Improvements on the distribution of maximal segmental scores in a Markovian sequence

Simona Grusea, Sabine Mercier

(7 March 2018)

IMPROVEMENTS ON THE DISTRIBUTION OF MAXIMAL SEGMENTAL SCORES IN A MARKOVIAN SEQUENCE

S. GRUSEA,* *Institut de Mathématiques de Toulouse, Université de Toulouse, INSA de Toulouse*

S. MERCIER,** *Institut de Mathématiques de Toulouse, Université de Toulouse, Jean Jaurès*

Abstract

Let $(A_i)_{i\geq 0}$ be a finite state irreducible aperiodic Markov chain and f a lattice score function such that the average score is negative and positive scores are possible. Define $S_0 := 0$ and $S_k := \sum_{i=1}^{k} f(A_i)$ the successive partial sums, S^+ the maximal non-negative partial sum, Q_1 the maximal segmental score of the first non-negative excursion and $M_n := \max_{0\leq k\leq \ell\leq n}(S_\ell - S_k)$ the *local score* first defined by Karlin and Altschul [8]. We establish recursive formulae for the exact distribution of S^+ and derive new approximations for the distributions of Q_1 and M_n. Computational methods are presented in a simple application case and comparison is performed between these new approximations and the ones proposed in [9] in order to evaluate improvements.

Keywords: local score; Markov theory; limit theorems; maximal segmental score

2010 Mathematics Subject Classification: Primary 60J10; 60F05; 60G70

Secondary 60F10; 60K15

1. Introduction

There is nowadays a huge amount of biological sequences available. The *local score* for one sequence analysis, first defined by Karlin and Altchul in [8] (see Equation (3) below for definition) quantifies the highest level of a certain quantity of interest, e.g.

* Postal address: Institut National des Sciences Appliquées, 135 avenue de Rangueil, 31400, Toulouse, France

** Postal address: Institut de Mathématiques de Toulouse, UMR5219, Université de Toulouse 2 Jean Jaurès, 5 allées Antonio Machado, 31058, Toulouse, Cedex 09, France

hydrophobicity, polarity, etc..., that can be found locally inside a given sequence. This allows for example to detect atypical segments in biological sequences. In order to distinguish significantly interesting segments from the ones that could have appeared by chance alone, it is necessary to evaluate the p-value of a given local score. Different results have already been established using different probabilistic models for sequences: independent and identically distributed variables model (i.i.d.) [2, 8, 9, 12], Markovian models [9, 7] or Hidden Markov Models [4]. In this article we will focus on the Markovian model.

An exact method was proposed by Hassenforder and Mercier [7] to calculate the distribution of the local score for a Markovian sequence, but this result is computationally time consuming for long sequences ($> 10^3$). Karlin and Dembo [9] established the limit distribution of the local score for a Markovian sequence and a random scoring scheme depending on the pairs of consecutive states in the sequence. They proved that the distribution of the local score is asymptotically a Gumble distribution, as in the i.i.d. case. In spite of its importance, their result in the Markovian case is unforfunately very little cited or used in the literature. A possible explanation could be the fact that the random scoring scheme defined in [9] is more general than the ones classically used in practical approaches. In [6] and [5], the authors verify by simulations that the local score in a certain dependence model follows a Gumble distribution, and use simulations to estimate the two parameters of this distribution.

In this article we study the Markovian case for a more classical scoring scheme. We propose a new approximation for the distribution of the local score of a Markovian sequence. We compare it to the one derived from the result of Karlin and Dembo [9] and illustrate the obtained improvement in a simple application case.

Mathematical framework Let $(A_i)_{i\geq 0}$ be an irreducible and aperiodic Markov chain taking its values in a finite set $\mathcal{A}$ containing r states denoted α, β, ... for simplicity. Let $\mathbf{P} = (p_{\alpha\beta})_{\alpha,\beta}$ be its transition probability matrix and $(\pi_\alpha)_\alpha$ its stationary frequency vector. In order to simplify the presentation, we suppose that $\mathbf{P}$ is positive ($\forall \alpha, \beta,\ p_{\alpha\beta} > 0$). We also suppose that the Markov chain is stationary, i.e. with initial distribution of A_0 given by π. $\mathbb{P}_\alpha$ will stand for the conditional probability given $\{A_0 = \alpha\}$. We consider a lattice score function $f : \mathcal{A} \to d\mathbb{Z}$, with $d \in \mathbb{N}$ being

the lattice step. Note that, since $\mathcal{A}$ is finite, we have a finite number of possible scores. Since the Markov chain $(A_i)_{i\geq 0}$ is supposed to be stationary, the distribution of A_i is π for every $i \geq 0$. We will simply denote $\mathbb{E}[f(A)]$ the average score.

In this article we make the hypothesis that the average score is negative, i.e.

$$\mathbb{E}[f(A)] = \sum_{\alpha} f(\alpha)\pi_\alpha < 0. \tag{1}$$

We will also suppose that for every $\alpha \in \mathcal{A}$ we have

$$\mathbb{P}_\alpha(f(A) > 0) > 0 \text{ and } \mathbb{P}_\alpha(f(A) < 0) > 0. \tag{2}$$

Let us introduce some definitions and notation. Let $S_0 := 0$ and denote

$$S_k := \sum_{i=1}^{k} f(A_i),$$

for $k \geq 1$ the successive partial sums. Let S^+ be the *maximal non-negative partial sum*

$$S^+ := \max\{0, S_k : k \geq 0\}.$$

Further, let $\sigma^- := \inf\{k \geqslant 1 : S_k < 0\}$ be the time of the first negative partial sum. Note that σ^- is an a.s.-finite stopping time due to (1).

Let $K_0 := 0$. For $i \geq 1$, we denote $K_i := \inf\{k > K_{i-1} : S_k - S_{K_{i-1}} < 0\}$ the successive decreasing ladder times of $(S_k)_{k\geq 0}$. Note that $K_1 = \sigma^-$.

Let us now consider the subsequence $(A_i)_{0\leq i\leq n}$ for a given length $n \in \mathbb{N} \setminus \{0\}$. Denote $m(n) := \max\{i \geq 0 : K_i \leq n\}$ the random variable corresponding to the number of decreasing ladder times arrived before n. For every $i = 1, \ldots, m(n)$, we call the sequence $(A_j)_{K_{i-1}<j\leq K_i}$ the i-th non-negative excursion.

Note that, due to the negative drift, we have $\mathbb{E}[K_1] < \infty$ (see Lemma 3.6) and $m(n) \to \infty$ *a.s.* when $n \to \infty$. To every non-negative excursions $i = 1, \ldots, m(n)$ we associate a *maximal segmental score* (called also *height*) Q_i defined by

$$Q_i := \max_{K_{i-1}\leq k<K_i} (S_k - S_{K_{i-1}}).$$

First introduced by Karlin and Altschul in [8], the *local score*, denoted M_n, is defined as the maximum segmental score for a sequence of length n:

$$M_n := \max_{0\leqslant k\leqslant \ell\leqslant n} (S_\ell - S_k). \tag{3}$$

Note that $M_n = \max(Q_1, \dots, Q_{m(n)}, Q^*)$, with Q^* being the maximal segmental score of the last incomplete non-negative excursion $(A_j)_{K_{m(n)}<j\leq n}$. Mercier and Daudin [12] give an alternative expression for M_n using the Lindley process $(W_k)_{k\geq 0}$ describing the excursions above zero between the successive stopping times $(K_i)_{i\geq 0}$. With $W_0 := 0$ and $W_{k+1} := \max(W_k + f(A_{k+1}), 0)$, we have $M_n = \max_{0\leqslant k\leqslant n} W_k$.

Remark 1.1. Karlin and Dembo [9] consider a random score function defined on pairs of consecutive states of the Markov chain: they associate to each transition $(A_{i-1}, A_i) = (\alpha, \beta)$ a bounded random score $X_{\alpha\beta}$ whose distribution depends on the pair (α, β). Moreover, they suppose that, for $(A_{i-1}, A_i) = (A_{j-1}, A_j) = (\alpha, \beta)$, the random scores $X_{A_{i-1}A_i}$ and $X_{A_{j-1}A_j}$ are independent and identically distributed as $X_{\alpha\beta}$. The framework of this article corresponds to the case when the score function is deterministic, with $X_{A_{i-1}A_i} = f(A_i)$.

Note also that in our case the hypotheses (1) and (2) assure the so-called cycle positivity, i.e. the existence of some state $\alpha \in \mathcal{A}$ satisfying

$$\mathbb{P}\left(\bigcap_{k=1}^{m-1} \{S_k > 0\} \mid A_0 = A_m = \alpha\right) > 0.$$

In [9], in order to simplify the presentation, the authors strengthen the assumption of cycle positivity by assuming that $\mathbb{P}(X_{\alpha\beta} > 0) > 0$ and $\mathbb{P}(X_{\alpha\beta} < 0) > 0$ for all $\alpha, \beta \in \mathcal{A}$ (see (1.19) of [9]), but precise that the cycle positivity is sufficient for their results to hold. Note that hypotheses (1) and (2) are usually verified in biological applications.

In Section 2 we first introduce few more definitions and notation. Then we present the main results: a recursive result for the exact distribution of the maximal non-negative partial sum S^+ for an infinite sequence in Theorem 2.1; based on the exact distribution of S^+, we further propose new approximations for the distribution of the height of the first non-negative excursion Q_1 in Theorem 2.3 and for the distribution of the local score M_n for a sequence of length n in Theorem 2.4. Section 3 contains the proofs of the results of Section 2 and of some useful lemmas which use techniques of Markov renewal theory and large deviations. In Section 4 we propose a computational method for deriving the quantities appearing in the main results. A simple scoring scheme is developed in Subsection 4.4, for which we compare our approximations to the ones proposed by Karlin and Dembo [9] in the Markovian case.

2. Statement of the main results

2.1. Definitions and notation

For every $\alpha, \beta \in \mathcal{A}$, we denote $q_{\alpha\beta} := \mathbb{P}_\alpha(A_{K_1} = \beta)$ and $\mathbf{Q} := (q_{\alpha\beta})_{\alpha,\beta}$. Define $\mathcal{A}^- = \{\alpha \in \mathcal{A} : f(\alpha) < 0\}$ and $\mathcal{A}^+ = \{\alpha \in \mathcal{A} : f(\alpha) > 0\}$. Note that the matrix $\mathbf{Q}$ is stochastic, with $q_{\alpha\beta} = 0$ for $\beta \in \mathcal{A} \setminus \mathcal{A}^-$. Its restriction $\tilde{\mathbf{Q}}$ to $\mathcal{A}^-$ is stochastic and irreducible. The states $(A_{K_i})_{i\geq 1}$ of the Markov chain at the end of the successive non-negative excursions define a Markov chain on $\mathcal{A}^-$ with transition probability matrix $\tilde{\mathbf{Q}}$. For every $i \geq 2$ we thus have $\mathbb{P}(A_{K_i} = \beta \,|A_{K_{i-1}} = \alpha) = q_{\alpha\beta}$ if $\alpha, \beta \in \mathcal{A}^-$ and 0 otherwise. Denote $\tilde{z} > 0$ the stationary frequency vector of the irreducible stochastic matrix $\tilde{\mathbf{Q}}$ and let $z := (z_\alpha)_{\alpha\in\mathcal{A}}$ with $z_\alpha = \tilde{z}_\alpha > 0$ for $\alpha \in \mathcal{A}^-$ and $z_\alpha = 0$ for $\alpha \in \mathcal{A}\setminus\mathcal{A}^-$. Note that z is invariant for the matrix Q i.e. $z\mathbf{Q} = z$.

Remark 2.1. Note that in Karlin and Dembo's Markovian model of [9] the matrix Q is irreducible, thanks to their random scoring function and to their hypotheses recalled in Remark 1.1.

Using the strong Markov property, conditionally on $(A_{K_i})_{i\geq 1}$ the r.v. $(Q_i)_{i\geq 1}$ are independent, with the distribution of Q_i depending only on $A_{K_{i-1}}$ and A_{K_i}.

For every $\alpha \in \mathcal{A}$, $\beta \in \mathcal{A}^-$ and $y \geq 0$, let

$$F_{\alpha\beta}(y) := \mathbb{P}_\alpha(Q_1 \leq y \,|A_{\sigma^-} = \beta) \quad \text{and} \quad F_\alpha(y) := \mathbb{P}_\alpha(Q_1 \leq y).$$

Note that for any $\alpha \in \mathcal{A}^-$ and $i \geq 1$, $F_{\alpha\beta}$ represents the cumulative distribution function (*cdf*) of the height Q_i of the i-th non-negative excursion given that it starts in state α and ends in state β, i.e. $F_{\alpha\beta}(y) = \mathbb{P}(Q_i \leq y \,|A_{K_i} = \beta, A_{K_{i-1}} = \alpha)$, whereas F_α represents the *cdf* of Q_i conditionally on the i-th non-negative excursion starting in state α, i.e. $F_\alpha(y) = \mathbb{P}(Q_i \leq y \,|A_{K_{i-1}} = \alpha)$.

We thus have

$$F_\alpha(y) = \sum_{\beta\in\mathcal{A}} F_{\alpha\beta}(y) q_{\alpha\beta} = \sum_{\beta\in\mathcal{A}^-} F_{\alpha\beta}(y) q_{\alpha\beta}.$$

We also introduce the stopping time $\sigma^+ := \inf\{k \geqslant 1 : S_k > 0\}$ with values in $\mathbb{N} \cup \{\infty\}$. Due to hypothesis (1) we have $\mathbb{P}_\alpha(\sigma^+ < \infty) < 1$, for all $\alpha \in \mathcal{A}$.

For every $\alpha, \beta \in \mathcal{A}$ and $\xi > 0$, let

$$L_{\alpha\beta}(\xi) := \mathbb{P}_\alpha(S_{\sigma^+} \leqslant \xi, \sigma^+ < \infty, A_{\sigma^+} = \beta).$$

Note that $L_{\alpha\beta}(\xi) = 0$ for $\beta \in \mathcal{A} \setminus \mathcal{A}^+$. We have $L_{\alpha\beta}(\infty) \leq \mathbb{P}_\alpha(\sigma^+ < \infty) < 1$, and hence

$$\int_0^\infty dL_{\alpha\beta}(\xi) = 1 - L_{\alpha\beta}(\infty) > 0. \tag{4}$$

Let us also denote

$$L_\alpha(\xi) := \sum_{\beta \in \mathcal{A}^+} L_{\alpha\beta}(\xi) = \mathbb{P}_\alpha(S_{\sigma^+} \leqslant \xi, \sigma^+ < \infty)$$

the conditional *cdf* of the first positive partial sum when it exists, given that the Markov chain starts in state α, and

$$L_\alpha(\infty) := \lim_{\xi \to \infty} L_\alpha(\xi) = \mathbb{P}_\alpha(\sigma^+ < \infty).$$

For any $\theta \in \mathbb{R}$ we introduce the following matrix

$$\Phi(\theta) := \left(p_{\alpha\beta} \cdot \exp(\theta f(\beta))\right)_{\alpha,\beta \in \mathcal{A}}.$$

Since the transition matrix $\mathbf{P}$ was supposed to be positive, by the Perron-Frobenius Theorem, the spectral radius $\rho(\theta) > 0$ of the matrix $\Phi(\theta)$ coincides with its dominant eigenvalue, for which there exists a unique positive right eigen vector $u(\theta) = (u_i(\theta))_{1 \leq i \leq r}$ (seen as a column vector) normalized so that $\sum_{i=1}^r u_i(\theta) = 1$. Moreover, $\theta \mapsto \rho(\theta)$ is differentiable and strictly log convex (see [11, 3, 10]). In Lemma 3.4 we prove that $\rho'(0) = \mathbb{E}[f(A)]$, hence $\rho'(0) < 0$ by Hypothesis (1). Together with the strict log convexity of ρ and the fact that $\rho(0) = 1$, this implies that there exists a unique $\theta^* > 0$ such that $\rho(\theta^*) = 1$ (see [3] for more details).

2.2. Main results. Improvements on the distribution of the local score

Let $\alpha \in \mathcal{A}$. We start by giving a result which allows to compute recursively the *cdf* of the maximal non-negative partial sum S^+. We denote by $F_{S^+,\alpha}$ the *cdf* of S^+ conditionally on starting in state α:

$$F_{S^+,\alpha}(\ell d) := \mathbb{P}_\alpha(S^+ \leq \ell d), \quad \forall \ell \in \mathbb{N}$$

and for every $k \in \mathbb{N} \setminus \{0\}$ and $\beta \in \mathcal{A}$:

$$L_{\alpha\beta}^{(k)} := \mathbb{P}_\alpha(S_{\sigma^+} = kd, \sigma^+ < \infty, A_{\sigma^+} = \beta).$$

Note that $L_{\alpha\beta}^{(k)} = 0$ for $\beta \in \mathcal{A} \setminus \mathcal{A}^+$ and $L_\alpha(\infty) = \sum_{\beta\in\mathcal{A}^+} \sum_{k=1}^{\infty} L_{\alpha\beta}^{(k)}$.

The following result gives a recurrence relation for the double sequence $(F_{S^+,\alpha}(\ell d))_{\alpha,\ell}$.

Theorem 2.1. (Exact result for the distribution of S^+.) *For all $\alpha \in \mathcal{A}$ and $\ell \geq 1$:*

$$F_{S^+,\alpha}(0) = \mathbb{P}_\alpha(\sigma^+ = \infty) = 1 - L_\alpha(\infty),$$
$$F_{S^+,\alpha}(\ell d) = 1 - L_\alpha(\infty) + \sum_{\beta\in\mathcal{A}^+} \sum_{k=1}^{\ell} L_{\alpha\beta}^{(k)}\ F_{S^+,\beta}((\ell-k)d).$$

The proof will be given in Section 3.

In Theorem 2.2 we obtain the asymptotic behavior of S^+ using Theorem 2.1 and ideas inspired from [9] and adapted to our framework (see also the discussion in Remark 1.1). Before stating this result, we need to introduce few more notations.

For every $\alpha, \beta \in \mathcal{A}$ and $\ell \in \mathbb{N}$ we denote

$$G_{\alpha\beta}^{(\ell)} := \frac{u_\beta(\theta^*)}{u_\alpha(\theta^*)} e^{\theta^*\ell d} L_{\alpha\beta}^{(\ell)}, \qquad G_{\alpha\beta}(\ell) := \sum_{k=0}^{\ell} G_{\alpha\beta}^{(k)}, \qquad G_{\alpha\beta}(\infty) := \sum_{k=0}^{\infty} G_{\alpha\beta}^{(k)}.$$

The matrix $\mathbf{G}(\infty) := (G_{\alpha\beta}(\infty))_{\alpha,\beta}$ is stochastic, using Lemma 3.3; the subset $\mathcal{A}^+$ is a recurrent class, whereas the states in $\mathcal{A} \setminus \mathcal{A}^+$ are transient. The restriction of $\mathbf{G}(\infty)$ to $\mathcal{A}^+$ is stochastic and irreducible; let us denote $\tilde{w} > 0$ the corresponding stationary frequency vector. Define $w = (w_\alpha)_{\alpha\in\mathcal{A}}$, with $w_\alpha = \tilde{w}_\alpha > 0$ for $\alpha \in \mathcal{A}^+$ and $w_\alpha = 0$ for $\alpha \in \mathcal{A} \setminus \mathcal{A}^+$. The vector w is invariant for $\mathbf{G}(\infty)$, i.e. $w\mathbf{G}(\infty) = w$.

Remark 2.2. Note that in Karlin and Dembo's Markovian model of [9] the matrix $\mathbf{G}(\infty)$ is positive, hence irreducible, thanks to their random scoring function and to their hypotheses recalled in Remark 1.1.

Remark 2.3. Note that the coefficients $L_{\alpha\beta}^{(k)}$ can be computed recursively (see Subsection 4.2). In Subsection 4.3 we present in detail a recursive procedure for computing the *cdf* $F_{S^+,\alpha}$, based on Theorem 2.1. Note also that, for every $\alpha, \beta \in \mathcal{A}$, there are a finite number of $L_{\alpha\beta}^{(k)}$ different from zero. Therefore, there are a finite number of non-null terms in the sum defining $G_{\alpha\beta}(\infty)$.

Theorem 2.2. (Asymptotic distribution of S^+.) *For every $\alpha \in \mathcal{A}$ we have*

$$\lim_{k\to+\infty} \frac{e^{\theta^* kd}\mathbb{P}_\alpha(S^+ > kd)}{u_\alpha(\theta^*)} = \frac{d}{c} \cdot \sum_{\gamma\in\mathcal{A}^+} \frac{w_\gamma}{u_\gamma(\theta^*)} \sum_{\ell\geq 0} (L_\gamma(\infty) - L_\gamma(\ell d)) e^{\theta^*\ell d} := c(\infty), \quad (5)$$

where $w = (w_\alpha)_{\alpha\in\mathcal{A}}$ is the stationary frequency vector of the matrix $\mathbf{G}(\infty)$ and

$$c := \sum_{\gamma,\beta\in\mathcal{A}^+} \frac{w_\gamma}{u_\gamma(\theta^*)} u_\beta(\theta^*) \sum_{\ell\geq 0} \ell d \cdot e^{\theta^*\ell d}\ L^{(\ell)}_{\gamma\beta}.$$

The proof is deferred to Section 3.

Remark 2.4. Note that there are a finite number of non-null terms in the above sums over ℓ. We also have the following alternative expression for $c(\infty)$:

$$c(\infty) = \frac{d}{c(e^{\theta^* d}-1)} \cdot \sum_{\gamma\in\mathcal{A}^+} \frac{w_\gamma}{u_\gamma(\theta^*)} \left\{ \mathbb{E}_\gamma \left[e^{\theta^* S_{\sigma^+}}; \sigma^+ < \infty \right] - L_\gamma(\infty) \right\}.$$

Indeed, by the summation by parts formula

$$\sum_{\ell=m}^{k} f_\ell (g_{\ell+1} - g_\ell) = f_{k+1} g_{k+1} - f_m g_m - \sum_{\ell=m}^{k} (f_{\ell+1} - f_\ell) g_{\ell+1},$$

we obtain

$$\begin{aligned}
&\sum_{\ell=0}^{\infty} (L_\gamma(\infty) - L_\gamma(\ell d)) e^{\theta^*\ell d} = \frac{1}{e^{\theta^* d}-1} \sum_{\ell=0}^{\infty} (L_\gamma(\infty) - L_\gamma(\ell d)) \left(e^{\theta^*(\ell+1)d} - e^{\theta^*\ell d} \right) \\
&= \frac{1}{e^{\theta^* d}-1} \\
&\quad \times \left\{ \lim_{k\to\infty} (L_\gamma(\infty) - L_\gamma(kd)) e^{\theta^* k d} - L_\gamma(\infty) - \sum_{\ell=0}^{\infty} (L_\gamma(\ell d) - L_\gamma((\ell+1)d)) e^{\theta^*(\ell+1)d} \right\} \\
&= \frac{1}{e^{\theta^* d}-1} \left\{ -L_\gamma(\infty) + \sum_{\ell=0}^{\infty} e^{\theta^*(\ell+1)d}\ \mathbb{P}_\gamma(S_{\sigma^+} = (\ell+1)d,\ \sigma^+ < \infty) \right\} \\
&= \frac{1}{e^{\theta^* d}-1} \left\{ \mathbb{E}_\gamma \left[e^{\theta^* S_{\sigma^+}}; \sigma^+ < \infty \right] - L_\gamma(\infty) \right\}.
\end{aligned}$$

Before stating the next results, let us denote for every integer $\ell < 0$ and $\alpha, \beta \in \mathcal{A}$,

$$Q^{(\ell)}_{\alpha\beta} := \mathbb{P}_\alpha(S_{\sigma^-} = \ell d, A_{\sigma^-} = \beta).$$

Note that $Q^{(\ell)}_{\alpha\beta} = 0$ for $\beta \in \mathcal{A} \setminus \mathcal{A}^-$. In Section 4 we give a recursive computational method for obtaining these quantities.

Using Theorem 2.2 we obtain the following

Theorem 2.3. (Asymptotic distribution of Q_1.) *We have the following asymptotic result on the distribution of the height of the first non-negative excursion: for every $\alpha \in \mathcal{A}$ we have*

$$\mathbb{P}_\alpha(Q_1 > kd) \underset{k\to\infty}{\sim} \mathbb{P}_\alpha(S^+ > kd) - \sum_{\ell<0} \sum_{\beta\in\mathcal{A}^-} \mathbb{P}_\beta\left(S^+ > (k-\ell)d\right) \cdot Q^{(\ell)}_{\alpha\beta}. \tag{6}$$

The proof will be given in Section 3.

Using now Theorems 2.2 and 2.3 we finally obtain the following result on the asymptotic distribution of the local score M_n for a sequence of length n.

Theorem 2.4. (Asymptotic distribution of the local score M_n.) *For every $\alpha \in \mathcal{A}$:*

$$\mathbb{P}_\alpha\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right) \underset{n\to\infty}{\sim} \exp\left\{-\frac{n}{A^*}\sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{P}_\beta\left(S^+ > \left\lfloor \frac{\log(n)}{\theta^*} + x \right\rfloor\right)\right\}$$
$$\times \exp\left\{\frac{n}{A^*}\sum_{k<0}\sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma\left(S^+ > \left\lfloor \frac{\log(n)}{\theta^*} + x \right\rfloor - kd\right)\cdot \sum_{\beta\in\mathcal{A}^-} z_\beta Q^{(k)}_{\beta\gamma}\right\}, \quad (7)$$

where $z = (z_\alpha)_{\alpha\in\mathcal{A}}$ is the invariant probability measure of the matrix $\mathbf{Q}$ defined in Subsection 2.1 and

$$A^* := \lim_{m\to+\infty} \frac{K_m}{m} = \frac{1}{\mathbb{E}(f(A))}\sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{E}_\beta(S_{\sigma^-}) \ \textit{a.s.}$$

Remark 2.5.

- Note that the asymptotic equivalent in Equation (7) does not depend on the initial state α.
- We recall, for comparison, the asymptotic result of [9] (Equation (1.27)) for the distribution of M_n:

$$\lim_{n\to+\infty} \mathbb{P}_\alpha\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right) = \exp\left(-K^* \exp(-\theta^*)\right), \quad (8)$$

  with $K^* = v(\infty)\cdot c(\infty)$, where $c(\infty)$ given in Theorem 2.2 is related to the defective distribution of the first positive partial sum S_{σ^+} (see also Remark 2.4) and $v(\infty)$ is related to the distribution of the first negative partial sum S_{σ^-} (see Equations (5.1) and (5.2) of [9] for more details). A more explicit formula for K^* is given in Subsection 4.4 for an application in a simple case.
- Note that our asymptotic equivalent in Equation (7) keeps the dependence on n, whereas the approximation derived from Equation (8) does not.

3. Proofs of the main results

3.1. Proof of Theorem 2.1

We have

$$\begin{aligned} F_{S^+,\alpha}(\ell d) &= \mathbb{P}_\alpha(\sigma^+ = \infty) + \mathbb{P}_\alpha(S^+ \leq \ell d, \sigma^+ < \infty) \\ &= 1 - L_\alpha(\infty) + \sum_{\beta \in \mathcal{A}^+} \sum_{k=1}^{\ell} \mathbb{P}_\alpha(S^+ \leq \ell d, \sigma^+ < \infty, S_{\sigma^+} = kd, A_{\sigma^+} = \beta) \\ &= 1 - L_\alpha(\infty) + \sum_{\beta \in \mathcal{A}^+} \sum_{k=1}^{\ell} L^{(k)}_{\alpha\beta}\ \mathbb{P}_\alpha(S^+ \leq \ell d \mid \sigma^+ < \infty, S_{\sigma^+} = kd, A_{\sigma^+} = \beta). \end{aligned}$$

The last probability can further be written

$$\mathbb{P}_\alpha(S^+ - S_{\sigma^+} \leq (\ell - k)d \mid \sigma^+ < \infty, S_{\sigma^+} = kd, A_{\sigma^+} = \beta) = \mathbb{P}_\beta(S^+ \leq (\ell - k)d),$$

by the strong Markov property applied to the stopping time σ^+. The stated result easily follows. □

3.2. Proof of Theorem 2.2

We first prove some preliminary lemmas.

Lemma 3.1. *We have* $\lim_{k\to\infty} \mathbb{P}_\alpha(S^+ > kd) = 0$ *for every* $\alpha \in \mathcal{A}$.

Proof. With $F_{S^+,\alpha}$ defined in Theorem 2.1, we introduce for every α and $\ell \geq 0$:

$$b_\alpha(\ell d) := \frac{1 - F_{S^+,\alpha}(\ell d)}{u_\alpha(\theta^*)} e^{\theta^* \ell d}, \quad a_\alpha(\ell d) := \frac{L_\alpha(\infty) - L_\alpha(\ell d)}{u_\alpha(\theta^*)} e^{\theta^* \ell d}.$$

Theorem 2.1 allows to obtain the following renewal system for the family $(b_\alpha)_{\alpha\in\mathcal{A}}$:

$$\forall \ell > 0, \forall \alpha \in \mathcal{A}, \quad b_\alpha(\ell d) = a_\alpha(\ell d) + \sum_{\beta} \sum_{k=0}^{\ell} b_\beta((\ell - k)d) G^{(k)}_{\alpha\beta}.$$

Since the restriction of $\tilde{\mathbf{G}}(\infty)$ of $\mathbf{G}(\infty)$ to $\mathcal{A}^+$ is stochastic, its spectral radius equals 1 and a corresponding right eigenvector is the vector having all components equal to 1; a left eigenvector is the stationary frequency vector $\tilde{w} > 0$.

Step 1: For every $\alpha \in \mathcal{A}^+$, a direct application of Theorem 2.2 of Athreya and Murthy [1] gives the formula in Equation 5 for the limit $c(\infty)$ of $b_\alpha(\ell d)$ when $\ell \to \infty$, which implies the stated result.

Step 2: Consider now $\alpha \notin \mathcal{A}^+$. By Theorem 2.1 we have

$$\mathbb{P}_\alpha(S^+ > \ell d) = L_\alpha(\infty) - \sum_{\beta\in\mathcal{A}^+}\sum_{k=1}^{\ell} L_{\alpha\beta}^{(k)} \ \left\{1 - \mathbb{P}_\beta(S^+ > (\ell-k)d)\right\}.$$

Since $\mathbb{P}_\beta(S^+ > (\ell-k)d) = 1$ for $k > \ell$ and $L_\alpha(\infty) = \sum_{\beta\in\mathcal{A}^+}\sum_{k=1}^{\infty} L_{\alpha\beta}^{(k)}$, we deduce

$$\mathbb{P}_\alpha(S^+ > \ell d) = \sum_{\beta\in\mathcal{A}^+}\sum_{k=1}^{\infty} L_{\alpha\beta}^{(k)} \ \mathbb{P}_\beta(S^+ > (\ell-k)d). \tag{9}$$

Note that for fixed α and β, there are a finite number of non-null terms in the above sum over k. Using the fact that for fixed $\beta \in \mathcal{A}^+$ and $k \geq 1$ we have $\mathbb{P}_\beta(S^+ > (\ell-k)d) \longrightarrow 0$ when $\ell \to \infty$, as shown previously in Step 1, the stated result follows. □

Lemma 3.2. *Let $\theta > 0$. With $u(\theta)$ defined in Subsection 2.1, the sequence of random variables $(U_m(\theta))_{m\geq 0}$ defined by $U_0(\theta) := 1$ and*

$$U_m(\theta) := \prod_{i=0}^{m-1}\left[\frac{\exp(\theta f(A_{i+1}))}{u_{A_i}(\theta)} \cdot \frac{u_{A_{i+1}}(\theta)}{\rho(\theta)}\right] = \frac{\exp(\theta S_m)u_{A_m}(\theta)}{\rho(\theta)^m u_{A_0}(\theta)} \quad \text{, for } m \geq 1$$

is a martingale with respect to the canonical filtration $\mathcal{F}_m = \sigma(A_0, \ldots, A_m)$.

Proof. We have

$$U_{m+1}(\theta) = U_m(\theta)\frac{\exp(\theta f(A_{m+1}))u_{A_{m+1}}(\theta)}{u_{A_m}(\theta)\rho(\theta)}.$$

Since $U_m(\theta)$ and $u_{A_m}(\theta)$ are measurable with respect to $\mathcal{F}_m$, we have

$$\mathbb{E}[U_{m+1}(\theta)|\mathcal{F}_m] = U_m(\theta)\frac{\mathbb{E}[\exp(\theta f(A_{m+1}))u_{A_{m+1}}(\theta)|\mathcal{F}_m]}{u_{A_m}(\theta)\rho(\theta)}.$$

By the Markov property we further have

$$\mathbb{E}[\exp(\theta f(A_{m+1}))u_{A_{m+1}}(\theta)|\mathcal{F}_m] = \mathbb{E}[\exp(\theta f(A_{m+1}))u_{A_{m+1}}(\theta)|A_m]$$

and by definition of $u(\theta)$,

$$\begin{aligned}\mathbb{E}[\exp(\theta f(A_{m+1}))u_{A_{m+1}}(\theta)|A_m = \alpha] &= \sum_\beta \exp(\theta f(\beta))u_\beta(\theta)p_{\alpha\beta} \\ &= (\Phi(\theta)u(\theta))_\alpha = u_\alpha(\theta)\rho(\theta).\end{aligned}$$

We deduce

$$\mathbb{E}[\exp(\theta f(A_{m+1}))u_{A_{m+1}}(\theta)|A_m] = u_{A_m}(\theta)\rho(\theta),$$

hence $\mathbb{E}[U_{m+1}(\theta)|\mathcal{F}_m] = U_m(\theta)$, which finishes the proof. □

Lemma 3.3. *With θ^* defined at the end of Subsection 2.1 we have*

$$\forall \alpha \in \mathcal{A}: \qquad \frac{1}{u_\alpha(\theta^*)} \sum_{\beta \in \mathcal{A}^+} \sum_{\ell=1}^{\infty} L_{\alpha\beta}^{(\ell)}\, e^{\theta^* \ell d}\, u_\beta(\theta^*) = 1. \tag{10}$$

Proof. The proof uses Lemma 3.1 and ideas inspired from [9] (Lemma 4.2). First note that the above equation is equivalent to

$$\mathbb{E}_\alpha[U_{\sigma^+}(\theta^*); \sigma^+ < \infty] = 1,$$

with $U_m(\theta)$ defined in Lemma 3.2. By applying the optional sampling theorem to the bounded stopping time $\tau_n := \min(\sigma^+, n)$ and to the martingale $(U_m(\theta^*))_m$, we obtain

$$1 = \mathbb{E}_\alpha[U_0(\theta^*)] = \mathbb{E}_\alpha[U_{\tau_n}(\theta^*)] = \mathbb{E}_\alpha[U_{\sigma^+}(\theta^*); \sigma^+ \leq n] + \mathbb{E}_\alpha[U_n(\theta^*); \sigma^+ > n].$$

We will show that $\mathbb{E}_\alpha[U_n(\theta^*); \sigma^+ > n] \longrightarrow 0$ when $n \to \infty$. Passing to the limit in the previous relation will then give the desired result. Since $\rho(\theta^*) = 1$, we have

$$U_n(\theta^*) = \frac{\exp(\theta^* S_n) u_{A_n}(\theta^*)}{u_{A_0}(\theta^*)}$$

and it suffices to show that $\lim_{n\to\infty} \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n] = 0$.

For a fixed $a > 0$ we can write

$$\begin{aligned}\mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n] = {} & \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n,\ \exists k \leq n : S_k \leq -2a] \\ & + \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, -2a \leq S_k \leq 0,\ \forall 0 \leq k \leq n]. \end{aligned} \tag{11}$$

The first expectation in the right-hand side of Equation (11) can further be bounded as follows:

$$\begin{aligned}\mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n,\ \exists k \leq n : S_k \leq -2a] \leq {} & \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n \leq -a] \\ & + \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n > -a,\ \exists k < n : S_k \leq -2a]. \end{aligned} \tag{12}$$

We obviously have

$$\mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n \leq -a] \leq \exp(-\theta^* a). \tag{13}$$

Let us further define the stopping time $T := \inf\{k \geq 1 : S_k \leq -2a\}$. Note that $T < \infty$ *a.s.* since $S_n \longrightarrow -\infty$ *a.s.* when $n \to \infty$. Indeed, by the ergodic theorem we

have $S_n/n \longrightarrow \mathbb{E}[f(A)] < 0$ when $n \to \infty$. Therefore we have

$$\begin{aligned}\mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n > -a,\ \exists k < n : S_k \leq -2a] &\leq \mathbb{P}_\alpha(T \leq n, S_n > -a) \\ &= \sum_{\beta \in \mathcal{A}^-} \mathbb{P}_\alpha(T \leq n, S_n > -a \mid A_T = \beta)\mathbb{P}_\alpha(A_T = \beta) \\ &\leq \sum_{\beta \in \mathcal{A}^-} \mathbb{P}_\alpha(S_n - S_T > a \mid A_T = \beta)\mathbb{P}_\alpha(A_T = \beta) \\ &\leq \sum_{\beta \in \mathcal{A}^-} \mathbb{P}_\beta(S^+ > a)\mathbb{P}_\alpha(A_T = \beta),\end{aligned}$$

by the strong Markov property. For every $a > 0$ we thus have

$$\limsup_{n\to\infty} \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n > -a,\ \exists k < n : S_k \leq -2a] \leq \sum_{\beta \in \mathcal{A}^-} \mathbb{P}_\beta(S^+ > a). \tag{14}$$

Considering the second expectation in the right-hand side of Equation (11), we have

$$\lim_{n\to\infty} \mathbb{P}_\alpha(-2a \leq S_k \leq 0,\ \forall 0 \leq k \leq n) = \mathbb{P}_\alpha(-2a \leq S_k \leq 0,\ \forall k \geq 0) = 0, \tag{15}$$

again since $S_n \longrightarrow -\infty$ *a.s.* when $n \to \infty$.

Equations (11),(12),(13),(14) and (15) imply that for every $a > 0$ we have

$$\limsup_{n\to\infty} \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n] \leq \exp(-\theta^* a) + \sum_{\beta \in \mathcal{A}^-} \mathbb{P}_\beta(S^+ > a).$$

Using Lemma 3.1 and taking $a \to \infty$ we obtain $\lim_{n\to\infty} \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n] = 0$. □

We are now ready to prove the Theorem 2.2.

Proof of Theorem 2.2:

For $\alpha \in \mathcal{A}^+$ the formula has been already shown in Step 1 of the proof of Lemma 3.1. For $\alpha \notin \mathcal{A}^+$ we will prove the stated formula using Theorem 2.1. From Equation (9), we have

$$\mathbb{P}_\alpha(S^+ > \ell d) = \sum_{\beta \in \mathcal{A}^+} \sum_{k=1}^{\infty} L_{\alpha\beta}^{(k)}\ \mathbb{P}_\beta(S^+ > (\ell - k)d),$$

hence

$$\frac{e^{\theta^* \ell d}\mathbb{P}_\alpha(S^+ > \ell d)}{u_\alpha(\theta^*)} = \sum_{\beta \in \mathcal{A}^+} \sum_{k=1}^{\infty} \frac{e^{\theta^*(\ell-k)d}\mathbb{P}_\beta(S^+ > (\ell - k)d)}{u_\beta(\theta^*)} L_{\alpha\beta}^{(k)} e^{\theta^* k d} \frac{u_\beta(\theta^*)}{u_\alpha(\theta^*)}.$$

Note that for every α and β there are a finite number of non-null terms in the above sum over k. Moreover, as shown in Lemma 3.1

$$\forall \beta \in \mathcal{A}^+,\ \forall k \geq 0 : \quad \frac{e^{\theta^*(\ell-k)d}\mathbb{P}_\beta(S^+ > (\ell - k)d)}{u_\beta(\theta^*)} \underset{\ell\to\infty}{\longrightarrow} c(\infty).$$

We finally obtain that

$$\lim_{\ell\to+\infty} \frac{e^{\theta^*\ell d}\mathbb{P}_\alpha(S^+ > \ell d)}{u_\alpha(\theta^*)} = \frac{c(\infty)}{u_\alpha(\theta^*)} \sum_{\beta\in\mathcal{A}^+}\sum_{k=1}^{\infty} L^{(k)}_{\alpha\beta}\ e^{\theta^* kd}\ u_\beta(\theta^*),$$

which equals $c(\infty)$ as desired, by Lemma 3.3.

3.3. Proof of Theorem 2.3

Since $S^+ \geq Q_1$, for every $\alpha \in \mathcal{A}$ we have

$$\mathbb{P}_\alpha(S^+ > kd) = \mathbb{P}_\alpha(Q_1 > kd) + \mathbb{P}_\alpha(S^+ > kd, Q_1 \leq kd).$$

We will further decompose the last probability with respect to the values taken by S_{σ^-} and A_{σ^-}, as follows:

$$\begin{aligned}
\mathbb{P}_\alpha(S^+ > kd, Q_1 \leq kd) &= \sum_{\ell<0}\sum_{\beta\in\mathcal{A}^-} \mathbb{P}_\alpha(S^+ > kd, Q_1 \leq kd, S_{\sigma^-} = \ell d, A_{\sigma^-} = \beta)\\
&= \sum_{\ell<0}\sum_{\beta\in\mathcal{A}^-} \mathbb{P}_\alpha(S^+ - S_{\sigma^-} > (k-\ell)d \mid A_{\sigma^-} = \beta, Q_1 \leq kd, S_{\sigma^-} = \ell d)\\
&\quad \times \mathbb{P}_\alpha(Q_1 \leq kd, S_{\sigma^-} = \ell d, A_{\sigma^-} = \beta)\\
&= \sum_{\ell<0}\sum_{\beta\in\mathcal{A}^-} \mathbb{P}_\beta(S^+ > (k-\ell)d) \cdot \left\{ Q^{(\ell)}_{\alpha\beta} - \mathbb{P}_\alpha(Q_1 > kd, S_{\sigma^-} = \ell d, A_{\sigma^-} = \beta)\right\},
\end{aligned}$$

by applying the strong Markov property to the stopping time σ^-. We thus obtain

$$\begin{aligned}
&\mathbb{P}_\alpha(S^+ > kd) - \sum_{\ell<0}\sum_{\beta\in\mathcal{A}^-} \mathbb{P}_\beta(S^+ > (k-\ell)d) \cdot Q^{(\ell)}_{\alpha\beta} - \mathbb{P}_\alpha(Q_1 > kd)\\
&= -\sum_{\ell<0}\sum_{\beta\in\mathcal{A}^-} \mathbb{P}_\beta(S^+ > (k-\ell)d)\ \mathbb{P}_\alpha(Q_1 > kd, S_{\sigma^-} = \ell d, A_{\sigma^-} = \beta).
\end{aligned}$$

By Theorem 2.2 we have $\mathbb{P}_\beta(S^+ > kd) = O(e^{-\theta^* kd})$ as $k \to \infty$, for every $\beta \in \mathcal{A}^-$, from which we deduce that the left-hand side of the previous equation is $o(\mathbb{P}_\alpha(Q_1 > kd))$ when $k \to \infty$. The stated result then easily follows. □

3.4. Proof of Theorem 2.4

We will first prove some useful lemmas.

Lemma 3.4. *We have* $\rho'(0) = \mathbb{E}[f(A)] < 0$.

Proof. By the fact that $\rho(\theta)$ is an eigenvalue of the matrix $\Phi(\theta)$ with corresponding eigenvector $u(\theta)$, we have

$$\rho(\theta)u_\alpha(\theta) = (\Phi(\theta)u(\theta))_\alpha = \sum_\beta p_{\alpha\beta}e^{\theta f(\beta)}u_\beta(\theta).$$

When derivating the previous relation with respect to θ we obtain

$$\frac{d}{d\theta}(\rho(\theta)u_\alpha(\theta)) = \sum_\beta p_{\alpha\beta}\left(f(\beta)e^{\theta f(\beta)}u_\beta(\theta) + e^{\theta f(\beta)}u'_\beta(\theta)\right).$$

We have $\rho(0) = 1$ et $u(0) =^t (1/r, \ldots, 1/r)$. For $\theta = 0$, we then get

$$\sum_\alpha \pi_\alpha \frac{d}{d\theta}(\rho(\theta)u_\alpha(\theta))\bigg|_{\theta=0} = \frac{1}{r}\mathbb{E}[f(A)] + \sum_{\alpha,\beta}\pi_\alpha p_{\alpha\beta}u'_\beta(0) = \frac{1}{r}\mathbb{E}[f(A)] + \sum_\beta \pi_\beta u'_\beta(0). \tag{16}$$

On the other hand,

$$\sum_\alpha \pi_\alpha \frac{d}{d\theta}(\rho(\theta)u_\alpha(\theta)) = \frac{d}{d\theta}\left(\sum_\alpha \pi_\alpha\rho(\theta)u_\alpha(\theta)\right) = \rho'(\theta)\sum_\alpha \pi_\alpha u_\alpha(\theta) + \rho(\theta)\sum_\alpha \pi_\alpha u'_\alpha(\theta).$$

For $\theta = 0$ we get

$$\sum_\alpha \pi_\alpha \frac{d}{d\theta}(\rho(\theta)u_\alpha(\theta))\bigg|_{\theta=0} = \frac{\rho'(0)}{r} + \rho(0)\cdot\sum_\alpha \pi_\alpha u'_\alpha(0). \tag{17}$$

From Equations (16) and (17) we deduce

$$\frac{\rho'(0)}{r} + \sum_\alpha \pi_\alpha u'_\alpha(0) = \frac{1}{r}\mathbb{E}[f(A)] + \sum_\beta \pi_\beta u'_\beta(0),$$

from which the stated result easily follows. □

Lemma 3.5. *There exist $\mathcal{I} > 0$ and $n_0 \geq 0$ such that $\forall n \geq n_0$, $\mathbb{P}_\alpha(S_n \geq 0) \leq \exp(-n\mathcal{I})$ for every $\alpha \in \mathcal{A}$.*

Proof. By a large deviation principle for additive functionals of Markov chains (see Theorem 3.1.2. in [3]) we have

$$\limsup_{n\to+\infty}\frac{1}{n}\log\left(\mathbb{P}_\alpha\left(\frac{S_n}{n}\in\Gamma\right)\right) \leq -\mathcal{I},$$

with $\Gamma = [0, +\infty)$ and $\mathcal{I} = \inf_{x\in\bar{\Gamma}}\sup_{\theta\in\mathbb{R}}(\theta x - \log\rho(\theta))$. Since $\mathcal{A}$ is finite, it remains to prove that $\mathcal{I} > 0$.

For every $x \geq 0$, let us denote $g_x(\theta) := \theta x - \log \rho(\theta)$ and $I(x) := \sup_{\theta \in \mathbb{R}} g_x(\theta)$. We will first show that $I(x) = \sup_{\theta \in \mathbb{R}^+} g_x(\theta)$. Indeed, we have $g'_x(\theta) = x - \rho'(\theta)/\rho(\theta)$. By the strict convexity property of ρ (see [3, 10]) and the fact that $\rho'(0) = \mathbb{E}[f] < 0$ (by Lemma 3.4), we deduce that $\rho'(\theta) < 0$ for every $\theta \leq 0$, implying that $g'_x(\theta) > x \geq 0$ for $\theta \leq 0$. The function g'_x is therefore increasing on $\mathbb{R}^-$, and hence $I(x) = \sup_{\theta \in \mathbb{R}^+} g_x(\theta)$.

As a consequence, we deduce that $x \mapsto I(x)$ is non-decreasing on $\mathbb{R}^+$. We thus obtain $\mathcal{I} = \inf_{x \in \mathbb{R}^+} I(x) = I(0)$.

Further, we have $I(0) = \sup_{\theta \in \mathbb{R}} (-\log \rho(\theta)) = -\inf_{\theta \in \mathbb{R}^+} \log(\rho(\theta))$. Using again the fact that $\rho'(0) < 0$ (Lemma 3.4), the strict convexity of ρ and the fact that $\rho(0) = \rho(\theta^*) = 1$, we finally obtain $\mathcal{I} = -\log\left(\inf_{\theta \in \mathbb{R}^+} \rho(\theta)\right) > -\log \rho(0) = 0$. The statement then follows. □

Lemma 3.6. *We have $\mathbb{E}_\alpha(K_1) < \infty$ for every $\alpha \in \mathcal{A}$.*

Proof. Note that $\mathbb{P}_\alpha(K_1 > n) \leq \mathbb{P}_\alpha(S_n \geq 0)$. With $n_0 \in \mathbb{N}$ and $\mathcal{I} > 0$ defined in Lemma 3.5, using a well-known formula for the expectation, we get

$$\mathbb{E}_\alpha[K_1] = \sum_{n \geq 0} \mathbb{P}_\alpha(K_1 > n) \leq \sum_{n \geq 0} \mathbb{P}(S_n \geq 0) \leq C + \sum_{n \geq n_0} \exp(-n\mathcal{I}),$$

where $C > 0$ is a constant. The statement easily follows. □

Lemma 3.7. *We have*

$$\lim_{m \to +\infty} \frac{K_m}{m} = \sum_\beta z_\beta \mathbb{E}_\beta(K_1) \ \ a.s.$$

Proof. Recall that $K_1 = \sigma^-$. We can write

$$\frac{K_m}{m} = \frac{K_1}{m} + \frac{1}{m} \sum_{i=2}^{m} (K_i - K_{i-1}) = \frac{K_1}{m} + \sum_\beta \frac{1}{m} \sum_{i=2}^{m} (K_i - K_{i-1}) \mathbf{1}_{\{A_{K_{i-1}} = \beta\}}. \quad (18)$$

First note that $\dfrac{K_1}{m} \to 0$ *a.s.* when $m \to \infty$, since $K_1 < +\infty$ *a.s.* By the strong Markov property we have that, conditionally on $(A_{K_{i-1}})_{i \geq 2}$, the random variables $(K_i - K_{i-1})_{i \geq 2}$ are all independent and the distribution of $K_i - K_{i-1}$ depends only on $A_{K_{i-1}}$ and we have $\mathbb{P}(K_i - K_{i-1} = \ell \mid A_{K_{i-1}} = \alpha) = \mathbb{P}_\alpha(K_1 = \ell)$. Therefore, the couples $Y_i := (A_{K_{i-1}}, K_i - K_{i-1}), i \geq 2$ form a Markov chain on $\mathcal{A}^- \times \mathbb{N}$, with transition probabilities $\mathbb{P}(Y_i = (\beta, \ell) \mid Y_{i-1} = (\alpha, k)) = q_{\alpha\beta} \mathbb{P}_\beta(K_1 = \ell)$. Recall that the restriction $\tilde{\mathbf{Q}}$ of the matrix $\mathbf{Q}$ to the subspace $\mathcal{A}^-$ is irreducible. Therefore, the

Markov chain $(Y_i)_i$ is also irreducible and we can show that $\pi(\alpha,k) := z_\alpha \mathbb{P}_\alpha(K_1 = k)$ is its invariant distribution. Indeed, since z is invariant for $\mathbf{Q}$, we easily deduce that

$$\sum_{\alpha,k} \pi(\alpha,k) \cdot q_{\alpha\beta} \mathbb{P}_\beta(K_1 = \ell) = \pi(\beta,\ell).$$

For fixed β, when applying the ergodic theorem to the Markov chain $(Y_i)_i$ and the function $\varphi_\beta(\alpha,k) := k\mathbf{1}_{\{\alpha=\beta\}}$, we deduce

$$\frac{1}{m}\sum_{i=2}^{m}(K_i - K_{i-1})\mathbf{1}_{\{A_{K_{i-1}}=\beta\}} \longrightarrow \sum_{\alpha,k}\varphi_\beta(\alpha,k)\pi(\alpha,k) = z_\beta \mathbb{E}_\beta(K_1) \ \ a.s.$$

when $m \to \infty$. Taking the sum over β and using the relation (18) gives the desired result. □

Proof of Theorem 2.4:

The proof is inspired from [9].

Given $(A_{K_i})_{i\geq 0}$, the random variables $(Q_i)_{i\geq 1}$ are independent and the *cdf* of Q_i is $F_{A_{K_{i-1}}A_{K_i}}$. Therefore

$$\begin{aligned}\mathbb{P}_\alpha\left(M_{K_m} \leq y\right) &= \mathbb{E}_\alpha\left[\prod_{i=1}^{m} F_{A_{K_{i-1}}A_{K_i}}(y)\right] \\ &= \mathbb{E}_\alpha\left[\exp\left\{\sum_{\beta,\gamma\in\mathcal{A}} m\psi_{\beta\gamma}(m)\log(F_{\beta\gamma}(y))\right\}\right],\end{aligned}$$

with $\psi_{\beta\gamma}(m) := \#\{i : 1 \leq i \leqslant m, A_{K_{i-1}} = \beta, A_{K_i} = \gamma\}/m$. Given that $A_0 = \alpha \in \mathcal{A}^-$, the states $(A_{K_i})_{i\geq 0}$ form an irreducible Markov chain on $\mathcal{A}^-$ of transition matrix $\tilde{\mathbf{Q}} = (q_{\beta\gamma})_{\beta,\gamma\in\mathcal{A}^-}$ and stationary frequency vector $\tilde{z} = (z_\beta)_{\beta\in\mathcal{A}^-} > 0$.

Consequently, for $\beta,\gamma \in \mathcal{A}^-$ the ergodic theorem implies that $\psi_{\beta\gamma}(m) \longrightarrow z_\beta q_{\beta\gamma}$ *a.s.* when $m \to \infty$. On the other hand, for any $\alpha \in \mathcal{A}$, if $\beta \in \mathcal{A}\setminus\mathcal{A}^-$, then $\psi_{\beta\gamma}(m) \leqslant 1/m$ and thus $\psi_{\beta\gamma}(m) \longrightarrow 0$ *a.s.* when $m \to \infty$, for any $\gamma \in \mathcal{A}$. With $z_\beta = 0$ for $\beta \in \mathcal{A}\setminus\mathcal{A}^-$, we thus have $\psi_{\beta\gamma}(m) \longrightarrow z_\beta q_{\beta\gamma}$ *a.s.* when $m \to \infty$, for every $\beta,\gamma \in \mathcal{A}$.

Denoting $d_{\beta\gamma}(m) := m\left[1 - F_{\beta\gamma}\left(\frac{\log m}{\theta^*} + x\right)\right]$ and using the fact that $d_{\beta\gamma}(m)$ are uniformly bounded in β and γ, we have

$$\begin{aligned}\lim_{m\to\infty}\mathbb{P}_\alpha\left(M_{K_m} \leq \frac{\log m}{\theta^*} + x\right) &= \lim_{m\to\infty}\mathbb{E}_\alpha\left[\exp\left(-\sum_{\beta,\gamma\in\mathcal{A}}\psi_{\beta\gamma}(m)d_{\beta\gamma}(m)\right)\right] \\ &= \lim_{m\to\infty}\exp\left(-\sum_{\beta,\gamma\in\mathcal{A}} z_\beta q_{\beta\gamma} d_{\beta\gamma}(m)\right).\end{aligned}$$

Since

$$\sum_{\gamma\in\mathcal{A}} q_{\beta\gamma} d_{\beta\gamma}(m) = m\left[1 - F_\beta\left(\frac{\log m}{\theta^*} + x\right)\right],$$

$$\lim_{m\to\infty} \mathbb{P}_\alpha\left(M_{K_m} \leq \frac{\log m}{\theta^*} + x\right) = \lim_{m\to\infty} \exp\left(-m \sum_{\beta\in\mathcal{A}^-} z_\beta\left[1 - F_\beta\left(\frac{\log m}{\theta^*} + x\right)\right]\right).$$

But

$$1 - F_\beta\left(\frac{\log m}{\theta^*} + x\right) = \mathbb{P}_\beta\left(Q_1 > \frac{\log(m)}{\theta^*} + x\right) = \mathbb{P}_\beta\left(Q_1 > \left\lfloor \frac{\log(m)}{\theta^*} + x \right\rfloor\right),$$

and hence, with $y = y(m) := \dfrac{\log(m)}{\theta^*} + x$ we get, using Theorem 2.3:

$$1 - F_\beta\left(\frac{\log m}{\theta^*} + x\right) \underset{m\to\infty}{\sim} \mathbb{P}_\beta\left(S^+ > \left\lfloor \frac{\log(m)}{\theta^*} + x \right\rfloor\right) - \sum_{k<0} \sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma(S^+ > \lfloor y \rfloor - kd) \times \mathbb{P}_\beta(S_{\sigma^-} = kd, A_{\sigma^-} = \gamma).$$

This further leads to

$$\lim_{m\to\infty} \mathbb{P}_\alpha\left(M_{K_m} \leq \frac{\log m}{\theta^*} + x\right) = \lim_{m\to\infty} \exp\left\{-\sum_{\beta\in\mathcal{A}^-} m z_\beta \mathbb{P}_\beta\left(S^+ > \left\lfloor \frac{\log(m)}{\theta^*} + x \right\rfloor\right)\right\} \times \exp\left\{\sum_{k<0} \sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma(S^+ > \lfloor y \rfloor - kd) \cdot \sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{P}_\beta\left(S_{\sigma^-} = kd, A_{\sigma^-} = \gamma\right)\right\}.$$

Since $K_{m(n)} \leq n \leq K_{m(n)+1}$ and $m(n) \longrightarrow \infty$ *a.s.*, Lemma 3.7 implies that $\dfrac{n}{m(n)} \longrightarrow A^*$ *a.s.* Moreover, since $M_{K_{m(n)}} \leq M_n \leq M_{K_{m(n)+1}}$, we finally obtain

$$\lim_{n\to\infty} \mathbb{P}_\alpha\left(M_n \leq \frac{\log n}{\theta^*} + x\right) = \lim_{n\to\infty} \mathbb{P}_\alpha\left(M_{K_{\lfloor n/A^* \rfloor}} \leq \frac{\log n}{\theta^*} + x\right)$$

$$= \lim_{n\to\infty} \exp\left\{-\frac{n}{A^*} \sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{P}_\beta\left(S^+ > \left\lfloor \frac{\log(n)}{\theta^*} + x \right\rfloor\right)\right\}$$

$$\times \exp\left\{\sum_{k<0} \sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma\left(S^+ > \left\lfloor \frac{\log(n)}{\theta^*} + x \right\rfloor - kd\right) \cdot \sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{P}_\beta(S_{\sigma^-} = kd, A_{\sigma^-} = \gamma)\right\}.$$

It remains to prove the stated expression for $A^* := \lim_{m\to+\infty} \frac{K_m}{m}$ *a.s.* in order to finish the proof. Recall that $\sigma^- = K_1$. In Lemma 3.7 we proved that

$$A^* = \sum_\alpha z_\alpha \mathbb{E}_\alpha(\sigma^-).$$

Since $(U_m(\theta))_m$ is a martingale (see Lemma 3.2) and σ^- a stopping time, using the optional sampling theorem we get $\mathbb{E}_\alpha\left[U_{\sigma^-}(\theta)\right] = \mathbb{E}_\alpha\left[U_0(\theta)\right] = 1$. Consequently,

$$\begin{aligned}
1 &= \mathbb{E}_\alpha\left[\exp(\theta\cdot S_{\sigma^-})\frac{u_{A_{\sigma^-}}(\theta)}{u_{A_0}(\theta)}\frac{1}{\rho(\theta)^{\sigma^-}}\right] \\
&= \mathbb{E}_\alpha\left[\exp(\theta\cdot S_{\sigma^-})\frac{u_{A_{\sigma^-}}(\theta)}{u_\alpha(\theta)}\frac{1}{\rho(\theta)^{\sigma^-}}\right] \\
&= \sum_\beta \mathbb{E}_\alpha\left[\exp(\theta\cdot S_{\sigma^-})\frac{u_\beta(\theta)}{u_\alpha(\theta)}\frac{1}{\rho(\theta)^{\sigma^-}}\big|A_{\sigma^-}=\beta\right]\cdot\mathbb{P}_\alpha(A_{\sigma^-}=\beta) \\
&= \sum_\beta \frac{u_\beta(\theta)}{u_\alpha(\theta)}\mathbb{E}_\alpha\left[\frac{\exp(\theta\cdot S_{\sigma^-})}{\rho(\theta)^{\sigma^-}}\big|A_{\sigma^-}=\beta\right]\cdot q_{\alpha\beta}.
\end{aligned}$$

We deduce

$$u_\alpha(\theta) = \sum_\beta \mathbb{E}_\alpha\left[\frac{\exp(\theta\cdot S_{\sigma^-})}{\rho(\theta)^{\sigma^-}}\big|A_{\sigma^-}=\beta\right]\cdot u_\beta(\theta)q_{\alpha\beta}.$$

Derivating the above relation leads to

$$\begin{aligned}
&u'_\alpha(\theta) = \\
&\sum_\beta q_{\alpha\beta}u_\beta(\theta)\mathbb{E}_\alpha\left[\frac{S_{\sigma^-}\exp(\theta\cdot S_{\sigma^-})\rho(\theta)^{\sigma^-} - \exp(\theta\cdot S_{\sigma^-})\sigma^-\rho(\theta)^{\sigma^- -1}\rho'(\theta)}{\rho(\theta)^{2\sigma^-}}\big|A_{\sigma^-}=\beta\right] \\
&\quad + \sum_\beta q_{\alpha\beta}u'_\beta(\theta)\mathbb{E}_\alpha\left[\frac{\exp(\theta\cdot S_{\sigma^-})}{\rho(\theta)^{\sigma^-}}\big|A_{\sigma^-}=\beta\right].
\end{aligned}$$

Since $\rho(0)=1$, we obtain for $\theta=0$:

$$u'_\alpha(0) = \sum_\beta q_{\alpha\beta}u_\beta(0)\left(\mathbb{E}_\alpha\left[S_{\sigma^-}\big|A_{\sigma^-}=\beta\right] - \rho'(0)\mathbb{E}_\alpha\left[\sigma^-\big|A_{\sigma^-}=\beta\right]\right) + \sum_\beta q_{\alpha\beta}u'_\beta(0).$$

By the fact that $u(0) =^t (1/r,\ldots,1/r)$, we further get

$$u'_\alpha(0) = \frac{1}{r}\mathbb{E}_\alpha[S_{\sigma^-}] - \frac{\rho'(0)}{r}\mathbb{E}_\alpha(\sigma^-) + \sum_\beta q_{\alpha\beta}u'_\beta(0).$$

From the last relation we deduce

$$\sum_\alpha z_\alpha u'_\alpha(0) = \frac{1}{r}\sum_\alpha z_\alpha\mathbb{E}_\alpha\left[S_{\sigma^-}\right] - \frac{\rho'(0)}{r}\sum_\alpha z_\alpha\mathbb{E}_\alpha(\sigma^-) + \sum_\alpha\sum_\beta z_\alpha q_{\alpha\beta}u'_\beta(0). \quad (19)$$

On the other hand, since z is the stationnary frequency vector of the matrix $\mathbf{Q}$, we have $z = z\cdot\mathbf{Q}$ and thus

$$\sum_\alpha z_\alpha u'_\alpha(0) =^t z\cdot u'(0) =^t (z\mathbf{Q})\cdot u'(0) = \sum_\beta {}^t(z\mathbf{Q})_\beta\cdot u'_\beta(0) = \sum_\beta\sum_\alpha z_\alpha q_{\alpha\beta}u'_\beta(0). \quad (20)$$

Equations (19) and (20) imply that $\sum_\alpha z_\alpha \mathbb{E}_\alpha [S_{\sigma^-}] = \rho'(0) \cdot \sum_\alpha z_\alpha \mathbb{E}_\alpha(\sigma^-)$ and thus $A^* = \sum_\alpha z_\alpha \mathbb{E}_\alpha(\sigma^-) = \frac{1}{\rho'(0)} \sum_\alpha z_\alpha \mathbb{E}_\alpha [S_{\sigma^-}]$. Using now the fact that $\rho'(0) = \mathbb{E}[f(A)]$ (see Lemma 3.4) gives the stated expression for A^*. □

4. Applications and computational methods

In order to simplify the presentation, we suppose in this section that $d = 1$. Let $-u, \dots, 0, \dots, v$ be the possible scores, with $u, v \in \mathbb{N}$.

For $-u \leq j \leq v$, we introduce the matrix $\mathbf{P^{(j)}}$ with entries

$$P^{(j)}_{\alpha\beta} := \mathbb{P}_\alpha(A_1 = \beta, f(A_1) = j)$$

for $\alpha, \beta \in \mathcal{A}$. Note that $P^{(f(\beta))}_{\alpha\beta} = p_{\alpha\beta}$, $P^{(j)}_{\alpha\beta} = 0$ if $j \neq f(\beta)$ and $\mathbf{P} = \sum_{j=-u}^{v} \mathbf{P^{(j)}}$, where $\mathbf{P} = (p_{\alpha\beta})_{\alpha,\beta}$ is the transition probability matrix of the Markov chain $(A_i)_i$.

In order to obtain the approximate distribution of Q_1 given in Theorem 2.3, we need to compute the quantities $Q^{(\ell)}_{\alpha\beta}$ for $-u \leq \ell \leq v, \alpha, \beta \in \mathcal{A}$. This is the topic of the next subsection. We denote $\mathbf{Q}^{(\ell)}$ the matrix $(Q^{(\ell)}_{\alpha\beta})_{\alpha,\beta\in\mathcal{A}}$.

4.1. Computation of $\mathbf{Q^{(\ell)}}$ for $-u \leq \ell \leq v$, and of $\mathbf{Q}$

Recall that $Q^{(\ell)}_{\alpha\beta} = \mathbb{P}_\alpha(S_{\sigma^-} = \ell, A_{\sigma^-} = \beta)$, and hence $Q^{(\ell)}_{\alpha\beta} = 0$ for $\ell \geq 0$ or $\beta \in \mathcal{A} \setminus \mathcal{A}^-$. Note also that $\sigma^- = 1$ if $f(A_1) < 0$. Let $-u \leq \ell \leq -1$. When decomposing with respect to the possible values j of $f(A_1)$, we obtain:

$$\begin{aligned} Q^{(\ell)}_{\alpha\beta} = \mathbb{P}_\alpha(A_1 = \beta, f(A_1) = \ell) + \mathbb{P}_\alpha(S_{\sigma^-} = \ell, A_{\sigma^-} = \beta, f(A_1) = 0) \\ + \sum_{j=1}^{v} \mathbb{P}_\alpha(S_{\sigma^-} = \ell, A_{\sigma^-} = \beta, f(A_1) = j). \end{aligned}$$

Note that the first term on the right hand side is exactly $P^{(\ell)}_{\alpha\beta}$ defined at the beginning of this section. We further have, by the law of total probability and the Markov property:

$$\begin{aligned} \mathbb{P}_\alpha(S_{\sigma^-} = \ell, A_{\sigma^-} = \beta, f(A_1) = 0) &= \sum_\gamma P^{(0)}_{\alpha\gamma}\ \mathbb{P}_\alpha(S_{\sigma^-} = \ell, A_{\sigma^-} = \beta\ |A_1 = \gamma, f(A_1) = 0) \\ &= \sum_\gamma P^{(0)}_{\alpha\gamma}\ \mathbb{P}_\gamma(S_{\sigma^-} = \ell, A_{\sigma^-} = \beta) = (\mathbf{P^{(0)}}\mathbf{Q}^{(\ell)})_{\alpha\beta}. \end{aligned}$$

Let $j \in \{1, \dots, v\}$ be fixed. We have

$$\mathbb{P}_\alpha(S_{\sigma^-} = \ell, A_{\sigma^-} = \beta, f(A_1) = j) = \sum_\gamma P^{(j)}_{\alpha\gamma}\ \mathbb{P}_\alpha(S_{\sigma^-} = \ell, A_{\sigma^-} = \beta\ |A_1 = \gamma, f(A_1) = j).$$

For every possible $s \geq 1$, we denote $\mathcal{T}_s$ the set of all possible s-tuples $t = (t_1, \ldots, t_s)$ verifying $-u \leq t_i \leq -1$ for $i = 1, \ldots, s$, $t_1 + \cdots + t_{s-1} \geq -j > 0$ and $t_1 + \cdots + t_s = \ell - j > 0$. Decomposing the possible paths from $-k$ to ℓ gives

$$Q_{\alpha\beta}^{(\ell)} = P_{\alpha\beta}^{(\ell)} + (\mathbf{P^{(0)}Q^{(\ell)}})_{\alpha\beta} + \sum_{j=1}^{v} \left(\mathbf{P^{(j)}} \sum_{s} \sum_{t \in \mathcal{T}_s} \prod_{i=1}^{s} \mathbf{Q^{(t_i)}} \right)_{\alpha\beta},$$

hence

$$\mathbf{Q^{(\ell)}} = \mathbf{P^{(\ell)}} + \mathbf{P^{(0)}Q^{(\ell)}} + \sum_{j=1}^{v} \mathbf{P^{(j)}} \sum_{s} \sum_{t \in \mathcal{T}_s} \prod_{i=1}^{s} \mathbf{Q^{(t_i)}}. \tag{21}$$

Recalling that $\mathbf{Q} = (q_{\alpha\beta})_{\alpha,\beta}$ with $q_{\alpha\beta} = \mathbb{P}_\alpha(A_{\sigma^-} = \beta) = \sum_{\ell<0} Q_{\alpha\beta}^{(\ell)}$, we have

$$\mathbf{Q} = \sum_{\ell<0} \mathbf{Q^{(\ell)}}. \tag{22}$$

Example: In the case where $u = v = 1$, we only have the possible values $\ell = -1$, $j = 1$, $s = 2$ and $t_1 = t_2 = -1$, thus

$$\mathbf{Q^{(-1)}} = \mathbf{P^{(-1)}} + \mathbf{P^{(0)}} \cdot \mathbf{Q^{(-1)}} + \mathbf{P^{(1)}}(\mathbf{Q^{(-1)}})^2 \text{ and } \mathbf{Q} = \mathbf{Q^{(-1)}}. \tag{23}$$

4.2. Computation of $L_{\alpha\beta}^{(\ell)}$ for $0 \leq \ell \leq v$, and of $L_\alpha(\infty)$

Recall that $L_{\alpha\beta}^{(\ell)} = \mathbb{P}_\alpha(S_{\sigma^+} = \ell, \sigma^+ < \infty, A_{\sigma^+} = \beta)$. Denote $\mathbf{L^{(\ell)}} := (L_{\alpha\beta}^{(\ell)})_{\alpha,\beta}$. First note that $L_{\alpha\beta}^{(\ell)} = 0$ for $\ell \leq 0$ or $\beta \in \mathcal{A} \setminus \mathcal{A}^+$. Using a similar method as the one used to obtain $Q_{\alpha\beta}^{(\ell)}$ in the previous subsection, we denote for every possible $s \geq 1$, $\mathcal{T}'_s$ the set of all s-tuples $t = (t_1, \ldots, t_s)$ verifying $1 \leq t_i \leq v$ for $i = 1, \ldots, s$, $t_1 + \cdots + t_{s-1} \leq k$ and $t_1 + \cdots + t_s = \ell + k > 0$.

For every $0 < \ell \leq v$ we then have

$$\mathbf{L^{(\ell)}} = \mathbf{P^{(\ell)}} + \mathbf{P^{(0)}L^{(\ell)}} + \sum_{k=1}^{u} \mathbf{P^{(-k)}} \sum_{s} \sum_{t \in \mathcal{T}'_s} \prod_{i=1}^{s} \mathbf{L^{(t_i)}} \tag{24}$$

Since $L_\alpha(\infty) = \mathbb{P}_\alpha(\sigma^+ < \infty) = \sum_\beta \sum_{\ell=1}^{v} L_{\alpha\beta}^{(\ell)}$, and denoting by $\mathbf{L}(\infty)$ the column vector containing all $L_\alpha(\infty)$ for $\alpha \in \mathcal{A}$, and by $\mathbb{1}_r$ the column vector of size r with all components equal to 1, we can write

$$\mathbf{L}(\infty) = \sum_{\ell=1}^{v} \mathbf{L^{(\ell)}} \cdot \mathbb{1}_r. \tag{25}$$

Example: In the case where $u = v = 1$, equation (24) gives

$$\mathbf{L^{(1)}} = \mathbf{P^{(1)}} + \mathbf{P^{(0)}} \cdot \mathbf{L^{(1)}} + \mathbf{P^{(-1)}} \cdot (\mathbf{L^{(1)}})^2, \tag{26}$$

$$\mathbf{L^{(\ell)}} = 0 \text{ for } \ell > 1, \text{ thus } \mathbf{L}(\infty) = \mathbf{L^{(1)}} \cdot \mathbb{1}_r. \tag{27}$$

4.3. Computation of $F_{S^+,\alpha}(\ell)$ for $\ell \geq 0$

For $\ell \geq 0$ let us denote $\mathbf{F}_{S^+,\cdot}(\ell) := (F_{S^+,\alpha}(\ell))_{\alpha\in\mathcal{A}}$, seen as a column vector of size r. From Theorem 2.1 we deduce that for $\ell = 0$ and every $\alpha \in \mathcal{A}$ we have

$$F_{S^+,\alpha}(0) = 1 - L_\alpha(\infty).$$

For $\ell = 1$ and every $\alpha \in \mathcal{A}$ we get

$$F_{S^+,\alpha}(1) = 1 - L_\alpha(\infty) + \sum_{\beta\in\mathcal{A}} L^{(1)}_{\alpha\beta}\ F_{S^+,\beta}(0).$$

With $\mathbf{L}(\infty) = (L_\alpha(\infty))_{\alpha\in\mathcal{A}}$, seen as a column vector, we can write

$$\begin{aligned}
\mathbf{F}_{S^+,\cdot}(1) &= 1 - \mathbf{L}(\infty) + \mathbf{L^{(1)}}\mathbf{F}_{S^+,\cdot}(0),\\
\mathbf{F}_{S^+,\cdot}(\ell) &= 1 - \mathbf{L}(\infty) + \sum_{k=1}^{\ell} \mathbf{L^{(k)}}\mathbf{F}_{S^+,\cdot}(\ell-k),\ \forall \ell \geq 1.
\end{aligned}$$

See Subsection 4.2 for how to compute $\mathbf{L^{(k)}}$ for $k \geq 1$ and $\mathbf{L}(\infty)$.

4.4. Application in a simple case

Let us consider the simple case where the possible score values are $-1, 0, 1$, corresponding to the case $u = v = 1$. We will use the results in the previous subsections (see Equations (23, 26, 27)) to derive the distribution of the maximal non-negative partial sum S^+. This distribution can be determined using the following matrix equalities:

$$\mathbf{L}(\infty) = \left(\sum_{\beta} L^{(1)}_{\alpha\beta}\right)_{\alpha} = \mathbf{L^{(1)}} \cdot \mathbb{1}_r, \tag{28}$$

with $\mathbf{L^{(1)}}$ given in Equation (24) and

$$\mathbf{F}_{S^+,\cdot}(0) = 1 - \mathbf{L}(\infty), \tag{29}$$

$$\mathbf{F}_{S^+,\cdot}(\ell) = 1 - \mathbf{L}(\infty) + \mathbf{L^{(1)}}\mathbf{F}_{S^+,\cdot}(\ell-1). \tag{30}$$

This allows to further derive the approximate distributions of Q_1 and M_n given in Theorems 2.3 and 2.4.

We present hereafter a numerical application for the local score of a DNA sequence. We suppose that we have a Markovian sequence whose possible letters are $\{A, C, G, T\}$

and whose transition probability matrix is given by

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/6 & 1/6 & 1/6 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/6 & 1/6 & 1/6 & 1/2 \\ 1/6 & 1/6 & 1/2 & 1/6 \end{pmatrix}.$$

We choose the respective scores $-1, -1, 0, 1$ for the letters A, C, G, T for which Hypothesis (1) and (2) are verified. We use the successive iteration methodology described in Equation (5.12) of [9] in order to compute $\mathbf{L^{(1)}}$ and $\mathbf{Q^{(-1)}}$, solutions of Equations (23) and (26), from which we derive the formulas proposed in our Theorems 2.1, 2.3 and 2.4 for the approximate distributions of S^+, Q_1 and M_n respectively. We also compute the different approximations proposed in Karlin and Dembo [9]. We then compare these results with the corresponding empirical distributions computed using a Monte Carlo approach based on 10^5 simulations. We can see in Figure 1, left panel, that for $n = 300$ the empirical *cdf* of S^+ and the one obtained using Theorem 2.1 match perfectly. We can also visualize the fact that Theorem 2.1 improves the approximation of Karlin and Dembo in Lemma 4.3 of [9] for the distribution of S^+. The right panel of Figure 1 allows to compare, for different values of the sequence length n, the empirical *cdf* of S^+ and the exact *cdf* given in Theorem 2.1: we can see that our formula performs very satisfactory even for sequence length $n = 100$.

In this simple example the approximation of the distribution of Q_1 given in Theorem 2.3 and the one given in Lemma 4.4 of [9] give quite similar numerical values.

In Figure 2 we compare three approximations for the *cdf* of M_n: the Karlin and Dembo's approximation given in Equation (1.27) of [9] (see also Equation (8)), our approximation proposed in Theorem 2.4, and a Monte Carlo approximation. For the simple scoring scheme of this application, the parameter K^* of the Karlin and Dembo's approximation for M_n is given by Equation (5.6) of [9]

$$K^* = (e^{-\theta^*} - e^{-2\theta^*}) \cdot \mathbb{E}[-f(A)] \cdot \sum_{\gamma} z_\gamma u_\gamma(\theta^*) \cdot \sum_{\gamma} w_\gamma / u_\gamma(\theta^*).$$

More precisely, in the left panel we plot the probability $p(n, x) := \mathbb{P}\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)$ as a function of n, for a fixed value $x = -8$. This illustrates the asymptotic behavior of this probability with growing n. We can also observe the fact that Karlin and

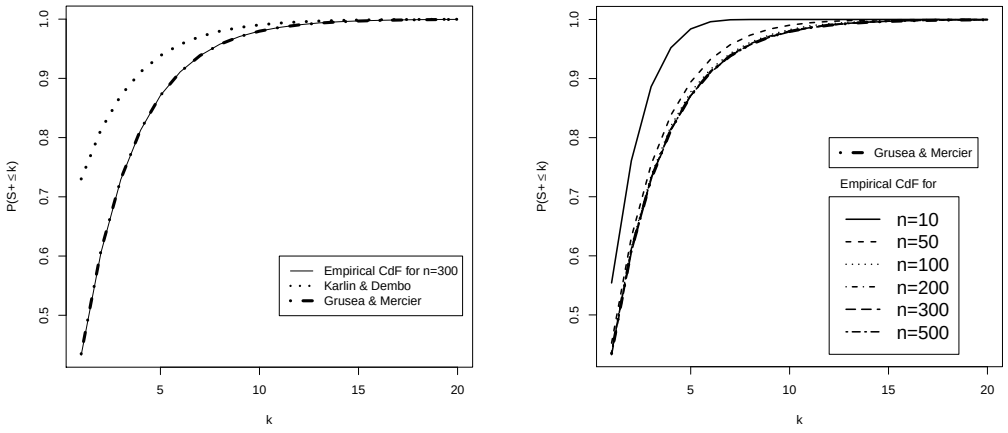

FIGURE 1: Cumulative distribution function of S^+ for the simple scoring scheme $(-1, 0, +1)$ and $A_0 =$"A". Left panel: Comparison between the approximation of Karlin and Dembo proposed in [9], a Monte Carlo estimation with sequences of length $n = 300$, and our exact formula proposed in Theorem 2.1. Right panel: Comparison, for different values of n, of the Monte Carlo empirical cumulative distribution function and the exact one given in Theorem 2.1.

Dembo's approximation does not depend on n. In Figure 2, right panel, we compare the approximation of Karlin and Dembo [9] for the same probability $p(n, x)$ with our approximation, for varying x and fixed $n = 100$. We observe that the improvement brought by our approximation is more significant for negative values of x. For fixed n and extreme deviations (large x) the two approximations are quite similar and accurate.

References

[1] ATHREYA, K. B. AND RAMA MURTHY, K. (1976). Feller's renewal theorem for systems of renewal equations. *J. Indian Inst. Sci.* **58(10),** 437–459.

[2] CELLIER D., CHARLOT, F. AND MERCIER, S. (2003). An improved approximation for assessing the statistical significance of molecular sequence features. J. Appl. Prob., **40,** 427–441.

[3] DEMBO, A. AND KARLIN, S. (1991). Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of markov variables. *Ann. Probab.*, **19(4),** 1756–1767.

[4] Durbin, R. and Eddy, S. and Krogh, A. and Mitchion, G. (1998). *Biological sequence analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press.

[5] FARIELLO M.-I. AND BOITARD S. AND MERCIER S. AND ROBELIN D. AND FARAUT T. AND ARNOULD C. AND LE BIHAN-DUVAL E. AND RECOQUILLAY J. AND SALIN G. AND DAHAIS P. AND PITEL F.

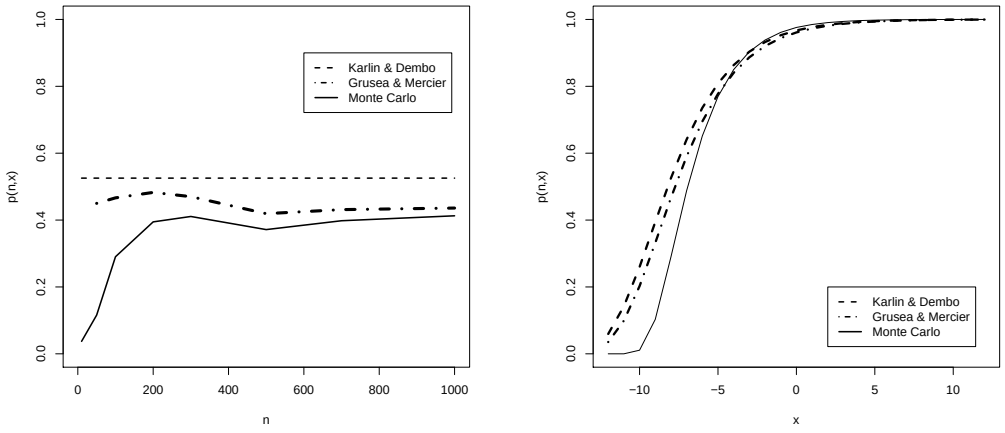

FIGURE 2: Comparison of the different approximations for $p(n,x) = \mathbb{P}\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)$ with the simple scoring scheme $(-1, 0, +1)$: Karlin and Dembo's result [9] (see Equation (8)), our approximation proposed in Theorem 2.4 and Monte Carlo estimation. Left panel: $p(n,x)$ as a function of n, for fixed $x = -8$. Right panel: $p(n,x)$ as a function of x, for fixed $n = 100$.

AND LETERRIER C. AND SANCRISTOBAL M. (2017). A new local score based method applied to behavior-divergent quail lines sequenced in pools precisely detects selection signatures on genes related to autism. *Molecular Ecology* **26(14),** 3700–3714.

[6] GUEDJ, M. AND ROBELIN,D. AND HOEBEKE, M. AND LAMARINE, M. AND WOJCIK, J. AND NUEL, G. (2006). Detecting local high-scoring segments: a first-stage approach for genome-wide association studies,*Stat. Appl. Genet. Mol. Biol.*, **5(1)**.

[7] HASSENFORDER, C. AND MERCIER, S. (2007). Exact Distribution of the Local Score for Markovian Sequences. *Ann. Inst. Stat. Math.*, **59(4),** 741–755.

[8] KARLIN, S. AND ALTSCHUL, S.-F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Nat. Acad. Sci. USA, **87,** 2264–2268.

[9] KARLIN, S. AND DEMBO, A. (1992). Limit distributions of maximal segmental score among Markov-dependent partial sums. Adv. Appl. Prob., **24,** 113–140.

[10] KARLIN, S. AND OST, F. (1987). Counts of long aligned word matches among random letter sequences. Adv. Appl. Prob., **19,** 293–351.

[11] LANCASTER, P. (1969). *Theory of Matrices*, Academic Press, New York.

[12] MERCIER, S. AND DAUDIN, J.J. (2001). Exact distribution for the local score of one i.i.d. random sequence. *J. Comp. Biol.*, **8(4),** 373–380.