Discrete Mathematics and Theoretical Computer Science **6**, 2004, 191–214

Rare Events and Conditional Events on Random Strings

Mireille Régnier[1] and Alain Denise[2] †

[1] *INRIA,78153 Le Chesnay, France*
[2] *LRI-Université Paris-Sud, UMR CNRS 8623, 91405 Orsay, France*

received Feb 7, 2003, *revised Dec 18, 2003*, *accepted Feb 11, 2004*.

Some strings -the texts- are assumed to be randomly generated, according to a probability model that is either a Bernoulli model or a Markov model. A rare event is the over or under-representation of a word or a set of words. The aim of this paper is twofold. First, a single word is given. We study the tail distribution of the number of its occurrences. Sharp large deviation estimates are derived. Second, we assume that a given word is overrepresented. The conditional distribution of a second word is studied; formulae for the expectation and the variance are derived. In both cases, the formulae are precise and can be computed efficiently. These results have applications in computational biology, where a genome is viewed as a text.

Keywords: large deviations, combinatorics, generating fumctions, words, genome

1 Introduction

In this paper, we study the distribution of the number of occurrences of a word or a set of words in random texts. So far, the first moments, e.g. the mean and the variance, have been extensively studied by various authors under different probability models and different counting schemes [Wat95, Rég00, Szp01]. Moreover, it is well known that the random variable that counts the number of occurrences converges, in law, to the normal law [BK93, PRdT95, RS97a, NSF99, FGSV01] when the size n of the text grows to infinity. Nevertheless, very few results are known out of the convergence domain, also called the central domain. This paper aims at filling this gap, as rare events occur out of the convergence domain.

First, we study the *tail distribution.* We consider a single given word, H_1. In [RS97a, NSF99], a large deviation principle is established; in [RS97a] the rate function is implicitely defined, but left unsolved. In [RS98], the authors *approximate* the exact distribution by the so-called compound Poisson distribution, and compute the tail distribution of this approximate distribution. We provide a precise expansion of the exact probability out of the convergence domain. More precisely, we derive a computable formula for the rate function, and two more terms in the asymptotic expansion. This accuracy is made possible by

†This research was partially supported by IST Program of the EU under contract number 99-14186 (ALCOM-FT) and French Bioinformatics and IMPG Programs.

the combinatorial structure of the problem. Second, we rely on these results to address the *conditional counting* problem. The overrepresentation (or the under-representation) of a word H_1 modifies the distribution of the number of occurrences of the other words. In this paper, we study the expectation and the variance of the number of occurrences of a word H_2, when an other word, H_1, is exceptional, that is either overrepresented or under-represented. Our results on the tail distribution of H_1 allow to show that the conditional expectation and variance of H_2 are linear functions of the size n of the text. We derive explicit formulae for the linearity constants.

The complexity to compute the *tail distribution* or the *conditional counting* moments is low. As a matter of fact, it turns out that the problem reduces to the solution of a polynomial equation the degree of which is equal to the length of the overrepresented word. The approach is valid for various counting models.

These results have applications in computational biology, where a genome is viewed as a text. Available data on the genome(s) are increasing continuously. To extract relevant information from this huge amount of data, it is necessary to provide efficient tools for *"in silico"* prediction of potentially interesting regions. The statistical methods, now widely used [BJVU98, GKM00, BLS00, LBL01, EP02, MMML02] rely on a simple basic assumption: an exceptional word, *i.e.* a word which occurs significantly more (or less) in real sequences than in random ones, may denote a biological functionality. The conditional counting problem is adressed when one wants to detect a weak biological signal, the word H_2, hidden by a stronger signal, the word H_1 [BFW$^+$00, DRV01].

Section 2 is devoted to the introduction of some preliminary notions and results. The tail distribution of a single word is studied in section 3. Conditional events are addressed in Section 4.

2 Preliminary notions

2.1 Probability model

Our assumption is that the languages are generated on some alphabet $\mathcal{S}$ of size V by an ergodic and stationary source. The models we handle are either the *Bernoulli model* or the *Markov model*.

In the Markov model, a text T is a realization of a *stationary* Markov process of order K where the probability of the next symbol occurrence depends on the K previous symbols. Given two K-uples $(\alpha_1,\cdots,\alpha_K)$ and $(\beta_1,\cdots,\beta_K)$ from $\mathcal{S}^K$, the probability that a β-ocurrence ends at position l, when an α-occurrence ends at position $l-1$, does not depend on the position l in the text. E.g., we denote

$$p_{\alpha,\beta} = P((T_{l-K+1},\cdots,T_l)=\beta|(T_{l-K}\cdots T_{l-1})=\alpha)$$

These probabilities define a $V^K\times V^K$ matrix $\mathbb{P}=\{p_{\alpha,\beta}\}$ that is called the *transition matrix*. As the probability $p_{\alpha,\beta}$ is 0 if $(\alpha_2,\cdots,\alpha_K)\neq(\beta_1,\cdots,\beta_{K-1})$, the transition matrix $\mathbb{P}$ is sparse when $K>1$. Vector $\pi=(\pi_1,\ldots,\pi_{V^K})$ denotes the stationary distribution satisfying $\pi\mathbb{P}=\pi$, and Π is the stationary matrix that consists of V^K identical rows equal to π. Finally, $\mathbb{Z}$ is the **fundamental matrix** $\mathbb{Z}=(\mathbb{I}-(\mathbb{P}-\Pi))^{-1}$ where $\mathbb{I}$ is the identity matrix.

Definition 2.1 *Given a word z of length $|z|$ greater than or equal to K, we denote $P(w|z)$ the* conditional probability *that a w occurrence starts at a given position l in the text, knowing that a z occurrence starts at position $l-|z|+1$.*

Given a word w of size $|w|, |w| \geq K$, we denote $f(w)$ and $l(w)$ the w-prefix and the w-suffix of length K. For i in $\{1, \cdots, |w| - K + 1\}$, we denote $w[i]$ the i-th factor of length K. That is

$$w[i] = w_i \cdots w_{i+K-1}$$

We denote $P(w)$ the stationary probability *that the word w occurs in a random text. That is*

$$P(w) = \pi_{f(w)} \prod_{i=1}^{|w|-K} \mathbb{P}_{w[i],w[i+1]}$$

It will appear that all counting results depend on the Markovian process through submatrices of the matrix $\mathbb{F}(z)$ defined below.

Definition 2.2 *Given a Markovian model of order K, let $\mathbb{F}(z)$ be the $V^K \times V^K$ matrix*

$$\mathbb{F}(z) = (\mathbb{P} - \Pi)(\mathbb{I} - (\mathbb{P} - \Pi)z)^{-1} \quad . \tag{1}$$

It is worth noticing that $\mathbb{F}(z)$ can be reexpressed as a power series in $\mathbb{Z}$.

In the Bernoulli model, one assumes that the text is randomly generated by a memoryless source. Each letter s of the alphabet has a given probability p_s to be generated at any step. Generally, the p_s are not equal and the model is said to be *biased*. When all p_s are equal, the model is said to be *uniform*. The Bernoulli model can be viewed as a Markovian model of order $K = 0$.

2.2 The correlation polynomials and matrices

Finding a word in a random text is, in a certain sense, correlated to the previous occurrences of the same word or other words. For example, the probability to find $\mathrm{H}_1 = \mathtt{ATT}$, knowing that one has just found $\mathrm{H}_2 = \mathtt{TAT}$, is - intuitively - rather good since a $\mathtt{T}$ just after H_2 is enough to give H_1. The correlation polynomials and the correlation matrices give a way to formalize this intuitive observation. At first, let us define the overlapping set and the correlation set [GO81] of two words.

Definition 2.3 *The* overlapping set *of two words* $\mathrm{H_i}$ *and* $\mathrm{H_j}$ *is the set of suffixes of* $\mathrm{H_i}$ *which are prefixes of* $\mathrm{H_j}$. *The* correlation set *is the set of* $\mathrm{H_i}$*-suffixes in the associated* $\mathrm{H_j}$*-factorizations. It is denoted by $\mathcal{A}_{i,j}$. When* $\mathrm{H_i} = \mathrm{H_j}$, *the* correlation set *is called the* autocorrelation set *of* $\mathrm{H_i}$.

For example, the overlapping set of $\mathrm{H}_1 = \mathtt{ATT}$ and $\mathrm{H}_2 = \mathtt{TAT}$ is $\{\mathtt{T}\}$. The associated factorization of H_2 is $T \cdot AT$. The correlation set is $\mathcal{A}_{1,2} = \{\mathtt{AT}\}$. The overlapping set of H_2 with itself is $\{TAT, T\}$. The associated factorizations are $TAT \cdot \varepsilon$ and $T \cdot AT$, where ε is the empty string. The autocorrelation set of H_2 is $\{\varepsilon, \mathtt{AT}\}$. As any string belongs to its overlapping set, the empty string belongs to any autocorrelation set.

Definition 2.4 *In the Markov model, the* correlation polynomial *of two words* $\mathrm{H_i}$ *and* $\mathrm{H_j}$ *is defined as follows:*

$$A_{i,j}(z) = \sum_{w \in \mathcal{A}_{i,j}} P(w|l(\mathrm{H_i}))\mathrm{z}^{|\mathrm{w}|} \quad .$$

In the Bernoulli model, the correlation polynomial *is*

$$A_{i,j}(z) = \sum_{w \in \mathcal{A}_{i,j}} P(w) z^{|w|} \quad .$$

When $\mathrm{H_i} = \mathrm{H_j}$, *this polynomial is called the* autocorrelation polynomial *of* $\mathrm{H_i}$.

Given two words $\mathrm{H_1}$ *and* $\mathrm{H_2}$, *the matrix*

$$\mathbb{A}(z) = \begin{bmatrix} A_{1,1}(z) & A_{1,2}(z) \\ A_{2,1}(z) & A_{2,2}(z) \end{bmatrix}$$

is called the correlation matrix.

Definition 2.5 *Given two ordered sets* $\mathcal{H}_1 = \{\mathrm{H_1^1}, \cdots, \mathrm{H_1^q}\}$ *and* $\mathcal{H}_2 = \{\mathrm{H_2^1}, \cdots, \mathrm{H_2^r}\}$, *let* $\mathbb{G}_{\mathcal{H}_1,\mathcal{H}_2}(z)$ *be the* $q \times r$ *matrix*

$$(\mathbb{G}_{\mathcal{H}_1,\mathcal{H}_2}(z))_{i,j} = \mathbb{F}(z)_{l(\mathrm{H_1^i}),\mathrm{f}(\mathrm{H_2^j})} \cdot \frac{1}{\pi_{f(\mathrm{H_2^j})}} \quad .$$

2.3 Word counting

There are several ways to count word occurrences, that depend on the possible application. Let $\mathrm{H_1}$ and $\mathrm{H_2}$ be two words on the same alphabet. In the *overlapping counting model* [Wat95], any occurrence of each word is taken into account. Assume, for example, that $\mathrm{H_1} = \mathtt{ATT}, \mathrm{H_2} = \mathtt{TAT}$ and that the text is TTATTATATATT. This text contains 2 occurrences of $\mathrm{H_1}$ and 4 overlapping occurrences of $\mathrm{H_2}$ at positions $2, 5, 7$ and 9. In other models, such as the *renewal model* [TA97], some overlapping occurrences are not counted. Although our approach is valid for different counting models, we restrict here to the most commonly used, e.g. the *overlapping model* [Wat95].

When several words are searched simultaneously, we need some additional conditions on this set of words, $\mathcal{H}$. It is generally assumed that the set $\mathcal{H}$ is *reduced*.

Definition 2.6 *[BK93] A set of words is* reduced *if no word in this set is a proper factor of another word.*

The two words $\mathrm{H_1}$ and $\mathrm{H_2}$ do not play the same role in the conditional counting problem. We can partially relax the reduction condition.

Definition 2.7 *A couple of words* $(\mathrm{H_1}, \mathrm{H_2})$ *is reduced iff the set* $\{\mathrm{H_1}, \mathrm{H_2}\}$ *is reduced or* $\mathrm{H_1}$ *is a proper prefix of* $\mathrm{H_2}$.

Remark that, in the case where the set of words is given by a regular expression, this regular expression must be unambiguous. A discussion on ambiguity in counting problems and algorithmic issues can be found in [KM97].

2.4 Multivariate Probability Generating Functions

Our basic tools are the *multivariate probability generating functions*. Let $\mathcal{L}$ be some language that is randomly generated according to one of the models described above. For any integer n, let $\mathcal{L}_n$ be the set of words of size n that belong to $\mathcal{L}$. Given two words $\mathrm{H_1}$ and $\mathrm{H_2}$, we denote $X_{i,n}$ with $i \in \{1,2\}$, the random variable which counts the occurrences of H_i in a text from this set $\mathcal{L}_n$; we denote $P(X_{i,n} = k)$ the probability that $\mathrm{H_i}$ occurs k times. The *probability generating function* of the random variable $X_{i,n}$ is denoted $P_{i,n}$. We have

$$P_{i,n}(u) = \sum_{k \geq 0} P(X_{i,n} = k) u^k \quad .$$

Definition 2.8 *Given a language $\mathcal{L}$, the multivariate generating function that counts* H_1 *and* H_2 *occurrences in the texts that belong to this language $\mathcal{L}$ is*

$$L(z,u_1,u_2) = \sum_{n\geq 0} z^n \sum_{k_1+k_2\geq 0} P(X_{1,n}=k_1 \text{ and } X_{2,n}=k_2)u_1^{k_1}u_2^{k_2} \quad .$$

The multivariate generating function that counts H_1*-occurrences (only) is*

$$L_1(z,u_1) = \sum_{n\geq 0} z^n \sum_{k_1\geq 0} P(X_{1,n}=k_1)u_1^{k_1} = \sum_{n\geq 0} z^n P_{1,n}(u_1) \quad . \tag{2}$$

Remark: These multivariate generating functions satisfy the equation

$$L_1(z,u_1) = L(z,u_1,1) \quad .$$

Moreover, $L_1(z,1) = L_1(z,1,1)$ is the *ordinary generating function* of the language $\mathcal{L}$.

One important language is the set of all possible words on the alphabet $\mathcal{S}$, denoted below by $\mathcal{T}$. Language $\mathcal{T}$ is also named the language of texts. A general expression for its multivariate generating function $T(z,u_1,u_2)$ is derived in [Rég00]. For a single word H_1 of sixe m_1, it depends on H_1 through the entire series of the variable z defined as follows:

$$D_1(z) = (1-z)A_1(z) + P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1} + \mathbb{F}(\mathrm{z})_{\mathrm{l}(\mathrm{H}_1),\mathrm{f}(\mathrm{H}_1)} \cdot \frac{1}{\pi_{\mathrm{f}(\mathrm{H}_1)}} \quad . \tag{3}$$

In the Bernoulli model, this series $D_1(z)$ is a polynomial.

Proposition 2.1 *[RS97a] The multivariate generating function that counts the occurrences of a single word* H_1 *of sixe m_1, in a Bernoulli or a Markov model, satisfies the equation*

$$T_1(z,u_1) = T(z,u_1,1) = \frac{u_1}{1-u_1M_1(z)} \frac{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}}{\mathrm{D}_1(\mathrm{z})^2} \tag{4}$$

where

$$M_1(z) = \frac{D_1(z)+z-1}{D_1(z)} \quad . \tag{5}$$

As a consequence, our counting results only depend on this series $D_1(z)$. Similarly, for two words counts, all the results depend on H_1 and H_2 through the matrix $\mathbb{D}(z)$ defined below.

Definition 2.9 *Given a reduced couple of words* H_1 *and* H_2 *of size m_1 and m_2, let $\mathbb{D}(z)$ be the 2×2 matrix*

$$\mathbb{D}(z) = (1-z)\mathbb{A}(z) + \begin{bmatrix} P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1} & \mathrm{P}(\mathrm{H}_2)\mathrm{z}^{\mathrm{m}_2} \\ P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1} & \mathrm{P}(\mathrm{H}_2)\mathrm{z}^{\mathrm{m}_2} \end{bmatrix} + \mathbb{G}_{\{\mathrm{H}_1\},\{\mathrm{H}_2\}}(z) \quad . \tag{6}$$

We denote, for i,j in $\{1,2\}$,

$$D_{i,j}(z) = \mathbb{D}(z)_{i,j}$$

3 Significance of an Exceptional Word

In this section, we study the *tail distribution* of the number of occurrences of a single word H_1 in a random text $\mathcal{T}$. In [RS97a], a large deviation principle is established by the Gartner-Ellis Theorem. We derive below an explicit formula for the rate function and a precise expansion of the probabilities in the large deviation domain. These results should be compared to [Hwa98] although the validity domains in [Hwa98] are closer to the central region.

3.1 Sharp large deviations estimates

Definition 3.1 *The fundamental equation is the equation* (E_a)

$$D_1(z)^2-(1+(a-1)z)D_1(z)-az(1-z)D_1'(z)=0\ , \tag{7}$$

where a *is a real positive number satisfying* $0\le a\le 1$.

Lemma 3.1 *Assume that* $a>P(\mathrm{H}_1)$. *When* H_1 *is selfoverlapping or when* $\frac{1}{m_1}>a$, *there exists a largest real positive solution of the fundamental equation that satisfies* $0<z_a<1$. *It is called the fundamental root of* (E_a) *and denoted* z_a.

Proof: Define the function of the real variable z: $r_a(z)=D_1(z)^2-(1+(a-1)z)D_1(z)-az(1-z)D_1'(z)$. It satisfies $r_a(0)=0$ and $r_a(1)=P(\mathrm{H}_1)(\mathrm{P}(\mathrm{H}_1)-\mathrm{a})$ that is negative if $a>P(\mathrm{H}_1)$. Moreover, $r_a'(0)=(1-a)(D_1'(0)+1)$. This derivative is strictly positive if $A_1(z)\neq 1$. If $A_1(z)=1$, that is if H_1 is not selfoverlapping, then $r_a(z)=pz^{m_1}[1-am-z(1+a-am)+pz^{m_1}]$ and $r_a^{(m)}(0)>0$ if $a<\frac{1}{m_1}$. Hence, $r_a(z)$ has a zero in $]0,1[$. □

We are now ready to state the main result of this section.

Theorem 3.1 *Let* H_1 *be a given word and* a *be some real number such that* $a\neq P(\mathrm{H}_1)$. *In a Bernoulli and a Markov model, the random variable* $X_{1,n}$ *satisfies*

$$P(X_{1,n}=na)=\frac{1}{\sigma_a\sqrt{2\pi n}}e^{-nI(a)+\delta_a}(1+O(\frac{1}{n}))\ , \tag{8}$$

where

$$\begin{aligned} I(a) &= a\ln\left(\frac{D_1(z_a)}{D_1(z_a)+z_a-1}\right)+\ln z_a\ , & (9)\\ \sigma_a^2 &= a(a-1)-a^2z_a\left(\frac{2D_1'(z_a)}{D_1(z_a)}-\frac{(1-z_a)D_1''(z_a)}{D_1(z_a)+(1-z_a)D_1'(z_a)}\right)\ , & (10)\\ \delta_a &= \ln[\frac{P(\mathrm{H})\mathrm{z_a^{m_1}}}{\mathrm{D_1(z_a)+(1-z_a)D_1'(z_a)}}] & (11)\end{aligned}$$

and z_a *is the fundamental root of* (E_a).

Remark: $\frac{D_1(z)}{1-z}$ is the generating function of a language [RS97a]. It satisfies $D_1(0)=1$. Hence, it has positive coefficients and cannot be 0 at a real value. It follows that $D_1(z_a)\neq 0$ and that $D_1(z_a)+z_a-1\neq 0$.

Remark: It follows from (8) that $-\frac{\ln P(X_{1,n}\geq na)}{n}$ has a finite limit, $I(a)$, when n tends to ∞. This limit is the *rate function* of the large deviation theory [DZ92]. Equation (8) provides two additional terms in the asymptotic expansion and a correction to the result claimed in [RS97a].

Remark: When $a = P(\mathrm{H}_1)$, Equation (8) still provides the probability in the central domain. As a matter of fact, the fundamental root z_a is equal to 1. The rate function is $I(a) = 0$, as expected in the central domain, and $\delta_a = 0$. One can check that

$$\sigma_a^2 = P(\mathrm{H}_1)\left(2\mathrm{A}_1(1) - 1 + (1-2\mathrm{m})\mathrm{P}(\mathrm{H}_1) + 2\mathrm{P}(\mathrm{H}_1)\mathbb{F}(1)_{\mathrm{l}(\mathrm{H}_1),\mathrm{f}(\mathrm{H}_1)} \cdot \frac{1}{\pi_{\mathrm{f}(\mathrm{H}_1)}}\right) .$$

This is the variance previously computed in the Bernoulli case by various authors [Wat95] and in the Markov case in [RS97a].

The next proposition provides a local expansion of the rate function.

Proposition 3.1 *The rate function I satisfies, for any $\tilde{a}$ in a neighbourhood of a,*

$$I(\tilde{a}) = I(a) + I'(a)(\tilde{a}-a) + \frac{1}{2}I''(a)(\tilde{a}-a)^2 + O((\tilde{a}-a)^3) \tag{12}$$

where

$$I'(a) = \ln\left(\frac{D_1(z_a)+z_a-1}{D_1(z_a)}\right) , \tag{13}$$

$$I''(a) = -\frac{1}{\sigma_a^2} . \tag{14}$$

3.2 Technical results

Our proof of Theorem 3.1 is purely analytic. It follows from the definition of $T_1(z,u)$ in (2) that

$$P(X_{1,n} = na) = [z^n][u^{na}]T_1(z,u) .$$

Using the expression (4) this is

$$P(X_{1,n} = na) = [z^n]\frac{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}}{D_1(z)^2}M_1(z)^{na-1} .$$

Let us denote $\frac{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}}{D_1(z)^2}M_1(z)^{na-1}$ by $f_a(z)$. When na is an integer, this function is an analytic function. Let us show that the radius of convergence is strictly greater than 1. The generating function $M_1(z)$ is the probability generating function of a language; hence, all its coefficients are positive and the radius of convergence is at least $R = 1$. It follows from the equation $M_1(z) = 1 + \frac{z-1}{D_1(z)}$ that $M_1(1) = 1$: hence, the radius of convergence of M_1 is strictly greater than 1. Now, this equation implies that the singularities of M_1 are the singularities of $D_1(z)$ and the roots of $D_1(z) = 0$. Hence, these singularities and these roots are necessarily greater than 1. Finally, all singularities of $f_a(z)$ are greater than 1.

Let us observe that there exists a direct proof in the Bernoulli model. The series $\frac{D_1(z)}{1-z} = A_1(z) + \frac{1}{1-z} \cdot P(H)z^{m_1}$ has only positive coefficients; hence, the root with smallest modulus is real positive. As $A_1(z)$ and $P(H)z^{m_1}$ have positive coefficients, a real positive root of $D_1(z)$ is greater than 1.

Cauchy formula for univariate series can be written as

$$P(X_{1,n}=na)=\frac{1}{2i\pi}\oint\frac{1}{z^{n+1}}\frac{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}}{D_1(z)^2}M_1(z)^{na-1}dz\ ,$$

where the integration is done along any contour around 0 included in the convergence circle. We define the function $h_a(z)$ of the complex variable z by the equation

$$h_a(z)=a\ln M_1(z)-\ln z\ .$$

The integral above can be expressed in the form $J_g(a)=\frac{1}{2i\pi}\oint e^{nh_a(z)}g(z)dz$ where $g(z)$ is an analytic function. Here, $g(z)$ is set to be $\frac{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1-1}}{D_1(z)^2}\frac{1}{M_1(z)}=\frac{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}}{zD_1(z)(D_1(z)+z-1)}$. We need to establish an asymptotic expansion of this integral.

Theorem 3.2 *Given an analytic function g, let $J_g(a)$ be the integral*

$$J_g(a)=\frac{1}{2i\pi}\oint e^{nh_a(z)}g(z)dz\ . \tag{15}$$

If g is such that $g(0)\neq 0$, then the integral $J_g(a)$ satisfies

$$J_g(a)=\frac{e^{-nh_a(z_a)}g(z_a)}{2\tau_a\sqrt{\pi n}}\left[1+\frac{1}{n}\left(-\frac{g''(z_a)}{g(z_a)}\frac{1}{2\tau_a^2}+\beta_a\frac{g'(z_a)}{g(z_a)}\frac{3}{\tau_a}+3\gamma_a\right)+O(\frac{1}{n^2})\right]\ , \tag{16}$$

where

$$\begin{aligned}\tau_a &= \frac{\sigma_a}{az_a}\ ,\\ \beta_a &= \frac{h_a^{(3)}(z_a)}{3!\tau_a^3}\ ,\\ \gamma_a &= \frac{h_a^{(4)}(z_a)}{4!\tau_a^4}\end{aligned}$$

and z_a is the fundamental root of (7). If there exists an integer l such that $G(z)=z^{-l}g(z)$ is analytic at $z=0$, with $G(0)\neq 0$, then

$$J_g(a)=J_G(a)z_a^l\cdot\left[1-\frac{1}{2z_a^2\tau_a^2}\cdot\frac{l^2}{n}+\left(\frac{1}{2z_a^2\tau_a^2}+\frac{3\beta_a}{\tau_a z_a}-\frac{1}{\tau_a^2 z_a}\frac{G'(z_a)}{G(z_a)}\right)\frac{l}{n}+O(\frac{1}{n^2})\right]\ . \tag{17}$$

Before dealing with the proof of Theorem 3.2, we observe that $h_a(z_a)$ is the function $I(a)$ defined in (9) and that the dominating term is $\frac{G(z_a)z_a^l}{\tau_a}=g(z_a)\cdot\frac{az_a}{\sigma_a}=\frac{e^{\delta_a}}{\sigma_a}$. This is Equation (8). The following terms in the expansion will be necessary to deal with conditional events in Section 4

Proof of Theorem 3.2: We prove (16) by the saddle point method [Hen77]. We need to establish a technical lemma.

Lemma 3.2 *Let a be a real number. The function $h_a(z) = a \ln M_1(z) - \ln z$ and all its derivatives are rational functions of D_1 and its derivatives. They satisfy the following equalities:*

$$\begin{aligned} h_a(z_a) &= -I(a) \ , \\ h'_a(z_a) &= 0 \ , \\ h''_a(z_a) &= \tau_a^2 \ . \end{aligned}$$

Moreover, there exists a neighbourhood of z_a, included in the convergence domain, and a positive integer η such that

$$\mathcal{R}(h_a(z) - h_a(z_a)) \geq \eta \ . \tag{18}$$

Proof: A differentiation of Equation (5) shows that the derivatives of $h_a(z)$ are rational functions of D_1 and its derivatives. The values at point z_a follow from the Fundamental Equation (E_a). As $h''(z_a) > 0$, the second derivative h'' is strictly positive in some neighbourhood of z_a; this establishes the lower bound on the real part. □

Let us chose a suitable contour of integration for (15). A Taylor expansion of $h_a(z)$ and $g(z)$ around $z = z_a$ yields:

$$\begin{aligned} h_a(z_a + y) &= h_a(z_a) + \frac{y^2}{2} h\text{''}_a(z_a) + \frac{y^3}{3!} h_a^{(3)}(z_a) + \frac{y^4}{4!} h_a^{(4)}(z_a) + O(y^5) \ , \\ g(z_a + y) &= g(z_a) + y g'(z_a) + y^2 \frac{g''(z_a)}{2} + O(y^3) \ . \end{aligned}$$

With the change of variable $y = \frac{x}{\tau_a \sqrt{n}}$, the integrand rewrites, when ny^3 is small,

$$e^{-nI(a)} g(z_a) \left[1 + \frac{g'(z_a)}{g(z_a)} \cdot \frac{x}{\tau_a \sqrt{n}} + \frac{g''(z_a)}{g(z_a)} \frac{x^2}{2\tau_a^2 n} + \beta_a \frac{x^3}{\sqrt{n}} + \frac{\beta_a}{\tau_a} \frac{g'(z_a)}{g(z_a)} \frac{x^4}{n} + \gamma \frac{x^4}{n} + O(\frac{1}{n^{3/2}}) \right] \ .$$

We choose as a first part of the contour a vertical segment $[z_1, z_2] = [z_a - \frac{i}{n^\alpha}, z_a + \frac{i}{n^\alpha}]$. In order to keep ny^3 small when ny^2 tends to ∞, we choose $\frac{1}{3} < \alpha < \frac{1}{2}$. In that case, each term x^k provides a contribution $\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} x^k dx = F_k \sqrt{2\pi}$. These integrals satisfy $F_{2p} = \frac{\Gamma(2p)}{2^{p-1}\Gamma(p)}$ and $F_{2p+1} = 0$. Hence, the odd terms do not contribute to the integral. This yields an asymptotic expansion of $P(X_{1,n} = na)$ in $\frac{1}{n^{p+1/2}}$.

We now close our contour in order to get an exponentially negligible contribution. The bound (18) implies that the contributions of the segments $[0, z_1]$ and $[0, z_2]$ are exponentially smaller than $e^{-nI(a)}$.

We need now establish (17). In order to use (16), we rewrite

$$[z^n] e^{nh_a(z)} g(z) = [z^{n-l}] e^{nh_a(z)} G(z) = [z^{n-l}] e^{(n-l)h_{\tilde{a}}(z)} G(z)$$

where $\tilde{a}$ is defined by the equation

$$na = (n-l)\tilde{a} \ .$$

It follows that $\tilde{a} = a + \frac{al}{n} + a\frac{l^2}{n^2} + O(\frac{1}{n^2})$. We substitute $(\tilde{a}, n-l)$ to (a, n) in Equation (16) and compute a Taylor expansion of all parameters : the fundamental root $z_{\tilde{a}}$, the rate function $I(\tilde{a})$, the variance term τ_a and the constant term $g(z_a)$.

Fundamental root The function $\psi(a,z) = h_a(z)$ satisfies the functional equation

$$\frac{\partial\psi}{\partial z}(a,z_a) = \phi(a,z_a) = h'_a(z_a) = 0$$

where $\phi(a,z) = \frac{\partial\psi}{\partial z}(a,z)$. It implicitely defines z_a as a function $z(a)$ of the variable a. Two differentaiations with respect to a yield the derivatives of $z(a)$ at point a. More precisely,

$$\frac{\partial\phi}{\partial a} + \frac{\partial\phi}{\partial z} z'(a) = 0$$

From $\frac{\partial\phi}{\partial a}(a,z_a) = \frac{M'_1(z_a)}{M_1(z_a)} = \frac{1}{az_a}$ and $\frac{\partial\phi}{\partial z}(a,z_a) = h''(z_a) = \tau_a^2 = \frac{\sigma_a^2}{a^2 z_a^2}$, we get

$$z'(a) = -\frac{az_a}{\sigma_a^2} = -\frac{1}{\tau_a^2 a z_a} \quad .$$

Hence, $z(\tilde{a}) = z_a - \frac{1}{\tau_a^2 z_a} \cdot \frac{l}{n} + O(\frac{1}{n^2})$.

Rate function We need here a local expansion of the rate function $I(a)$ around the point a that is interesting in its own.

$$\begin{aligned}
\psi(\tilde{a}, z(\tilde{a})) &= \psi(a,z_a) + (\tilde{a}-a)\left(\frac{\partial\psi}{\partial a} + \frac{\partial\psi}{\partial z}\cdot z'(a)\right) \\
&+ \frac{(\tilde{a}-a)^2}{2}\left(\frac{\partial^2\psi}{\partial a^2} + 2\frac{\partial^2\psi}{\partial a\partial z}\cdot z'(a) + \frac{\partial\psi}{\partial z}\cdot z''(a) + \frac{\partial^2\psi}{\partial z^2}\cdot z'(a)^2\right) + O((\tilde{a}-a)^3)
\end{aligned}$$

We have the following equalities:

$$\begin{aligned}
\frac{\partial\psi}{\partial z}(a,z_a) &= h'_a(z_a) = 0 \ , \\
\frac{\partial\psi}{\partial a}(a,z) &= \ln M_1(z) \qquad \Rightarrow \frac{\partial^2\psi}{\partial a^2}(a,z) \ = \ 0 \ , \\
\frac{\partial^2\psi}{\partial z^2}(a,z_a) &= h''_a(z_a) = \tau_a^2 \ , \\
\frac{\partial^2\psi}{\partial a\partial z}(a,z_a) &= \frac{\partial\phi}{\partial a}(a,z_a) = \frac{1}{az_a} \ .
\end{aligned}$$

The coefficient of $(\tilde{a}-a)$ reduces to $\frac{\partial\psi}{\partial a} = \ln M_1(z)$. The coefficient of $(\tilde{a}-a)^2$ rewrites

$$\frac{z'(a)}{2}\left(\frac{2}{az_a} + \tau_a^2 z'(a)\right) = \frac{z'(a)}{2}\left(\frac{2}{az_a} - \frac{1}{az_a}\right) = \frac{z'(a)}{2az_a} = -\frac{1}{2\tau_a^2 a^2 z_a^2} = -\frac{1}{2\sigma_a^2}$$

and (12) follows.

From the equation $\tilde{a} - a = a\frac{l}{n} + a\frac{l^2}{n^2}$, it follows that $(n-l)(\tilde{a}-a) = al + O(\frac{1}{n^2})$ and $(n-l)(\tilde{a}-a)^2 = \frac{a^2 l^2}{n} + O(\frac{1}{n^2})$, and we get the rate function

$$(n-l)I(\tilde{a}) = -nI(a) + l(I(a) + a\ln M_1(z_a)) - \frac{1}{2\tau_a^2 z_a^2}\frac{l^2}{n} + O(\frac{1}{n^2}) \quad .$$

As $I(a) + a \ln M_1(z_a) = \ln z_a$ and $G(z_a) z_a^l = g(z_a)$, this term provides the correcting term

$$e^{-\frac{1}{2\tau_a^2 z_a^2} \frac{l^2}{n} + O(\frac{1}{n^2})} = 1 - \frac{1}{2\tau_a^2 z_a^2} \frac{l^2}{n} + O(\frac{1}{n^2}) \quad .$$

Variance We now compute the contribution of $\sqrt{\tau_{\tilde{a}}^2 (n-l)}$. We have:

$$(n-l)\tau_{\tilde{a}}^2 = n\tau_a^2 \left(1 - \frac{l}{n} + 2\frac{\tau_a'}{\tau_a}(\tilde{a} - a) + O(\frac{1}{n^2})\right)$$

The equality $\tau_a^2 = h_a''(z) = \frac{\partial^2 \psi}{\partial z^2}(a, z_a)$ above implies that

$$2\tau_a \tau_a' \;=\; \frac{\partial}{\partial a}\frac{\partial^2 \psi}{\partial z^2}(a, z_a) \;=\; h^{(3)}(z_a) z'(a) + \frac{\partial^2 \phi}{\partial z \partial a} \;=\; h^{(3)}(z_a) z'(a) + \frac{\partial}{\partial z}\left(\frac{M_1'(z)}{M_1(z)}\right) \quad .$$

Hence,

$$2\frac{\tau_a'}{\tau_a^2} = \frac{1}{\tau_a^2}\left(-\frac{3!\tau_a^3 \beta_a}{\tau_a^2 a z_a} + \frac{1}{a}(h_a''(z) - \frac{1}{z^2})\right) = -\frac{3!\beta_a}{\tau_a a z_a} + \frac{1}{a}(1 - \frac{1}{\tau_a^2 z_a^2}) \quad .$$

Finally, $(n-l)\tau_{\tilde{a}}^2 = n\tau_a^2 (1 - \frac{l}{n}(\frac{1}{\tau_a^2 z_a^2} + \frac{3!\beta_a}{\tau_a z_a})$ and the contribution is

$$\frac{1}{\tau_{\tilde{a}}\sqrt{n-l}} = \frac{1}{\tau_a \sqrt{n}}\left(1 + \frac{l}{n}(\frac{1}{2\tau_a^2 z_a^2} + \frac{3\beta_a}{\tau_a z_a})\right) \quad .$$

Constant term We now compute the contribution of $G(z_{\tilde{a}})$. We have

$$\begin{aligned} G(z_{\tilde{a}}) &= G(z_a)\left(1 + \frac{G'(z_a)}{G(z_a)} z'(a)(\tilde{a} - a) + O(\frac{1}{n^2})\right) \\ &= G(z_a)\left(1 - \frac{l}{n}\frac{G'(z_a)}{G(z_a)}\frac{1}{z_a \tau_a^2} + O(\frac{1}{n^2})\right) \quad . \end{aligned}$$

This is Equation (17).

□

4 Conditional Events

We consider here the *conditional counting* problem. The conditional expectation and variance can be expressed as functions of the coefficients of the multivariate generating function of the language of texts $\mathcal{T}$. More precisely, it follows from the equation $P(X_{2,n} = k_2 | X_{1,n} = k_1) = \frac{P(X_{1,n} = k_1 \text{ and} X_{2,n} = k_2)}{P(X_{1,n} = k_1)}$, that

$$E(X_{2,n} | X_{1,n} = k_1) = \frac{\sum_{k_2 \geq 0} k_2 P(X_{1,n} = k_1 \text{ and } X_{2,n} = k_2)}{P(X_{1,n} = k_1)} \quad .$$

Definition (2) implies that

$$P(X_{1,n} = k_1) = [z^n u_1^{k_1}] T_1(z, u_1) = [z^n u_1^{k_1}] T(z, u_1, 1) \quad .$$

Moreover:

$$\sum_{k_2} k_2 P(X_{1,n}=k_1 \text{ and } X_{2,n}=k_2)) = \sum_{k_2} k_2 [z^n u_1^{k_1} u_2^{k_2}] T(z,u_1,u_2)$$
$$= [z^n u_1^{k_1}] \sum_{k_2} k_2 [u_2^{k_2}] T(z,u_1,u_2) = [z^n u_1^{k_1}] \frac{\partial T}{\partial u_2}(z,u_1,1) \ .$$

It follows that

$$E(X_{2,n}|X_{1,n}=k_1) = \frac{[z^n u_1^{k_1}] \frac{\partial T}{\partial u_2}(z,u_1,1)}{[z^n u_1^{k_1}] T(z,u_1,1)} \ . \tag{19}$$

Similarly, we can prove

$$Var(X_{2,n} \mid X_{1,n} = na) = \frac{[z^n u_1^{k_1}] \left(\frac{\partial^2 T(z,u_1,u_2)}{\partial u_2^2} + \frac{\partial T(z,u_1,u_2)}{\partial u_2} \right)}{[z^n u_1^{k_1}] T(z,u_1)} - E((X_{2,n} \mid X_{1,n} = na)^2 \ . \tag{20}$$

Given two words, the software *RegExpCount* allows to compute and derive $T(z,u_1,u_2)$. The shift-of-the-mean method allows to compute the linearity constant for the mean and the expectation in [Nic00]. This step is costly; notably, it must be repeated when n varies.

Our closed formulae provide an efficient alternative. The general expression for $T(z,u_1,u_2)$ given in [Rég00] is a matricial expression that is not suitable for the computation of the partial derivatives that occur in (19) and (20). In 4.1 below, we provide a new expression that is suitable for a partial derivative. At point $u_2 = 1$, the partial derivatives rewrite as $\frac{u_1^{m_1} \psi(z)}{(1-u_1 M_1(z))^k}$ where ψ is analytic in z in a larger domain than $\frac{1}{1-u_1 M_1(z)}$. Hence, the methodolny of Section 3 applies.

4.1 Multivariate Generating Functions for Word Counting

Our enumeration method follows the scheme developed in [Rég00]. More details on this formalism can be found in [Rég00, Szp01]. In this paper, a set of *basic languages*, the *initial, minimal* and *tail* languages, is defined and any counting problem is rewritten as a problem of text decomposition over these basic languages. This is in the same vein as the general decomposition of combinatorial stuctures over basic data structures presented in [FS96]. Such basic languages satisfy equations that depend on the counting model. These equations translate into equations for corresponding generating functions, and multivariate generating functions for the counting problem are rewritten over this set of basic generating functions.

We briefly present this formalism when two words H_1 and H_2 are counted. The initial languages $\tilde{\mathcal{R}}_i$ (for $i = 1$ or 2) are defined as the languages of words ending with H_i and containing no other occurrence of H_1 or H_2. The minimal language $\mathcal{M}_{i,j}$ (for $i \in \{1,2\}$ and $j \in \{1,2\}$) contains the words w which end with H_j and such that $H_i w$ contains exactly two occurrences of $\{H_1, H_2\}$: the one at the beginning and the one at the end. The tail language $\tilde{\mathcal{U}}_i$ is the language of words w such that $H_i w$ contains exactly one occurrence of H_i and no other $\{H_1, H_2\}$-occurrence. For example, let us assume that $H_1 =$ `ATT` and $H_2 =$ `TAT`. The text `TTATTATATATT` can be decomposed as follows:

$$\underbrace{\mathtt{TTATT}}_{\in \mathcal{R}_1}\underbrace{\mathtt{ATATATT}}_{\in \mathcal{M}_1}\underbrace{\mathtt{TT}}_{\in \mathcal{U}_1} \text{ and } \underbrace{\mathtt{TTAT}}_{\in \tilde{\mathcal{R}}_2}\underbrace{\mathtt{T}}_{\in \mathcal{M}_{2,1}}\underbrace{\mathtt{AT}}_{\in \mathcal{M}_{1,2}}\underbrace{\mathtt{AT}}_{\in \mathcal{M}_{2,2}}\underbrace{\mathtt{AT}}_{\in \mathcal{M}_{2,2}}\underbrace{\mathtt{T}}_{\in \mathcal{M}_{2,1}}\underbrace{\mathtt{TT}}_{\in \tilde{\mathcal{U}}_1}$$

Among the many decompositions of $\mathcal{T}$ according to these languages, the following new one is of particular interest for conditional counting.

Theorem 4.1 *Let $\mathcal{T}_+ \subset \mathcal{T}$ be the set of words on the alphabet $\mathcal{S}$ which contain at least one occurrence of* H_1 *or at least one occurrence of* H_2*. It satisfies the language equation*

$$\mathcal{T}_+ = \tilde{\mathcal{R}}_2 \mathcal{M}_{2,2}^* \tilde{\mathcal{U}}_2 + \mathcal{R}_1 \mathcal{M}_1^* \mathcal{U}_1 \tag{21}$$

that translates into the functional equation on the generating functions

$$T(z,u_1,u_2) = \frac{u_2 \tilde{R}_2(z) \tilde{U}_2(z)}{1 - u_2 M_{2,2}(z)} + \frac{u_1 R_1(z,u_2) U_1(z,u_2)}{1 - u_1 M_1(z,u_2)} \quad . \tag{22}$$

Proof: The first term of the right member is the set of words of $\mathcal{T}_+$ which do not contain any occurrence of H_1 ; such a text can be decomposed according to H_2 occurrences, using basic languages $\tilde{\mathcal{R}}_2, \mathcal{M}_{2,2}$, $\tilde{\mathcal{U}}_2$. The second term is the set of words of $\mathcal{T}_+$ that contain at least one occurrence of H_1; such a text can be decomposed according to H_1 occurrences, using basic languages $\mathcal{R}_1, \mathcal{M}_1, \mathcal{U}_1$. □

The proposition below establishes a decomposition of the basic languages for a single pattern onto the basic languages for several words. The bivariate generating functions that count H_2-occurrences in these basic languages follow.

Proposition 4.1 *Given a reduced couple of words* $(\mathrm{H}_1, \mathrm{H}_2)$*, the* basic languages *satisfy the following equations:*

$$\begin{aligned} \mathcal{R}_1 &= \tilde{\mathcal{R}}_1 + \tilde{\mathcal{R}}_2 \mathcal{M}_{2,2}^* \mathcal{M}_{2,1} \\ \mathcal{U}_1 &= \tilde{\mathcal{U}}_1 + \mathcal{M}_{1,2} \mathcal{M}_{2,2}^* \tilde{\mathcal{U}}_2 \\ \mathcal{M}_1 &= \mathcal{M}_{1,1} + \mathcal{M}_{1,2} \mathcal{M}_{2,2}^* \mathcal{M}_{2,1} \quad . \end{aligned}$$

The multivariate generating functions that count H_2*-occurrences in these languages are:*

$$R_1(z,u_2) = \tilde{R}_1(z) + \frac{u_2 \tilde{R}_2(z) M_{2,1}(z)}{1 - u_2 M_{2,2}(z)} \quad , \tag{23}$$

$$U_1(z,u_2) = \tilde{U}_1(z) + \frac{u_2 M_{1,2}(z) \tilde{U}_2(z)}{1 - u_2 M_{2,2}(z)} \quad , \tag{24}$$

$$M_1(z) = M_{1,1}(z) + \frac{u_2 M_{1,2}(z) M_{2,1}(z)}{1 - u_2 M_{2,2}(z)} \quad . \tag{25}$$

Proof: The proof of the first equation relies on a very simple observation: a word w in $\mathcal{R}_1$ is not in $\tilde{\mathcal{R}}_1$ iff it contains k occurrences of H_2 *before* H_1, with $k \geq 1$. Hence, such a word rewrites in a unique manner: $w = r_2 w_1 ... w_{k-1} m_{2,1}$ where $r_2 \in \tilde{\mathcal{R}}_2$, $w_i \in \mathcal{M}_{2,2}$ and $m_{2,1} \in \mathcal{M}_{2,1}$. A similar reasoning leads to the second and third equations. □

4.2 Partial derivatives

The proof of our main theorems, Theorem 4.2 and Theorem 4.3, relies on a suitable computation of the partial derivatives of the bivariate generating function. Notably, $\frac{\partial T}{\partial u_2}(z,u_1,1)$ yields the generating function of conditional expectations.

Proposition 4.2 *Let* $(\mathrm{H}_1,\mathrm{H}_2)$ *be a couple of words. The bivariate generating function of the* H_1*-conditional expectation of* H_2*-occurrences is, in Bernoulli and Markov models:*

$$\frac{\partial T}{\partial u_2}(z,u_1,1)=\phi_0(z)+\frac{u_1\phi_1(z)}{(1-u_1M_1(z,1))}+\frac{u_1^2\phi_2(z)}{(1-u_1M_1(z,1))^2} \tag{26}$$

where

$$\phi_0(z) = \frac{(-P(\mathrm{H}_1)\mathrm{D}_{1,2}(\mathrm{z})\mathrm{z}^{\mathrm{m}_1}+\mathrm{P}(\mathrm{H}_2)\mathrm{D}_1(\mathrm{z})\mathrm{z}^{\mathrm{m}_2})(-D_{2,1}(z)+D_1(z))}{(1-z)^2D_1(z)^2}\ , \tag{27}$$

$$\phi_1(z) = \frac{-2P(\mathrm{H}_1)\mathrm{D}_{2,1}(\mathrm{z})\mathrm{D}_{1,2}(\mathrm{z})\mathrm{z}^{\mathrm{m}_1}+\mathrm{P}(\mathrm{H}_2)\mathrm{D}_1(\mathrm{z})\mathrm{D}_{2,1}(\mathrm{z})\mathrm{z}^{\mathrm{m}_2}+\mathrm{P}(\mathrm{H}_1)\mathrm{D}_{1,2}(\mathrm{z})\mathrm{D}_1(\mathrm{z})\mathrm{z}^{\mathrm{m}_1}}{(1-\mathrm{z})\mathrm{D}_1(\mathrm{z})^3}\ , \tag{28}$$

$$\phi_2(z) = \frac{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}\mathrm{D}_{1,2}(\mathrm{z})\mathrm{D}_{2,1}(\mathrm{z})}{\mathrm{D}_1(\mathrm{z})^4}\ . \tag{29}$$

Proof: Deriving with respect to u_2 yields:

$$\begin{aligned}\frac{\partial T}{\partial u_2}(z,u_1,u_2) &= \frac{\tilde{R}_2(z)\tilde{U}_2(z)}{(1-u_2M_{2,2}(z))^2}+\frac{u_1}{1-u_1M_1(z,u_2)}\frac{\partial R_1(z,u_2)U_1(z,u_2)}{\partial u_2}\\ &+ \frac{u_1^2}{(1-u_1M_1(z,u_2))^2}\times R_1(z,u_2)U_1(z,u_2)\frac{\partial M_1(z,u_2)}{\partial u_2}\end{aligned}$$

Equations (23)-(25) allow for an easy derivation of (30). The partial derivatives of probability generating functions of languages $\mathcal{R}_1$, $\mathcal{U}_1$ and $\mathcal{M}_1$ satisfy the following equations:

$$\begin{aligned}\frac{\partial R_1}{\partial u_2}(z,u_2) &= \frac{\tilde{R}_2(z)M_{2,1}(z)}{(1-u_2M_{2,2}(z))^2}\ ,\\ \frac{\partial U_1}{\partial u_2}(z,u_2) &= \frac{M_{1,2}(z)\tilde{U}_2(z)}{(1-u_2M_{2,2}(z))^2}\ ,\\ \frac{\partial M_1}{\partial u_2}(z,u_2) &= \frac{M_{1,2}(z)M_{2,1}(z)}{(1-u_2M_{2,2}(z))^2}\ .\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\partial T}{\partial u_2}(z,u_1,u_2) &= \frac{\tilde{R}_2(z)\tilde{U}_2(z)}{(1-u_2M_{2,2}(z))^2}+\frac{u_1}{1-u_1M_1(z,u_2)}\frac{\tilde{R}_2(z)M_{2,1}(z)U_1(z,u_2)+R_1(z,u_2)M_{1,2}(z)\tilde{U}_2(z)}{(1-u_2M_{2,2}(z))^2}\\ &+ \frac{u_1^2}{(1-u_1M_1(z,u_2))^2}\frac{R_1(z,u_2)U_1(z,u_2)M_{1,2}(z)M_{2,1}(z)}{(1-u_2M_{2,2}(z))^2}\end{aligned} \tag{30}$$

To complete the proof, we rely on the results proved in [RS97b, Rég00], where the monovariate generating functions of the basic languages are expressed in terms of the coefficients of $\mathbb{D}(z)$. More precisely:

Proposition 4.3 *The matrix* $\mathbb{D}(z)$ *is regular when* $|z|<1$. *The generating functions of the basic languages are defined by the following equations:*

$$(\tilde{R}_1(z),\tilde{R}_2(z)) = (P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1},\mathrm{P}(\mathrm{H}_2)\mathrm{z}^{\mathrm{m}_2})\mathbb{D}(\mathrm{z})^{-1}\ , \tag{31}$$

$$\mathbb{I}-\mathbb{M}(z) = (1-z)\mathbb{D}(z)^{-1}\ , \tag{32}$$

$$\begin{bmatrix}\tilde{U}_1(z)\\ \tilde{U}_2(z)\end{bmatrix} = \frac{1}{1-z}\mathbb{D}(z)^{-1}\begin{bmatrix}1\\1\end{bmatrix}\ . \tag{33}$$

The classical inversion formulae in dimension 2 lead to the equation

$$\mathbb{D}(z)^{-1} = \frac{1}{determinant(\mathbb{D}(z))}\begin{bmatrix} D_{2,2}(z) & -D_{1,2}(z) \\ -D_{2,1}(z) & D_1(z) \end{bmatrix} .$$

Setting $u_2 = 1$ in (30) and substituting the expressions given in (31-33) yield (26). □

4.3 Conditional expectation

Our main theorem is Theorem 4.2 below. We introduce a few notations.

Notation: Let us denote

$$g(z) = \frac{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}}{zD_1(z)(D_1(z)+z-1)} , \tag{34}$$

$$\bar{g}(z) = \frac{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}}{z}\cdot\frac{D_{1,2}(z)D_{2,1}(z)}{D_1(z)^2(D_1(z)+z-1)^2} . \tag{35}$$

Let us denote l and $\bar{l}$ the orders at $z=0$ of g and $\bar{g}$, respectively, and let

$$\bar{\theta}(z) = \frac{\bar{g}(z)}{g(z)} = \frac{D_{1,2}(z)D_{2,1}(z)}{D_1(z)(D_1(z)+z-1)} , \tag{36}$$

$$\Theta(z) = z^{l-\bar{l}}\bar{\theta}(z) \tag{37}$$

Theorem 4.2 *Let $\mathcal{T}$ be the language of all possible words on an alphabet $\mathcal{S}$. Assume that $\mathcal{T}$ is randomly generated by a Bernoulli or a Markov process. Given a reduced couple of words $(\mathrm{H}_1, \mathrm{H}_2)$, we denote $X_{1,n}$ and $X_{2,n}$ the two random variables that count the number of occurrences of H_1 and H_2, respectively. The conditional expectation of $X_{2,n}$, knowing that $\frac{X_{1,n}}{n} = a$ is*

$$E(X_{2,n} \mid X_{1,n} = na) \sim n\mu(a) + \lambda(a) \tag{38}$$

where μ and λ are functions of the autocorrelation polynomials at the point z_a that is the solution of Equation (7). With the notations of (8) and (34)-(37), these functions are

$$\mu(a) = a\bar{\theta}(z_a) , \tag{39}$$

$$\begin{aligned} \lambda(a) &= -\bar{\theta}(z_a) + 3a\frac{\beta_a}{\tau_a}\Theta'(z_a)z_a^{\bar{l}-l} - a\frac{1}{2\tau_a^2}\left(\Theta''(z_a) + 2\Theta'(z_a)\cdot\frac{\partial \ln z^{-l}g(z)}{\partial z}(z_a)\right)z_a^{\bar{l}-l} \\ &+ \frac{D_1(z_a)^2\phi_1(z_a)}{P(\mathrm{H}_1)\mathrm{z}_\mathrm{a}^{\mathrm{m}_1}} + \mu(a)\frac{1}{2z_a^2\tau_a^2}(\bar{l}^2 - l^2) \\ &+ \mu(a)(\bar{l}-l)\left(\frac{1}{2\tau_a^2 z_a^2} + \frac{3\beta_a}{\tau_a z_a}\right) - \frac{\mu(a)}{z_a\tau_a^2}\left(\bar{l}\frac{\partial \ln \bar{g}(z)}{\partial z} - l\frac{\partial \ln g(z)}{\partial z}\right)(z_a) , \end{aligned} \tag{40}$$

with

$$\beta_a = \frac{1}{3!}\cdot\frac{1}{\tau_a^3}\cdot\frac{\partial^3}{\partial z^3}\left(a\ln\frac{D_1(z)+z-1}{D_1(z)} - \ln z\right)(z_a) .$$

Remark: In the central region, the substitutions $z_a = 1$ and $a = P(\mathrm{H}_1)$ in (39) steadily give that $\mu(a) = P(\mathrm{H}_2)$.

Proof: We are ready to compute (19). To get the *linear term*, we observe that

$$[z^n u_1^{k_1}]\frac{u_1^2}{(1-u_1M_1(z,1))^2} = [z^n](k_1-1)M_1(z,1)^{k_1-2} \quad .$$

We observe that $M_1(z,1)$ is equal to $M_1(z)$ in the previous section and that $D_{1,1}(z) = D_1(z)$. With $k_1 = na$, the ratio (19) to be computed becomes

$$(na-1)\frac{[z^n]\phi_2(z)M_1(z)^{na-2}}{P(X_{1,n}=na)} = (na-1)\frac{[z^n]\phi_2(z)M_1(z)^{na-2}}{J_g(a)} \quad .$$

In this ratio, the computation of the numerator contribution is similar to the computation of (8). The integrand rewrites

$$J_{\bar{g}}(a) = \frac{1}{2i\pi}\oint e^{nh_a(z)}\bar{g}(z)dz \quad ,$$

where $\bar{g}(z) = \frac{\phi_2(z)}{zM_1(z)^2} = \frac{P(\mathrm{H}_1)z^{m_1}}{z}\cdot\frac{D_{1,2}(z)D_{2,1}(z)}{D_1(z)^2(D_1(z)+z-1)^2}$ and the ratio (19) becomes $(na-1)\frac{J_{\bar{g}}(a)}{J_g(a)}$. Using (17), this is

$$(na-1)\frac{J_{\bar{G}}(a)}{J_G(a)}z_a^{\bar{l}-l}\left[1-\frac{1}{2z_a^2\tau_a^2}\cdot\frac{\bar{l}^2-l^2}{n}+\left(\frac{1}{2z_a^2\tau_a^2}+\frac{3\beta_a}{\tau_a z_a}\right)\frac{\bar{l}-l}{n}-\frac{1}{\tau_a^2 z_a}\left(\frac{\bar{G}'(z_a)}{\bar{G}(z_a)}\frac{\bar{l}}{n}-\frac{G'(z_a)}{G(z_a)}\frac{l}{n}\right)+O(\frac{1}{n^2})\right] \quad .$$

We use (16) to compute $\frac{J_{\bar{G}}(a)}{J_G(a)}z_a^{\bar{l}-l}$. The exponential terms simplify and the γ-terms cancel. The ratio (19) becomes

$$\begin{aligned}(na-1)\frac{\bar{G}(z_a)}{G(z_a)}z_a^{\bar{l}-l} \quad \times \quad & \left[1+\frac{1}{n}\cdot\frac{3\beta_a}{\tau_a}\left(\frac{\bar{G}'(z_a)}{\bar{G}(z_a)}-\frac{G'(z_a)}{G(z_a)}\right)-\frac{1}{n}\cdot\frac{1}{2\tau_a^2}\left(\frac{\bar{G}''(z_a)}{\bar{G}(z_a)}-\frac{G''(z_a)}{G(z_a)}\right)+O(\frac{1}{n^2})\right] \\ \times \quad & \left[1-\frac{1}{2z_a^2\tau_a^2}\cdot\frac{\bar{l}^2-l^2}{n}+\left(\frac{1}{2z_a^2\tau_a^2}+\frac{3\beta_a}{\tau_a z_a}\right)\frac{\bar{l}-l}{n}\right. \\ & \left.-\frac{1}{\tau_a^2 z_a}\left(\frac{\bar{G}'(z_a)}{\bar{G}(z_a)}\frac{\bar{l}}{n}-\frac{G'(z_a)}{G(z_a)}\frac{l}{n}\right)+O(\frac{1}{n^2})\right] \quad .\end{aligned}$$

With the notations above, we have $\frac{\bar{G}(z)}{G(z)}z^{\bar{l}-l} = \frac{\bar{g}(z)}{g(z)} = \bar{\theta}(z)$. It follows that the *linear term* is

$$a\bar{\theta}(z_a) = \mu(a) \quad .$$

Let us compute now the *constant term*. First, $-\frac{\bar{g}(z_a)}{g(z_a)}$ yields a contribution $-\bar{\theta}(z_a)$. Second, we observe that $\frac{f'(z)}{f(z)} = \frac{\partial \ln f(z)}{\partial z}$. This yields a contribution $a\cdot\bar{\theta}(z_a)\cdot\frac{3\beta_a}{\tau_a}\cdot\frac{\partial\ln\Theta(z)}{\partial z}(z_a) = a\frac{3\beta_a}{\tau_a}\Theta'(z_a)z_a^{\bar{l}-l}$. Third, the general equation

$$\frac{f''(z)}{f(z)} = \frac{\partial^2 \ln f(z)}{\partial z^2}+\left(\frac{\partial \ln f(z)}{\partial z}\right)^2$$

implies that

$$\begin{aligned}
\frac{\bar{G}''(z)}{\bar{G}(z)} - \frac{G''(z)}{G(z)} &= \frac{\partial^2 \ln \Theta(z)}{\partial z^2} + \left(\frac{\bar{G}'(z)}{\bar{G}(z)} - \frac{G'(z)}{G(z)}\right)\left(\frac{\bar{G}'(z)}{\bar{G}(z)} + \frac{G'(z)}{G(z)}\right) \\
&= \frac{\Theta''(z_a)}{\Theta(z_a)} - \frac{\Theta'(z_a)^2}{\Theta(z_a)^2} + \frac{\Theta'(z_a)}{\Theta(z_a)} \cdot \left(\frac{\Theta'(z_a)}{\Theta(z_a)} + 2\frac{G'(z)}{G(z)}\right) \\
&= \frac{1}{\Theta(z_a)}\left(\Theta''(z_a) + 2\Theta'(z_a)\frac{\partial \ln G(z)}{\partial z}\right)
\end{aligned}$$

which contributes

$$-\frac{a z_a^{\bar{l}-l}}{2\tau_a^2}\left(\Theta''(z_a) + 2\Theta'(z_a)\frac{\partial \ln z^{-l} g(z)}{\partial z}(z_a)\right) .$$

The next contribution is $\mu(a)(\bar{l}^2 - l^2)(\frac{-1}{2z_a^2\tau_a^2})$. The last term in the product contributes

$$\mu(a)(\bar{l} - l)\left(\frac{1}{2\tau_a^2 z_a^2} + \frac{3\beta_a}{\tau_a z_a}\right) - \frac{\mu(a)}{z_a \tau_a^2}\left(\bar{l}\frac{\partial \ln z^{-\bar{l}}\bar{g}(z)}{\partial z} - l\frac{\partial \ln z^{-l} g(z)}{\partial z}\right)(z_a) .$$

Furthermore, we have

$$-\frac{\mu(a)}{z_a\tau_a^2}\left(\bar{l}\frac{\partial \ln z^{-\bar{l}}\bar{g}(z)}{\partial z} - l\frac{\partial \ln z^{-l} g(z)}{\partial z}\right)(z_a) = -\frac{\mu(a)}{z_a\tau_a^2}\left(\bar{l}\frac{\partial \ln \bar{g}(z)}{\partial z} - l\frac{\partial \ln g(z)}{\partial z}\right)(z_a) + \frac{\mu(a)}{z_a^2\tau_a^2}(\bar{l}^2 - l^2)$$

Finally, the last contribution to the constant term comes from

$$\frac{[z^n u_1^{k_1}] u_1 \phi_1(z)(1 - u_1 M_1(z))^{-1}}{P(X_{1,n} = na)} = \frac{[z^n]\phi_1(z) M_1(z)^{na-1}}{P(X_{1,n} = na)} .$$

This coefficient is

$$\frac{\phi_1(z_a)}{z_a M_1(z_a)} \cdot \frac{1}{g(z_a)} = \frac{\phi_1(z_a) D_1(z_a)^2}{P(\mathrm{H}_1) \mathrm{z_a^{m_1}}} .$$

□

4.4 Conditional variance

We prove here that the variance is a linear function of n, except for a few degenerate cases. We provide the linearity constant.

Theorem 4.3 *With the same conditions as Theorem 4.2, the* conditional variance *of $X_{2,n}$, when $X_{1,n}$ is known and equal to na, is a linear function of n. More precisely,*

$$Var(X_{2,n} \mid X_{1,n} = na) \sim n\nu(a)$$

where

$$\nu(a) = \mu(a)\left[1 + 2\frac{M_{2,2}(z_a)}{1 - M_{2,2}(z_a)} - \bar{\theta}(z_a) - \frac{\mu(a)}{\tau_a^2}\left(\frac{\partial \ln \bar{\theta}(z_a)}{\partial z}\right)^2\right] . \tag{41}$$

Remark: It follows from (41) that the expectation $n\mu(a)$ is a tight approximation of the variance, when $M_{2,2}(z)$ is small. This is the usual case, but the contribution of $2\frac{M_{2,2}(z_a)}{1-M_{2,2}(z_a)}$ may be significant, for instance when $\mathrm{H}_2 = x^*$, where x is some character of the alphabet.

The linearity constant may also be 0 in some degenerate cases. For example, with an alphabet of size 2, the choices of the two words AB and BA leads to $M_{2,2}(z) = 0$ and $\bar{\theta}(z) = 1$. The variance is 0. As a matter of fact, the difference between the number of occurrences of AB and the number of occurrences of BA is at most 1.

We now use the formula (20). We only need to compute the second partial derivative of T. We proceed with a second differentiation of (30), using again the partial derivatives $\frac{\partial R_1}{\partial u_2}(z,u_2)$, $\frac{\partial U_1}{\partial u_2}(z,u_2)$ and $\frac{\partial M_1}{\partial u_2}(z,u_2)$. This yields notably $\frac{\partial^2 M_1}{\partial u_2^2} = \frac{M_{1,2}(z)M_{2,1}(z)M_{2,2}(z)}{(1-u_2M_{2,2}(z))^3}$ and, finally, we get Proposition 4.4 below.

Proposition 4.4 *With the same hypotheses as Proposition 4.2, we have*

$$\frac{\partial^2 T}{\partial u_2^2}(z,u_1,1) = \psi_0(z) + \frac{u_1\psi_1(z)}{(1-u_1M_1(z,1))} + \frac{u_1^2\psi_2(z)}{(1-u_1M_1(z,1))^2} + \frac{u_1^3\psi_3(z)}{(1-u_1M_1(z,1))^3} \tag{42}$$

where

$$\begin{aligned}
\psi_0(z) &= 2\frac{\phi_0(z)M_{2,2}(z)}{1-M_{2,2}(z)} \quad ,\\
\psi_1(z) &= 2\phi_0(z)\phi_2(z)\cdot\frac{D_1(z)^2}{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}} \quad ,\\
\psi_2(z) &= 2\phi_2(z)[\phi_1(z)\frac{D_1(z)^2}{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}} + \frac{M_{2,2}(z)}{1-M_{2,2}(z)}] \quad ,\\
\psi_3(z) &= 2\phi_2(z)^2\cdot\frac{D_1(z)^2}{P(\mathrm{H}_1)\mathrm{z}^{\mathrm{m}_1}} \quad .
\end{aligned}$$

Proof of Theorem 4.3: As a consequence of (42), we get

$$[u_1^{k_1}]\frac{\partial^2 T(z,u_1,u_2)}{\partial u_2^2} = \left[\frac{\psi_3(z)}{M_1(z)^3}\cdot\frac{(k_1-1)(k_1-2)}{2} + (k_1-1)\frac{\psi_2(z)}{M_1(z)^2} + \frac{\psi_1(z)}{M_1(z)}\right]M_1(z)^{k_1} \quad .$$

Let us denote:

$$\tilde{g} = \frac{\psi_3(z)}{z\cdot M_1(z)^3}, \quad \tilde{f} = \frac{\psi_2(z)}{z\cdot M_1(z)^2} \quad .$$

It follows that

$$[z^n u_1^{k_1}]\frac{\partial^2 T(z,u_1,u_2)}{\partial u_2^2} = \frac{(na)^2-3na}{2}J_{\tilde{g}}(a) + naJ_{\tilde{f}}(z_a) + O(e^{-nI(a)}) \quad .$$

Hence,

$$Var(X_{2,n} \mid X_{1,n} = na) = n^2\frac{a^2}{2}\frac{J_{\tilde{g}}(a)}{J_g(a)} + n\left(-\frac{3}{2}a\cdot\frac{\tilde{g}(z_a)}{g(z_a)} + a\frac{\tilde{f}(z_a)}{g(z_a)} + \mu(a)\right) - (n\mu(a)+\lambda(a))^2 + O(1) \quad .$$

To achieve the derivation, we need to establish relationships between $\tilde{g}(z), g(z)$ and $\bar{g}(z)$. We check that

$$\frac{\tilde{g}(z)}{g(z)} = 2\left(\frac{\bar{g}(z)}{g(z)}\right)^2 = 2\bar{\theta}(z)^2 \quad . \tag{43}$$

It follows that the quadratic terms $\frac{1}{2}\frac{\tilde{g}(z_a)}{g(z_a)}$ and $\left(\frac{\bar{g}(z_a)}{g(z_a)}\right)^2$ cancel. Consequently, the *variance is a linear function of n* . In a few degenerate cases, it is a constant function. Let us compute now the linearity coefficient. First of all, the sum $-\frac{3}{2}a\cdot\frac{\tilde{g}(z_a)}{g(z_a)}+\mu(a)$ contributes by $\mu(a)(1-3\bar{\theta}(z_a))$. The term $-\bar{\theta}(z_a)$ in $\lambda(a)$ yields the contribution $2\mu(a)\bar{\theta}(z_a)$. Then, we consider in turn the terms in (16) and (17) that contribute to $n^2\frac{a^2}{2}\frac{J_{\tilde{g}}(a)}{J_g(a)}+na\frac{\tilde{f}(z_a)}{g(z_a)}$ and $-2n\mu(a)\lambda(a)$. The first term is $a\frac{\tilde{f}(z_a)}{g(z_a)}-2\mu(a)\cdot\frac{D_1(z_a)^2\phi_1(z_a)}{P(\mathrm{H}_1)\mathrm{z_a^{m_1}}}$. As $a\frac{\bar{g}(z_a)}{g(z_a)}=\mu(a)$ and $\frac{\tilde{f}(z)}{\bar{g}(z)}=\frac{\psi_2(z)}{\phi_2(z)}$, this difference simplifies into

$$\mu(a)\left(\frac{\psi_2(z)}{\phi_2(z)}-2\frac{D_1(z_a)^2\phi_1(z_a)}{P(\mathrm{H}_1)\mathrm{z_a^{m_1}}}\right)=2\mu(a)\frac{M_{2,2}(z_a)}{1-M_{2,2}(z_a)} \quad .$$

We now observe that other contributions to $\frac{\partial^2 T}{\partial u_2^2}$ and $E(X_{2,n})^2$ have a common multiplicative factor: $\frac{a^2}{2}\frac{\tilde{g}(z_a)}{g(z_a)}=a^2(\frac{\bar{g}(z_a)}{g(z_a)})^2=\mu(a)^2$ or $-2\mu(a)\cdot a\frac{\bar{g}(z_a)}{g(z_a)}=-2\mu(a)^2$.

The next terms are the $(\frac{1}{2\tau_a^2 z_a^2}+\frac{3\beta_a}{\tau_a z_a})$ terms; the contributions are $\mu(a)^2(\tilde{l}-l)$ and $-2\mu(a)^2(\bar{l}-l)$. Equation (43) implies that $\tilde{l}-l=2(\bar{l}-l)$. Hence, these two contributions cancel. Similarly, the $(\tilde{l}^2-l^2)$ and $(\bar{l}^2-l^2)$-terms contribute

$$n\cdot\mu(a)^2\cdot\frac{(-1)}{2\tau_a^2 z_a^2}[(\tilde{l}^2-l^2)-2(\bar{l}^2-l^2)] \quad .$$

As $\tilde{l}-l=2(\bar{l}-l)$, we have $\tilde{l}-\bar{l}=\bar{l}-l$. Hence, $(\tilde{l}^2-l^2)-2(\bar{l}^2-l^2)$ rewrites $2(\bar{l}-l)^2$ and this yields $-\frac{\mu(a)^2}{\tau_a^2 z_a^2}(\bar{l}-l)^2$.

The two terms $\mu(a)^2\cdot\frac{3\beta_a}{\tau_a}\cdot\frac{\partial\ln\frac{\tilde{G}(z_a)}{G(z_a)}}{\partial z}$ and $2\mu(a)^2\cdot\frac{3\beta_a}{\tau_a}\cdot\frac{\partial\ln\frac{\bar{G}(z_a)}{G(z_a)}}{\partial z}$ contribute

$$\frac{3\beta_a}{\tau_a}\cdot\mu(a)^2\cdot[\frac{\partial\ln\frac{\tilde{G}(z_a)}{G(z_a)}}{\partial z}-2\frac{\partial\ln\frac{\bar{G}(z_a)}{G(z_a)}}{\partial z}] \quad .$$

Using again (43), we get

$$\frac{\partial\ln\frac{\tilde{G}}{G(z)}}{\partial z}=2\frac{\partial\ln\frac{\bar{G}}{G(z)}}{\partial z} \tag{44}$$

and these terms cancel. Now, we have

$$\frac{\tilde{G}''(z)}{\tilde{G}(z)}-\frac{G''(z)}{G(z)}=\frac{\partial^2\ln\frac{\tilde{G}(z)}{G(z)}}{\partial z^2}+\frac{\partial\ln\frac{\tilde{G}(z)}{G(z)}}{\partial z}\cdot\frac{\partial\ln(\tilde{G}(z)G(z))}{\partial z} \quad .$$

Deriving (44), we get $\frac{\partial^2 \ln \frac{\tilde{G}}{G(z)}}{\partial z^2} = 2\frac{\partial^2 \ln \frac{\bar{G}}{G(z)}}{\partial z^2}$. Hence, the two terms

$$\mu(a)^2 \left(\frac{\partial^2 \ln \frac{\tilde{G}(z)}{G(z)}}{\partial z^2} + \frac{\partial \ln \frac{\tilde{G}(z)}{G(z)}}{\partial z} \cdot \frac{\partial \ln(\tilde{G}(z)G(z))}{\partial z} \right)(z_a)$$

and

$$-2\mu(a)^2 \left(\frac{\partial^2 \ln \frac{\bar{G}(z)}{G(z)}}{\partial z^2} + \frac{\partial \ln \frac{\bar{G}(z)}{G(z)}}{\partial z} \cdot \frac{\partial \ln(\bar{G}(z)G(z))}{\partial z} \right)(z_a)$$

contribute

$$\mu(a)^2 \cdot \left(\frac{\partial \ln \frac{\tilde{G}(z)}{G(z)}}{\partial z} \cdot \frac{\partial \ln(\tilde{G}(z)G(z))}{\partial z} \right)(z_a) - 2\left(\frac{\partial \ln \frac{\bar{G}(z)}{G(z)}}{\partial z} \cdot \frac{\partial \ln(\bar{G}(z)G(z))}{\partial z} \right)(z_a) \quad .$$

We can factorize $2\frac{\partial \ln \frac{\bar{G}(z)}{G(z)}}{\partial z}$ and rewrite :

$$\frac{\partial \ln(\tilde{G}(z)G(z))}{\partial z} - \frac{\partial \ln(\bar{G}(z)G(z))}{\partial z} = \frac{\partial \ln(\frac{\tilde{G}(z)}{G})}{\partial z} - \frac{\partial \ln(\frac{\bar{G}(z)}{G})}{\partial z} = \frac{\partial \ln(\frac{\bar{G}(z)}{G(z)})}{\partial z} \quad .$$

Finally, the contribution of these two terms is

$$2\mu(a)^2 (\frac{\partial \ln \frac{\bar{G}(z)}{G(z)}}{\partial z})(z_a)^2 \cdot \frac{(-1)}{2\tau_a^2} = -\frac{\mu(a)^2}{\tau_a^2} \cdot (\frac{\partial \ln \Theta}{\partial z}(z_a))^2 \quad .$$

The last contribution is

$$\mu(a)^2 \cdot \frac{-1}{\tau_a^2 z_a} \left[\left(\tilde{l}\frac{\partial \ln \tilde{G}(z)}{\partial z} - l\frac{\partial \ln G(z)}{\partial z} \right)(z_a) - 2\left(\bar{l}\frac{\partial \ln \bar{G}(z)}{\partial z} - l\frac{\partial \ln G(z)}{\partial z} \right)(z_a) \right] \quad .$$

The third factor can be expressed as $\tilde{l}\frac{\partial \ln(\frac{\tilde{G}(z)}{G})}{\partial z} - 2\bar{l}\frac{\partial \ln(\frac{\bar{G}(z)}{G})}{\partial z} = 2(\tilde{l}-\bar{l})\frac{\partial \ln(\frac{\bar{G}(z)}{G})}{\partial z} = 2(\bar{l}-l)\frac{\partial \ln(\frac{\bar{G}(z)}{G})}{\partial z}$. Now, the overall contribution of

$$-\frac{\mu(a)^2}{\tau_a^2} \left[\frac{(\bar{l}-l)^2}{z_a^2} + \left(\frac{\partial \ln \Theta}{\partial z}(z_a) \right)^2 + 2(\bar{l}-l)\frac{1}{z_a}\frac{\partial \ln(\frac{\bar{G}(z)}{G})}{\partial z}(z_a) \right]$$

is $-\frac{\mu(a)^2}{\tau_a^2}\left(\frac{\partial \ln \bar{\theta}}{\partial z}(z_a) \right)^2$.

□

5 Conclusion

Our formulae apply for both Bernoulli and Markov models for random texts generation and provide sharp large deviation estimates. This approach needs much less computations than exact methods, in the domain where such methods are computable. Experimental evidence is presented in [DRV01], where our results are compared to others ([BFW$^+$00] and *RSA-tools*). Other applications, and a comparison with other methods [RS98, Nue01, RS01], are presented in [Rég03] and will be extended in a forthcoming paper. Maple procedures that implement a part of our results are available on request. An extension to under-represented words is possible, and related results are presented in [VM03].

A slight modification allows for the extension of these formulae to other counting models, such as the *renewal model* [Wat95, TA97]. A natural –and useful– generalisation of this work would be to give similar formulae for sets of motifs. In particular, computing expectation and variance conditioned by several overrepresented motifs would be useful to detect new significant information in biolnical sequences.

Acknowledgements

The authors are very grateful to the anonymous referee for his precise and valuable comments.

References

[BFW$^+$00] E. Beaudoing, S. Freier, J. Wyatt, J.M. Claverie, and D. Gautheret. Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Research.*, 10:1001–1010, 2000.

[BJVU98] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Prediction of Regulatory Elements in Silico on a Genomic Scale. *Genome Research*, 8:1202–1215, 1998.

[BK93] Edward A. Bender and Fred Kochman. The Distribution of Subwords Counts is Usually Normal. *European Journal of Combinatorics*, 14:265–275, 1993.

[BLS00] H. Bussemaker, H. Li, and E. Siggia. Building a Dictionary for Genomes: Identification of Presumptive Regulatory Sites by Statistical Analysis. *PNAS*, 97(18):10096–101000, 2000.

[DRV01] A. Denise, M. Régnier, and M. Vandenbogaert. Assessing Statistical Significance of Over-represented Oligonucleotides. In *WABI'01*, pages 85–97. Springer-Verlag, 2001. in Proc. First Intern. Workshop on Algorithms in Bioinformatics, Aarhus, Denmark, August 20001.

[DZ92] A. Dembo and O. Zeitouni. *Large Deviations Techniques*. Jones and Bartlett, Boston, 1992.

[EP02] E. Eskin and P. Pevzner. Finding Composite Regulatory Patterns in DNA Sequences. *Bioinformatics*, 1(1):1–9, 2002.

[FGSV01] P. Flajolet, Y. Guivarch, W. Szpankowski, and B. Vallee. Hidden Patterns Statistics. In *ICALP'01*, volume 2076 of *Lecture Notes in Computer Science*, pages 152–165. Springer-Verlag, 2001. Proc. ICALP'01.

[FS96] Ph. Flajolet and R. Sedgewick. *An Introduction to the Analysis of Algorithms*. Addison-Wesley, 1996.

[GKM00] M. Gelfand, E. Koonin, and A. Mironov. Prediction of Transcription Regulatory Sites in A*rchaea* by a Comparative Genome Approach. *Nucleic Acids Research*, 28:695–705, 2000.

[GO81] L. Guibas and A.M. Odlyzko. String Overlaps, Pattern Matching and Nontransitive Games. *Journal of Combinatorial Theory,* Series A, 30:183–208, 1981.

[Hen77] P. Henrici. *Applied and Computational Complex Analysis*. John Wiley, New York, 1977.

[Hwa98] H. K. Hwang. Large Deviations of Combinatorial Distributions II. Local Limit Theorems. *Annals of Applied Probability*, 8(1):163–181, 1998.

[KM97] S. Kurtz and G. Myers. Estimating the Probability of Approximate Matches. In *CPM'97*, Lecture Notes in Computer Science. Springer-Verlag, 1997.

[LBL01] X. Liu, D.L. Brutlag, and J. Liu. Bioprospector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-expressed Genes. In *6-th Pacific Symposium on Biocomputing*, pages 127–138, 2001.

[MMML02] M. Markstein, P. Markstein, V. Markstein, and M. Levine. Genome-wide Analysis of Clustered Dorsal Binding Sites Identifies Putative Target Genes in the drosophila Embryo. *PNAS*, 99(2):763–768, 2002.

[Nic00] P. Nicodème. The Symbolic Package RegExpCount. In *GCB'00*, 2000. presented at GCB'00, Heidelberg, October 2000; available at http://algo.inria.fr/libraries/software.html.

[NSF99] Pierre Nicodème, Bruno Salvy, and Philippe Flajolet. Motif Statistics. In *ESA'99*, volume 1643 of *Lecture Notes in Computer Science*, pages 194–211. Springer-Verlag, 1999. Proc. European Symposium on Algorithms-ESA'99, Prague; to appear in TCS.

[Nue01] G. Nuel. *Grandes déviations et chaines de Markov pour l'étude des mots exceptionnels dans les séquences biologiques*. Phd thesis, Université René Descartes, Paris V, 2001. defended in July,2001.

[PRdT95] B. Prum, F. Rodolphe, and E. de Turckheim. Finding Words with Unexpected Frequencies in DNA sequences. *J. R. Statist. Soc. B.*, 57:205–220, 1995.

[Rég00] M. Régnier. A Unified Approach to Word Occurrences Probabilities. *Discrete Applied Mathematics*, 104(1):259–280, 2000. Special issue on Computational Biology;preliminary version at RECOMB'98.

[Rég03] M. Régnier. Mathematical Tools for Regulatory Signals Extraction, 2003. presented at BGRS'02; to appear in ComPlexUs, Kluwer series.

[RS97a] M. Régnier and W. Szpankowski. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica*, 22(4):631–649, 1997. preliminary draft at ISIT'97.

[RS97b] M. Régnier and W. Szpankowski. On the Approximate Pattern Occurrences in a Text. In IEEE Computer Society, editor, *Compression and Complexity of SEQUENCES 1997*, pages 253–264, 1997. In Proceedings SEQUENCE'97,Positano, Italy.

[RS98] G. Reinert and S. Schbath. Compound Poisson Approximation for Occurrences of Multiple Words in Markov Chains. *Journal of Computational Biology*, 5(2):223–253, 1998.

[RS01] S. Robin and S. Schbath. Numerical Comparison of Several Approximations on the Word Count Distribution in Random Sequences. *Journal of Computational Biology*, 8(4):349–359, 2001.

[Szp01] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley and Sons, New York, 2001.

[TA97] M.S. Tanushev and R. Arratia. Central Limit Theorem for Renewal Theory for Several Patterns. *Journal of Computational Biology*, 4(1):35–44, 1997.

[VM03] M. Vandenbogaert and V. Makeev. Analysis of Bacterial RM-systems through Genome-scale Analysis and Related Taxonomic Issues. *In Silico Biology*, 3:12, 2003. preliminary version at BGRS'02.

[Wat95] M. Waterman. *Introduction to Computational Biology*. Chapman and Hall, London, 1995.