ESAIM: PS December 2010, Vol. 14, p. 435–455
DOI: 10.1051/ps/2009008

ESAIM: Probability and Statistics
www.esaim-ps.org

LARGE DEVIATIONS AND FULL EDGEWORTH EXPANSIONS FOR FINITE MARKOV CHAINS WITH APPLICATIONS TO THE ANALYSIS OF GENOMIC SEQUENCES

Pierre Pudlo[1]

Abstract. To establish lists of words with unexpected frequencies in long sequences, for instance in a molecular biology context, one needs to quantify the exceptionality of families of word frequencies in random sequences. To this aim, we study large deviation probabilities of multidimensional word counts for Markov and hidden Markov models. More specifically, we compute local Edgeworth expansions of arbitrary degrees for multivariate partial sums of lattice valued functionals of finite Markov chains. This yields sharp approximations of the associated large deviation probabilities. We also provide detailed simulations. These exhibit in particular previously unreported periodic oscillations, for which we provide theoretical explanations.

Mathematics Subject Classification. 60J10, 60F10, 60J55, 92D20, 60F05.

Received July 10, 2008. Revised March 3, 2009 and June 15, 2009.

Introduction

This paper is devoted to the determination of exact asymptotics of the probabilities of large deviations events for multidimensional additive functionals of finite Markov chains. A motivation that arises in molecular biology is the determination of under and over represented words in genomic sequences (DNA, RNA, and proteins), see Reinert *et al.* [28] for example. Words with unexpected frequencies in genomic sequences are natural candidates to represent biological signals. A well known example is the Chi motif in the sequence of *Escherichia coli*, namely the word `GCTGGTGG`, which is massively over represented and which plays a crucial role in the conservation of the genome of this bacterium. Of course, to detect words with unexpected frequencies, one must specify a stochastic model of the sequence. The most usual models are Markovian, that is, either one assumes that the sequence itself is Markov, or one uses hidden Markov models. Since hidden Markov chains can be represented as functionals of Markov chains, and since Markov chains of higher order are projections of simple Markov chains (that is, Markov chains of order 1) defined on product spaces, functionals of Markov chains and of hidden Markov chains of any order can be viewed as functionals of simple Markov chains.

The distribution of the number of visits to a given state by a Markov chain is a much studied subject, in particular in a molecular biology context. For instance, Robin and Daudin [29], Régnier [25] and Stefanov *et al.* [33] provide exact formulas for these distributions. Nuel [23] turns the occurrences of any pattern over a

Keywords and phrases. Markov chains, hidden Markov models, large deviations, edgeworth expansions, protein and DNA sequences.

[1] I3M, Université Montpellier 2, Place E. Bataillon, 34095 Montpellier Cedex, France; `pierre.pudlo@univ-montp2.fr`

Article published by EDP Sciences

Markovian model of order r into the occurrences of a subset of states over a Markov chain with minimal state space. See also Lladser *et al.* [16]. Approximations using the normal distribution and the Poisson distribution are in Schbath [31], Nicodème *et al.* [21], Reinert *et al.* [28] and Roquain and Schbath [30]. Large deviations asymptotics are in Régnier and Szpankowski [27], Régnier [25] and Régnier and Denise [26]. Numerical algorithms in Nuel [22] allow to study the significance of the count of one word or one pattern in a large deviations regime. The results of Nuel's software compare favorably with other asymptotic methods, when rare events are considered. Finally, for memoryless models, Flajolet *et al.* [7] provide asymptotics of occurrences of patterns with bounded or unbounded spacings between their letters.

In a statistical point of view, we want to test the null hypothesis that the observed frequency of a word is correctly predicted by the random model. When this hypothesis is rejected, we say that the word has unexpected frequency or that the observed counting is exceptional. The associated p-value is some probability in the tail of the distribution of the counting of this word in the random model. Moreover, given different words with unexpected frequency, we want to find the most unexpected one among them, that is the one with the lowest p-value. Thus, we must have good approximations of those different p-values to compare them.

In this context, large deviations principles (LDP) may seem appealing, because one compares probabilities of events on an exponential scale with respect to the length n of the sequence. Hence, one hopes for the events associated to small values of the rate function of the LDP to be massively more likely than events associated to large values of the rate function. While this conclusion is indeed correct in the limit $n \to \infty$, LDP are in fact used as a convenient tool to compare frequencies of words in real-world sequences, whose length can be large but is obviously finite. To give a caricature of an example where this can go awry, assume that the countings of the words w and w' both deviate from their theoretical values in the model under consideration and that the probability that the counting of w, respectively w', corresponds approximately to the observed relative counting in a sequence of length n is $c\exp\{-n\}$, respectively $c\exp\{-2n + 100n^{7/8}\}$, for a suitable positive constant c. Then, for any $n \ll 10^{16}$, that is, in quite a few concrete situations, the observed counting of w is much more exceptional than the observed counting of w', although the simple comparison of the rates $I(w) = 1$ and $I(w') = 2$ of the corresponding LDP would lead to the opposite prediction.

Another important point is that one wants to establish ordered lists of words with unexpected frequencies, rather than to show merely that one specific word is under or over represented. Hence, one must study joint laws of occurrences of different words in a sequence, we give an example in Section 3.2. To sum up the preceding, we are interested in precise asymptotics of large deviations probabilities for multidimensional functionals of Markov chains.

The mathematical setting is as follows. We are given a Markov chain $\{X_n\}_n$ on a finite state space E, with transition kernel Q. We write $\mathbb{P}_a$ for the probability measure $\mathbb{P}$ conditioned by the event $\{X_0 = a\}$. Let $F : E \to \mathbb{Z}^d$ and, for every $n \geq 1$,

$$S_n := \sum_{k=0}^{n-1} F(X_k).$$

For every $x = \{x_j\}_j$ and $y = \{y_j\}_j$ in $\mathbb{R}^d$, the inequality $x \geq y$ means that, for every j, $x_j \geq y_j$. Likewise, $x > y$ means that, for every j, $x_j > y_j$. The canonical basis of $\mathbb{R}^d$ are the vectors with a single one in each of the $j = 1 \dots d$ positions. We are interested in estimating the probabilities of events of the form $\{S_n \geq nv\}$, where v is in $\mathbb{R}^d$, or $\{S_n \geq \sigma_n\}$, where $\sigma_n = nv + o(n)$ and $\sigma_n \in \mathbb{Z}^d$. Let us introduce the following assumptions.

(H1) The Markov chain $\{X_n\}_n$ is irreducible and aperiodic, with invariant distribution π.

(H2) The additive process $\{(X_n, S_n)\}_n$ is aperiodic in the following sense: for every vector $\vec{e}$ of the canonical basis of $\mathbb{R}^d$, there exists a positive integer n, states a and b, and a point σ in $\mathbb{Z}^d$, such that $\mathbb{P}_a(X_n = b, S_n = \sigma)$ and $\mathbb{P}_a(X_n = b, S_n = \sigma + \vec{e})$ are both positive.

(H3) The point v lies in the interior of the convex hull of the range of F and

$$v > \mathbb{E}_\pi(F(X_0)) = \int F(x)\pi(\,\mathrm{d}x).$$

The large deviation probabilities $\mathbb{P}_a(S_n \geq nv)$ follows an exponential decrease with finite asymptotic rate $\Lambda^\star(v) = -\lim_{n\to\infty} n^{-1}\log\mathbb{P}_a(S_n \geq nv)$, that does not depend on the initial state a. Under (H1), (H2) and (H3), we obtain an expansion of $\mathbb{P}_a(S_n \geq nv)$ in Theorem 2.2, namely

$$\mathbb{P}_a(S_n \geq \sigma_n) = \frac{\mathrm{e}^{-J(v,n)}}{n^{d/2}}\left[\sum_{\ell=0}^{k} n^{-\ell/2}\Theta_a^\ell(n) + O\left(n^{-(k+1)/2}\right)\right], \qquad \text{as } n\to\infty, \tag{0.1}$$

where $J(v,n) = n\Lambda^\star(v) + t_v\cdot(\sigma_n - nv)$ for some $t_v \in \mathbb{R}^d$, and where $\Theta_a^\ell(n)$ are finite coefficients that may be explicitly calculated.

We use a classical method to prove this result, see Jensen [11]. The first step is an Edgeworth expansion of

$$g_a(n,\sigma) = \mathbb{E}_a(g(X_n)\mathbf{1}\{S_n = \sigma\})$$

when $g : E \to \mathbb{R}$, $\sigma \in \mathbb{Z}^d$ and $n\to\infty$. For this aim, we introduce $Q(z)$, an analytic perturbation of the transition kernel Q, such that

$$Q^n(z)g_a = \mathbb{E}_a\big(g(X_n)\exp(z\cdot S_n)\big)$$

and we obtain spectral results in Propositions 4.2 and 4.3.

In a second time, we normalize the kernel $Q(z)$, see (2.6), to build a new transition kernel $Q^{(z)}$, generally called the twisted kernel. For a suitable z in $\mathbb{R}^d$, under this twisted kernel, the additive process S_n has asymptotic mean v, *i.e.*, $\lim n^{-1}S_n = v$ almost surely, see (2.8). This change of measure and our Edgeworth expansion leads us to our main result (Thm. 2.2).

Our main contributions are: the proof of an Edgeworth expansion of any order and a complete treatment of the lattice case. Specifically, our assumption (H2) is milder than the classical ones in the literature, see Remark (ii) after Theorem 2.2. This yields the term $\exp\big(-t_v\cdot(\sigma_n - nv)\big)$ which is new, up to our knowledge, see Remark (iii) after Theorem 2.2.

The rest of the paper is organized as follows. In Section 1, we recall some known bounds on the deviations of functionals of Markov chains. In Section 2, we state the main results of the paper, namely local Edgeworth expansions in Theorem 2.1 and precise large deviations estimates in Theorem 2.2, and we prove Theorem 2.2 assuming Theorem 2.1. In Section 3.1, we perform some simulations and we explain the phenomena of periodic oscillations. In Section 3.2, we present numerical results on the genome of *Escherichia coli*. Section 4 is devoted to some preliminaries and Section 5 to the proof itself of Theorem 2.1. This uses Lemma 5.2, which we prove in Section 6.

1. State of the art

The seminal paper on large deviations of additive functionals of Markov chains is Miller [18]. The usual large deviations asymptotics of those processes have been refined in two ways, the authors proving either rigorous bounds, see Section 1.1, or exact asymptotic equivalents, see Section 1.2. Remarks (ii) and (iii) given after our Theorem 2.2 compare our results with the existing literature.

1.1. **Rigorous bounds**

Iscoe *et al.* [10] prove that there exists finite positive constants K_1 and K_2, such that, for every positive integer n and every state a,

$$K_1\, n^{-d/2}\mathrm{e}^{-n\Lambda^\star(v)} \leq \mathbb{P}_a(S_n \geq nv) \leq K_2\,\mathrm{e}^{-n\Lambda^\star(v)},$$

where $\Lambda^\star(v)$ is the exponential rate given by a LDP. We use some techniques of Iscoe *et al.* [10], namely the twisted kernel and the representation formula (see pp. 383–389 of this paper). In turn this transformation is a generalisation of the conjugate distribution, used to prove exact large deviations estimates for sums of i.i.d.

random variables, see for instance Bahadur and Rao [2], Ney [19] and pp. 110–113 of Dembo and Zeitouni [6]. A comprehensive reference on this method is Jensen [11].

Ney and Nummelin [20] improve on the tools of Iscoe *et al.* [10] and use the regenerative structure of the Markov chain, under a weaker recurrence hypothesis.

León and Perron [15] prove upper bounds of large deviations events for Markov-additive processes in dimension $d = 1$, when the underlying Markov chain is finite and reversible, with stationary distribution π. These authors get the upper bound

$$\mathbb{P}_\pi(S_n \geq nv) \leq \mathrm{e}^{-n\,K},$$

where K is a positive number that depends on the stationary mean $\mathbb{E}_\pi(F(X_0))$, on the end-points of the support of F, and on the second largest eigenvalue λ_2 of the transition kernel. When λ_2 is nonnegative, K is indeed the exponential rate given by the large deviation principle.

In the same spirit, there is the results of Kargin [12]. In this article, the author obtains a Bernstein-Hoeffding inequality for large deviation probabilities of multivariate additive functionals of a reversible, finite Markov chain.

1.2. Exact asymptotic equivalents

Chaganty and Sethuraman [4] provide equivalents of the probabilities of large deviations events for arbitrary random variables, making specific assumptions on the moment generating functions. To give a flavour of their results, let $\{T_n\}_n$ denote a sequence of lattice valued random variables with span 1, and $\{t_n\}_n$ a sequence of integers such that $t_n = O(n)$. For every n, let r_n denote the unique positive solution of $M_n'(r_n) = t_n$, where $M_n(z) := \log \mathbb{E}(\mathrm{e}^{z\,T_n})$, and let $\gamma_n := \mathbb{E}(\mathrm{e}^{r_n T_n})$. Then, assuming some technical conditions that we omit, Chaganty and Sethuraman [4] show that, when $n \to \infty$,

$$\mathbb{P}(T_n \geq t_n) = (2\pi M_n''(r_n))^{-1/2}(1-\mathrm{e}^{-r_n})^{-1}\mathrm{e}^{-r_n t_n}\,\gamma_n\,(1+o(1)). \tag{1.1}$$

In dimension $d = 1$ and for partial sums of i.i.d. random variables, full asymptotic expansions of $\mathbb{P}(S_n \geq nv)$ are in Bahadur and Rao [2]. In dimension $d \geq 2$, the behaviour of $\mathbb{P}(S_n \in nB)$ depends crucially on the geometry of the boundary of the Borel set B. As a consequence, the situation is much more complicated, even in the i.i.d. case. When B is convex and its interior is not empty, Ney [19] obtains asymptotics which depend on the geometry of the boundary of B around its so-called dominating point. Iltis [8] strengthens these results. Andriani and Baldi [1] recover this equivalent and focus on the geometric meaning of the terms involved. We mention that Barbe and Broniatowski [3] settle the question for sums of i.i.d. random variables and arbitrary Borel sets B.

Kontoyiannis and Meyn [14] prove an equivalent of the pre-exponent term for Markov-additive processes in dimension $d = 1$, when the underlying Markov chain lives on a general state space and is geometrically ergodic. Their main results are summed up in a multiplicative ergodic theorem, which gives asymptotics on the Fourier or Laplace transform of S_n. To this aim, they impose a Lyapounov condition on the Markov chain, which ensures the geometric ergodicity. These authors get the first two terms of the Edgeworth expansion of the distribution function and they prove that the error term is $o(n^{-1/2})$. No higher order Edgeworth expansion is given. Datta and McCormick [5] also get the first two terms of the Edgeworth expansion. In the multidimensional case, Iltis [9] gives an equivalent of $\mathbb{P}(S_n \in n\,B)$, thus reducing the problem to the evaluation of the asymptotics of certain integrals, whose behaviour is however not studied.

2. Results

2.1. Local Edgeworth expansions

Consider a nonzero function $g : E \to \mathbb{R}$. Our aim is to prove local Edgeworth expansions, of any order and uniform in σ, on

$$g_a(n,\sigma) = \mathbb{E}_a[g(X_n)\mathbf{1}\{S_n = \sigma\}].$$

In particular, when $g \equiv 1$, $g_a(n,\sigma)$ is the density of the distribution of S_n with respect to the counting measure of $\mathbb{Z}^d$.

Theorem 2.1 below provides such expansions, and uses a sequence of bounded functions $\{\psi^k\}_k$, which we now construct. For every z in $\mathbb{C}^d$, consider the kernel $Q(z)$ such that, for every states a and b,

$$Q(z)(a,b) := Q(a,b)\,\mathrm{e}^{z\cdot F(a)}.$$

Note that, for all $n \geq 0$, $Q^n(z)g$ in a vector indexed by the state space E, and that, for all state a,

$$\Big[Q^n(z)g\Big]_a = \mathbb{E}_a\big(g(X_n)\exp(z\cdot S_n)\big).$$

At least when z belongs to a suitable neighbourhood V of the origin of $\mathbb{C}^d$, since the state space E is finite, $Q(z)$ has a dominating eigenvalue, say $\mathrm{e}^{\Lambda(z)}$. Actually, for every $n \geq 1$, every z in V,

$$Q^n(z) = \mathrm{e}^{n\Lambda(z)}N(z) + R^n(z),$$

where $N(z)$ is a projection matrix and $R(z)$ a matrix such that $\lim_{n\to\infty}\mathrm{e}^{-n\Lambda(z)}\|R^n(z)\| = 0$. Thus,

$$G(z) := N(z)g \tag{2.1}$$

is a right eigenvector of $Q(z)$ of eigenvalue $\mathrm{e}^{\Lambda(z)}$ and

$$G(z) = \lim_{n\to\infty}\mathrm{e}^{-n\Lambda(z)}Q^n(z)g.$$

(See Sect. 4 for the existence of Λ, N, R and G, and more specifically Prop. 4.2.) Moreover, $z \mapsto G(z)$ and $z \mapsto \Lambda(z)$ are analytic functions on V and their Taylor expansions at the origin read

$$\Lambda(z) = \sum_{k\geq 1}\Lambda^{(k)}(z) \quad G_a(z) = \sum_{k\geq 0}G_a^{(k)}(z),$$

for every state a, where, for every nonnegative integer k, each of $\Lambda^{(k)}(z)$ and $G_a^{(k)}(z)$ is either zero or a homogeneous polynomial in z of degree k. For instance,

$$G_a(0) = \mathbb{E}_\pi(g(X_0)), \quad \Lambda^{(1)}(z) = m\cdot z, \quad \Lambda^{(2)}(z) = \frac{1}{2}z\cdot\Gamma z$$

where $m := \mathbb{E}_\pi(F(X_0))$ is the asymptotic mean and Γ is the asymptotic covariance matrix, hence, when $n \to \infty$, $S_n = nm + o(n)$ almost surely and

$$\Gamma = \frac{1}{n}\operatorname{Var}(S_n) + o(1). \tag{2.2}$$

Since, $\{n^{-1}\operatorname{Var}(S_n)\}$ is a sequence of positive semidefinite symmetric matrices, Γ is also a positive semidefinite symmetric matrix. Our assumptions imply that Γ is nonsingular, see Lemma 4.1.

Let $L := \Lambda - \Lambda^{(1)} - \Lambda^{(2)}$. For every u in $\mathbb{C}$ and z in $\mathbb{C}^d$ such that uz belongs to V, let

$$P(u,z) := \mathrm{e}^{L(uz)/u^2}\,G(uz). \tag{2.3}$$

For every z in $\mathbb{C}^d$, this defines a vector function $P(\cdot,z)$, analytic at the origin, whose expansion along the powers of u reads

$$P(u,z) = \sum_{k\geq 0}P^{(k)}(z)\,u^k. \tag{2.4}$$

For every nonnegative integer k and every state a, we introduce a function ψ_a^k given by

$$\psi_a^k := P_a^{(k)}(D)\,\varphi_\Gamma,$$

where φ_Γ is the density of the centered normal distribution on $\mathbb{R}^d$ with covariance matrix Γ and $P_a^{(k)}(D)$ is a differential operator defined as follows. For every vector $K := \{K_j\}_j$ of nonnegative integers, denote

$$z^K := \prod_{j=1}^d {z_j}^{K_j}, \quad D^K\varphi := \frac{\partial^{K_1+\cdots+K_d}\varphi}{\partial {t_1}^{K_1}\cdots\partial {t_d}^{K_d}}\cdot$$

Note that, if the K_j are all equal to zero, $D^K\varphi$ is φ itself. For every polynomial $P(z)$ in $\mathbb{C}[z_1,\ldots,z_d]$, the conventions above allow to define the effect of the differential operator $P(D)$ on every smooth function φ as

$$P(D)\varphi := \sum_K \beta_K D^K\varphi, \qquad P(z) := \sum_K \beta_K z^K.$$

The functions ψ_a^k are bounded. Each function $P_a^{(k)}(z)$ is a polynomial function of z, of degree at most $(3k)$, which involves a finite number of functions $\Lambda^{(\ell)}$ and $G_a^{(\ell)}$. For instance, $P^{(0)}(z) = G^{(0)}(z) = G(0)$, hence $P_a^{(0)}(z) = \mathbb{E}_\pi(g(X_0))$ does not depend on a, and

$$P_a^{(1)}(z) = \Lambda^{(3)}(z)\,G_a^{(0)}(z) + G_a^{(1)}(z).$$

Theorem 2.1. *Assume that (H1) and (H2) hold. Then, for every nonnegative integer k, there exists a positive constant C_k, which depends on (Q,F,a,g), but not on σ, such that, for every σ in $\mathbb{Z}^d$ and every integer n,*

$$\left| g_a(n,\sigma) - n^{-d/2}\sum_{\ell=0}^k n^{-\ell/2}\psi_a^\ell\left(n^{-1/2}(\sigma - nm)\right)\right| \le C_k n^{-(d+k+1)/2}.$$

The proof of Theorem 2.1 is postponed to Section 5.

Remarks. (i) The functions ψ_a^k depend on the transition kernel Q and on the functions F and g. What might be more surprising is that, for every positive integer k, ψ_a^k also depends on the starting point a of the Markov chain. On the other hand, ψ_a^0 does not depend on a. One can also observe that the constant C_k does not depend on σ.

(ii) Since Theorem 2.1 holds for any order k, one may be tempted to consider the limit $k \to \infty$, *i.e.* the infinite series

$$\sum_{\ell=0}^\infty n^{-\ell/2}\psi_a^\ell\left(n^{-1/2}(\sigma - nm)\right).$$

However, as is well know even in the i.d.d. case, the resulting infinite series needs not converge for any n. Actually, the sequence $\{C_k\}_k$ in Theorem 2.1 is not bounded and $C_k n^{-(d+k+1)/2}$ does not go to zero when n is fixed and $k \to \infty$.

(iii) When σ is far away from nm (we recall that $m = \mathbb{E}_\pi(F(X_0))$), $g_a(n,\sigma)$ and its expansion might be both small with respect to $n^{-(d+k+1)/2}$. On the other hand, when σ is close to nm, $\psi_a^0(n^{-1/2}(\sigma - nm))$ is close to the positive real number

$$\langle\Gamma\rangle := \varphi_\Gamma(0) = ((2\pi)^d \det\Gamma)^{-1/2},$$

and $C_0\, n^{-(d+1)/2}$ is small with respect to the first term of the expansion.

Thus, Theorem 2.1 cannot be directly applied to get approximations of probabilities which quantify the exceptionality of words in a biological sequence. Indeed, the difference between the observed frequency and the expected frequency is, at least in interesting cases, large. What we need are asymptotics in the regime of large deviations, *cf.* Theorem 2.2 below.

2.2. Precise large deviation expansions

When $n \to \infty$, $\log \mathbb{E}_a(\exp(t \cdot S_n)) = n\Lambda(t) + o(n)$ and $\Lambda(t)$ does not depend on the initial state a of the Markov chain. Furthermore, $\mathrm{e}^{\Lambda(t)}$ is the greatest eigenvalue of the positive and irreducible matrix $Q(t)$, see Dembo and Zeitouni [6] (p. 73), and the sequence $\{n^{-1}S_n\}_n$ satisfies a large deviation principle (LDP). The rate function $\Lambda^\star$ is given by

$$\Lambda^\star(v) = \sup_{t\in\mathbb{R}^d}\{t \cdot v - \Lambda(t)\}. \tag{2.5}$$

To state our second result, we fix v in $\mathbb{R}^d$. Without loss of generality, we can assume that $v > m$. Otherwise, we can change the sign of some components of F.

The right eigenvector $G(t)$ defined in (2.1) induces a twisted kernel $Q^{(t)}$, defined as follows: for every t in $\mathbb{R}^d$ and every states a and b,

$$Q^{(t)}(a, b) := G_a(t)^{-1}Q(a, b)\,\mathrm{e}^{t\cdot F(a)-\Lambda(t)}G_b(t). \tag{2.6}$$

Let $\mathbb{P}^{(t)}$ denote the probability measure such that $\{X_n\}_n$ is a Markov chain of kernel $Q^{(t)}$, and $\mathbb{E}^{(t)}$ denote the expectation with respect to $\mathbb{P}^{(t)}$. Assumption (H3) implies that there exists one and only one root t to the equation $\Lambda'(t) = v$, see Ney and Nummelin [20]. Then $\Lambda^\star(v) = t \cdot v - \Lambda(t)$ and one gets the following representation formula: for every function h and every state a,

$$\mathbb{E}_a(h(X_n, S_n)) = \mathrm{e}^{-n\Lambda^\star(v)}G_a(t)\mathbb{E}^{(t)}_a\Big(G_{X_n}(t)^{-1}\mathrm{e}^{-t\cdot(S_n-nv)}h(X_n, S_n)\Big). \tag{2.7}$$

The main interest of this transformation is that, under $\mathbb{P}^{(t)}$, the asymptotic mean of S_n is v, namely

$$\lim_{n\to\infty} n^{-1}S_n = \mathbb{E}^{(t)}_{\pi(t)}(F(X_0)) = v \quad \mathbb{P}^{(t)}\text{-a.s.}, \tag{2.8}$$

where $\pi(t)$ denotes the stationary distribution of $Q^{(t)}$, see Ney and Nummelin [20].

Using Theorem 2.1, this transformation yields an expansion of

$$p_a(n, \sigma) := \exp\big(n\Lambda^\star(v) + t \cdot (\sigma - nv)\big)\,\mathbb{P}_a(S_n = \sigma),$$

which, in turn, yields a precise approximation of $\mathbb{P}_a(S_n = \sigma)$ when σ is around nv. Summing these, one also gets an expansion of $q_a(n, \sigma)$ where, for every σ in $\mathbb{Z}^d$,

$$q_a(n, \sigma) := \exp\big(n\Lambda^\star(v) + t \cdot (\sigma - nv)\big)\,\mathbb{P}_a(S_n \geq \sigma).$$

For every nonnegative integer k, introduce

$$\vartheta^k_a(z) := G_a(t)\psi^k_a(z),$$

where $\{\psi^k_a\}_k$ denotes the sequence of Theorem 2.1 for the twisted kernel $Q^{(t)}$ and the function $g(a) = G_a(t)^{-1}$. Also, we consider

$$\Theta^k_a(n) := \sum_{y\geq 0} \mathrm{e}^{-t\cdot y}\,\vartheta^k_a\Big(n^{-1/2}(\sigma_n + y - nv)\Big),$$

where the sum is taken over all $y \geq 0$ in $\mathbb{Z}^d$. Note that $v > m$ implies that $t > 0$. Moreover the functions ψ^k_a are bounded. Thus $\Theta^k_a(n)$ is finite. With these notations in mind, one gets the following result.

Theorem 2.2. *Assume that (H1), (H2) and (H3) hold. Then, for every nonnegative integer k, there exists a positive constant C'_k, which depends on (Q, F, a, g) but not on σ, such that, for every σ in $\mathbb{Z}^d$ and every positive integer n,*

$$\left| p_a(n,\sigma) - n^{-d/2} \sum_{\ell=0}^{k} n^{-\ell/2} \vartheta_a^\ell \left(n^{-1/2}(\sigma - nv) \right) \right| \le C'_k n^{-(d+k+1)/2}.$$

Moreover, for every sequence $\{\sigma_n\}_n$ of vectors in $\mathbb{Z}^d$ such that $\sigma_n = nv + o(\sqrt{n})$ and every integer k, there exists a positive constant C''_k, which depends on (Q, F, a, v) but not on the sequence $\{\sigma_n\}_n$, such that, for every positive integer n,

$$\left| q_a(n,\sigma_n) - n^{-d/2} \sum_{\ell=0}^{k} n^{-\ell/2} \Theta_a^\ell(n) \right| \le C''_k n^{-(d+k+1)/2}.$$

Remarks. (i) The probabilities $\mathbb{P}_a(S_n = \sigma_n)$ and $\mathbb{P}_a(S_n \ge \sigma_n)$ are equivalent up to a multiplicative constant, in other words their ratios converge to a positive, finite number. Thus, the local behaviour of S_n near σ_n is of the same order than its behaviour on $\{x \in \mathbb{Z}^d : x \ge \sigma_n\}$. This is a classical phenomenon of large deviation estimates in the logarithmic scale: a key principle is that any large deviation is done in the least unlikely of all the available unlikely ways. See, for instance, the definition of a dominating point in Ney [19] as the least unlikely point of a Borel set.

(ii) We need to introduce the sequence $\{\sigma_n\}_n$ because nv might not belong to $\mathbb{Z}^d$ for every value of n, in which case $\mathbb{P}_a(S_n = nv)$ would be zero. Even the behaviour of the tail probability $\mathbb{P}_a(S_n \ge nv)$ depends on the fact that nv is in $\mathbb{Z}^d$ or not. This is one of the difficulty in the lattice case. For instance, Dembo and Zeitouni [6] need some restrictive assumptions to state the theorem of Bahadur and Rao [2] on p. 110. In the lattice case and for i.i.d. random variables, this yields an approximation of $\mathbb{P}(S_n \ge nv)$, provided $\mathbb{P}(F(X_0) = v)$ is neither 0 nor 1. Alas, when studying the number of visits to a given state, F denotes an indicator function, hence this assumption on $\mathbb{P}(F(X_0) = v)$ implies that $v = 0$ or $v = 1$, while one wants to get asymptotics when v equals the observed frequency of the given set, which is typically neither 0 nor 1.

(iii) Once we understand that we should replace nv by some sequence σ_n in $\mathbb{Z}^d$, a second issue arise. How to get a tractable formula for $\mathbb{P}_a(S_n \ge \sigma_n)$ when $\sigma_n = nv + o(n^{-1/2})$? For instance, the practical use of the expansion of Chaganty and Sethuraman [4], given in (1.1) depends on one's ability to get asymptotics for the sequences of general term $M''_n(r_n)$, γ_n and r_n. It might require heavy numerical calculations. On the contrary, our results provide directly asymptotics of those probabilities. In particular, their equivalent does not explain directly the oscillations observed in our simulations (see Sect. 3.1) whereas the term $\exp(-t \cdot (\sigma_n - nv))$ in our results is a transparent explanation.

The equivalent of Kontoyiannis and Meyn [14] on Markov-additive processes in dimension $d = 1$ does not explain directly the oscillations observed in our simulations. More precisely, Theorem 6.5 of [14] gives an equivalent of $\mathbb{P}(S_n \ge nv_n)$ in dimension $d = 1$ and for v_n lying in the support of S_n. However, the reasoning that gives a simpler equivalent below this theorem, when v_n converges to some v, misses the term $\exp(-t \cdot (nv_n - nv))$.

(iv) In the biological context described in the introduction, v is a vector of observed frequencies. Thus the coordinates of v are rational numbers. When n is the length of the observed sequence, the coordinates of nv are of course integer numbers since these are counting numbers. But for many other values of n, nv does not belong to $\mathbb{Z}^d$. Since $\{S_n \ge nv\} = \{S_n \ge \lceil nv \rceil\}$, we use $\sigma_n = \lceil nv \rceil$ in biological applications. Note that, when v is rational, the sequence $\{\lceil nv \rceil - nv\}_n$ is periodic and the factor $\exp(-t \cdot (\sigma_n - nv))$ leads to periodic oscillations. See Section 3.1 for more on this.

Proof of Theorem 2.2. Assume that Theorem 2.1 holds. The representation formula of equation (2.7) gives

$$p_a(n,\sigma) = G_a(t) \mathbb{E}_a^{(t)} \left[G_{X_n}(t)^{-1} \mathbf{1}\{S_n = \sigma\} \right].$$

Note that, if a, b are two states, $Q(a,b) > 0$ if and only if $Q^{(t)}(a,b) > 0$. Thus, if (H1) and (H2) hold for the original Markov chain, they hold for the twisted Markov chain. Hence, Theorem 2.1 applied to the twisted kernel, with $g(a) = G_a(t)^{-1}$, completes the proof of the first part.

Summing the previous inequalities, one gets the second part of the theorem. Indeed,

$$q_a(n,\sigma_n) = G_a(t)\sum_{y\geq 0} \mathrm{e}^{-t\cdot y}\, g_a^{(t)}(n,\sigma_n+y),$$

where the sum is taken over all $y \geq 0$ in $\mathbb{Z}^d$, and

$$g_a^{(t)}(n,\sigma_n+y) = \mathbb{E}_a^{(t)}\left[G_{X_n}(t)^{-1}\mathbf{1}\{S_n=\sigma_n+y\}\right].$$

Theorem 2.1 applied uniformly on $y \geq 0$ then concludes the proof of Theorem 2.2. □

3. Numerical illustrations

3.1. Simulations

In this section $\{\sigma_n\}_n$ denotes a sequence in $\mathbb{Z}^d$ such that $\sigma_n = nv + o(\sqrt{n})$. We study the behaviour of $\mathbb{P}_\pi(S_n \geq \sigma_n)$ through simulations. Let $\Gamma(t)$ denote the $d\times d$ covariance matrix and $\pi(t)$ the stationary distribution of the Markov chain with respect to the twisted kernel $Q^{(t)}$. Lemma 4.1 shows that $\Gamma(t)$ is nonsingular. Let

$$\langle\Gamma(t)\rangle = \varphi_{\Gamma(t)}(0) = \big((2\pi)^d\,\det\Gamma(t)\big)^{-1/2}.$$

Theorem 2.2 implies that

$$\mathbb{P}_a(S_n=\sigma_n) = \vartheta^0(0)\,n^{-d/2}\mathrm{e}^{-t\cdot(\sigma_n-nv)-n\Lambda^\star(v)}\,(1+o(1)),$$

where the positive constant $\vartheta^0(0)$ is $\vartheta^0(0) = \langle\Gamma(t)\rangle G_a(t)$. If we assume furthermore that $\sigma_n = nv + O(1)$, then

$$\mathbb{P}_a(S_n=\sigma_n) = \vartheta^0(0)\,\big(1+O(\sqrt{n})\big)n^{-d/2}\mathrm{e}^{-t\cdot(\sigma_n-nv)-n\Lambda^\star(v)}.$$

Since σ_n belongs to $\mathbb{Z}^d$, the hypothesis $\sigma_n = nv+O(1)$ is the tighter control that one can impose sensibly upon the behaviour of σ_n. The second part of Theorem 2.2 then implies that

$$\mathbb{P}_a(S_n\geq\sigma_n) = \Theta^0(0)n^{-d/2}\mathrm{e}^{-t\cdot(\sigma_n-nv)-n\Lambda^\star(v)}\,(1+o(1)),$$

where $\Theta^0(0) = \vartheta^0(0)\tau(t)^{-1}$ and $\tau(t) := \prod_{j=1}^{d}(1-\mathrm{e}^{-t_j})$.

Thanks to the representation formula in equation (2.7), it is enough to simulate

$$R_n := \mathbb{E}_{\pi(t)}^{(t)}\big(\mathrm{e}^{-t\cdot(S_n-nv)}\mathbf{1}\{S_n\geq\sigma_n\}\big),$$

for the value of t such that $v = \mathbb{E}_{\pi(t)}(F(X_0))$ is the asymptotic expected value of $n^{-1}S_n$ under $\mathbb{P}^{(t)}$. Specifically, we check the following predictions.

(i) The behaviour of $\mathbb{P}(S_n\geq nv)$ depends on d through the factor $n^{d/2}$.

(ii) The factor $\mathrm{e}^{-t\cdot(\sigma_n-nv)}$ may lead to periodic oscillations.

(iii) The factor $\langle\Gamma(t)\rangle\,\tau(t)^{-1}$ is correct.

Additionally, we observe that the convergence of R_n to its limit points is slow when t is close to the origin. Indeed,

$$R_n = \mathrm{e}^{-t\cdot(\sigma_n-nv)}\sum_{y\geq 0}\mathrm{e}^{-t\cdot y}R_n(y) \quad \text{where } R_n(y) := \mathbb{P}_{\pi(t)}^{(t)}(S_n=y+\sigma_n)\cdot$$

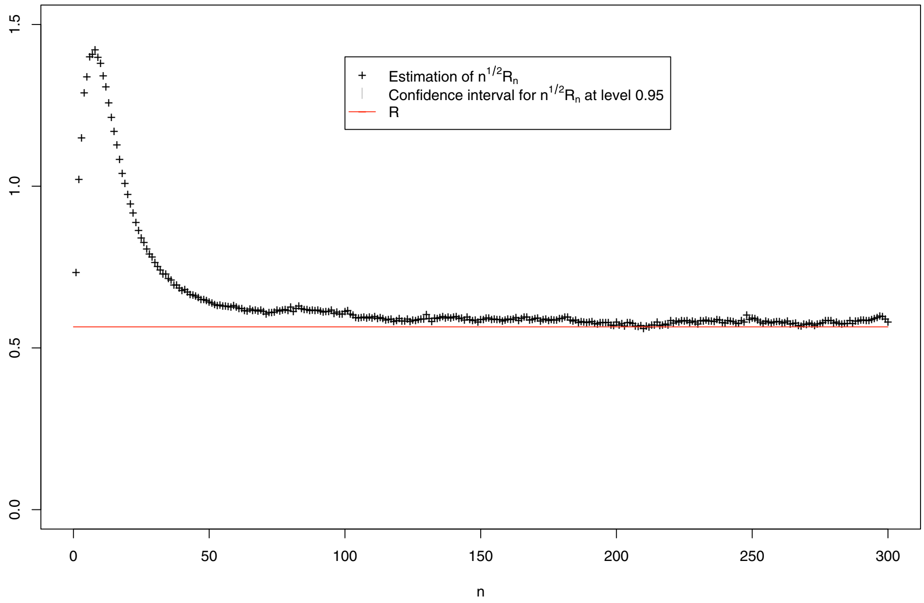

FIGURE 1. Here $F = \mathbf{1}_0 - \mathbf{1}_1$ and $X_{n+1} = X_n + U_n$ modulo 7 for an i.i.d. sequence $\{U_n\}_n$ with distribution $\frac{7}{10}\delta_0 + \frac{3}{10}\delta_1$, $t = 2$, and $R_n = \mathbb{E}(\mathrm{e}^{-t\cdot S_n}\mathbf{1}\{S_n \geq 0\})$. Then $n^{1/2}\,R_n$ converges to $R \approx .5650774$.

From Theorem 2.1, $R_n(y)$ is approximated by $\widetilde{R}_n(y)$ uniformly over y, with

$$\widetilde{R}_n(y) := \varphi_{\Gamma(t)}\left(\frac{y + s_n - nv}{\sqrt{n}}\right).$$

Every $\widetilde{R}_n(y)$ converges to $\langle\Gamma(t)\rangle$ when $n \to \infty$ and the convergence is fast when y is close to the origin, otherwise it is slow. Moreover, when t is close to the origin, many terms contribute to the overall sum. Thus, the overall convergence is significantly slower when t is close to the origin. Such phenomena do not appear for $\mathbb{P}^{(t)}(S_n = \sigma_n)$ because only $R_n(0)$ gets involved.

As regards the technical side of the simulations, we used the R language and environment [24]. We used different random number generators, such as the Mersenne-Twister generator whose period is around 2^{19937}. We obtained the same results with other, widely tested, random generators, such as the Marsaglia-Multicarry generator.

Figures 1, 2 and 3 exhibit a convergence for $d = 1$, $d = 3$ and $d = 2$ respectively. In Figures 1 and 2, $v = 0$ is in $\mathbb{Z}^d$, $\sigma_n = nv$ is in $\mathbb{Z}^d$, and $\mathrm{e}^{-t\,(\sigma_n - nv)} = 1$ for every n. By contrast, in Figures 3 and 4, the factor $\mathrm{e}^{-t\cdot(\sigma_n - nv)}$ leads to oscillations. To see why, note that $\{S_n \geq n\,v\} = \{S_n \geq \lceil nv\rceil\}$ where, for every $x = \{x_j\}_j$ in $\mathbb{R}^d$, the jth component of $\lceil x\rceil$ is defined as the unique integer y_j such that $x_j \leq y_j < x_j + 1$. Furthermore, when the components of v are rational, the sequence $\lceil nv\rceil - nv$ is periodic, with period 6 in the case of Figure 4. Indeed, the simulations in Figure 4 use the $\lceil nv\rceil$ round-off and the results are clearly 6 periodic. In Figure 3, we use the other possible round-off, namely $\sigma_n = \lfloor nv\rfloor$. We obtain similar oscillations whose period is 4, that is, precisely the period of the sequence $\lfloor nv\rfloor - nv$.

The values of the limit points in each figures are easy to compute since we know the law of the Markov chain under $Q^{(t)}$ in each simulation. Indeed, we compute $\Gamma(t)$ and use Theorem 2.2 to get numerical values of the limits.

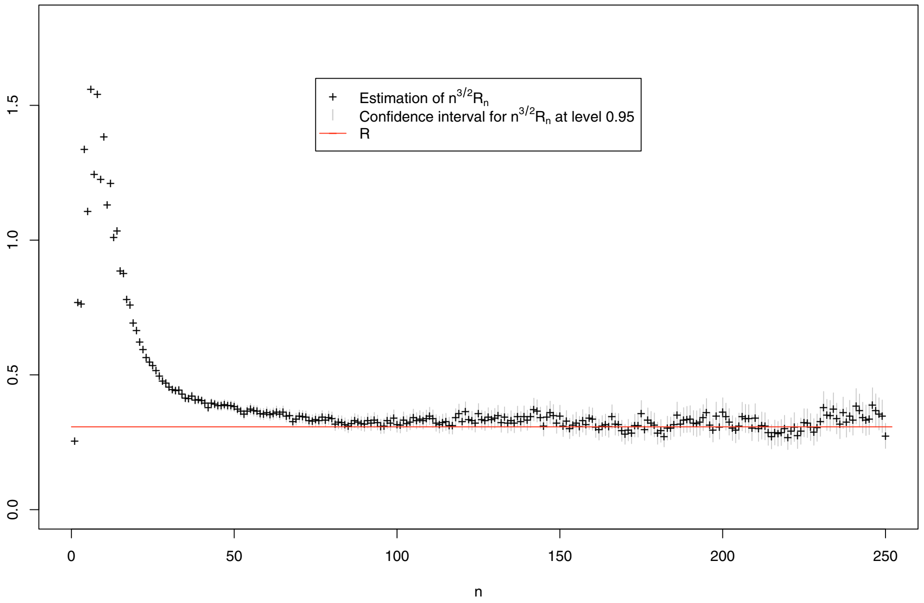

FIGURE 2. Here $F = (\mathbf{1}_0 - \mathbf{1}_1, \mathbf{1}_2 - \mathbf{1}_3, \mathbf{1}_4 - \mathbf{1}_5)$ and $X_{n+1} = X_n + U_n$ modulo 7 for an i.i.d. sequence $\{U_n\}_n$ with distribution $\frac{7}{10}\delta_0 + \frac{3}{10}\delta_1$, $t = (1, 3, 1)$, and $R_n = \mathbb{E}(\mathrm{e}^{-t \cdot S_n} \mathbf{1}\{S_n \geq 0\})$. Then $n^{3/2} R_n$ converges to $R \approx .3072178$.

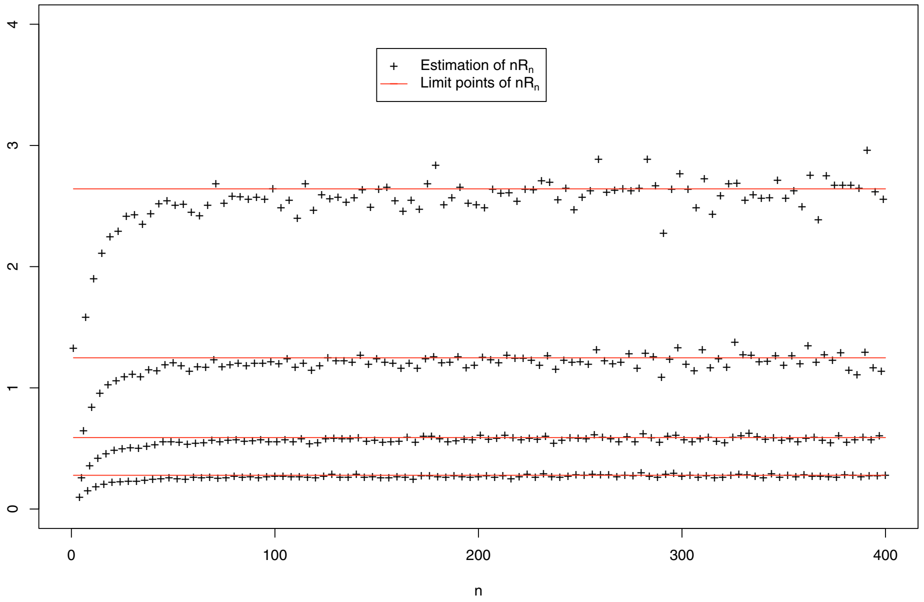

FIGURE 3. Here $F = (\mathbf{1}_0, \mathbf{1}_1)$ and $X_{n+1} = X_n + U_n$ modulo 4 for an i.i.d. sequence $\{U_n\}_n$ with distribution $\frac{4}{5}\delta_0 + \frac{1}{10}\delta_1 + \frac{1}{10}\delta_2$, $t = (1, 2)$, $v = (\frac{1}{4}, \frac{1}{4})$, and $R_n = \mathbb{E}(\mathrm{e}^{-t \cdot (S_n - nv)} \mathbf{1}\{S_n \geq \lfloor nv \rfloor\})$. Then $u(n) = n\, R_n$ exhibits periodic oscillations. Indeed, $u(4n + k)$ converges to $R\,\mathrm{e}^{\beta_k}$ when $n \to \infty$, with $R \approx .2784627$, $\beta_0 = 0$, $\beta_1 = \frac{3}{4}$, $\beta_2 = \frac{3}{2}$ and $\beta_3 = \frac{9}{4}$.

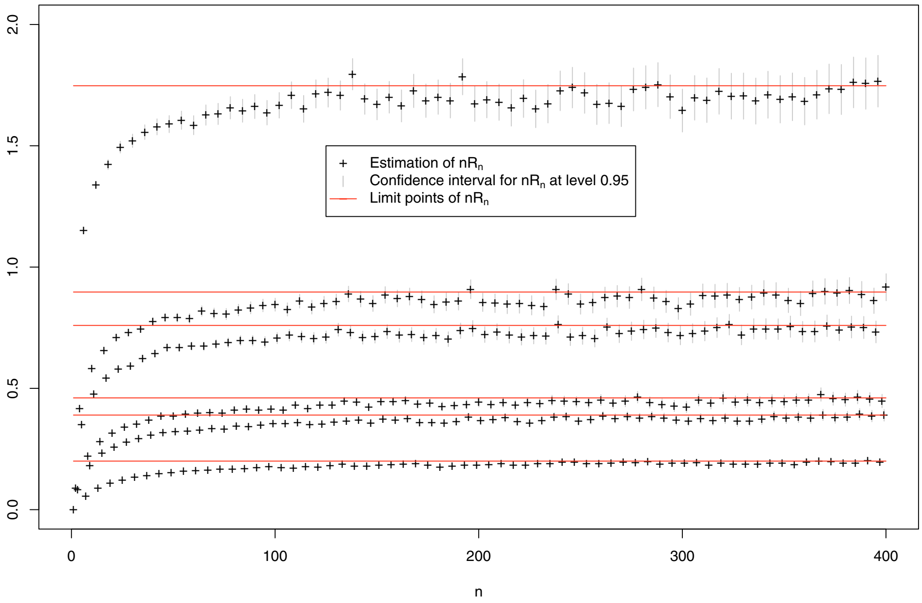

FIGURE 4. Here $F = (\mathbf{1}_0, \mathbf{1}_1)$ and $\{X_n\}_n$ is i.i.d. with distribution $\frac{1}{2}\delta_0 + \frac{1}{6}(\delta_1 + \delta_2 + \delta_3)$, $t = (1, 2)$, $v = (\frac{1}{2}, \frac{1}{6})$, and $R_n = \mathbb{E}(\mathrm{e}^{-t\cdot(S_n - nv)}\mathbf{1}\{S_n \geq nv\})$. Then $u(n) = n\,R_n$ exhibits periodic oscillations. Indeed, $u(6n + k)$ converges to $R\,\mathrm{e}^{\beta_k}$ when $n \to \infty$, with $R \approx 1.747124$, $\beta_0 = 0$, $\beta_1 = -\frac{13}{6}$, $\beta_2 = -\frac{4}{3}$, $\beta_3 = -\frac{3}{2}$, $\beta_4 = -\frac{2}{3}$, and $\beta_5 = -\frac{5}{6}$.

3.2. Genomic sequences

Assume we want to study the frequency of the word TCCAA in the genome of *Escherichia coli*. The random model is Markovian of order 2. The parameters of this model are estimated by maximum likelihood on the sequencing of the commensal K-12 strain, whose length is $n \approx 4.6 \times 10^6$. The observed frequency of TCCAA is approximately 2.7631×10^{-4}, which is much smaller than the one predicted by the random model, namely 9.053×10^{-4}. Let us denote $p(\texttt{TCCAA})$ the p-value associated with this word. This p-value is

$$p(\texttt{TCCAA}) = \mathbb{P}(S_n \leq nv),$$

where S_n denotes the counting of TCCAA in the random sequence. Assume that, for all integer k, X_k is the variable that stands for the word of length 5 beginning at position k in the random sequence. (Remark that X_k and X_{k+1} overlap.) Since the model is Markovian of order 2, $\{X_k\}_{k\geq 1}$ is a simple Markov chain, whose state space is the set of words of length 5. Then, $S_n = \sum_{k=1}^{n} \mathbf{1}_{\texttt{TCCAA}}(X_k)$ is clearly a Markov-additive process.

Up to a change of sign, Theorem 2.2 yields

$$\log p(\texttt{TCCAA}) = -n\Lambda^\star(v) + t_v \cdot \big(\lfloor nv \rfloor - nv\big) - \frac{1}{2}\log n + \log \vartheta^0(0) + O(n^{-1/2}),$$

where v is the observed frequency of TCCAA.

Let us denote

$$\chi_n = \sum_{k=1}^{n} \delta_{X_k},$$

where δ_x is the Dirac measure at x. The rate function $\Lambda^\star$ of S_n is given by (2.5) and has no simpler expression. However, the empirical measure χ_n of all words of length 5 has an explicit rate function, say $\mathcal{J}$. See, for instance, Dembo and Zeitouni [6] (pp. 78–82), specifically exercise 3.1.20. The counting S_n of $\texttt{TCCAA}$ is equal to $n\chi_n(\texttt{TCCAA})$. Hence, the use of the contraction principle, see Theorem 4.2.1 (p. 126) of Dembo and Zeitouni [6] leads us to a numerical value by minimizing $\mathcal{J}(\nu)$ under the constrain that $\nu(\texttt{TCCAA})$ is equal to the observed frequency, say v:

$$\Lambda^\star(v) = \inf\big\{\mathcal{J}(\nu) : \nu(\texttt{TCCAA}) = v\big\}. \tag{3.1}$$

The parameter $t = t_v$ that appears in our theorems is such that $\Lambda'(t) = v$. Since $\Lambda^\star$ is the convex-conjugate function of Λ, we have $\Lambda^{\star\prime}(v) = t$. And (3.1) can also give a numerical value for t. This yields $\Lambda^\star(v) \approx 3.183 \times 10^{-4}$ and $t \approx 3.285 \times 10^{-4}$ for the p-value of the frequency of $\texttt{TCCAA}$. Once $\Lambda^\star(v)$ and t are known, we can compute the twisted kernel. Moreover, the other coefficients of our approximation are given by the first terms of power expansions and can be numerically computed, as in Section 3.1.

Actually, doing this study on all other necessary words, one sees that $\texttt{TCCAA}$ is the word with the smallest p-value, that is the most exceptional word, among the words of length 5 in the genome of *Escherichi coli*. To find the second most exceptional word, we propose to condition the model by the observed frequency of $\texttt{TCCAA}$. In an equivalent way, one seeks the word w with a minimal $p(w, \texttt{TCCAA})$, where $p(w, \texttt{TCCAA})$ is the p-value of the null hypothesis that the counting of w as well as the counting of $\texttt{TCCAA}$ are correctly predicted by the Markovian model of order 2. This second most exceptional word is $\texttt{GGCCG}$.

4. Preliminaries to the proof of the main results

4.1. A technical lemma

Lemma 4.1. *If (H1) and (H2) hold, the matrix* Γ *is nonsingular.*

Proof. We will prove that, if $v \in \mathbb{R}^d$ is nonzero, then $v \cdot \Gamma v > 0$. We must extend the notation slightly insofar as we will explicitly introduce the dependency on the function F in the Markov-additive process. This simply means that, instead of writing S_n and Γ, we will use $S_n(F)$ and $\Gamma(F)$ for the remainder of the proof.

Easy algebra shows that $v{\cdot}\Gamma v = \lim n^{-1} \operatorname{Var}(S_n(v{\cdot}F)) = \Gamma(v{\cdot}F)$. Fix j so that $v_j \neq 0$. Apply assumption (H2) with $\vec{e}$ being the vector with one single one in the j position. We get a positive integer n, states a, b and a point $\sigma \in \mathbb{Z}^d$ such that $\mathbb{P}_a(X_n = b, S_n(v \cdot F) = v \cdot \sigma)$ and $\mathbb{P}_a(X_n = b, S_n(v \cdot F) = v \cdot \sigma + v_j)$ are both positive. In other words, $S_n(v \cdot F)$ takes at least two different values, depending on the realization of the Markov chain. Thus, $v \cdot F$ is not constant. And, since $v \cdot F$ is a one-dimensional function, the real number $\Gamma(v \cdot F)$, which is the asymptotic variance of a one-dimensional Markov additive process, is positive. □

4.2. Asymptotics on the Fourier transform of S_n

In this section, we fix a function $g : E \to \mathbb{C}$ and we study the asymptotics of $\widehat{g}_a(n, \mathrm{i}t)$ when $n \to \infty$, for a fixed value of t in $\mathbb{R}^d$, where

$$\widehat{g}_a(n, \mathrm{i}t) := \mathbb{E}_a\big(g(X_n)\exp(\mathrm{i}t \cdot S_n)\big)$$

is the Fourier transform of $g_a(n, \sigma) = \mathbb{E}_a\left(g(X_n)\mathbf{1}\{S_n = \sigma\}\right)$. Then the vector $\{\widehat{g}_a(n, \mathrm{i}t)\}_a$ indexed by the states of the Markov chain is equal to $Q(\mathrm{i}t)^n g$, where g is seen as a column vector. The leading term of the asymptotics of $\widehat{g}_a(n, \mathrm{i}t)$ when t is close to the origin is due to Mann [17] (pp. 24–33), and is given in Proposition 4.2 below.

Call λ a dominating eigenvalue of a given matrix if λ is a simple eigenvalue and if the modulus of every other eigenvalue is strictly smaller than $|\lambda|$. Then there exists a neighbourhood U of the origin in $\mathbb{R}^d$ and a function ρ_0 defined on U such that ρ_0 has an entire expansion in U and for every t in U, $\rho_0(t)$ is a dominating eigenvalue of $Q(\mathrm{i}t)$. See Mann [17] for the proof.

We establish the uniform convergence of $\widehat{g}_a(n, \mathrm{i}t)$ in a neighbourhood of the origin. Propositions 4.2 and 4.3 below and every other result in this paper are independent of Mann's result. Let R denote any real number, greater than the second greatest modulus of the eigenvalues of the transition kernel Q. One can and we will assume that $R < 1$.

Proposition 4.2. *There exists a positive constant C_1, a neighbourhood V of the origin in $\mathbb{C}^d$, a holomorphic function $z \mapsto \rho(z)$ on V with values in $\mathbb{C}$ and a holomorphic function $z \mapsto N(z)$ on V with values in the space of projection matrices, such that, for every positive integer n and every z in V,*

$$\|\rho(z)^{-n} Q^n(z) - N(z)\| \leq C_1 \, R^n.$$

Proof of Proposition 4.2. In this proof, $\mathbb{C}^E$ is equipped with the Hermitian product

$$G \cdot H = \sum_{a \in E} \bar{H}_a G_a, \quad \text{if } G = \{G_a\}_a,\ H = \{H_a\}_a \text{ are in } \mathbb{C}^E.$$

First step. We first show that there exists a neighbourhood V of the origin in $\mathbb{C}^d$ and some functions $z \mapsto \rho(z)$, $z \mapsto H(z)$ and $z \mapsto K(z)$, defined on V with values in $\mathbb{C}$, $\mathbb{C}^E$ and $\mathbb{C}^E$ respectively, such that

1. $H(0) = \{1\}_{a \in E}$;
2. for every z in V, $\rho(z)$ is a dominating eigenvalue of $Q(z)$;
3. $H(z)$ is a right eigenvector of $Q(z)$ for the eigenvalue $\rho(z)$; and
4. $K(z)$ is a left eigenvector of $Q(z)$ for the eigenvalue $\rho(z)$.

Indeed, the characteristic polynomial of $Q(z)$ is an analytic perturbation of the characteristic polynomial of $Q = Q(0)$, and 1 is a dominating eigenvalue of Q. Hence, at least on a neighbourhood of the origin in $\mathbb{C}^d$, an eigenvalue of $Q(z)$ is dominating, say $\rho(z)$, and $z \mapsto \rho(z)$ is analytic at the origin because $\rho(0)$ is simple.

Likewise, a right eigenvector (respectively, a left eigenvector) of $Q(z)$ for the eigenvalue $\rho(z)$ solves an analytic perturbation of the linear system which gives a right eigenvector (respectively, a left eigenvector) of Q for the simple eigenvalue 1. Since $I = \{1\}_{a \in E}$ is a right eigenvector of Q, one can choose $H(0) = I$.

Second step. Next, we show that, if z is in some neighbourhood of the origin, $Q(z)$ is conjugate to some matrix, that depends analytically on z and is diagonal by block. Let $K(-z)^\perp$ denote the orthogonal space of the vector $K(-z)$ in $\mathbb{C}^E$. Since $K(-z)$ is a left eigenvector of $Q(-z) = \overline{Q(z)}$, one obtains for every vector G in $K(-z)^\perp$,

$$K(-z) \cdot Q(z) G = K(-z) \overline{Q(z)} \cdot G = \rho(-z) \, K(-z) \cdot G = 0,$$

i.e., $K(-z)^\perp$ is mapped into itself by $Q(z)$.

By Perron-Frobenius theorem, the coordinates of $H(0)$ and $K(0)$ are positive, hence the Hermitian product of $K(0)$ and $H(0)$ is positive. By continuity, the modulus of the Hermitian product of $K(-z)$ and $H(z)$ is positive for all z in a neighbourhood of the origin. Hence $H(z)$ is not in $K(-z)^\perp$ and the space $\mathbb{C}^E$ decomposes into a pair of supplementary subspaces as

$$\mathbb{C}^E = K(-z)^\perp \oplus \mathbb{C} H(z).$$

Moreover, $Q(z)$ maps each of those subspaces into itself.

Next, we choose a basis $B(z)$ of $K(-z)^\perp$ which depends analytically on z on a neighbourhood of the origin. Then $B(z) \cup \{H(z)\}$ is a basis of $\mathbb{C}^E$ and the transfer matrix $T(z)$ from the canonical basis of $\mathbb{C}^E$ to $B(z) \cup \{H(z)\}$ depends analytically on z on a neighbourhood of the origin. Furthermore, the matrix of the endomorphism $Q(z)$ in the basis $B(z) \cup \{H(z)\}$ is diagonal by blocks because $\mathbb{C}\, H(z)$ and $K(-z)^\perp$ are stable by $Q(z)$.

Thus, there exists holomorphic applications $z \mapsto T(z)$ and $z \mapsto M(z)$ from V to the spaces of square matrices of dimension d and $(d-1)$ respectively such that, for every z in V, $T(z)$ is nonsingular and

$$\rho(z)^{-1} Q(z) = T(z) \begin{pmatrix} 1 & 0 \\ 0 & M(z) \end{pmatrix} T(z)^{-1}.$$

Conclusion. Let r_M denote the spectral radius of $M(0)$. Since R is greater than r_M, for any r in (r_M, R), Householder's theorem, see Serre [32] (p. 66), provides the existence of a norm $\|\cdot\|_\dagger$ on $\mathbb{C}^E$ such that the norm of $M(0)$ with respect to the associated norm on matrices, namely

$$\|M(0)\|_\dagger = \sup\left\{\left\|\begin{pmatrix}0 & 0\\ 0 & M(0)\end{pmatrix} G\right\|_\dagger : G \in \mathbb{C}^E, \|G\|_\dagger = 1\right\},$$

is at most r. By continuity of $z \mapsto M(z)$, the norm $\|M(z)\|_\dagger$ is at most $R > r$ for every z in a neighbourhood of the origin. By continuity of $z \mapsto T(z)$ and $z \mapsto T(z)^{-1}$, the norms of $T(z)$ and $T(z)^{-1}$ are uniformly bounded for z in a neighbourhood of the origin, say by C_3. Introduce the matrix

$$N(z) = T(z)\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix} T(z)^{-1}.$$

Then the norm $\|\rho(z)^{-n}\, Q^n(z) - N(z)\|_\dagger$ is at most $C_3^2\, R^n$. And, since $\mathbb{C}^E$ is a finite dimensional vector space, there exists C_4 such that $\|\cdot\| \leq C_4\|\cdot\|_\dagger$. Choosing $C_1 = C_4\sqrt{C_3}$ concludes the proof of Proposition 4.2. □

The results above describe the behaviour of the characteristic function of S_n near the origin. We also need to control the characteristic functions of S_n when $n \to +\infty$, outside a neighbourhood of 0. Hypothesis (H2) is crucial in the proof of this proposition.

Proposition 4.3.

1. *If λ is an eigenvalue of $Q(\mathrm{i}t)$ where t belongs to $\mathbb{T} := [-\pi; +\pi]^d$ and $t \neq 0$, then $|\lambda| < 1$;*
2. *For every positive real number ε, there exists positive constants C_2 and A, with $A < 1$, such that, for every t in $\mathbb{T}$ such that $\|t\| \geq \varepsilon$,*

$$|\mathbb{E}_a(\mathrm{e}^{\mathrm{i}t\cdot S_n})| \leq C_2 A^n.$$

The proof of Proposition 4.3 uses the following, classical lemma.

Lemma 4.4. *Let p_j, x_j and y_j, $j = 1, \ldots, K$, be real numbers such that, $0 < p_j \leq 1$, $0 \leq x_j \leq 1$ for all j and $\sum_j p_j = 1$. The equality*

$$\left|\sum_{j=1}^{K} p_j x_j \exp(\mathrm{i}y_j)\right| = 1$$

implies that, first, all x_j are equal to 1 and, second, $y_1 = y_2 = \ldots = y_K$ modulo 2π.

Proof of Proposition 4.3. 1. The modulus of an eigenvalue of $Q(\mathrm{i}t)$ is at most 1. Indeed, for every state a,

$$\sum_b |Q(\mathrm{i}t)(a,b)| = \sum_b Q(a,b) = 1.$$

We assume that a complex number of modulus 1, say $\exp(\mathrm{i}s)$, $s \in \mathbb{R}$, is an eigenvalue of $Q(\mathrm{i}t)$, and we want to show that $t = 0$. Let Y denote a right eigenvector of $Q(\mathrm{i}t)$ for the eigenvalue $\mathrm{e}^{\mathrm{i}s}$. Fix to 1 the maximum of $|Y_a|$ over the states a. We will first prove that $|Y_b| = 1$ for all state b, and then that $t = 0$.

Choose a state a_0 such that $|Y_{a_0}| = 1$. Since we assume (H1), the transition kernel is irreducible. Thus, we can choose n_0 such that $Q^{n_0}(a_0, b) > 0$ for all state b. Then

$$1 = \left|\mathrm{e}^{\mathrm{i}n_0 s}\, Y_{a_0}\right| = \left|\mathrm{e}^{\mathrm{i}n_0 t\cdot F(a_0)} \sum_b Q^{n_0}(a_0,b)\, Y_b\right| = \left|\sum_b Q^{n_0}(a_0,b)\, Y_b\right|.$$

Since the sum over the states b of $Q^{n_0}(a_0, b)$ is 1, every $Q^{n_0}(a_0, b) > 0$ and every $|Y_b| \leq 1$, Lemma 4.4 shows that $|Y_b| = 1$ for every state b. Therefore every Y_a can be written $Y_a = \exp(\mathrm{i}\,y(a))$ for some $y(a)$ in $\mathbb{R}$.

For every nonnegative integer n, Y is a right eigenvector of $Q(\mathrm{i}t)^n$ for the eigenvalue $\exp(\mathrm{i}ns)$, that is,

$$\mathbb{E}_a \exp\big(\mathrm{i}t\cdot S_n + \mathrm{i}y(X_n)\big) = \exp\big(\mathrm{i}ns + \mathrm{i}y(a)\big).$$

Lemma 4.4 implies that, $\mathbb{P}_a$ almost surely,

$$t\cdot S_n + y(X_n) = ns + y(a) \quad \text{modulo } 2\pi.$$

By hypothesis (H2), for every $\vec{e}$ in the canonical basis of $\mathbb{Z}^d$, there exists a, b, n and x, such that the events $\{S_n = x,\, X_n = b\}$ and $\{S_n = x+\vec{e},\, X_n = b\}$ both have positive $\mathbb{P}_a$ probabilities. Hence

$$t\cdot x + y(b) = ns + y(a) = t\cdot(x+\vec{e}) + y(b) \quad \text{modulo } 2\pi,$$

that is, $t\cdot\vec{e} = 0$ modulo 2π, for every $\vec{e}$. Finally, since $t\in\mathbb{T}$, $t=0$.

2. Let $r(t)$ denote the spectral radius of $Q(\mathrm{i}t)$. Since $Q(\mathrm{i}t)$ depends continuously on t, $r(t)$ depends continuously on t. The first item of Proposition 4.3 implies that $r(t)<1$ for every t in $\mathbb{T}$ such that $\|t\|\geq\varepsilon$. Hence the supremum of the function r on this compact set is at most $A'<1$. Recall that, for every matrix M of spectral radius r_M, the norm of M^n is $(r_M)^{n+o(n)}$ when $n\to\infty$. Thus, for every $A>A'$, the sequence $A^{-n}|\mathbb{E}_a(\mathrm{e}^{\mathrm{i}t\cdot S_n})|$ is bounded. This concludes the proof of Proposition 4.3. □

5. Proof of Theorem 2.1

We are ready to prove an asymptotic expansion of the distribution of S_n. We write the two terms involved in the left hand side of Theorem 2.1 as Fourier transforms. On the one hand,

$$\mathbb{E}_a\big[g(X_n)\mathbf{1}\{S_n=\sigma\}\big] = (2\pi)^{-d}\int_{\mathbb{T}} \widehat{g}_a(n,\mathrm{i}t)\,\mathrm{e}^{-\mathrm{i}t\cdot\sigma}\,\mathrm{d}t,$$

where we recall that $\mathbb{T} := [-\pi;\pi]^d$. On the other hand, for every nonnegative integer k,

$$\begin{aligned} n^{-d/2}\,\psi_a^k(y) &= (2\pi)^{-d}\int_{\mathbb{R}^d} P_a^{(k)}(\mathrm{i}t)\,\exp\left(-\frac{1}{2}t\cdot\Gamma t - \mathrm{i}t\cdot y\right)\frac{\mathrm{d}t}{n^{d/2}} \\ &= (2\pi)^{-d}\int_{\mathbb{R}^d} P_a^{(k)}(\mathrm{i}t\sqrt{n})\exp\left(-\frac{1}{2}n\,t\cdot\Gamma t - \mathrm{i}t\cdot y\sqrt{n}\right)\mathrm{d}t. \end{aligned}$$

(Note that $P_a^{(k)}$ is *not* homogeneous of degree k.) Using this for $y := (\sigma - n\,m)/\sqrt{n}$, one sees that $(2\pi)^d$ times the left-hand side of the theorem is bounded by the sum of three terms $I_1(n)$, $I_2(n)$, and $I_3(n)$, defined as

$$\begin{aligned} I_1(n) &:= \int_{\|t\|\leq\varepsilon}\left|\widehat{g}_a(n,\mathrm{i}t) - \sum_{\ell=0}^{k} n^{-\ell/2}\,P_a^{(\ell)}(\mathrm{i}t\sqrt{n})\,\mathrm{e}^{-\frac{1}{2}n\,t\cdot\Gamma t + \mathrm{i}n\,t\cdot m}\right|\mathrm{d}t, \\ I_2(n) &:= \int_{\|t\|\geq\varepsilon,\, t\in\mathbb{T}} |\widehat{g}_a(n,\mathrm{i}t)|\,\mathrm{d}t, \\ I_3(n) &:= \int_{\|t\|\geq\varepsilon}\left|\sum_{\ell=0}^{k} n^{-\ell/2}\,P_a^{(\ell)}(\mathrm{i}t\sqrt{n})\,\mathrm{e}^{-\frac{1}{2}n\,t\cdot\Gamma t}\right|\mathrm{d}t, \end{aligned}$$

where we cancelled unnecessary factors such as $\mathrm{e}^{-\mathrm{i}t\cdot\sigma}$ in $I_1(n)$ and $I_2(n)$ and $\mathrm{e}^{-\mathrm{i}t\cdot y\sqrt{n}}$ in $I_3(n)$.

In the next steps of the proof, we first show that $I_2(n)$ and $I_3(n)$ are exponentially small, then we use the approximation of $\widehat{g}(n,\mathrm{i}t)$ given by Proposition 4.2 to replace $\widehat{g}(n,\mathrm{i}t)$ in $I_1(n)$ by $G(\mathrm{i}t)\,\mathrm{e}^{n\,\Lambda(\mathrm{i}t)}$, and finally we show that $I_1(n)$ is bounded by a power of n.

5.1. Bounding $I_2(n)$ and $I_3(n)$

Proposition 4.3 yields $|\widehat{g}_a(n, \mathrm{i}t)| \leq C_2\, A^n$ with $A < 1$. Hence $I_2(n)$ is exponentially small, more precisely

$$I_2(n) \leq C_2\, A^n\, \mathrm{Vol}(\mathbb{T}) = C_2\, A^n\, (2\pi)^d.$$

As regards $I_3(n)$, $P_a^{(\ell)}$ is a polynomial of degree at most (3ℓ), hence, for every $\|z\| \geq \varepsilon$, $|P_a^{(\ell)}(z)| \leq c_{\ell,a,\varepsilon}\, \|z\|^{3\ell}$. Using this for $z := \mathrm{i}t\sqrt{n}$, one gets

$$I_3(n) \leq \sum_{\ell=0}^{k} c_{\ell,a,\varepsilon} n^{\ell} \int_{\|t\|\geq\varepsilon} \mathrm{e}^{-\frac{1}{2}n\, t\cdot\Gamma t}\, \|t\|^{3\ell}\, \mathrm{d}t.$$

Since Γ is positive definite, there exists a positive γ such that $t\Gamma\, t \geq \gamma\, \|t\|^2$, hence for every positive integer n,

$$\int_{\|t\|\geq\varepsilon} \mathrm{e}^{-\frac{1}{2}n\, t\cdot\Gamma t}\, \|t\|^{3\ell}\, \mathrm{d}t \leq \mathrm{e}^{-\frac{1}{4}n\gamma\varepsilon^2} \int \mathrm{e}^{-\frac{1}{4}t\cdot\Gamma t}\, \|t\|^{3\ell}\, \mathrm{d}t.$$

The last integral above converges, hence $I_3(n)$ is exponentially small.

5.2. Approximating $\widehat{g}_a(n, \mathrm{i}t)$ in $I_1(n)$

Proposition 4.2 gives an approximation of $\widehat{g}_a(n, z)$ for all z in a neighbourhood V of the origin. If we choose ε small enough so that $\{\mathrm{i}t, \|t\| \leq \varepsilon\} \subset V$, then for every $\|t\| \leq \varepsilon$, one has

$$\left|\widehat{g}_a(n, \mathrm{i}t) - G_a(\mathrm{i}t)\, \mathrm{e}^{n\,\Lambda(\mathrm{i}t)}\right| \leq C_1\, R^n \left|\mathrm{e}^{n\,\Lambda(\mathrm{i}t)}\right|.$$

Hence $I_1(n) \leq I_4(n) + C_1\, R^n\, I_5(n)$, with

$$I_4(n) := \int_{\|t\|\leq\varepsilon} \left| G_a(\mathrm{i}t)\, \mathrm{e}^{n\,\Lambda(\mathrm{i}t)} - \sum_{\ell=0}^{k} n^{-\ell/2}\, P_a^{(\ell)}(\mathrm{i}t\,\sqrt{n})\, \mathrm{e}^{-\frac{1}{2}n\, t\cdot\Gamma t + \mathrm{i}n t\cdot m} \right| \mathrm{d}t$$

$$I_5(n) := \int_{\|t\|\leq\varepsilon} \left|\mathrm{e}^{n\,\Lambda(\mathrm{i}t)}\right|\, \mathrm{d}t.$$

Since $\Lambda^{(0)}(z) = 0$, $\Lambda^{(1)}(z) = m \cdot z$ and $\Lambda^{(2)}(z) = \frac{1}{2} z \cdot \Gamma\, z$, one gets

$$n\,\Lambda(\mathrm{i}t) = -\frac{1}{2} n\, t \cdot \Gamma t + \mathrm{i}n\, t \cdot m + n\, L(\mathrm{i}t),$$

where we recall that $L := \Lambda - \Lambda^{(1)} - \Lambda^{(2)}$. One may cancel the terms $\mathrm{e}^{\mathrm{i}n\, t\cdot m}$ in $I_4(n)$ and $I_5(n)$, and factorise the terms $\mathrm{e}^{-\frac{1}{2}n\, t\cdot\Gamma t}$ in $I_4(n)$. This yields

$$I_4(n) = \int_{\|t\|\leq\varepsilon} \left| G_a(\mathrm{i}t)\, \mathrm{e}^{n\, L(\mathrm{i}t)} - \sum_{\ell=0}^{k} n^{-\ell/2}\, P_a^{(\ell)}(\mathrm{i}t\sqrt{n}) \right| \mathrm{e}^{-\frac{1}{2}n\, t\cdot\Gamma t}\, \mathrm{d}t$$

$$I_5(n) = \int_{\|t\|\leq\varepsilon} \left|\mathrm{e}^{n\, L(\mathrm{i}t)}\right|\, \mathrm{e}^{-\frac{1}{2}n\, t\cdot\Gamma t}\, \mathrm{d}t.$$

5.3. Bounding $I_1(n)$

We use the following elementary lemma of complex analysis, see Knopp [13] (p. 77) for instance.

Lemma 5.1. *Assume that the function h is analytic at the origin in $\mathbb{C}^d$, that the Taylor expansion of h around the origin reads*

$$h(z) = \sum_{k\geq 0} h^{(k)}(z),$$

where each $h^{(k)}$ is zero or a homogeneous polynomial of degree k, and that this expansion converges absolutely for $\|z\| \leq \alpha$, for some positive constant α. Then, there exists a positive constant β such that the following properties hold.

(i) For every nonnegative integer k and every z in $\mathbb{C}^d$ such that $\|z\| \leq \alpha$,

$$\left|h^{(k)}(z)\right| \leq \beta\,(\alpha^{-1}\,\|z\|)^k.$$

(ii) For every nonnegative integer k and every z in $\mathbb{C}^d$ such that $\|z\| \leq \frac{1}{2}\alpha$,

$$\left|h(z) - \sum_{\ell=0}^{k} h^{(\ell)}(z)\right| \leq 2\beta(\alpha^{-1}\|z\|)^{k+1}.$$

We apply part *(ii)* of Lemma 5.1 to $h = \Lambda$ and $k = 2$. If ε is small enough, this yields $|L(\mathrm{i}t)| \leq C_4\,\|t\|^3$ for every t such that $\|t\| \leq \varepsilon$, for a finite C_4. Since Γ is definite positive, there exists a positive γ such that $t\cdot\Gamma t \geq \gamma\,\|t\|^2$ for every t. Hence, if $4C_4\varepsilon \leq \gamma$ and $\|t\| \leq \varepsilon$, then $|L(\mathrm{i}t)| \leq \frac{1}{4}t\cdot\Gamma t$. This implies that

$$I_5(n) \leq \int_{\|t\|\leq\varepsilon} \mathrm{e}^{-\frac{1}{4}n\gamma\|t\|^2}\,\mathrm{d}t = O(n^{-d/2}).$$

As regards $I_4(n)$, equations (2.3) and (2.4) read

$$G_a(\mathrm{i}t)\,\mathrm{e}^{n\,L(\mathrm{i}t)} = \sum_{\ell\geq 0} n^{-\ell/2}\,P_a^{(\ell)}(\mathrm{i}t\,\sqrt{n}).$$

Hence,

$$I_4(n) = \int_{\|t\|\leq\varepsilon} \left|\sum_{\ell\geq k+1} n^{-\ell/2}\,P_a^{(\ell)}(\mathrm{i}t\,\sqrt{n})\right| \mathrm{e}^{-\frac{1}{2}n\,t\cdot\Gamma t}\,\mathrm{d}t.$$

We use Lemma 5.2 below, whose proof is postponed to Section 6, to control each $P_a^{(\ell)}$.

Lemma 5.2. *There exists positive constants α and β such that, for every ℓ and every z such that $\|z\| \leq \frac{1}{2}\alpha$,*

$$|P_a^{(\ell)}(z)| \leq \beta\,(2y)^\ell \sum_{j=0}^{\ell} \frac{(\beta\,y^2)^j}{j!}, \qquad \text{where } y := \alpha^{-1}\|z\|.$$

Applying this to $I_4(n)$, we use the fact that $t\cdot\Gamma t \geq \gamma\|t\|^2$ and the change of variables $s = \alpha^{-1}\|t\|\sqrt{n}$. Considering separately the indexes j in the upper bound of $P^{(\ell)}$ of Lemma 5.2 such that $j \leq k$ and the indexes j such that $j > k$, one gets

$$I_4(n) \leq (2\alpha)^d\omega_{d-1}\,\beta\,n^{-d/2}(I_6(n) + I_7(n)),$$

where ω_{d-1} denotes the surface of the unit sphere in $\mathbb{R}^d$,

$$I_6(n) := \sum_{j=0}^{k} I_8(j,n), \quad I_7(n) := \sum_{j=k+1}^{\infty} I_8(j,n),$$

and, for every nonnegative integer j,

$$I_8(j,n) := \sum_{i \geq \max\{k+1,j\}} \int_0^{\alpha^{-1}\varepsilon\sqrt{n}} \left(\frac{2s}{\sqrt{n}}\right)^i \frac{(\beta\, s^2)^j}{j!}\, s^{d-1}\, \mathrm{e}^{-\frac{1}{2}\gamma\alpha^2 s^2}\, \mathrm{d}s.$$

From now on, we assume that $4\varepsilon \leq \alpha$. Then $4s \leq \sqrt{n}$ uniformly on the integration interval $0 \leq s \leq \alpha^{-1}\varepsilon\sqrt{n}$. Hence, for any I,

$$\sum_{i \geq I} \left(\frac{2s}{\sqrt{n}}\right)^i \leq 2\left(\frac{2s}{\sqrt{n}}\right)^I. \tag{5.1}$$

Using (5.1) with $I = k+1$ to bound the terms $I_8(j,n)$ such that $j \leq k$, one gets

$$I_6(n) \leq 2^{k+2} n^{-(k+1)/2} \sum_{j=0}^{k} \int_0^{+\infty} \frac{(\beta s^2)^j}{j!} s^{k+d}\, \mathrm{e}^{-\frac{1}{2}\alpha^2\gamma s^2}\, \mathrm{d}s.$$

The last sum involves a finite number of finite integrals, hence there exists a finite C_6, independent of n, such that

$$I_6(n) \leq C_6\, n^{-(k+1)/2}.$$

As regards $I_7(n)$, one uses (5.1) with $I = j$ to bound the terms $I_8(j,n)$ such that $j > k$. Furthermore,

$$\sum_{j>k} 2\left(\frac{2s}{\sqrt{n}}\right)^j \frac{(\beta\, s^2)^j}{j!} \leq 2\left(\frac{2s}{\sqrt{n}}\right)^{k+1} (\beta\, s^2)^{k+1}\, \mathrm{e}^{(2s/\sqrt{n})\beta\, s^2}.$$

Since the last exponential above is at most $\mathrm{e}^{2\alpha^{-1}\varepsilon\beta\, s^2}$,

$$I_7(n) \leq 2\left(\frac{2\beta}{\sqrt{n}}\right)^{k+1} \int_0^{+\infty} s^{3k+d+1}\, \mathrm{e}^{-\gamma'\, s^2}\, \mathrm{d}s$$

where $\gamma' := \frac{1}{2}\gamma\alpha^2 - 2\alpha^{-1}\varepsilon\beta$. If ε is small enough, γ' is positive and the last integral above is finite. Hence, $I_7(n) \leq C_7\, n^{-(k+1)/2}$.

Finally, $I_4(n)$, hence $I_1(n)$, hence the left hand side of the inequality in Theorem 2.1, are bounded by multiples of $n^{-(k+d+1)/2}$. This concludes the proof of Theorem 2.1. □

6. Proof of Lemma 5.2

From part *(ii)* of Lemma 5.1, there exists positive constants α and β such that, for every integer k,

$$|G_a^{(k)}(z)| \leq \beta\, y^k, \quad |\Lambda^{(k)}(z)| \leq \beta\, y^k, \quad \text{where } y := \alpha^{-1}\|z\|.$$

In equation (2.3), one can expand the exponential as

$$\mathrm{e}^{L(uz)/u^2} = \sum_{j \geq 0} \left(\frac{L(uz)}{u^3}\right)^j \frac{u^j}{j!}.$$

Furthermore, $L(uz) = \sum_{k\geq 3} \Lambda^{(k)}(uz)$, and $\Lambda^{(k)}$ is a homogeneous polynomial of degree k, thus,

$$\left(\frac{L(uz)}{u^3}\right)^j = \sum_{i_1,\dots,i_j\geq 0} u^{i_1+\cdots+i_j}\,\lambda(i_1,\dots,i_j)(z),$$

where

$$\lambda(i_1,\dots,i_j)(z) := \Lambda^{(i_1+3)}(z)\cdots\Lambda^{(i_j+3)}(z).$$

Evaluating the u^ℓ term in the expansion of equation (2.4), one gets

$$P_a^{(\ell)}(z) = \sum_* G_a^{(i)}(z)\,\lambda(i_1,\dots,i_j)(z)\frac{1}{j!},$$

where the summation $*$ runs over every non-negative integers i, j, i_1, ..., i_j, such that

$$i + j + i_1 + \cdots + i_j = \ell.$$

Since $|\Lambda^{(k)}(z)| \leq \beta\, y^k$ for every positive integer k, one gets

$$|\lambda(i_1,\dots,i_j)(z)| \leq \beta^j\, y^{i_1+\cdots+i_j+3j} = \beta^j\, y^{\ell-i+2j}.$$

Thus,

$$|P_a^{(\ell)}(z)| \leq \sum_{j\leq\ell} \beta^{j+1}\, y^{\ell+2j}\,\frac{n_\ell(j)}{j!},$$

where $n_\ell(j)$ denotes the number of $j$uples $(i_1,\dots,i_j)$ such that there exists a positive integer i such that $i + j + i_1 + \cdots + i_j = \ell$, that is, such that

$$j + i_1 + \cdots + i_j \leq \ell.$$

Lemma 6.1 below shows that $n_\ell(j) \leq 2^\ell$. This concludes the proof of Lemma 5.2.

Lemma 6.1. *For every nonnegative integers $j \leq \ell$, $n_\ell(j) = \binom{\ell}{j}$.*

Proof of Lemma 6.1. Say that $n_\ell(j)$ enumerates a set $N_\ell(j)$ of $j$uples. Assume that j is positive. Consider the map which associates to any $j$uple $(i_1,\dots,i_j)$ such that $i_1 = 0$ the $(j-1)$uple $(i_2,\dots,i_j)$, and to any $j$uple such that $i_1 \geq 1$ the $j$uple $(i_1-1,\dots,i_j)$. The images of the $j$uples in $N_\ell(j)$ of the first kind span $N_{\ell-1}(j-1)$, while the images of the $j$uples of the second kind span $N_{\ell-1}(j)$. Since the map is injective, this shows that $n_\ell(j) = n_{\ell-1}(j-1) + n_{\ell-1}(j)$. Since $n_0(j) = 1$ for every nonnegative integer j, the proof of Lemma 6.1 is complete. □

Acknowledgements. We would like to thank D. Piau for carefully reading many versions of this paper and for suggesting essential contributions. We also would like to thank the two referees for their very detailed reports which have helped to improve this paper.

References

[1] C. Andriani and P. Baldi, Sharp estimates of deviations of the sample mean in many dimensions. *Ann. Inst. H. Poincaré Probab. Statist.* **33** (1997) 371–385.

[2] R.R. Bahadur and R.R. Rao, On deviations of the sample mean. *Ann. Math. Statist.* **31** (1960) 1015–1027.

[3] P. Barbe and M. Broniatowski, Large-deviation probability and the local dimension of sets, in *Proceedings of the 19th Seminar on Stability Problems for Stochastic Models, Vologda, 1998, Part I.* (2000), Vol. 99, pp. 1225–1233.

[4] N.R. Chaganty and J. Sethuraman, Strong large deviation and local limit theorems. *Ann. Probab.* **21** (1993) 1671–1690.

[5] S. Datta and W.P. McCormick, On the first-order Edgeworth expansion for a Markov chain. *J. Multivariate Anal.* **44** (1993) 345–359.

[6] A. Dembo and O. Zeitouni, *Large deviations techniques and applications.* Volume 38 of *Appl. Math.* (New York). Second edition. Springer-Verlag, New York (1998).

[7] P. Flajolet, W. Szpankowski and B. Vallée, Hidden word statistics. *J. ACM* **53** (2006) 147–183 (electronic).

[8] M. Iltis, Sharp asymptotics of large deviations in $\mathbf{R}^d$. *J. Theoret. Probab.* **8** (1995) 501–522.

[9] M. Iltis, Sharp asymptotics of large deviations for general state-space Markov-additive chains in $\mathbf{R}^d$. *Statist. Probab. Lett.* **47** (2000) 365–380.

[10] I. Iscoe, P. Ney and E. Nummelin, Large deviations of uniformly recurrent Markov additive processes. *Adv. Appl. Math.* **6** (1985) 373–412.

[11] J.L. Jensen, *Saddlepoint approximations.* The Clarendon Press Oxford University Press, New York (1995).

[12] V. Kargin, A large deviation inequality for vector functions on finite reversible Markov chains. *Ann. Appl. Probab.* **17** (2007) 1202–1221.

[13] K. Knopp, *Theory of Functions, Part I. Elements of the General Theory of Analytic Functions.* Dover Publications, New York (1945).

[14] I. Kontoyiannis and S.P. Meyn, Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.* **13** (2003) 304–362.

[15] C.A. León and F. Perron, Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.* **14** (2004) 958–970.

[16] M.E. Lladser, M.D. Betterton and R. Knight, Multiple pattern matching: a Markov chain approach. *J. Math. Biol.* **56** (2008) 51–92.

[17] B. Mann, *Berry-Esseen Central Limit Theorems For Markov Chains.* Ph.D. thesis, Harvard University, 1996.

[18] H.D. Miller, A convexivity property in the theory of random variables defined on a finite Markov chain. *Ann. Math. Statist.* **32** (1961) 1260–1270.

[19] P. Ney, Dominating points and the asymptotics of large deviations for random walk on $\mathbf{R}^d$. *Ann. Probab.* **11** (1983) 158–167.

[20] P. Ney and E. Nummelin, Markov additive processes, Part I. Eigenvalue properties and limit theorems. *Ann. Probab.* **15** (1987) 561–592.

[21] P. Nicodème, B. Salvy and P. Flajolet, Motif statistics. In *Algorithms – ESA '99, Prague. Lect. Notes Comput. Sci.* **1643**. Springer, Berlin (1999), pp 194–211.

[22] G. Nuel, Numerical solutins for Patterns Statistics on Markov chains. *Stat. Appl. Genet. Mol. Biol.* **5** (2006).

[23] G. Nuel, Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *J. Appl. Probab.* **45** (2008) 226–243.

[24] R Development Core Team, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria (2003). ISBN 3-900051-00-3.

[25] M. Régnier, A unified approach to word occurrence probabilities. *Discrete Appl. Math.* **104** (2000) 259–280, Combinatorial molecular biology.

[26] M. Régnier and A. Denise, Rare events and conditional events on random strings. *Discrete Math. Theor. Comput. Sci.* **6** (2004) 191–213 (electronic).

[27] M. Régnier and W. Szpankowski, On pattern frequency occurrences in a Markovian sequence. *Algorithmica* **22** (1998) 631–649.

[28] G. Reinert, S. Schbath and M.S. Waterman, Applied Combinatorics on Words. In *Encyclopedia of Mathematics and its Applications*, Vol. 105, chap. Statistics on Words with Applications to Biological Sequences. Cambridge University Press (2005).

[29] S. Robin and J.-J. Daudin, Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probab.* **36** (1999) 179–193.

[30] E. Roquain and S. Schbath, Improved compound Poisson approximation for the number of occurrences of any rare word family in a stationary Markov chain. *Adv. Appl. Probab.* **39** (2007) 128–140.

[31] S. Schbath, Compound Poisson approximation of word counts in DNA sequences. *ESAIM: PS* **1** (1997) 1–16.

[32] D. Serre, Matrices, volume 216 of *Graduate Texts Math.*. Springer-Verlag, New York (2002). Theory and applications, translated from the 2001 French original.

[33] V.T. Stefanov, S. Robin and S. Schbath, Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Appl. Math.* **155** (2007) 868–880.