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

Simona Grusea, Sabine Mercier

(24 September 2019)

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 excursion above 0 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 a new approximation for the tail behaviour of Q_1, together with an asymptotic equivalence for the distribution of M_n. Computational methods are explicitly presented in a simple application case. Comparison is performed between the new approximations and the ones proposed by Karlin and Dembo [9] in order to evaluate improvements, both in the simple application case and on the real data examples considered in [8].

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

2010 Mathematics Subject Classification: Primary 60J10; 60F05; 60G70
Secondary 60F10; 60K15

* 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

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. 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 [7, 9] and 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, in the case of a non-lattice scoring scheme, the distribution of the local score is asymptotically a Gumble distribution, as in the i.i.d. case. In the lattice case, they give asymptotic lower and upper bounds of Gumbel type for the local score distribution. 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 [5] and [6], 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, given as an asymptotic equivalence when the length of the sequence tends to infinity, for the distribution of the local score of a Markovian sequence. We compare it to the asymptotic bounds of Karlin and Dembo [9] and illustrate the improvements both in a simple application case and on the real data

examples proposed in [8].

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 $\mathbf{P} = (p_{\alpha\beta})_{\alpha,\beta\in\mathcal{A}}$ be its transition probability matrix and π its stationary frequency vector. In this work we suppose that $\mathbf{P}$ is positive ($\forall\alpha,\beta,\ p_{\alpha\beta} > 0$). We also suppose that the initial distribution of A_0 is given by π, hence the Markov chain is stationary. $\mathbb{P}_\alpha$ 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 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.

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

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

$$\text{Hypothesis (2):} \quad \mathbb{P}_\alpha(f(A_1) > 0) > 0. \tag{2}$$

Note that, thanks to the assumption $p_{\alpha\beta} > 0, \forall\alpha,\beta$, Hypothesis (2) is equivalent to the existence of $\beta \in \mathcal{A}$ such that $f(\beta) > 0$.

Let us introduce some definitions and notation. Let $S_0 := 0$ and $S_k := \sum_{i=1}^{k} f(A_i)$, 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 Hypothesis (1), and let

$$Q_1 := \max_{0\leq k<\sigma^-} S_k.$$

First introduced by Karlin and Altschul [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, in order to study the distributions of the variables S^+, Q_1 and M_n, which all take values in $d\mathbb{N}$, it suffices to focus on the case $d = 1$. We will thus consider $d = 1$ throughout the article.

Remark 1.1. Karlin and Dembo [9] consider a more general model, with 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 $X_{\alpha\beta}$. Their model is more general also in that the scores are not restricted to the lattice case and may be continuous random variables.

The framework of this article corresponds to the case where the score function is deterministic and lattice, 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 $\alpha \in \mathcal{A}$ and of some $m \geq 2$ 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.

In Section 2 we first introduce few more definitions and notation. We then 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 a new approximation for the tail behaviour of the height of the first excursion Q_1 in Theorem 2.3. We also establish, in Theorem 2.4, an asymptotic equivalence result for the distribution of the local score M_n when the length n of the sequence tends to infinity. 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. In Subsection 4.5 we also show the improvements brought by the new approximations on the real data examples of [8].

2. Statement of the main results

2.1. Definitions and notation

Let $K_0 := 0$ and for $i \geq 1$, $K_i := \inf\{k > K_{i-1} : S_k - S_{K_{i-1}} < 0\}$ 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 excursion above 0.

Note that, due to the negative drift, we have $\mathbb{E}[K_1] < \infty$ (see Lemma 3.7) and $m(n) \to \infty$ *a.s.* when $n \to \infty$. To every excursion $i = 1, \ldots, m(n)$ we associate its *maximal segmental score* (called also *height*) Q_i defined by

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

Note that $M_n = \max(Q_1, \ldots, Q_{m(n)}, Q^*)$, with Q^* being the maximal segmental score of the last incomplete 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 $M_n = \max_{0\leqslant k\leqslant n} W_k$.

For every $\alpha, \beta \in \mathcal{A}$, we denote $q_{\alpha\beta} := \mathbb{P}_\alpha(A_{K_1} = \beta)$ and $\mathbf{Q} := (q_{\alpha\beta})_{\alpha,\beta\in\mathcal{A}}$. 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, since $q_{\alpha\beta} \geq p_{\alpha\beta} > 0$, $\forall \alpha, \beta \in \mathcal{A}^-$. The states $(A_{K_i})_{i\geq 1}$ of the Markov chain at the end of the successive excursions define a Markov chain on $\mathcal{A}^-$ with transition probability matrix $\tilde{\mathbf{Q}}$.

For every $i \geq 2$ we thus have $\mathbb{P}(A_{K_i} = \beta \mid A_{K_{i-1}} = \alpha) = q_{\alpha\beta}$ 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 $\mathbf{Q}$ 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

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

Note that for any $\alpha \in \mathcal{A}^-$ and $i \geq 1$, $F_{Q_1,\alpha,\beta}$ represents the cumulative distribution function (*cdf*) of the height Q_i of the i-th excursion given that it starts in state α and ends in state β, i.e. $F_{Q_1,\alpha,\beta}(y) = \mathbb{P}(Q_i \leq y \mid A_{K_i} = \beta, A_{K_{i-1}} = \alpha)$, whereas $F_{Q_1,\alpha}$ represents the *cdf* of Q_i conditionally on the i-th excursion starting in state α, i.e. $F_{Q_1,\alpha}(y) = \mathbb{P}(Q_i \leq y \mid A_{K_{i-1}} = \alpha)$. We thus have $F_{Q_1,\alpha}(y) = \sum_{\beta\in\mathcal{A}^-} F_{Q_1,\alpha,\beta}(y) q_{\alpha\beta}$.

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 $\beta \in \mathcal{A} \setminus \mathcal{A}^+$, and $L_{\alpha\beta}(\infty) \leq \mathbb{P}_\alpha(\sigma^+ < \infty) < 1$, therefore $\int_0^\infty dL_{\alpha\beta}(\xi) = L_{\alpha\beta}(\infty) < 1$.

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 $\mathbf{\Phi}(\theta) := \left(p_{\alpha\beta} \cdot \exp(\theta f(\beta))\right)_{\alpha,\beta\in\mathcal{A}}$. Since the transition matrix $\mathbf{P}$ is positive, by the Perron-Frobenius Theorem, the spectral radius $\rho(\theta) > 0$ of the matrix $\mathbf{\Phi}(\theta)$ 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 [3, 10, 11]). In Lemma 3.5 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: $F_{S^+,\alpha}(\ell) := \mathbb{P}_\alpha(S^+ \leq \ell)$, $\forall \ell \in \mathbb{N}$ and for every

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

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

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 $(F_{S^+,\alpha}(\ell))_{\alpha,\ell}$, involving the coefficients $L_{\alpha\beta}^{(k)}$ which can be computed recursively (see Subsection 4.2).

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

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

The proof will be given in Section 3.

In Theorem 2.2 we obtain an asymptotic result for the tail behavior of S^+ using Theorem 2.1 and ideas inspired from [9] 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 $k \in \mathbb{N}$ we denote

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

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. In Subsection 4.3 we 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 $L_{\alpha\beta}^{(k)}$ terms different from zero. Therefore, there are a finite number of non-null terms in the sum defining $G_{\alpha\beta}(\infty)$.

The following result is the analogous, in our settings, of Lemma 4.3 of Karlin and Dembo [9].

Theorem 2.2. (Asymptotics for the tail behaviour of S^+.) *For every $\alpha \in \mathcal{A}$ we have*

$$\lim_{k\to+\infty} \frac{e^{\theta^* k}\mathbb{P}_\alpha(S^+ > k)}{u_\alpha(\theta^*)} = \frac{1}{c}\cdot \sum_{\gamma\in\mathcal{A}^+} \frac{w_\gamma}{u_\gamma(\theta^*)} \sum_{\ell\geq 0}(L_\gamma(\infty) - L_\gamma(\ell))e^{\theta^*\ell} := c(\infty), \qquad (4)$$

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

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

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 $c(\infty)$:

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

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

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

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

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

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

Using Theorem 2.2 we obtain the following result, where the notation $f_k \underset{k\to\infty}{\sim} g_k$ means $f_k - g_k = o(g_k)$, or equivalently $\dfrac{f_k}{g_k} \underset{k\to\infty}{\longrightarrow} 1$.

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

$$\mathbb{P}_\alpha(Q_1 > k) \underset{k\to\infty}{\sim} \mathbb{P}_\alpha(S^+ > k) - \sum_{\ell<0} \sum_{\beta\in\mathcal{A}^-} \mathbb{P}_\beta\left(S^+ > k-\ell\right) \cdot Q^{(\ell)}_{\alpha\beta}. \tag{5}$$

The proof will be given in Section 3.

Remark 2.5. Note that, as a straightforward consequence of Theorems 2.2 and 2.3, we recover the following limit result of Karlin and Dembo [9] (Lemma 4.4):

$$\lim_{k\to+\infty} \frac{e^{\theta^* k}\mathbb{P}_\alpha(Q_1 > k)}{u_\alpha(\theta^*)} = c(\infty)\left\{1 - \sum_{\beta\in\mathcal{A}^-} \frac{u_\beta(\theta^*)}{u_\alpha(\theta^*)} \sum_{\ell<0} e^{\theta^*\ell} Q^{(\ell)}_{\alpha\beta}\right\}.$$

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 $\alpha \in \mathcal{A}$ and $x \in \mathbb{R}$ we have:*

$$\mathbb{P}_\alpha\left(M_n \le \frac{\log(n)}{\theta^*} + x\right) \underset{n\to\infty}{\sim} \exp\left\{-\frac{n}{A^*} \sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{P}_\beta\left(S^+ > \lfloor \log(n)/\theta^* + x \rfloor\right)\right\}$$
$$\times \exp\left\{\frac{n}{A^*} \sum_{k<0} \sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma\left(S^+ > \lfloor \log(n)/\theta^* + x \rfloor - k\right) \cdot \sum_{\beta\in\mathcal{A}^-} z_\beta Q^{(k)}_{\beta\gamma}\right\}, \tag{6}$$

where $z = (z_\alpha)_{\alpha\in\mathcal{A}}$ is the invariant probability measure of the matrix $\mathbf{Q}$ defined in Subsection 2.1 and $A^* := \lim_{m\to+\infty} \dfrac{K_m}{m} = \dfrac{1}{\mathbb{E}(f(A))} \sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{E}_\beta[S_{\sigma^-}]$ a.s.

Remark 2.6.

- Note that the asymptotic equivalent in Equation (6) does not depend on the initial state α.
- We recall, for comparison, the asymptotic lower and upper bounds of Karlin and Dembo [9] for the distribution of M_n:

$$\liminf_{n\to+\infty} \mathbb{P}_\alpha\left(M_n \le \frac{\log(n)}{\theta^*} + x\right) \ge \exp\left\{-K^+ \exp(-\theta^* x)\right\}, \tag{7}$$

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

with $K^+ = K^* \exp(\theta^*)$ and $K^* = v(\infty) \cdot c(\infty)$, where $c(\infty)$ is given in Theorem 2.2 and 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.

- Even if the expression of our asymptotic equivalent in Equation (6) seems more cumbersome than the asymptotic bounds of Karlin and Dembo recalled in Equations (7) and (8), the practical implementations are equivalent.

3. Proofs of the main results

3.1. Proof of Theorem 2.1

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

It then suffices to note that

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

by the strong Markov property applied to the stopping time σ^+. □

3.2. Proof of Theorem 2.2

We first prove some preliminary lemmas.

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

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

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

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

$$\forall \ell > 0, \forall \alpha \in \mathcal{A}, \quad b_\alpha(\ell) = a_\alpha(\ell) + \sum_{\beta} \sum_{k=0}^{\ell} b_\beta(\ell - k) G_{\alpha\beta}^{(k)}. \tag{9}$$

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 (4) for the limit $c(\infty)$ of $b_\alpha(\ell)$ when $\ell \to \infty$, which implies $\lim_{k\to\infty} \mathbb{P}_\alpha(S^+ > k) = 0$.

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

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

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

$$\mathbb{P}_\alpha(S^+ > \ell) = \sum_{\beta\in\mathcal{A}^+} \sum_{k=1}^{\infty} L_{\alpha\beta}^{(k)} \ \mathbb{P}_\beta(S^+ > \ell - k). \tag{10}$$

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 $\mathbb{P}_\beta(S^+ > \ell - k) \longrightarrow 0$ 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. For every $m \in \mathbb{N}$ and $\theta > 0$, $U_m(\theta)$ is clearly measurable with respect to $\mathcal{F}_m$ and integrable, since $\mathcal{A}$ is finite. We can write

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

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 $u(\theta)$,

$$\mathbb{E}[\exp(\theta f(A_{m+1}))u_{A_{m+1}}(\theta)|A_m = \alpha] = \sum_{\beta} \exp(\theta f(\beta))u_\beta(\theta)p_{\alpha\beta} = u_\alpha(\theta)\rho(\theta).$$

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*

$$\forall \alpha \in \mathcal{A}: \qquad \frac{1}{u_\alpha(\theta^*)} \sum_{\beta\in\mathcal{A}^+} \sum_{\ell=1}^{\infty} L_{\alpha\beta}^{(\ell)}\ e^{\theta^*\ell}\ u_\beta(\theta^*) = 1. \tag{11}$$

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

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

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

$$\begin{aligned}\mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n] = \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n,\ \exists k \le n : S_k \le -2a] \\ + \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, -2a \le S_k \le 0,\ \forall 0 \le k \le n].\end{aligned} \tag{12}$$

The first expectation in the right-hand side of Equation (12) can further be bounded:

$$\begin{aligned}\mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n,\ \exists k \le n : S_k \le -2a] \le \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n \le -a] \\ + \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n > -a,\ \exists k < n : S_k \le -2a].\end{aligned} \tag{13}$$

We obviously have

$$\mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n \leq -a] \leq \exp(-\theta^* a). \tag{14}$$

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 $S_n/n \longrightarrow \mathbb{E}[f(A)] < 0$ *a.s.* 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

$$\limsup_{n\to\infty} \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n > -a,\ \exists k < n : S_k \leq -2a] \leq \sum_{\beta \in \mathcal{A}^-} \mathbb{P}_\beta(S^+ > a). \tag{15}$$

Considering the second expectation in the right-hand side of Equation (12), we have

$$\lim_{n\to\infty} \mathbb{P}_\alpha(-2a \leq S_k \leq 0,\ \forall 0 \leq k \leq n) = \mathbb{P}_\alpha(-2a \leq S_k \leq 0,\ \forall k \geq 0) = 0, \tag{16}$$

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

Equations (12),(13),(14),(15) and (16) 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 $\lim_{n\to\infty} \mathbb{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n] = 0$. □

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 prove the stated formula using Theorem 2.1. Equation (10) implies the formula in Equation (9).

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 *Step 1* of the proof of Lemma 3.1, we have

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

We finally obtain

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

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

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

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

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

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

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

By Theorem 2.2 we have $\mathbb{P}_\beta(S^+ > k) = O(e^{-\theta^* k})$ as $k \to \infty$, for every $\beta \in \mathcal{A}^-$, from which we deduce that the left-hand side of the previous equation is $o(\mathbb{P}_\alpha(Q_1 > k))$ 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. *There exists a constant $C > 0$ such that, for every $\alpha \in \mathcal{A}$, $\beta \in \mathcal{A}^-$ and $y > 0$, we have $\mathbb{P}_\alpha(Q_1 > y | A_{\sigma^-} = \beta) \leq Ce^{-\theta^* y}$.*

Proof. The proof is partly inspired from [9]. Let $y > 0$ and denote $\sigma(y)$ the first exit time of S_n from the interval $[0, y]$. Applying the optional sampling theorem to the martingale $(U_m(\theta^*))_m$ (see Lemma 3.2) and to the stopping time $\sigma(y)$, we get

$$\mathbb{E}_\alpha \left[U_{\sigma(y)}(\theta^*)\right] = \mathbb{E}_\alpha \left[U_0(\theta^*)\right] = 1. \tag{17}$$

The applicability of the optional sampling theorem is guaranteed by the fact that there exists $\tilde{C} > 0$ such that, for every $n \in \mathbb{N}$, we have $0 < U_{\min(\sigma(y),n)}(\theta^*) \leq \tilde{C}$ *a.s.* Indeed, this follows from the fact that, when $\sigma(y) > n$ we have $0 \leq S_n \leq y$, and when $\sigma(y) \leq n$, either $S_{\sigma(y)} < 0$ or $y < S_{\sigma(y)} < y + \max\{f(\alpha) : \alpha \in \mathcal{A}^+\}$.

We deduce from Equation (17) that, for some constant $K > 0$, we have:

$$1 = \mathbb{E}_\alpha \left[e^{\theta^* S_{\sigma(y)}} \frac{u_{A_{\sigma(y)}}(\theta^*)}{u_{A_0}(\theta^*)} \right] \geq K e^{\theta^* y}\ \mathbb{E}_\alpha \left[e^{\theta^*(S_{\sigma(y)} - y))} \middle| S_{\sigma(y)} > y \right] \cdot \mathbb{P}_\alpha(S_{\sigma(y)} > y)$$
$$\geq K e^{\theta^* y}\ \mathbb{P}_\alpha(S_{\sigma(y)} > y) \geq K e^{\theta^* y}\ \mathbb{P}_\alpha(S_{\sigma(y)} > y | A_{\sigma^-} = \beta) q_{\alpha\beta}.$$

Note further that, $\mathcal{A}$ being finite, there exists $c > 0$ such that for all $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{A}^-$ we have $q_{\alpha\beta} = \mathbb{P}_\alpha(A_{\sigma^-} = \beta) \geq p_{\alpha\beta} \geq c$. In order to obtain the bound in the statement, it remains to note that $\mathbb{P}_\alpha(Q_1 > y | A_{\sigma^-} = \beta) = \mathbb{P}_\alpha(S_{\sigma(y)} > y | A_{\sigma^-} = \beta)$. □

Lemma 3.5. *We have* $\rho'(0) = \mathbb{E}[f(A)] < 0$.

Proof. By the fact that $\rho(\theta)$ is an eigenvalue of the matrix $\mathbf{\Phi}(\theta)$ with corresponding eigenvector $u(\theta)$, we have $\rho(\theta)u_\alpha(\theta) = (\mathbf{\Phi}(\theta)u(\theta))_\alpha = \sum_\beta p_{\alpha\beta} e^{\theta f(\beta)} u_\beta(\theta)$.

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

$$\sum_\alpha \pi_\alpha \frac{d}{d\theta}(\rho(\theta)u_\alpha(\theta))\bigg|_{\theta=0} = \frac{1}{r}\mathbb{E}[f(A)] + \sum_{\alpha,\beta} \pi_\alpha p_{\alpha\beta} u'_\beta(0) = \frac{1}{r}\mathbb{E}[f(A)] + \sum_\beta \pi_\beta u'_\beta(0). \tag{18}$$

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

$$\sum_\alpha \pi_\alpha \frac{d}{d\theta}(\rho(\theta)u_\alpha(\theta))\bigg|_{\theta=0} = \frac{\rho'(0)}{r} + \rho(0) \cdot \sum_\alpha \pi_\alpha u'_\alpha(0). \tag{19}$$

From Equations (18) and (19) 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 3.6. *There exists $n_0 \geq 0$ such that $\forall n \geq n_0$ and $\forall \alpha \in \mathcal{A}$ we have*

$$\mathbb{P}_\alpha(S_n \geq 0) \leq \left(\inf_{\theta\in\mathbb{R}^+} \rho(\theta)\right)^n, \text{ with } 0 < \inf_{\theta\in\mathbb{R}^+} \rho(\theta) < 1.$$

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 $\rho'(0) = \mathbb{E}[f(A)] < 0$ (by Lemma 3.5), 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 3.5), the strict convexity of ρ and the fact that $\rho(0) = \rho(\theta^*) = 1$, we finally obtain $\mathcal{I} = -\log\left(\inf_{\theta\in\mathbb{R}^+}\rho(\theta)\right) > -\log\rho(0) = 0$. □

Lemma 3.7. *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 $n_0 \in \mathbb{N}$ defined in Lemma 3.6, using a well-known alternative formula for the expectation, we get

$$\mathbb{E}_\alpha[K_1] = \sum_{n\geq 0} \mathbb{P}_\alpha(K_1 > n) \leq \sum_{n\geq 0} \mathbb{P}(S_n \geq 0) \leq C + \sum_{n\geq n_0}\left(\inf_{\theta\in\mathbb{R}^+}\rho(\theta)\right)^n,$$

where $C > 0$ is a constant and $0 < \inf_{\theta\in\mathbb{R}^+}\rho(\theta) < 1$. The statement easily follows. □

Lemma 3.8. *The sequence $\left(\frac{K_m}{m}\right)_{m\geq 1}$ converges a.s. when $m \to \infty$. Therefore, $A^* := \lim_{m\to\infty}\frac{K_m}{m}$ appearing in the statement of Theorem 2.4 is well defined. Moreover, we have $A^* = \sum_\beta z_\beta \mathbb{E}_\beta[K_1]$ a.s.*

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

$$\frac{K_m}{m} = \frac{K_1}{m} + \frac{1}{m}\sum_{i=2}^m (K_i - K_{i-1}) = \frac{K_1}{m} + \sum_\beta \frac{1}{m}\sum_{i=2}^m (K_i - K_{i-1})\mathbf{1}_{\{A_{K_{i-1}}=\beta\}}. \quad (20)$$

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, the distribution of $K_i - K_{i-1}$ depends only on $A_{K_{i-1}}$, and $\mathbb{P}(K_i - K_{i-1} = \ell \mid A_{K_{i-1}} = \alpha) = \mathbb{P}_\alpha(K_1 = \ell)$. 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 subset $\mathcal{A}^-$ is irreducible. Since z is invariant for $\mathbf{Q}$, we easily deduce that $\sum_{\alpha,k} \pi(\alpha, k) \cdot q_{\alpha\beta}\mathbb{P}_\beta(K_1 = \ell) = \pi(\beta, \ell)$, and hence the Markov chain $(Y_i)_i$ is also irreducible, with invariant distribution $\pi(\alpha, k) := z_\alpha \mathbb{P}_\alpha(K_1 = k)$.

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

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

Taking the sum over β and using Equation (20) gives the result in the statement. □

Proof of Theorem 2.4:

Step 1: The proof of this step is partly inspired from [9]. We will prove that for any convergent sequence $(x_m)_m$ we have

$$\mathbb{P}_\alpha\left(M_{K_m} \leq \frac{\log(m)}{\theta^*} + x_m\right) \underset{m\to\infty}{\sim} \exp\left\{-m \sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{P}_\beta\left(S^+ > \lfloor \log(m)/\theta^* + x_m \rfloor\right)\right\}$$
$$\times \exp\left\{m \sum_{k<0}\sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma\left(S^+ > \lfloor \log(m)/\theta^* + x_m \rfloor - k\right) \cdot \sum_{\beta\in\mathcal{A}^-} z_\beta Q^{(k)}_{\beta\gamma}\right\}.$$

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, for any $y > 0$:

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

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 $\psi_{\beta\gamma}(m)$ equals either 0 or $1/m$, 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 $\beta,\gamma \in \mathcal{A}$.

We will further use a Taylor series expansion of the log function. Let us denote $d_{\beta\gamma}(m) := m\left[1 - F_{\beta\gamma}\left(\frac{\log(m)}{\theta^*} + x_m\right)\right]$ for every $m \geq 1$. Thanks to Lemma 3.4, $d_{\beta\gamma}(m)$ are uniformly bounded in m, β and γ. Since, $0 \leq \psi_{\beta\gamma}(m) \leq 1$, we obtain

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

Since

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

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

But

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

and using Theorem 2.3 we get:

$$\begin{aligned}1 - F_\beta\left(\frac{\log(m)}{\theta^*} + x_m\right) &\underset{m\to\infty}{\sim} \mathbb{P}_\beta\left(S^+ > \lfloor \log(m)/\theta^* + x_m \rfloor\right) \\ &\quad - \sum_{k<0}\sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma\left(S^+ > \lfloor \log(m)/\theta^* + x_m \rfloor - k\right) \cdot Q^{(k)}_{\beta\gamma}.\end{aligned}$$

This further leads to

$$\begin{aligned}&\mathbb{P}_\alpha\left(M_{K_m} \leq \frac{\log(m)}{\theta^*} + x_m\right) \underset{m\to\infty}{\sim} \exp\left\{-m\sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{P}_\beta\left(S^+ > \lfloor \log(m)/\theta^* + x_m \rfloor\right)\right\} \\ &\times \exp\left\{m\sum_{k<0}\sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma\left(S^+ > \lfloor \log(m)/\theta^* + x_m \rfloor - k\right) \cdot \sum_{\beta\in\mathcal{A}^-} z_\beta Q^{(k)}_{\beta\gamma}\right\}.\end{aligned}$$

Step 2: We now deduce the stated asymptotic equivalent for the distribution of M_n. Since going from the distribution of M_{K_m} to the distribution of M_n is more delicate in our case than in [9], we present in details the proof of this step.

Let $x \in \mathbb{R}$. Since $K_{m(n)} \leq n \leq K_{m(n)+1}$ and $(M_n)_n$ is non decreasing, we have

$$\mathbb{P}_\alpha\left(M_{K_{m(n)+1}} \leq \frac{\log(n)}{\theta^*} + x\right) \leq \mathbb{P}_\alpha\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right) \leq \mathbb{P}_\alpha\left(M_{K_{m(n)}} \leq \frac{\log(n)}{\theta^*} + x\right). \tag{21}$$

Since $m(n) \longrightarrow \infty$ *a.s.*, Lemma 3.8 implies that $\dfrac{m(n)}{n} \longrightarrow \dfrac{1}{A^*}$ *a.s.*, with $A^* = \lim_{m\to\infty} \frac{K_m}{m}$.

Fix now $\varepsilon > 0$. We have

$$\begin{aligned}
&\mathbb{P}_\alpha\left(M_{K_{m(n)}} \leq \frac{\log(n)}{\theta^*} + x\right) \\
&\leq \mathbb{P}_\alpha\left(M_{K_{m(n)}} \leq \frac{\log(n)}{\theta^*} + x,\ \left|\frac{m(n)}{n} - \frac{1}{A^*}\right| \leq \varepsilon\right) + \mathbb{P}_\alpha\left(\left|\frac{m(n)}{n} - \frac{1}{A^*}\right| > \varepsilon\right) \\
&\leq \mathbb{P}_\alpha\left(M_{K_{\lceil n/A^* - n\varepsilon\rceil}} \leq \frac{\log(n)}{\theta^*} + x\right) + \mathbb{P}_\alpha\left(\left|\frac{m(n)}{n} - \frac{1}{A^*}\right| > \varepsilon\right).
\end{aligned} \tag{22}$$

Using the result of *Step 1*, we obtain

$$\frac{\mathbb{P}_\alpha\left(M_{K_{\lceil n/A^* - n\varepsilon\rceil}} \leq \frac{\log(n)}{\theta^*} + x\right)}{E_n} \underset{n\to\infty}{\sim} R_n(\varepsilon), \tag{23}$$

where E_n is the asymptotic equivalent given in the statement

$$\begin{aligned}
E_n := \exp&\left\{-\frac{n}{A^*}\sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{P}_\beta\left(S^+ > \lfloor \log(n)/\theta^* + x\rfloor\right)\right\} \\
&\times \exp\left\{\frac{n}{A^*}\sum_{k<0}\sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma\left(S^+ > \lfloor \log(n)/\theta^* + x\rfloor - k\right)\cdot \sum_{\beta\in\mathcal{A}^-} z_\beta Q^{(k)}_{\beta\gamma}\right\}
\end{aligned}$$

and

$$\begin{aligned}
R_n(\varepsilon) := \exp&\left\{\varepsilon\cdot n\sum_{\beta\in\mathcal{A}^-} z_\beta \mathbb{P}_\beta\left(S^+ > \lfloor \log(n)/\theta^* + x\rfloor\right)\right\} \\
&\times \exp\left\{-\varepsilon\cdot n\sum_{k<0}\sum_{\gamma\in\mathcal{A}^-} \mathbb{P}_\gamma\left(S^+ > \lfloor \log(n)/\theta^* + x\rfloor - k\right)\cdot \sum_{\beta\in\mathcal{A}^-} z_\beta Q^{(k)}_{\beta\gamma}\right\}.
\end{aligned}$$

Using Theorem 2.2 we obtain

$$\limsup_{n\to\infty} R_n(\varepsilon) \leq \exp\left\{\varepsilon\cdot c(\infty) e^{-\theta^* x} D^*\right\}, \tag{24}$$

with

$$D^* := e^{\theta^*} \sum_{\beta \in \mathcal{A}^-} z_\beta u_\beta(\theta^*) - \sum_{\beta,\gamma \in \mathcal{A}^-} z_\beta u_\gamma(\theta^*) \sum_{k<0} e^{k\theta^*} Q^{(k)}_{\beta\gamma}.$$

Equations (21), (22), (23) and (24), together with the fact that $\frac{m(n)}{n} \longrightarrow \frac{1}{A^*}$ *a.s.* imply that

$$\limsup_{n\to\infty} \frac{\mathbb{P}_\alpha\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)}{E_n} \leq \exp\left\{\varepsilon \cdot c(\infty) e^{-\theta^* x} D^*\right\}. \tag{25}$$

In a similar manner, we can show that

$$\liminf_{n\to\infty} \frac{\mathbb{P}_\alpha\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)}{E_n} \geq \exp\left\{-\varepsilon \cdot c(\infty) e^{-\theta^* x} G^*\right\}, \tag{26}$$

with

$$G^* := \sum_{\beta \in \mathcal{A}^-} z_\beta u_\beta(\theta^*) - e^{\theta^*} \sum_{\beta,\gamma \in \mathcal{A}^-} z_\beta u_\gamma(\theta^*) \sum_{k<0} e^{k\theta^*} Q^{(k)}_{\beta\gamma}.$$

Taking now the limit $\varepsilon \to 0$ in Equations (25) and (26) gives

$$1 \leq \liminf_{n\to\infty} \frac{\mathbb{P}_\alpha\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)}{E_n} \leq \limsup_{n\to\infty} \frac{\mathbb{P}_\alpha\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)}{E_n} \leq 1,$$

and hence $\mathbb{P}_\alpha\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right) \underset{n\to\infty}{\sim} E_n$, with E_n the asymptotic equivalent given in the statement.

Step 3: The last step is to prove the stated expression for A^*. Recall that $\sigma^- = K_1$. In Lemma 3.8 we proved that $A^* = \sum_\alpha z_\alpha \mathbb{E}_\alpha(\sigma^-)$.

Let $n \in \mathbb{N}$. By applying the optional sampling theorem to the martingale $(U_m(\theta))_m$ and to the bounded stopping time $\min(\sigma^-, n)$, we get $\mathbb{E}_\alpha\left[U_{\min(\sigma^-,n)}(\theta)\right] = \mathbb{E}_\alpha\left[U_0(\theta)\right] = 1$. Furthermore, we have

$$1 = \mathbb{E}_\alpha\left[U_{\sigma^-}(\theta); \sigma^- \leq n\right] + \mathbb{E}_\alpha\left[U_n(\theta); \sigma^- > n\right]. \tag{27}$$

We will show that $\mathbb{E}_\alpha\left[U_n(\theta); \sigma^- > n\right] \longrightarrow 0$ when $n \to \infty$. It suffices to prove that $\mathbb{E}_\alpha\left[\frac{e^{\theta S_n}}{\rho(\theta)^n}; \sigma^- > n\right] \longrightarrow 0$. By the Cauchy-Schwartz inequality, we have

$$\mathbb{E}_\alpha\left[\frac{e^{\theta S_n}}{\rho(\theta)^n}; \sigma^- > n\right] \leq \left(\mathbb{E}_\alpha\left[e^{2\theta S_n}\right]\right)^{1/2} \left(\frac{\mathbb{P}_\alpha(\sigma^- > n)}{\rho(\theta)^{2n}}\right)^{1/2}.$$

Further, using Theorem 2.2, we can easily see that $\mathbb{E}_\alpha\left[e^{2\theta S^+}\right] < \infty$ if $0 \leq \theta < \frac{\theta^*}{2}$. Moreover, by Lemma 3.6, we have $\mathbb{P}_\alpha(\sigma^- > n) \leq \mathbb{P}_\alpha(S_n \geq 0) \leq \left(\inf_{\tilde{\theta} \in \mathbb{R}^+} \rho(\tilde{\theta})\right)^n$.

Since $\rho(\theta) \to 1$ when $\theta \to 0$, for sufficiently small θ we will both have $\theta < \dfrac{\theta^*}{2}$ and $\rho(\theta)^2 > \inf_{\theta \in \mathbb{R}^+} \rho(\theta)$, implying that $\mathbb{E}_\alpha \left[\dfrac{e^{\theta S_n}}{\rho(\theta)^n}; \sigma^- > n \right] \longrightarrow 0$ when $n \to \infty$.

When passing to the limit as $n \to \infty$ in Equation (27), we deduce that, for θ sufficiently small, we have $\mathbb{E}_\alpha [U_{\sigma^-}(\theta)] = \mathbb{E}_\alpha [U_0(\theta)] = 1$. 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 that, for θ sufficiently small, we have

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

For θ sufficiently small, by derivating the above relation, we obtain:

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

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

$$\sum_\alpha z_\alpha u'_\alpha(0) = \frac{1}{r} \sum_\alpha z_\alpha \mathbb{E}_\alpha [S_{\sigma^-}] - \frac{\rho'(0)}{r} \sum_\alpha z_\alpha \mathbb{E}_\alpha[\sigma^-] + \sum_\alpha \sum_\beta z_\alpha q_{\alpha\beta} u'_\beta(0). \quad (28)$$

On the other hand, since z is invariant for $\mathbf{Q}$, we obtain

$$\sum_\alpha z_\alpha u'_\alpha(0) =^t z \cdot u'(0) =^t (z\mathbf{Q}) \cdot u'(0) = \sum_\beta {}^t(z\mathbf{Q})_\beta \cdot u'_\beta(0) = \sum_\beta \sum_\alpha z_\alpha q_{\alpha\beta} u'_\beta(0). \quad (29)$$

Equations (28) and (29) 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 3.5) gives the stated expression for A^*. □

4. Applications and computational methods

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 asymptotic result on the tail 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 $Q^{(\ell)}_{\alpha\beta} = 0$ if $\ell \geq 0$ or $\beta \in \mathcal{A}\backslash\mathcal{A}^-$. 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:

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

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 over all the possible paths from $-j$ 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

$$\mathbf{Q^{(\ell)}} = \mathbf{P^{(\ell)}} + \mathbf{P^{(0)}Q^{(\ell)}} + \sum_{j=1}^{v} \mathbf{P^{(j)}} \sum_{s} \sum_{t \in \mathcal{T}_s} \prod_{i=1}^{s} \mathbf{Q^{(t_i)}}. \tag{30}$$

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

$$\mathbf{Q} = \sum_{\ell<0} \mathbf{Q^{(\ell)}}. \tag{31}$$

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

$$\mathbf{Q^{(-1)}} = \mathbf{P^{(-1)}} + \mathbf{P^{(0)}} \cdot \mathbf{Q^{(-1)}} + \mathbf{P^{(1)}}(\mathbf{Q^{(-1)}})^2 \text{ and } \mathbf{Q} = \mathbf{Q^{(-1)}}. \tag{32}$$

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

$$\mathbf{L^{(\ell)}} = \mathbf{P^{(\ell)}} + \mathbf{P^{(0)}L^{(\ell)}} + \sum_{k=1}^{u} \mathbf{P^{(-k)}} \sum_{s} \sum_{t \in \mathcal{T}'_s} \prod_{i=1}^{s} \mathbf{L^{(t_i)}} \tag{33}$$

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

$$\mathbf{L}(\infty) = \sum_{\ell=1}^{v} \mathbf{L^{(\ell)}} \cdot \mathbb{1}_r. \tag{34}$$

Example: In the case where $u = v = 1$, equation (33) gives

$$\mathbf{L^{(1)}} = \mathbf{P^{(1)}} + \mathbf{P^{(0)}} \cdot \mathbf{L^{(1)}} + \mathbf{P^{(-1)}} \cdot (\mathbf{L^{(1)}})^2, \tag{35}$$

$$\mathbf{L^{(\ell)}} = 0 \text{ for } \ell > 1, \text{ thus } \mathbf{L}(\infty) = \mathbf{L^{(1)}} \cdot \mathbb{1}_r. \tag{36}$$

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. Numerical 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 (32, 35, 36)) to derive the distribution of the maximal non-negative partial sum S^+. This distribution can be determined using the following matrix equalities:

$$\mathbf{L}(\infty) = \left(\sum_\beta L^{(1)}_{\alpha\beta}\right)_\alpha = \mathbf{L^{(1)}} \cdot \mathbb{1}_r, \tag{37}$$

with $\mathbf{L^{(1)}}$ given in Equation (33) and

$$\begin{aligned}
\mathbf{F}_{S^+,\cdot}(0) &= 1 - \mathbf{L}(\infty), &(38)\\
\mathbf{F}_{S^+,\cdot}(\ell) &= 1 - \mathbf{L}(\infty) + \mathbf{L^{(1)}}\mathbf{F}_{S^+,\cdot}(\ell-1). &(39)
\end{aligned}$$

This allows to further derive the approximation results on the 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 Hypotheses (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 (32) and (35), from which we derive the approximate formulas proposed in our Theorems 2.1, 2.3 and 2.4 for the 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^+ (see Theorem 2.2 for the analogous formula in our settings). 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 in this example, even for sequence length $n = 100$.

In this simple example, the approximate formula for the tail 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 Figures 2 and 3 we compare three approximations for the *cdf* of M_n: the Karlin and Dembo's [9] asymptotic bounds (the lower bound, depending on K^+ and recalled in Equation (7), and the upper bound, depending on K^* and recalled in Equation (8)), our approximation proposed in Theorem 2.4, and a Monte Carlo estimation. For the simple scoring scheme of this application, the parameter K^* appearing in the asymptotic bounds of Karlin and Dembo [9] is given by their Equation (5.6):

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

More precisely, in Figure 2 we plot the probability $p(n, x) := \mathbb{P}\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)$ as

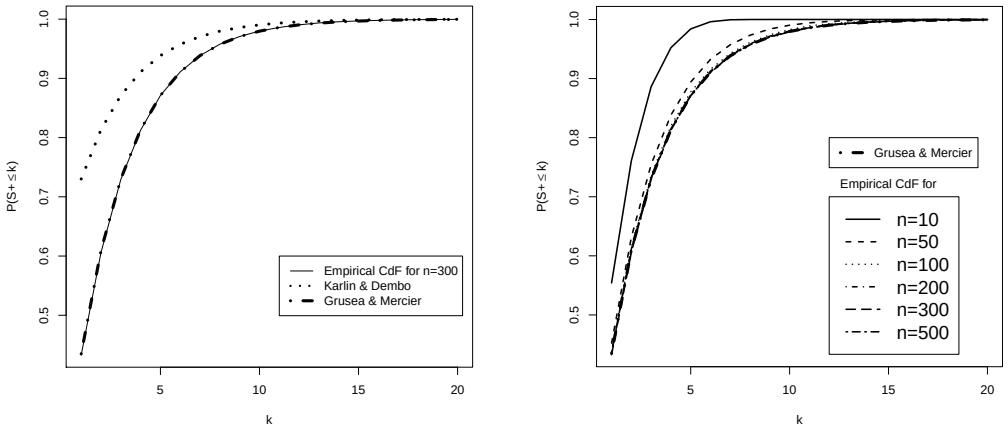

FIGURE 1: Cumulative distribution function of S^+ for the simple scoring scheme $(-1, 0, +1)$ and $A_0 =$"A". Left panel: Comparison between the approximation of Karlin and Dembo [9], a Monte Carlo estimation with sequences of length $n = 300$, and our exact formula proposed in Theorem 2.1. Right panel: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

a function of n, for two fixed values $x = -5$ and -8. This illustrates the asymptotic behavior of this probability with growing n. We can also observe the fact that Karlin and Dembo's asymptotic bounds do not depend on n. In Figure 3, we compare the asymptotic bounds 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.

4.5. Numerical applications on real data

We consider the examples presented in [8] for which we could recover the given sequences. On each sequence separately, we learn the score frequencies f_x for each possible score x, as well as the transition probability matrix P, for which we give each row P_x. For each example, we also show the corresponding invariant probability π,

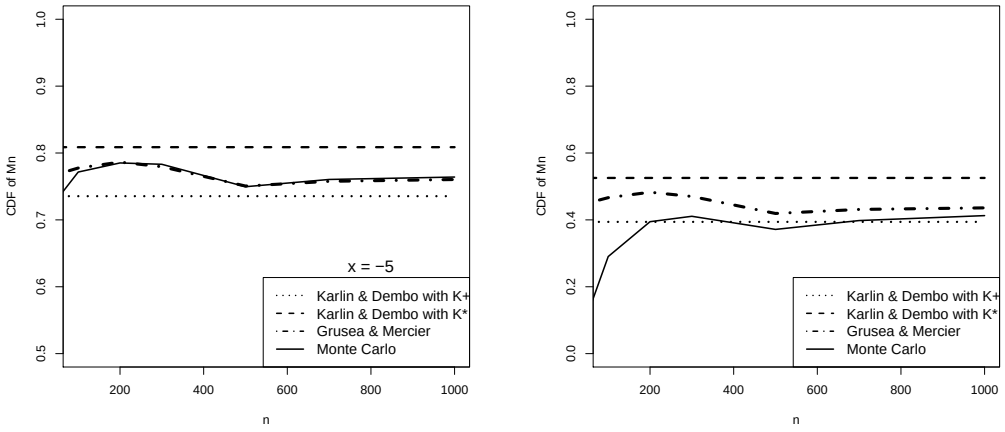

FIGURE 2: Comparison of the different approximations for $p(n,x) = \mathbb{P}\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)$ as a function of n, for fixed x and for the simple scoring scheme $(-1, 0, +1)$: the asymptotic lower and upper bounds of Karlin and Dembo's [9] (see Equations (7) and (8)), the approximation we propose in Theorem 2.4 and Monte Carlo estimation. Left panel: $p(n,x)$ for $x = -5$. Right panel: $p(n,x)$ for $x = -8$.

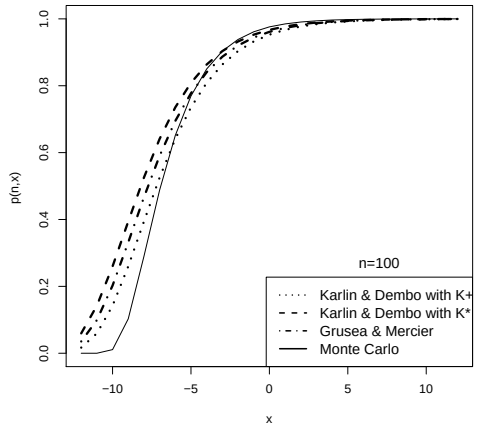

FIGURE 3: Comparison of the different approximations for $p(n,x) = \mathbb{P}\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)$ $p(n,x)$ as a function of x, for fixed $n = 100$, and for the simple scoring scheme $(-1, 0, +1)$: the asymptotic lower and upper bounds of Karlin and Dembo's [9] (see Equations (7) and (8)), our approximation proposed in Theorem 2.4 and Monte Carlo estimation.

which is in general close to the score frequencies, as expected. Biologists awared us that since 1990 the sequences referred by [8] may have a little bit changed due to the evolution of sequencing, which can explain some small differences in score frequencies between our sequences and the ones in [8]. Note that our Hypotheses (1) and (2) are both verified in all the following applications.

For each example, we computed the corresponding p-values of the observed local score using the asymptotic lower and upper bounds of Karlin and Dembo [9] (p_{KDe} refers to the bound with K^* based on Equation (8), and p_{KDe-K^+} refers to the bound with K^+ based on Equation (7)), the approximation we propose in Theorem 2.4 (p_{GMe}), and an empirical Monte Carlo estimation (p_{MC}) based on 10^5 simulations of sequences of the given length. Note that in all examples we have $p_{MC} \leq p_{GMe} \leq p_{KDe} \leq p_{KDe-K^+}$, except in Example d)ii), where we have $p_{GMe} \leq p_{MC} \leq p_{KDe} \leq p_{KDe-K^+}$. In order to simplify the presentation, in what follows we only show the results based on the best of the two bounds of Karlin and Dembo, which is p_{KDe}. We also compute the percentage of relative error for both theoretical methods:

$$RE(KDe) = 100 \cdot \frac{p_{KDe} - p_{MC}}{p_{MC}}, \quad RE(GMe) = 100 \cdot \frac{p_{GMe} - p_{MC}}{p_{MC}}. \tag{40}$$

The p-value given by [8] in the i.i.d. model ($p_{KDe-iid}$) is recalled.

We also computed two classical measures of dissimilarity between the theoretical approximate distribution of the local score (the one we propose, denoted GMe, respectively the one given by the asymptotic upper bound of Karlin and Dembo [9], denoted KDe), and the empirical distribution obtained by Monte Carlo simulations, denoted MC:
- the Kolmogorov-Smirnov distance:

$$d_{KS}(GMe) := \max_x(|P_{GMe}(M_n \leq x) - P_{MC}(M_n \leq x)|). \tag{41}$$

- the Kullback-Leibler divergence:

$$d_{KL}(GMe) := \sum_x P_{MC}(M_n = x) \cdot \log\left(\frac{P_{MC}(M_n = x)}{P_{GMe}(M_n = x)}\right). \tag{42}$$

We define similarly $d_{KS}(KDe)$ and $d_{KL}(KDe)$ using the asymptotic upper bound of Karlin and Dembo [9] for the distribution of the local score (see Equation (8)).

We gather the relative errors and the two distance computations in Table 1.

TABLE 1: Numerical comparison between our approximation for the local score distribution and the one of [9], using relative errors (see Equation (40)) and two classical dissimilarity measures recalled in Equations (41) and (42).

	$d_{KS}(KDe)$	$d_{KS}(GMe)$	$d_{KL}(KDe)$	$d_{KL}(GMe)$	$RE(KDe)$	$RE(GMe)$
a)i)	0.44	0.03	1.14	< 0.01	259%	7%
a)ii)	0.48	0.06	1.32	0.02	307%	12%
b)	0.81	0.01	12.85	$\simeq 10^{-3}$	1043%	3%
c)ii)	0.80	0.13	11.6	0.07	562%	5%
d)i)	0.66	0.06	4.78	0.01	870%	22%
d)ii)	0.84	0.20	5.64	0.29	307%	−18%
d)iii)	0.69	0.06	5.37	0.01	1061%	64%

Examples c)i) and c)iii) have not been considered, since we did not recover the sequences presented in [8]. Note that the sequence a)i) has one supplementary amino acid than the one referenced in [8] and the local score is equal to 21 instead of 24 in their article. Example e) has not been studied because one of the transition probabilities is equal to 0 and does not verify our hypotheses.

Exemple a), Mixed charge segment: $s = 2$ for the amino acids aspartate (D), glutamate (E), histidine (H), lysine (K) and arginine (R), and $s = -1$ otherwise.

i) Human keratin cytoskeletal type II (UniProtKB-P04264): $n = 644$, $M_n = 24$, positions 238-292. $f_{-1} = 82.2\%; f_2 = 17.8\%$. $P_{-1} = (0.784, 0.216); P_2 = (0.821, 0.179)$. $p_{KDe} = 5.06 \cdot 10^{-3}; p_{GMe} = 1.51 \cdot 10^{-3}; p_{MC} = 1.41 \cdot 10^{-3}$. $\pi = [0.792; 0.208]$.

ii) Human c-jun, nuclear transcription factor (UniProtKB-P05412): $n = 331$, $M_n = 29$, positions 246-285. $f_{-1} = 79.5\%; f_2 = 20.5\%$. $P_{-1} = (0.805, 0.195); P_2 = (0.754, 0.246)$. $p_{KDe} = 2.2 \cdot 10^{-3}; p_{GMe} = 6.03 \cdot 10^{-4}; p_{MC} = 5.4 \cdot 10^{-4}; p_{KDe-iid} < 2 \cdot 10^{-4}$. $\pi = [0.795; 0.205]$.

Exemple b), Acidic charge segments: $s = 2$ for aspartate (D) and glutamate (E); $s = -2$ for lysine (K) and arginine (R) ; and $s = -1$ otherwise.

Zeste protein (UniProtKB-P09956): $n = 575$, $M_n = 11$, positions 194-209. $f_{-2} = 8.0\%; f_{-1} = 82.8\%; f_2 = 9.2\%$. $P_{-2} = (0.109, 0.696, 0.195); P_{-1} = (0.078, 0.853, 0.069)$;

$P_2 = (0.075, 0.717, 0.208)$. $p_{KDe} = 5.76 \cdot 10^{-1}; p_{GMe} = 5.21 \cdot 10^{-2}$;
$p_{MC} = 5.04 \cdot 10^{-2}; p_{KDe-iid} = 3.7 \cdot 10^{-3}$. $\pi = [0.080; 0.828; 0.092]$.

Exemple c), High-scoring basic charge segments: $s = 2$ for lysine (K), arginine (R) and histidine (H); $s = -2$ for aspartate (D) and glutamate (E); $s = -1$ otherwise.

ii) Zeste protein (UniProtKB-P09956): $n = 575$, $M_n = 12$, positions 78-86. $f_{-2} = 9.2\%; f_{-1} = 79.7\%; f_2 = 11.1\%$. $P_{-2} = (0.208, 0.698, 0.094); P_{-1} = (0.068, 0.827, 0.105)$;
$P_2 = (0.172, 0.656, 0.172)$. $p_{KDe} = 13.9 \cdot 10^{-2}; p_{GMe} = 2.2 \cdot 10^{-2}$;
$p_{MC} = 2.1 \cdot 10^{-2}; p_{KDe-iid} = 4 \cdot 10^{-3}$. $\pi = [0.093; 0.796; 0.111]$.

Exemple d), Strong Hydrophobic segments: $s = 1$ for isoleucine (I), leucine (L), valine (V), phenylalanine (F), methionine (M), cysteine (C), alanine(A); $s = -1$ for glycine (G), serine (S), threonine (T), tryptophan (W), tyrosine (Y), proline (P); $s = -2$ for arginine (R), lysine (K), aspartate (D), glutamate (E), histidine (H), asparagine (N), glutamine (Q).

i) Drosophila engrailed (UniProtKB-P02836): $n = 552$, $M_n = 17$, positions 63-88. $f_{-2} = 34.6\%; f_{-1} = 33.7\%; f_1 = 31.7\%$. $P_{-2} = (0.466, 0.230, 0.304)$;
$P_{-1} = (0.254, 0.449, 0.297)$; $P_1 = (0.314, 0.337, 0.349)$. $p_{KDe} = 5.82 \cdot 10^{-4}$; $p_{GMe} = 7.31 \cdot 10^{-5}; p_{MC} = 6 \cdot 10^{-5}; p_{KDe-iid} = 1.8 \cdot 10^{-5}$. $\pi = [0.346; 0.338; 0.316]$.

ii) Human c-mas, angiotensin receptor factor (UniProtKB-P04201): $n = 325$, $M_n = 15$, positions 186-212. $f_{-2} = 23.4\%; f_{-1} = 29.8\%; f_1 = 46.8\%$. $P_{-2} = (0.381, 0.316, 0.303)$;
$P_{-1} = (0.206, 0.289, 0.505); P_1 = (0.179, 0.298, 0.523)$. $p_{KDe} = 8.77 \cdot 10^{-1}; p_{GMe} = 1.77 \cdot 10^{-1}; p_{MC} = 2.15 \cdot 10^{-1}; p_{KDe-iid} = 0.80 \cdot 10^{-1}$. $\pi = [0.234; 0.3; 0.466]$.

iii) Cystic Fibrosis (UniProtKB-P13569): $n = 1480$, $M_n = 21$, positions 986-1029. $f_{-2} = 31.55\%; f_{-1} = 26.9\%; f_1 = 41.55\%$. $P_{-2} = (0.355, 0.270, 0.375)$;
$P_{-1} = (0.322, 0.271, 0.407)$; $P_1 = (0.282, 0.267, 0.451)$. $p_{KDe} = 22.5 \cdot 10^{-3}; p_{GMe} = 3.19 \cdot 10^{-3}; p_{MC} = 1.94 \cdot 10^{-3}; p_{KDe-iid} = 10^{-3}$. $\pi = [0.316; 0.269; 0.415]$.

Aknowledgements: We thank the referee for very useful comments, which led to an improvement of the manuscript.

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. 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.