Rozenberg, G. Subword complexities of various classes of deterministic developmental languages without interactions. *Theor. Comput. Sci.* **1975**, *1*, 59–75.
2. Morse, M.; Hedlund, G.A. Symbolic Dynamics. *Am. J. Math.* **1938**, *60*, 815–866.
3. Jacquet, P.; Szpankowski, W. *Analytic Pattern Matching: From DNA to Twitter*; Cambridge University Press: Cambridge, UK, 2015.
4. Bell, T.C.; Cleary, J.G.; Witten, I.H. *Text Compression*; Prentice-Hall: Upper Saddle River, NJ, USA, 1990.
5. Burge, C.; Campbell, A.M.; Karlin, S. Over-and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci.* **1992**, *89*, 1358–1362.
6. Fickett, J.W.; Torney, D.C.; Wolf, D.R. Base compositional structure of genomes. *Genomics* **1992**, *13*, 1056–1064.
7. Karlin, S.; Burge, C.; Campbell, A.M. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* **1992**, *20*, 1363–1370.
8. Karlin, S.; Mrázek, J.; Campbell, A.M. Frequent Oligonucleotides and Peptides of the Haemophilus Influenzae Genome. *Nucleic Acids Res.* **1996**, *24*, 4263–4272.
9. Pevzner, P.A.; Borodovsky, M.Y.; Mironov, A.A. Linguistics of Nucleotide Sequences II: Stationary Words in Genetic Texts and the Zonal Structure of DNA. *J. Biomol. Struct. Dyn.* **1989**, *6*, 1027–1038.
10. Chen, X.; Francia, B.; Li, M.; Mckinnon, B.; Seker, A. Shared information and program plagiarism detection. *IEEE Trans. Inf. Theory* **2004**, *50*, 1545–1551.
11. Chor, B.; Horn, D.; Goldman, N.; Levy, Y.; Massingham, T. Genomic DNA k-mer spectra: models and modalities. *Genome Biol.* **2009**, *10*, R108.
12. Price, A.L.; Jones, N.C.; Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **2005**, *21*, i351–i358.
13. Janson, S.; Lonardi, S.; Szpankowski, W. On the Average Sequence Complexity. In *Annual Symposium on Combinatorial Pattern Matching*. Springer: Berlin/Heidelberger, Germany, 2004; pp. 74–88.
14. Jacquet, P.; Szpankowski, W. Autocorrelation on words and its applications: Analysis of suffix trees by string-ruler approach. *J. Comb. Theory Ser. A* **1994**, *66*, 237–269.
15. Liang, F.M. *Word Hy-phen-a-tion by Com-put-er*. Technical report, Stanford University: Stanford, CA, USA, 1983.
16. Weiner, P. Linear pattern matching algorithms. In *14th Annual Symposium on Switching and Automata Theory (swat 1973)*. IEEE: Piscataway, NJ, USA, 1973; pp. 1–11.

17. Gheorghiciuc, I.; Ward, M.D. On correlation Polynomials and Subword Complexity. *Discrete Math. Theor. Comput. Sci.* **2007**, *AofA 07*, 1–18.
18. Bassino, F.; Clément, J.; Nicodème, P. Counting occurrences for a finite set of words: Combinatorial methods. *ACM Trans. Algorithms* **2012**, *8*, 31.
19. Park, G.; Hwang, H.K.; Nicodème, P.; Szpankowski, W. Profile of Tries. In *Latin American Symposium on Theoretical Informatics*. Springer: Berlin/Heidelberger, Germany, 2008; pp. 1–11.
20. Flajolet, P.; Sedgewick, R. *Analytic Combinatorics*; Cambridge University Press: Cambridge, UK, 2009.
21. Lothaire, M. *Applied Combinatorics on Words*; Cambridge University Press: Cambridge, UK, 2005; Volume 105.
22. Szpankowski, W. *Average Case Analysis of Algorithms on Sequences*; John Wiley & Sons: Chichester, UK, 2011; Volume 50.
23. Widder, D.V. *The Laplace Transform (PMS-6)*; Princeton University Press: Princeton, NJ, USA, 2015.
24. van der Waerden, B.L. On the method of saddle points. *Appl. Sci. Res.* **1952**, *2*, 33–45.