ON PATTERN FREQUENCY OCCURRENCES IN A MARKOVIAN SEQUENCE*†

Mireille Régnier‡
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France
`Mireille.Regnier@inria.fr`

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

Abstract

Consider a given pattern H and a random text T generated by a Markovian source. We study the frequency of pattern occurrences in a random text when overlapping copies of the pattern are counted separately. We present exact and asymptotic formulæ for moments (including the variance), and probability of r pattern occurrences for three different regions of r, namely: (i) $r = O(1)$, (ii) central limit regime, and (iii) large deviations regime. In order to derive these results, we first construct certain language expressions that characterize pattern occurrences which are later translated into generating functions. We then use analytical methods to extract asymptotic behaviors of the pattern frequency from the generating functions. These findings are of particular interest to molecular biology problems (e.g., finding patterns with unexpectedly high or low frequencies, and gene recognition), information theory (e.g., second-order properties of the relative frequency), and pattern matching algorithms (e.g., q-gram algorithms).

Key Words: Frequency of pattern occurrences, Markov source, autocorrelation polynomials, languages, generating functions, asymptotic analysis, large deviations.

*This paper was presented in part at the 1997 *Intern. Symp. on Information Theory*, Ulm, Germany.

†This research was supported by NATO Collaborative Grant CRG.950060. Part of this work was done during authors visits at Purdue University and at INRIA, Rocquencourt.

‡This work was additionally supported by ESPRIT LTR Project No. 20244 (ALCOM-IT) and GREG "Motifs dans les Sequences".

§This research was additionally supported by NSF Grants CCR-9201078, NCR-9206315 and NCR-9415491.

1 Introduction

Repeated patterns and related phenomena in words (also called sequences or strings) are known to play a central role in many facets of computer science, telecommunications, and molecular biology. One of the most fundamental questions arising in such studies is the frequency of pattern occurrences in another string known as the *text*. Applications of these results include wireless communications (cf. [1]), approximate pattern matching (cf. [23, 37]), molecular biology (cf. [32]), code synchronization (cf. [18, 19, 20]), and source coding (cf. [8]). In fact, this work and the one by Fudos *et al.* [14] were motivated by problems arising in approximate pattern matching by q-grams (cf. [23, 37]), developing performance models for database systems in wireless communications (cf. [1]), and gene recognition in a DNA sequence (cf. [32]), respectively. Actually, one of the earliest applications appears to be in code synchronization (cf. [18]).

We study the problem in a probabilistic framework in which the text is generated randomly either by a memoryless source (the so called *Bernoulli model*) or by a Markovian source (the so called *Markovian model*). In the former, every symbol of a finite alphabet is created independently of the other symbols, and the probabilities of symbol generation are not the same (if all probabilities of symbol generation are the same, the model is called *symmetric* Bernoulli model). In the Markovian model, the next symbol depends on a finite number of previous symbols.

The problem of pattern occurrences in a random string is a classical one. Feller already in 1968 suggested a solution in his book [11]. Several other authors also contributed to this problem: e.g., see [3, 6, 22, 27] and references there. However, the most important recent contributions belong to Guibas and Odlyzko, who in a series of papers (cf. [18, 19, 20]) laid the foundations of the analysis for the symmetric Bernoulli model. In particular, the authors of [20] computed the moment generating function for the number of strings of length n that do *not* contain any one of a given set of patterns. Certainly, this suffices to estimate the probability of at least one pattern occurrence in a random string generated by the symmetric Bernoulli model. Furthermore, Guibas and Odlyzko [20] in a passing remark also presented some basic results for several pattern occurrences in a random text for the symmetric Bernoulli model, and for the probability of no occurrence of a given pattern in the asymmetric Bernoulli model. Recently, Fudos *et al.* [14] computed the probability of exactly r occurrences of a pattern in a random text in the *asymmetric* Bernoulli model, just directly extending the results of Guibas and Odlyzko. The Markovian model was tackled by Li [27], Chrysaphinou and Papastavridis [6] who extended the Guibas and Odlyzko results of no pattern occurrence to Markovian texts. Prum *et al.* [33] (see also [36]) obtained the limiting distribution for the number of pattern occurrences in the Markovian model but without an explicit computation of the variance. Recently, Flajolet *et al.* [13] considered pattern occurrences in a random tree. Some other contributions are [3, 16, 24, 25, 30, 32, 39].

In this paper, we provide a complete characterization of the frequency of pattern occurrences in a random text generated according either to the Bernoulli model or the Markovian model. Our method of analysis treats uniformly both models, and therefore we concentrate on discussing the Markovian model. Let O_n denote the number of occurrences of a given pattern H in a random text when *overlapping* copies of the pattern are counted separately. In Theorem 2.1 we present the generating function of O_n which can be used to compute exactly the probability of r pattern occurrences in the text. Furthermore, this allows for an easy computation of all moments, using for instance a symbolic computation system. In this paper we present explicit formulæ for the mean and the variance of O_n. We observe that the evaluation of the variance was quite challenging as pointed out in [32] and [33]. It turns out that the variance depends on the internal structure of the pattern through the so called *autocorrelation polynomial*. We should point out that Prum *et al.* [33] proposed two statistical methods to estimate the variance which should be compared with our computations (cf. Theorem 2.2, and Section 3).

We also estimate asymptotically the probability of exactly r occurrences of the pattern for three different ranges of r (cf. Theorem 2.2). Namely, (i) $r = O(1)$, (ii) $r = EO_n + x\sqrt{n}$ for $x = O(1)$ (i.e., central limit regime), and (iii) $r = (1 + \delta)EO_n$ (i.e., large deviations regime). For our results to hold we assume that $nP(H) \to \infty$ (see [16] for other regimes of $nP(H)$). However, for a *given* pattern H it is natural to assume that the length of the pattern is constant with respect to n (and we adopt this assumption throughout).

Our results should be of particular interest to molecular biology, pattern matching algorithms, and information theory (e.g., relative frequency, code synchronization, coding, etc.). Two problems of molecular biology can benefit from these results, namely: finding patterns with unexpectedly (high or low) frequencies (the so called contrast words) [15], and recognizing genes by using statistical properties [10]. Statistical methods have been successfully used from the early 80's to extract information from sequences of DNA. In particular, identifying deviant short motifs, the frequency of which is either too high or too low, might point out unknown biological information (cf. [10] and others for the analysis of functions of contrast words in DNA texts). From this perspective, our results give estimates for the statistical significance of deviations of word occurrences from the expected values and allow a biologist to build a dictionary of contrast words in genetic texts. Recently, Coward [5] used our results to the search of exceptional patterns in the yeast genome.

Another biological problem for which our results might be useful is gene recognition. Most gene recognition techniques rely on the observation that the statistics of patterns (motifs/codon) occurrences in coding and non-coding regions are different. Our findings allow the estimation of the statistical significance of such differences, and the construction of the confidence interval for pattern occurrences.

One can also use these results to recognize statistical properties of various other information sources such as images, text, etc. In information theory, *relative frequency* defined as $\Delta_n = O_n/(n-m+1)$, where m is the length of the pattern, is often used to assess statistics of information sources. It is well known [8, 29] that Δ_n converges almost surely to the probability $P(H)$ of the pattern H, but much less is known about second-order properties of Δ_n such as the limiting distribution, large deviations, and rate of convergence. The rate of convergence to the source entropy – which is related to the rate of convergence of the relative frequency [29] – has recently appeared in the formulation of some results on data compression (cf. [28, 38, 41]). Marton and Shields [29] proved that Δ_n converges exponentially fast to $P(H)$ for sources satisfying the so called *blow-up property* (e.g., Markov sources, hidden Markov, etc). Our results characterize precisely such a convergence in the central limit regime and the large deviations regime for Markovian sources.

In the accompanying paper [35], we extended our results to *approximate* pattern occurrences or a set of pattern occurrences. Such extension is vital to some approximate pattern matching algorithms. Recently, Sutinen and Szpankowski [37] used these results to the performance evaluation of q-gram filtration algorithms.

This paper is organized as follows. In the next section, we present our main results and their consequences. The proofs are delayed until the last section. Our derivation in Section 3.1 use a language approach, thus is also valid for Markovian models since no probabilistic assumption is made. In Section 3.2 we translate language relationships into associated generating functions, and finally we use analytical tools in Section 3.3 to derive asymptotic results.

2 Main Results

Let us consider two strings, a pattern string $H = h_1 h_2 \dots h_m$ and a text string $T = t_1 t_2 \dots t_n$ of respective lengths equal to m and n over an alphabet $\mathcal{S}$ of size V. We shall write $\mathcal{S} = \{1, 2, \dots, V\}$ to simplify the presentation. Throughout, we assume that the pattern string is *fixed* and given, while the text string is random. More precisely, we consider the following two probabilistic models of text generation:

(B) Bernoulli Model

The text is a realization of a sequence of independently, identically distributed random variables (i.i.d.), such that a symbol $s \in \mathcal{S}$ occurs with probability $P(s)$.

(M) Markovian Model

The text is a realization of a *stationary* Markov sequence of order K, that is, the probability of the next symbol occurrence depends on the K previous symbols. In most derivations we deal only with the first order Markov chain ($K = 1$), and then we define the transition matrix $\mathbf{P} = \{p_{i,j}\}_{i,j \in \mathcal{S}}$ where $p_{i,j} = \Pr\{t_{k+1} = j | t_k = i\}$. By $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_V)$ we denote the stationary distribution satisfying $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$, and we write $\mathbf{\Pi}$ for the stationary matrix that consists of V identical rows equal to $\boldsymbol{\pi}$. Finally, by $\mathbf{Z}$ we denote the **fundamental matrix** $\mathbf{Z} = (\mathbf{I} - (\mathbf{P} - \mathbf{\Pi}))^{-1}$ where $\mathbf{I}$ is the identity matrix.

Throughout the paper we use systematically the following notation: lower-case bold letters represents row-vectors (e.g., $\boldsymbol{\pi}$), while upper-case bold letters denote matrices (e.g., $\mathbf{\Pi}$). To extract a particular element, say with index (i, j), from a matrix, say $\mathbf{P}$, we shall write $[\mathbf{P}]_{i,j} = p_{i,j}$. Finally, we recall that $(\mathbf{I} - \mathbf{P})^{-1} = \sum_{k \geq 0} \mathbf{P}^k$ provided the inverse matrix exists (i.e., $\det(\mathbf{I} - \mathbf{P}) \neq 0$ or $||\mathbf{P}|| < 1$ for any matrix norm $|| \cdot ||$). Below, we write $P(H_i^j) = \Pr\{T_{i+k}^{j+k} = H_i^j\}$ for the probability of an occurrence of substring $H_i^j = h_i \ldots h_j$ in the random text T_{i+k}^{j+k} between symbols $i + k$ and $j + k$ for any k (in particular, $P(H)$ denotes the probability of H appearing in the text).

Our goal is to estimate the frequency of multiple pattern occurrences in the text assuming either the Bernoulli or the Markovian model. We find it convenient and useful to express our findings in terms of languages. A language $\mathcal{L}$ is a collection of words satisfying some properties. We associate with a language $\mathcal{L}$ a generating function defined below.

Definition 1 *For any language $\mathcal{L}$ we define its generating function $L(z)$ as*

$$L(z) = \sum_{w \in \mathcal{L}} P(w) z^{|w|} \tag{1}$$

where $P(w)$ is the stationary probability of word w occurrence, $|w|$ is the length of w, and we adopt a usual convention that $P(\epsilon) = 1$, where ϵ is the empty word. In addition, we define H*-conditional generating function of $\mathcal{L}$ as*

$$L_{\mathrm{H}}(z) = \sum_{w \in \mathcal{L}} P(w | w_{-m} = h_1 \cdots w_{-1} = h_m) z^{|w|} = \sum_{w \in \mathcal{L}} P(w | w_{-m}^{-1} = H\} z^{|w|} \tag{2}$$

where w_{-i} stands for a symbol preceding the first character of w at distance i.

It turns out that several properties of pattern occurrences depend on the so called *autocorrelation polynomial* that we define next:

Definition 2 *Given a string H, we define the* autocorrelation set $\mathcal{A}$ *as:*

$$\mathcal{A} = \{H_{k+1}^m : \; H_1^k = H_{m-k+1}^m\} \; , \tag{3}$$

and by HH we denote the set of positions k satisfying $H_1^k = H_{m-k+1}^m$. The generating function of language $\mathcal{A}$ is denoted as $A(z)$ and we call it the autocorrelation polynomial. *Its H-conditional generating function is denoted $A_{\mathrm{H}}(z)$. In particular,*

$$A_H(z) = \sum_{k \in HH} P(H_{k+1}^m | H_k^k) z^{m-k} \tag{4}$$

for a Markov chain of order $K = 1$.

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

Example: *Autocorrelation Functions*

Let us assume that $H = 101$ over a binary alphabet $\mathcal{S} = \{0, 1\}$. Observe that $HH = \{1, 3\}$ and $\mathcal{A} = \{\epsilon, 01\}$, where ϵ is the empty word. Thus, for the symmetric Bernoulli model (both symbols occur with the same probability equal to 0.5) we have $A(z) = 1 + \frac{z^2}{4}$, while for the Markovian model of order one, we obtain $A_{101}(z) = 1 + p_{10}p_{01}z^2$. $\Box$

We now can formulate our main results. In the sequel, we denote by $O_n(\mathrm{H})$ (or simply by O_n) a random variable representing the number of occurrences of H in a random text T of size n. We introduce the generating function of the language $\mathcal{T}_r$ of words that contains exactly r occurrences of H, namely: $T^{(r)}(z) = \sum_{n \geq 0} \Pr\{O_n(\mathrm{H}) = r\} z^n$. We also define a bivariate generating function as follows:

$$T(z, u) = \sum_{r=1}^{\infty} T^{(r)}(z) u^r = \sum_{r=1}^{\infty} \sum_{n=0}^{\infty} \Pr\{O_n(\mathrm{H}) = r\} z^n u^r \tag{5}$$

for $|z| \leq 1$ and $|u| \leq 1$.

Our main results are summarized in the following two theorems. First theorem presents exact formulæ for the generating functions $T^{(r)}(z)$ and $T(z, u)$, and can be used to compute exactly all parameters related to the pattern occurrence $O_n(H)$. In the second theorem, we provide asymptotic formulæ for $\Pr\{O_n(H) = r\}$ for three regimes of r, namely: (i) $r = O(1)$, (ii) $r = EO_n + x\sqrt{\mathrm{Var}\; O_n}$ when $x = O(1)$ (i.e., local central limit), (iii) $r = (1 + \delta)EO_n$ for some δ (i.e., large deviations). All proofs are presented in the next section: section 3.2 contains the proof of Theorem 2.1 while the proof of Theorem 2.2 can be found in Section 3.3.

Theorem 2.1 *Let H be a given pattern of size m, and T be a random text of length n generated according to a stationary Markov chain of order 1 over a V-ary alphabet $\mathcal{S}$. The generating functions $T^{(r)}(z)$ and $T(z,u)$ become*

$$T^{(r)}(z) = R(z)M_{\mathrm{H}}^{r-1}(z)U_{\mathrm{H}}(z)\ , r \geq 1 \tag{6}$$

$$T(z,u) = R(z)\frac{u}{1-uM_{\mathrm{H}}(z)}U_{\mathrm{H}}(z)\ , \tag{7}$$

where

$$M_{\mathrm{H}}(z) = 1 + \frac{z-1}{D_H(z)}\ , \tag{8}$$

$$U_{\mathrm{H}}(z) = \frac{1}{D_H(z)}\ , \tag{9}$$

$$R(z) = z^m P(H)\frac{1}{D_{\mathrm{H}}(z)}\ . \tag{10}$$

with

$$D_H(z) = (1-z)A_{\mathrm{H}}(z) + z^m P(H)(1+(1-z)F(z)). \tag{11}$$

The function $F(z)$ is defined for $|z| \leq R$ where $R = \frac{1}{||\mathbf{P}-\mathbf{\Pi}||}$ as follows

$$F(z) = \frac{1}{\pi_{h_1}}[(\mathbf{P}-\mathbf{\Pi})(\mathbf{I}-(\mathbf{P}-\mathbf{\Pi})z)^{-1}]_{h_m,h_1}\ . \tag{12}$$

where h_1 and h_m are the first and the last symbols of H, respectively. In the Bernoulli model, $F(z) = 0$, and hence

$$D(z) = (1-z)A_{\mathrm{H}}(z) + z^m P(H)$$

with the other formulæ as above.

Theorem 2.1 is a starting point of our next finding that deals with asymptotics for $n \to \infty$ when $nP(H) \to \infty$. The results below are derived in Section 3.3 using analytical tools.

Theorem 2.2 *Let the hypotheses of Theorem 2.1 be fulfilled and $nP(H) \to \infty$.*

(i) Moments. *The expectation $EO_n(H)$ satisfies for $n \geq m$:*

$$EO_n(H) = P(H)(n-m+1)\ , \tag{13}$$

while the variance becomes for some $r > 1$

$$\mathrm{Var}\ O_n(H) = nc_1 + c_2 + O(r^{-n})\ , \tag{14}$$

where

$$\begin{aligned}
c_1 &= P(H)(2A_H(1) - 1 - (2m-1)P(H) + 2P(H)E_1)) \ , && (15)\\
c_2 &= P(H)((m-1)(3m-1)P(H) - (m-1)(2A_H(1)-1) - 2A'_{\mathrm{H}}(1)) \\
&- 2(2m-1)P(H)^2E_1 + 2E_2P(H)^2, && (16)
\end{aligned}$$

and the constants E_1, E_2 are

$$\begin{aligned}
E_1 &= \frac{1}{\pi_{h_1}}[(\mathbf{P}-\mathbf{\Pi})\mathbf{Z}]_{h_m,h_1}, && (17)\\
E_2 &= \frac{1}{\pi_{h_1}}[(\mathbf{P}^2-\mathbf{\Pi})\mathbf{Z}^2]_{h_m,h_1} && (18)
\end{aligned}$$

where $\mathbf{Z} = (\mathbf{I} - (\mathbf{P}-\mathbf{\Pi}))^{-1}$ is the fundamental matrix of the underlying Markov chain.
In the **Bernoulli model**, *$E_1 = E_2 = 0$ since $\mathbf{P} = \mathbf{\Pi}$, and (14) reduces to an equality for $n \geq 2m-1$. Thus*

$$\mathrm{Var}\ O_n(H) = nc_1 + c_2 \ . \tag{19}$$

with

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

(ii) Distribution: Case $r = O(1)$. *Let ρ_H be the root of $D_H(z) = 0$ of smallest modulus and multiplicity one. Then, ρ_H is real positive and lies outside the unit circle $|z| < 1$, and there exists $\rho > \rho_H$ such that:*

$$\Pr\{O_n(\mathrm{H}) = r\} = \sum_{j=1}^{r+1}(-1)^j a_j \binom{n}{j-1}\rho_{\mathrm{H}}^{-(n+j)} + O(\rho^{-n}) \ , \tag{20}$$

where

$$a_{r+1} = \frac{\rho_H^m P(\mathrm{H})\,(\rho_{\mathrm{H}}-1)^{r-1}}{(D'_{\mathrm{H}}(\rho_{\mathrm{H}}))^{r+1}} \ , \tag{21}$$

and the remaining coefficients can be computed according to

$$a_j = \frac{1}{(r+1-j)!}\lim_{z\to\rho_{\mathrm{H}}}\frac{d^{r+1-j}}{dz^{r+1-j}}\left(T^{(r)}(z)(z-\rho_{\mathrm{H}})^{r+1}\right) \tag{22}$$

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

(iii) Central Limit Theorem: Case $r = EO_n + x\sqrt{\mathrm{Var}\ O_n}$. *For $x = O(1)$ we have as $n \to \infty$*

$$\Pr\{O_n(H) = r\} = \frac{1}{\sqrt{2\pi c_1 n}}e^{-\frac{1}{2}x^2}\left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \ , \tag{23}$$

where c_1 is defined in (15) above.
(iv) LARGE DEVIATIONS: CASE $r = (1+\delta)EO_n$. *Let $a = (1+\delta)P(H)$ with $\delta \neq 0$. For complex t, define $\rho(t)$ to be the root of*

$$1 - e^t M_{\mathrm{H}}(e^{\rho}) = 0 \ , \tag{24}$$

while ω_a and σ_a are defined as

$$\begin{aligned} -\rho'(\omega_a) &= a, & (25)\\ -\rho''(\omega_a) &= \sigma_a^2. & (26) \end{aligned}$$

Then

$$\Pr\{O_n(H) = (1+\delta)EO_n\} = \frac{1}{\sigma_a\sqrt{2\pi(n-m+1)}} e^{-(n-m+1)I(a)} \left(1 + O\left(\frac{1}{n}\right)\right) \tag{27}$$

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

As mentioned before, the above results find applications in information theory and molecular biology. For example, the *relative frequency* is an important concept in information theory, and it is defined as

$$\Delta_n(H) = \frac{O_n(H)}{n-m+1} \ .$$

Relative frequency appears in the definition of types and typical types (cf. [8]), and is often used to estimate information source statistics. As an easy corollary to Theorem 2.2, we obtain the following second-order characterization of $\Delta_n(H)$:

Corollary 2.1 *Under hypotheses of Theorem 2.2, the following holds:*
(i) [CENTRAL LIMIT REGIME] *For $x = O(1)$*

$$\Pr\{\Delta_n(H) = P(H) + x\sqrt{c_1/(n-m+1)}\} = \frac{1}{\sqrt{2\pi c_1 n}} e^{-\frac{1}{2}x^2} \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \ . \tag{28}$$

(ii) [LARGE DEVIATIONS] *For $a = (1+\delta)P(H)$ with $\delta > 0$*

$$\Pr\{\Delta_n(H) \geq (1+\delta)P(H)\} = \frac{1}{\sigma_a\sqrt{2\pi(n-m+1)}(1-e^{-I(a)})} e^{-(n-m+1)I(a)} \left(1 + O\left(\frac{1}{n}\right)\right) \tag{29}$$

where ω_a and $I(a)$ are defined in Theorem 2.2 (iii).

3 Analysis

The key element of our analysis is a derivation of the generating function $T(z,u)$ presented in Theorem 2.1. The first part of the discussion below is quite general and works uniformly for both the Bernoulli model and the Markovian model. It is based on constructing some special languages and finding relationships among them. Later, in Section 3.2 we translate these relations into formulæ for generating functions.

3.1 Combinatorial Relationships on Certain Languages

A collection of words sharing a given property is commonly called a *language.* This section is devoted to presenting combinatorial relationships between some languages that help to derive some results in a uniform manner. In this section we do not make any probabilistic assumptions.

We start with some definitions:

Definition 3 *Given a pattern* H*:*

(i) Let $\mathcal{T}$ *be a language of words containing at least one occurrence of* H*, and for any integer* $r \geq 1$*, let* $\mathcal{T}_r$ *be the language of words containing exactly* r *occurrences of* H.

(ii) We define $\mathcal{R}$ *as the set of words containing only one occurrence of* H*, located at the right end. We also define* $\mathcal{U}$ *as*

$$\mathcal{U} = \{u: \ \mathrm{H} \cdot u \in \mathcal{T}_1\} \tag{30}$$

where the operation $\cdot$ *means concatenation of words. In other words, a word* $u \in \mathcal{U}$ *if* $\mathrm{H} \cdot u$ *has exactly one occurrence of* H *at the left end of* $\mathrm{H} \cdot u$*.*

(iii) Let $\mathcal{M}$ *be the language:*

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

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

We now can describe languages $\mathcal{T}$ and $\mathcal{T}_r$ in terms of $\mathcal{R}$, $\mathcal{M}$ and $\mathcal{U}$. This will further lead to a simple formula for the generating function of $O_n(\mathrm{H})$.

Theorem 3.1 *Language $\mathcal{T}$ satisfies the fundamental equation*

$$\mathcal{T} = \mathcal{R} \cdot \mathcal{M}^* \cdot \mathcal{U} \ . \tag{31}$$

Notably, language $\mathcal{T}_r$ can be represented for any $r \geq 1$ as follows

$$\mathcal{T}_r = \mathcal{R} \cdot \mathcal{M}^{r-1} \cdot \mathcal{U} \ . \tag{32}$$

Here, by definition $\mathcal{M}^0 := \{\epsilon\}$ and $\mathcal{M}^* := \bigcup_{r=0}^{\infty} \mathcal{M}^r$.

Proof: We first prove (32) and obtain our decomposition of $\mathcal{T}_r$ as follows: The first occurrence of H in a word belonging to $\mathcal{T}_r$ determines a prefix p that is in $\mathcal{R}$. Then, one concatenates a non-empty word w that creates the second occurrence of H. Hence, w is in $\mathcal{M}$. This process is repeated $r-1$ times. Finally, one adds after the last H occurrence a suffix u that does not create a new occurrence of H. Equivalently, Hu is such that u is in $\mathcal{U}$, and w is a proper subword of Hu. Finally, a word belongs to $\mathcal{T}$ if for some $1 \leq r < \infty$ it belongs to $\mathcal{T}_r$. The set union $\bigcup_{r=1}^{\infty} \mathcal{M}^{r-1}$ yields precisely $\mathcal{M}^*$. ■

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

Theorem 3.2 *The languages $\mathcal{M}, \mathcal{U}$ and $\mathcal{R}$ satisfy*

$$\bigcup_{k \geq 1} \mathcal{M}^k = \mathcal{W} \cdot \mathrm{H} + \mathcal{A} - \{\epsilon\} \ , \tag{33}$$

$$\mathcal{U} \cdot \mathcal{S} = \mathcal{M} + \mathcal{U} - \{\epsilon\} \ , \tag{34}$$

$$\mathrm{H} \cdot \mathcal{M} = \mathcal{S} \cdot \mathcal{R} - (\mathcal{R} - \mathrm{H}) \ , \tag{35}$$

where $\mathcal{W}$ is the set of all words, $\mathcal{S}$ is the alphabet set and $+$ and $-$ are disjoint union and subtraction of languages. In particular, a combination of (34) and (35) gives

$$\mathrm{H} \cdot \mathcal{U} \cdot \mathcal{S} - \mathrm{H} \cdot \mathcal{U} = (\mathcal{S} - \epsilon)\mathcal{R} \ . \tag{36}$$

Additionally, we have:

$$\mathcal{T}_0 \cdot \mathrm{H} = \mathcal{R} \cdot \mathcal{A} \ . \tag{37}$$

Proof: All the relations above are proved in a similar fashion. We first deal with (33). Let k be the number of H occurrences in $\mathcal{W} \cdot \mathrm{H}$. By definition, $k \geq 1$ and the last occurrence is on the right: this implies that $\mathcal{W} \cdot \mathrm{H} \subseteq \bigcup_{k \geq 1} \mathcal{M}^k$. Furthermore, a word w in $\bigcup_{k \geq 1} \mathcal{M}^k$ is not in $\mathcal{W} \cdot \mathrm{H}$ iff its size $|w|$ is smaller than $|\mathrm{H}|$. Then, the second H occurrence in Hw overlaps with H, which means that w is in $\mathcal{A} - \epsilon$.

Let us turn now to (34). When one adds a character s right after a word u from $\mathcal{U}$, two cases may occur. Either Hus still does not contain a second occurrence of H, which means that us is a non-empty word of $\mathcal{U}$. Or a new H appears, clearly at the right end. Then, us is in $\mathcal{M}$. Furthermore, the whole set $\mathcal{M} + (\mathcal{U} - \epsilon)$ is attained, i.e., a strict prefix of $\mathcal{M}$ cannot contain a new H occurrence. Hence, it is in $\mathcal{U}$, and a strict prefix of a $\mathcal{U}$-word is in $\mathcal{U}$.

We now prove (35). Let $x = sw$ be a word in $\mathrm{H} \cdot \mathcal{M}$ where s is a symbol from $\mathcal{S}$. As x contains exactly two occurrences of H located at its left and right ends, w is in $\mathcal{R}$ and x is in $\mathcal{S} \cdot \mathcal{R} - \mathcal{R}$. Reciprocally, if a word swH from $\mathcal{S} \cdot \mathcal{R}$ is not in $\mathcal{R}$, then swH contains a second H occurrence starting in sw. As wH is in $\mathcal{R}$, the only possible position is at the left end, and then x is in $\mathrm{H} \cdot \mathcal{M}$. We now rewrite:

$$\mathcal{S} \cdot \mathcal{R} - \mathcal{R} = \mathcal{S} \cdot \mathcal{R} - (\mathcal{R} \cap \mathcal{S} \cdot \mathcal{R}) = \mathcal{S} \cdot \mathcal{R} - (\mathcal{R} - \mathrm{H})$$

which yields $\mathrm{H} \cdot \mathcal{M} - \mathrm{H} = (\mathcal{S} - \epsilon) \cdot \mathcal{R}$.

Deriving (37) is only a little more intricate. Let t be some word in $\mathcal{T}_0$. We consider the factorization $t = w_1 w_2$ such that w_2 is the largest suffix that also is a $(m-k)$-prefix of H, with $k \in HH$ and $m = |H|$. In other words, w_2 is the largest suffix satisfying the equation $w_2 \cdot H = H \cdot a$, where a is in $\mathcal{A}$. If w_1H were not in $\mathcal{R}$, a second occurrence of H would occur in w_1H starting in w_1. As $w_1\mathrm{H}a = w_1 w_2 \mathrm{H}$, this contradicts the maximal property of w_2. Therefore, $\mathcal{T}_0 \cdot H \subseteq \mathcal{R} \cdot \mathcal{A}$. Finally, we consider a word $w_1 H a$ in $\mathcal{R} \cdot \mathcal{A}$. We may rewrite it as $H \cdot a = w_2 \cdot H$. It suffices now to show that $w_1 w_2 \in \mathcal{T}_0$. Indeed, since $|w_2| < |H|$, any occurrence of H would go across w_1 and $w_1 H$ would contain two occurrences of H, which contradicts the definition of $\mathcal{R}$. This proves $R \cdot \mathcal{A} \subseteq \mathcal{T}_0 \cdot H$, and completes the proof of Theorem 3.2. ■

3.2 Associated Generating Functions

In the previous section we did not make any probabilistic assumption. Thus, Theorem 3.2 is true for any model, including Bernoulli and Markovian ones. In this section, we translate the language relationships into generating functions. Therefore, we need to return to our probabilistic assumptions. Most of our derivations deal with the Markovian model.

To transfer our language relations into generating functions, we need a few rules associated with two operations on languages. Namely: disjoint union $+$ and concatenation $\cdot$ become the sum operation and the multiplication operation on generating functions. We start with the following simply property that is true for both probabilistic models:

(P1) Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two arbitrary languages with generating functions $L_1(z)$ and $L_2(z)$, respectively. Then, the language $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ is transferred into the generating function $L(z)$ such that

$$L(z) = L_1(z) + L_2(z) \ .$$

To translate the concatenation operation, one needs to consider the Bernoulli and the Markovian models separately. We start with the **Bernoulli model**:

(P2) Let us now consider a new language $\mathcal{L}$ which is constructed from the concatenation of two other languages, say $\mathcal{L}_1$ and $\mathcal{L}_2$, that is $\mathcal{L} = \mathcal{L}_1 \cdot \mathcal{L}_2$. In the *Bernoulli model*, the generating function $L(z)$ of $\mathcal{L}$ becomes

$$L(z) = L_1(z)L_2(z)$$

since $P(wv) = P(w)P(v)$ for $w \in \mathcal{L}_1$ and $v \in \mathcal{L}_2$. In particular, the generating function $L(z)$ of $\mathcal{L} = \mathcal{S} \cdot \mathcal{L}_1$ is $L(z) = zL_1(z)$, where $\mathcal{S}$ is the alphabet set, since $S(z) = \sum_{p\in\mathcal{S}} P(s)z = z$.

In the **Markovian model** $P(wv) \neq P(w)P(v)$, thus property (P2) is not any longer true. We have to replace it by a more sophisticated one. We have to condition $\mathcal{L}_2$ on symbols preceding a word from $\mathcal{L}_2$ (i.e., belonging to $\mathcal{L}_1$). In general, for a K order Markov chain, one must distinguish V^K ending states for $\mathcal{L}_1$ and V^K initial states for $\mathcal{L}_2$. For simplicity of presentation, we only consider first-order Markov chains (i.e., $K = 1$), and we write $\ell(w)$ for the last symbol of a word w. In particular, to re-write property (P2) we must introduce the following conditional generating function for a language $\mathcal{L}$:

$$L_i^j(z) = \sum_{w\in\mathcal{L}} P(w, \ell(w) = j | w_1 = i) z^{|w|}$$

Then property (P2) becomes for the Markovian model:

(P2) Let $\mathcal{L} = \mathcal{W} \cdot \mathcal{V}$. Then,

$$L_k^l(z) = \sum_{i,j\in\mathcal{S}} p_{ji} W_k^j(z) V_i^l(z) \ . \tag{38}$$

where $W_k^j(z)$ and $V_i^l(z)$ are conditional generating functions for $\mathcal{W}$ and $\mathcal{V}$ respectively. To prove this, let $w \in \mathcal{W}$ and $v \in \mathcal{V}$. Observe that

$$\begin{aligned} P(wv) &= \sum_{j\in\mathcal{S}} P(wv,\ \ell(w) = j) \\ &= \sum_{j\in\mathcal{S}} P(w,\ \ell(w) = j) P(v|\ell(w) = j) \\ &= \sum_{j\in\mathcal{S}} \sum_{i\in\mathcal{S}} P(w,\ \ell(w) = j) p_{ji} P(v|v_1 = i) \ . \end{aligned}$$

After conditioning on the first symbol of $\mathcal{W}$ and the last symbol of $\mathcal{V}$, we prove (38).

In passing, we observe that in the Markovian model of our problem one actually must deal only with two kinds of words: We have words w for which no assumption is made on the preceding words (e.g., these are the words in language $\mathcal{R}$ with generating function $R(z)$); and we deal with words for which the preceding word admits H as a suffix (e.g., words in $\mathcal{U}$ and $\mathcal{M}$ whose H-conditional generating functions are $U_H(z)$ and $M_H(z)$, respectively).

The lemma below together with Theorem 3.1 proves (6) and (7) of Theorem 2.1.

Lemma 3.1 *The generating functions associated with languages $\mathcal{M}$,$\mathcal{U}$ and $\mathcal{R}$ satisfy*

$$\frac{1}{1-M_{\mathrm{H}}(z)} = A_{\mathrm{H}}(z) + P(H)z^m\left(\frac{1}{1-z}+F(z)\right) \tag{39}$$

$$U_{\mathrm{H}}(z) = \frac{M_{\mathrm{H}}(z)-1}{z-1}\ , \tag{40}$$

$$R(z) = P(H)z^m \cdot U_{\mathrm{H}}(z) \tag{41}$$

provided the underlying Markov chain is stationary.

Proof: We first prove (40). Interestingly, it does not need the stationarity assumption. Let us consider the language relationship (34) from Theorem 3.2 which we re-write as $\mathcal{U}\cdot\mathcal{S}-\mathcal{U}=\mathcal{M}-\epsilon$. Observe that $\sum_{j\in\mathcal{S}} p_{i,j}z = z$. Hence, set $\mathcal{U}\cdot\mathcal{S}$ yields (conditioning on the left occurrence of H):

$$\sum_{w\in\mathcal{U}}\sum_{j\in\mathcal{S}} P(wj|H)z^{|wj|} = \sum_{i\in\mathcal{S}}\sum_{w\in\mathcal{U},\ell(w)=i} P(w|H)z^{|w|}\sum_{j\in\mathcal{S}} p_{i,j}z = U_{\mathrm{H}}(z)\cdot z\ .$$

Of course, $\mathcal{M}-\epsilon$ and $\mathcal{U}$ translate into $M_H(z)-1$ and $U_{\mathrm{H}}(z)$, and (40) is proved.

We now turn our attention to (41), and we use relationship (35) of Theorem 3.2. Observe that $\mathcal{S}\cdot\mathcal{R}$ can be rewritten as

$$\sum_{j,i\in\mathcal{S}^2}\sum_{iw\in\mathcal{R}} P(jiw)z^{|jiw|} = z^2\sum_{j\in\mathcal{S}}\sum_{i\in\mathcal{S}}\pi_j p_{j,i}\sum_{iw\in\mathcal{R}} P(w|w_{-1}=i)z^{|w|}\ .$$

But, due to the stationarity of the underlying Markov chain $\sum_j \pi_j p_{j,i} = \pi_i$. As $\pi_i P(w|w_{-1} = i) = P(iw)$, we get $zR(z)$. Furthermore, in (35) $H\cdot\mathcal{M}-H$ translates into $P(H)z^m\cdot(M_{\mathrm{H}}(z)-1)$. But, by (40), this becomes $P(H)z^m\cdot U_{\mathrm{H}}(z)(z-1)$, and after a simplification, we prove (41).

Finally, we deal with (39), and prove it using (33) from Theorem 3.2. Left-hand side of (33) involves language $\mathcal{M}$, hence we must condition on the left occurrence of H. In particular, $\bigcup_{r\geq 1}\mathcal{M}^r+\epsilon$ of (33) translates into $\frac{1}{1-M_{\mathrm{H}}(z)}$. Now we deal with $\mathcal{W}\cdot H$ of the right-hand side of (33). *Conditioning* on the left occurrence of H, the generating function $W_(z)H(z)$ of $\mathcal{W}\cdot H$ becomes

$$\begin{aligned}
W_H(z)H(z) &= \sum_{n\geq 0}\sum_{|w|=n} z^{n+m}P(wH|w_{-1}=\ell(\mathrm{H})) \\
&= \sum_{n\geq 0}\sum_{|w|=n} z^{n}P(wh_1|w_{-1}=\ell(H))P(v=h_2\dots h_m|v_{-1}=h_1)z^m \quad .
\end{aligned}$$

We have $P(v=h_2\dots h_m|v_{-1}=h_1)z^m = \frac{1}{\pi_{h_1}}z^mP(H)$, and for $n\geq 0$:

$$\sum_{|w|=n} P(wh_1|w_{-1}=\ell(\mathrm{H})) = [\mathbf{P}^{n+1}]_{\ell(\mathrm{H}),h_1}$$

where, we recall, $\ell(H)=h_m$ is the last character of H. In summary: language $\mathcal{W}\cdot\mathrm{H}$ contributes $P(H)z^m\left[\frac{1}{\pi_{h_1}}\sum_{n\geq 0}\mathbf{P}^{n+1}z^n\right]_{\ell(H),h_1}$, while language $\mathcal{A}-\{\epsilon\}$ introduces $A_{\mathrm{H}}(z)-1$. We now observe that for any symbols i and j

$$[\frac{1}{\pi_j}\sum_{n\geq 0}\mathbf{\Pi}z^n]_{i,j} = \sum_{n\geq 0} z^n = \frac{1}{1-z} \quad .$$

Using the equality $\mathbf{P}^{n+1}-\mathbf{\Pi}=(\mathbf{P}-\mathbf{\Pi})^{n+1}$ (which follows from a consecutive application of the identity: $\mathbf{\Pi}\mathbf{P}=\mathbf{\Pi}$), we finally obtain the sum in (39). This completes the proof of the theorem. ■

Remark. The generating function of language $\mathcal{T}_0^j$ (i.e., no H occurrence with the last symbol of a word from $\mathcal{T}_0^j$ being j) in the Markov case was previously derived by Chrysaphinou and Papastavridis in [6]. We observe that the generating function $T^{(0)}(z)$ of $\mathcal{T}_0$ easily follows from (6) and the equation $T^{(0)}(z)=\frac{1}{1-z}-\sum_{r\geq 1}T^{(r)}(z)$.

3.3 Moments and Limiting Distribution

In this final subsection, we derive the first two moments of O_n as well as asymptotics for $\Pr\{O_n=r\}$ for different ranges of r, that is, we prove Theorem 2.2. Actually, we should mention that using general results on Markov chains and renewal theory one immediately guesses that the limiting distribution must be normal for $r=EO_n+O(\sqrt{n})$. However, here the challenge is to estimate precisely the variance. Our approach offers an easy, uniform, and precise derivation of all moments, including the variance, as well as local limit distributions (including the convergence rate) for the central and large deviations regimes.

A. Moments

First of all, from Theorem 2.1 we conclude that

$$\begin{aligned} T_u(z,1) &= \frac{z^m P(H)}{(1-z)^2} \ , \\ T_{uu}(z,1) &= \frac{2z^m P(H) M_{\mathrm{H}}(z) D_{\mathrm{H}}(z)}{(1-z)^3} \end{aligned}$$

where $T_u(z,1)$ and $T_{uu}(z,1)$ are the first and the second derivatives of $T(z,u)$ at $u=1$. Now, we observe that both expressions admit as a numerator a function that is analytic beyond the unit circle. This allows for a very simple computation of the expectation and variance based on the following basic formula

$$[z^n](1-z)^{-p} = \frac{\Gamma(n+p)}{\Gamma(p)\Gamma(n+1)} \tag{42}$$

where $[z^n]$ means the coefficient of z^n. To obtain EO_n we proceed as follows, for $n \geq m$:

$$EO_n = [z^n]T_u(z,1) = P(H)[z^{n-m}](1-z)^{-2} = (n-m+1)P(H) \ .$$

We denote

$$\Phi(z) = 2z^m P(H) M_{\mathrm{H}}(z) D_{\mathrm{H}}(z) \ ,$$

which is a polynomial in the Bernoulli case. We use the Taylor expansion

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

where $f(z)$ is a polynomial of degree $2m-2$. It follows that $[z^n](z-1)f(z)$ is 0 for $n \geq 2m-1$ and, using formula (42), we get:

$$EO_n(O_n-1) = [z^n]T_{uu}(z,1) = \Phi(1)\frac{(n+2)(n+1)}{2} - \Phi'(1)(n+1) + \frac{1}{2}\Phi''(1).$$

Observing that $M_{\mathrm{H}}(z)D_{\mathrm{H}}(z) = D_{\mathrm{H}}(z) + (1-z)$, we use MAPLE to obtain a precise formula for the variance (cf. (14) of Theorem 2.2). In the Markov case, we have to compute the additional term

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

where $F(z)$ is analytic beyond the unit circle for $|z| \leq R$, with $R > 1$. The Taylor expansion of $F(z)$ is $E_1 + (1-z)E_2$ and, applying (42) again yields the result. In a similar manner, we can compute all the moments of O_n.

B. DISTRIBUTION: CASE $r = O(1)$

Now, we prove part (ii) of Theorem 2.2, that is, we establish an asymptotic expression for $\Pr\{O_n = r\}$ for $r = O(1)$. We first re-write the formula on $T^{(r)}(z)$ as follows

$$T^{(r)}(z) = \frac{z^m P(H)(D_H(z) + z - 1)^{r-1}}{D_H^{r+1}(z)} \; . \tag{43}$$

Observe that $\Pr\{O_n = r\}$ is the coefficient at z^n of $T^{(r)}(z)$. By Hadamard's theorem (cf. [34]), the asymptotics of the coefficients of $T^{(r)}(z)$ depend on the singularities of $T^{(r)}(z)$. In our case, the generating function is a rational function, thus we can only expect poles (for which the denominator $D_H(z)$ vanishes). The next lemma establishes the existence and properties of such a pole.

Lemma 3.2 *The equation $D_{\mathrm{H}}(z) = 0$ has at least one root, and all its roots are of modulus greater than* 1*.*

Proof: A root of $D_{\mathrm{H}}(z) = (1-z)/(1-M_H(z))$ is clearly a pole of $\frac{1}{1-M_{\mathrm{H}}(z)}$. As $\frac{1}{1-M_{\mathrm{H}}(z)}$ is the generating function of a language, it converges for $|z| < 1$ and has no pole of modulus smaller than 1. Since $D_{\mathrm{H}}(1) \neq 0$, $z = 1$ is a simple pole of $1/(1 - M_H(z))$. As all its coefficients are real and positive, there is no other pole of modulus $|z| = 1$. It follows that all roots of $D_{\mathrm{H}}(z)$ are of modulus greater than 1. The existence of a root is guaranteed since $D_{\mathrm{H}}(z)$ is either a polynomial (Bernoulli model) or a ratio of polynomials (Markov model). ∎

In view of the above, the generating function $T^{(r)}(z)$ can be expanded around its root of smallest modulus, say ρ_{H}, as Laurent's series (cf. [34, 40]):

$$T^{(r)}(z) = \sum_{j=1}^{r+1} \frac{a_j}{(z-\rho_{\mathrm{H}})^j} + \widetilde{T}^{(r)}(z) \tag{44}$$

where $\widetilde{T}^{(r)}(z)$ is analytical in $|z| < \rho'$ and ρ' is defined as $\rho' = \inf\{|\rho| : \ \rho > \rho_H \text{ and } D_{\mathrm{H}}(\rho) = 0\}$. The constants a_j satisfy the formulæ (22). This formula simplifies into (21) for the leading constant a_{-r-1}. As a consequence of analyticity [40] we have for $1 < \rho_H < \rho < \rho'$: $[z^n]\widetilde{T}^{(r)}(z) = O(\rho^{-n})$. Hence, the term $\widetilde{T}^{(r)}(z)$ contributes only to the lower terms in the asymptotic expansion of $T^{(r)}(z)$.

We need an asymptotic expansion for the first terms in (43). This is rather a standard computation (cf. [40]), but for completeness we provide a short proof. The following chain of identities is easy to justify for any $\rho > 0$:

$$\begin{aligned}
\sum_{j=1}^{r+1} \frac{a_j}{(z-\rho)^j} &= \sum_{j=1}^{r+1} \frac{a_j(-1)^j}{\rho^j(1-(z/\rho)^j} \\
&= \sum_{j=1}^{r+1}(-1)^j a_j \rho^{-j} \sum_{n=0}^{\infty} \binom{n+j-1}{n} \left(\frac{z}{\rho}\right)^n \\
&= \sum_{n=1}^{\infty} z^n \sum_{j=1}^{\min\{r+1,n\}} (-1)^j a_j \binom{n}{j-1} \rho^{-(n+j)} .
\end{aligned}$$

After some algebra, we prove part (ii) of Theorem 2.2.

C. Central Limit Theorem: Case $r = EO_n + xO(\sqrt{n})$

We now establish part (iii) of Theorem 2.2, that is, we compute $\Pr\{O_n = r\}$ for $r = EO_n + x\sqrt{\text{Var } O_n}$ when $x = O(1)$. Let $\mu_n = EO_n(H) = (n-m+1)P(H)$ and $\sigma_n^2 = \text{Var } O_n(H) \sim c_1 n$. To establish normality of $(O_n(H) - \mu_n)/\sigma_n$, it suffices, according to Lévy's theorem, to prove the following

$$\lim_{n\to\infty} e^{-\tau\mu_n/\sigma_n} T_n(e^{\tau/\sigma_n}) = e^{\tau^2/2} \tag{45}$$

for complex τ. Again, by Cauchy's theorem

$$T_n(u) = \frac{1}{2\pi i} \oint \frac{T(z,u)}{z^{n+1}} dz = \frac{1}{2\pi i} \oint \frac{uP(H)}{D_H^2(z)(1-uM_H(z))z^{n+1-m}} dz ,$$

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

$$1 - e^t M_{\mathrm{H}}(e^\rho) = 0. \tag{46}$$

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

$$T_n(e^t) = C(t)e^{-(n+1-m)\rho(t)} + O(R^{-n}) \tag{47}$$

where

$$C(t) = \frac{P(H)}{D_H^2(\rho(t))M'_H(\rho(t))}.$$

To study some properties of $\rho(t)$, we observe that the cumulant formula implies $EO_n(H) = [t]\log T_n(e^t)$ and $\sigma_n^2 = [t^2]\log T_n(e^t)$ where, we recall, $[t^r]f(t)$ denotes the coefficient of $f(t)$ at t^r. In our case, $\mu_n \sim -n\rho'(0)$ as well as $\sigma_n^2 \sim -n\rho''(0)$. Set now in (47) $t = \tau/\sigma_n \to 0$ for some complex τ. Since uniformly in t we have $\rho(t) = t\rho'(0) + \rho''(0)t^2/2 + O(t^3)$ for $t \to 0$, our estimate (47) leads to

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

which completes the proof of Theorem 2.2(iii).

Actually, we can proceed as in Greene and Knuth [17] or Hwang [21] to obtain much more refined local limit result. For example, a direct application of results from [17] (cf. Chap. 4.3.3) leads to the following for $x = o(n^{1/6})$

$$\Pr\{O_n = EO_n + x\sqrt{nc_1}\} = \frac{1}{\sqrt{2\pi n c_1}} e^{-\frac{1}{2}x^2}\left(1 - \frac{\kappa_3}{2c_1^{3/2}\sqrt{n}}\left(x - \frac{x^3}{3}\right)\right) + O(n^{-3/2}) , \quad (48)$$

where κ_3 a constant (i.e., the third cumulant).

D. Large deviations: Case $r = (1+\delta)EO_n$

Finally, we consider the large deviations result. From (47) we conclude that

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

Thus, directly from the Gärtner-Ellis theorem [4, 9] we prove that

$$\lim_{n\to\infty} \frac{\log \Pr\{O_n > na\}}{n} = -I(a) ,$$

where

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

with ω_a being a solution of $-\rho'(t) = a$, A stronger version of the above follows directly from Theorem 3.1 of [4]. To derive our result of Theorem 2.2, we shall use (48) and the "shift of mean" technique as discussed below (cf. [4, 17, 21]).

As in the central limit regime, we could use Cauchy's formula to compute the probability $\Pr\{O_n = r\}$ for $r = EO_n + xO(\sqrt{n})$. But, formula (48) is only good for $x = O(1)$. To expand its validity, we shift the mean of the generating function $T_n(u)$ to a new value, say $m = an = (1+\delta)P(H)(n-m+1)$, so we can again apply the central limit formula (48) around the new mean. To accomplish this, let us re-write (47) as

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

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

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

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

$$\begin{aligned} E\widetilde{X} &= \frac{g'(\omega)}{g(\omega)} = -\rho'(\omega) \ , \\ \mathrm{Var}\ \widetilde{X} &= \frac{g''(\omega)}{g(\omega)} - \left(\frac{g'(\omega)}{g(\omega)}\right)^2 = -\rho''(\omega) \ . \end{aligned}$$

Let us now choose ω_a such that

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

Then, the new sum $\widetilde{S}_n = Y + \widetilde{X}_1 + \ldots + \widetilde{X}_{n-m+1}$ has a new mean $(1+\delta)P(H)(n-m+1) = a(n-m+1)$, and hence we can apply to $\widetilde{S}_n$ the central limit result (48). To translate from $\widetilde{S_n}$ to S_n we use the following simple formula

$$[e^{tN}]\left(C(t)g^n(t)\right) = \frac{g^n(\omega)}{e^{\omega N}}[e^{tN}]\left(\frac{C(t)g^N(t+\omega)}{g^N(\omega)}\right) \tag{49}$$

where $N = a(n-m+1)$ and $[e^{tn}]g(t)$ denotes the coefficient of $g(t)$ at e^{tn}. Now, we can apply (48) to the right-hand side of the above to obtain

$$[e^{tN}]\left(\frac{C(t)g^N(t+\omega)}{g^N(\omega)}\right) = \frac{1}{\sigma_a\sqrt{2\pi(n-m+1)}}(1+O(n^{-1})) + O(n^{-5/2}) \ .$$

Finally, using (49) and the above, we prove Theorem 2.2(iv).

ACKNOWLEDGEMENT

It is our pleasure to acknowledge several discussions with A. Dembo, A. Odlyzko, P. Pevzner, E. Coward and E. Sutinen on the topic of this paper. Our deepest appreciation goes to Philippe Flajolet for his inspirations and support over last decade.

References

[1] D. Barbara, and T. Imielinski, Sleepers and Workoholics - Caching in Mobile Wireless Environments, *Proc. ACM SIGMOD*, 1-15, Minneapolis 1994.

[2] E. Bender, Central and Local Limit Theorems Applied to Asymptotic Enumeration, *J. Combin. Theory, Ser. A*, 15, 91-111, 1973.

[3] S. Breen, M. Waterman and N. Zhang, Renewal Theory for Several Patterns, *J. Appl. Prob.*, 22, 228-234, 1985.

[4] J. Bucklew, and J. Sadowsky, A Contribution to the Theory of Chernoff Bounds, *IEEE Trans. Information Theory*, 39, 249-254, 1993.

[5] E. Coward, Exact Calculation of Word Occurrence Probabilities and Its Applications to the Search of Repeats, *Proc. Mathematical Analysis of Biological Sequences*, Rouen 1997.

[6] C. Chrysaphinou, and S. Papastavridis, The Occurrence of Sequence of Patterns in Repeated Dependent Experiments, *Theory of Probability and Applications*, 79, 167-173, 1990.

[7] M. Crochemore and W. Rytter, *Text Algorithms*, Oxford University Press, New York 1995.

[8] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York 1981.

[9] R. Ellis, Large Deviations for a General Class of Random Vectors, *Ann. Probab.*, 12, 1-12, 1984.

[10] J. Fickett, Recognition of Protein Coding Regions in DNA Sequences, *Nucleic Acids Res.*, 10, 5303-5318, 1982.

[11] W. Feller, *An Introduction to Probability and its Applications*, Vol. 1, John Wiley & Sons, New York 1968.

[12] P. Flajolet and M. Soria, General Combinatorial Schemas: Gaussian Limit Distributions and Exponential Tails, *Discrete Mathematics*, 114, 159-180, 1993.

[13] P. Flajolet, X. Gourdon, C. Martinez, Patterns in Random Binary Search Trees, *Random Structures and Algorithms*, 11, 223-244, 1997.

[14] I. Fudos, E. Pitoura and W. Szpankowski, On Pattern Occurrences in a Random Text, *Information Processing Letters*, 57, 307-312, 1996.

[15] M.S. Gelfand, Prediction of Function in DNA Sequence Analysis, *J. Comput. Biol.*, 2, 87-117, 1995.

[16] M. Geske, A. Godbole, A. Schafner, A. Skolnick, G. Wallstrom, Compound Poisson Approximations for World Patterns Under Markovian Hypotheses, *J. Appl. Prob.*, 32, 877-892, 1995.

[17] D. Greene and D. E. Knuth, *Mathematics for the Analysis of Algorithms*, Birkhäuser, Boston 1990.

[18] L. Guibas and A. Odlyzko, Maximal Prefix-Synchronized Codes, *SIAM J. Appl. Math*, 35, 401-418, 1978.

[19] L. Guibas and A. Odlyzko, Periods in Strings, *J. Combin. Theory Ser. A*, 30, 19-43, 1981.

[20] L. Guibas and A. W. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *J. Combin. Theory Ser. A*, 30, 183-208, 1981.

[21] H-K. Hwang, *Théorèmes Limites pour les Structures Combinatoires et les Fonctions Arithmétiques*, Thèse de Doctorat de l'Ecole Polytechnique, 1994.

[22] P. Jacquet and W. Szpankowski, Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, *J. Combin. Theory Ser. A*, 66, 237-269, 1994.

[23] P. Jokinen and E. Ukkonen, Two Algorithms for Approximate String Matching in Static Texts, *Proc. MFCS 91*, *Lecture Notes in Computer Science* 520, 240-248, Springer Verlag 1991.

[24] S. Karlin, C. Bruge and A. Campbell, Statistical Analysis of Counts and Distributions of Restriction Sites in DNA Sequences, *Nucl. Acids Res.*, 20, 1363-1370, 1992.

[25] S. Karlin and F. Ost, Counts of Long Aligned Word Matches Among Random Letter Sequences, *Ann. Prob.*, 19, 293-351, 1987.

[26] D.E. Knuth, *The Art of Computer Programming: Fundamental Algorithms*, vol. 1., Addison-Wesley, Reading 1973.

[27] S. R. Li, A Martingale Approach to the Study of Occurrences of Sequence Patterns in Repeated Experiments, *Ann. Probab.*, 8, 1171-1176, 1980.

[28] T. Luczak and W. Szpankowski, A Suboptimal Lossy Data Compression Based on Approximate Pattern Matching, *IEEE Trans. Information Theory*, 43, 1439–1451, 1997.

[29] K. Marton and P. Shields, The Positive-Divergence and Blowing-up Properties, *Israel J. Math*, 80, 331-348 (1994).

[30] P. T. Nielsen, On the Expected Duration of a Search for Fixed Pattern in Random Data, *IEEE Trans. Information Theory*, 702-704, 1973.

[31] A. Odlyzko, Asymptotic Enumeration, in *Handbook of Combinatorics*, Vol. II, pp. 1063–1229, (Eds. R. Graham, M. Götschel and L. Lovász), Elsevier Science, 1995.

[32] P. Pevzner, M. Borodovsky, and A. Mironov, Linguistic of Nucleotide Sequences: The Significance of Deviations from Mean: Statistical Characteristics and Prediction of the Frequency of Occurrence of Words, *J. Biomol. Struct. Dynam.*, 6, 1013-1026, 1991.

[33] B. Prum, F. Rodolphe, and E. Turckheim, Finding Words with Unexpected Frequencies in Deoxyribonucleic Acid Sequence, *J.R. Stat. Soc. B*, 57, 205-220, 1995.

[34] R. Remmert, *Theory of Complex Functions*, Springer Verlag, New York 1991.

[35] M. Régnier and W. Szpankowski, On the Approximate Pattern Occurrence in a Text, *Proc. SEQUENCE'97*, Positano 1997.

[36] S. Schbath, *Etude Asymptotique du Nombre d'Occurrences d'un Mot dans une Chaîne de Markov et Application à la Recherche de Mots de Fréquence Exceptionnelle dans les Séquences d'ADN*, Thèse Université René Descartes Paris V, 1995.

[37] E. Sutinen and W. Szpankowski, On the Collapse of q-Gram Filtration, preprint.

[38] W. Szpankowski, Asymptotic Properties of Data Compression and Suffix Trees, *IEEE Trans. Information Theory*, 39, 1647-1659, 1993.

[39] M. Waterman, *Introduction to Computational Biology*, Chapman & Hall, New York 1995.

[40] H. Wilf, *generatingfunctionology*, Academic Press, Boston 1990.

[41] Z. Zhang and E. Yang, An On-Line Universal Lossy Data Compression Algorithm via Continuous Codebook Refinement – Part II: Optimality for Phi-Mixing Source Models, *IEEE Trans. Information Theory*, 42, 822-836, 1996.