J. Appl. Prob. **57,** 29–52 (2020)
*doi:*10.1017/jpr.2019.75

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

S. GRUSEA* ** AND
S. MERCIER,*** *Institut de Mathématiques de Toulouse*

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 (1990). 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. The new approximations are compared with those proposed by Karlin and Dembo (1992) in order to evaluate improvements, both in the simple application case and on the real data examples considered by Karlin and Altschul (1990).

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

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

1. Introduction

A large number of biological sequences are currently available. The *local score* in sequence analysis, first defined in [8], 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 us, for example, to detect atypical segments in biological sequences. In order to distinguish significantly interesting segments from ones that may 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 (i.i.d.) variables model [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 in [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

Received 7 March 2018; revision received 22 September 2019.

* Postal address: Institut National des Sciences Appliquées, 135 avenue de Rangueil, 31400, Toulouse, France.
** Email address: grusea@insa-toulouse.fr
*** Postal address: Institut de Mathématiques de Toulouse, UMR5219, Université de Toulouse 2 Jean Jaurès, 5 allées Antonio Machado, 31058, Toulouse, Cedex 09, France.

lattice case, they gave 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 unfortunately very little cited or used in the literature. A possible explanation could be 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 [9] and illustrate the improvements both in a simple application case and on the real data examples proposed in [8].

1.1. 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 $\alpha, \beta, \ldots$ 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 (for all $\alpha, \beta,\ p_{\alpha\beta}>0$). We also suppose that the initial distribution of A_0 is given by π, so that the Markov chain is stationary. P_α 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 by $\mathrm{E}[f(A)]$ the average score.

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

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

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

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

Note that, thanks to the assumption $p_{\alpha\beta}>0$ for all α, β, 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 $k\geq 1$ be the successive partial sums. Let S^+ be the *maximal non-negative partial sum*: $S^+:=\max\{0, S_k : k\geq 0\}$.

Further, let $\sigma^-:=\inf\{k\geq 1 : S_k<0\}$ be the time of the first negative partial sum. Note that σ^- is an almost surely (a.s.) finite stopping time due to Hypothesis (1), and let $Q_1:=\max_{0\leq k<\sigma^-} S_k$.

First introduced in [8], the *local score*, denoted M_n, is defined as the maximum segmental score for a sequence of length n: $M_n:=\max_{0\leq k\leq \ell\leq n}(S_\ell-S_k)$.

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] considered a more general model, with a random score function defined on pairs of consecutive states of the Markov chain: they associated with 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 supposed that, for $(A_{i-1}, A_i)=(A_{j-1}, A_j)=(\alpha, \beta)$, the random

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

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 also more general 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 Hypotheses (1) and (2) assure so-called cycle positivity, i.e. the existence of some state $\alpha \in \mathcal{A}$ and of some $m \geq 2$ satisfying $\mathrm{P}(\bigcap_{k=1}^{m-1}\{S_k > 0\} \mid A_0 = A_m = \alpha) > 0$. In [9], in order to simplify the presentation, the authors strengthened the assumption of cycle positivity by assuming that $\mathrm{P}(X_{\alpha\beta} > 0) > 0$ and $\mathrm{P}(X_{\alpha\beta} < 0) > 0$ for all $\alpha, \beta \in \mathcal{A}$ (see (1.19) of [9]), but in fact cycle positivity is sufficient for their results to hold.

In Section 2 we first introduce a few more definitions and some notation. We then present the main results: in Theorem 2.1, we propose a recursive result for the exact distribution of the maximal non-negative partial sum S^+ for an infinite sequence, and in Theorem 2.3, 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. 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 Section 4.4, for which we compare our approximations to those proposed by Karlin and Dembo [9] in the Markovian case. In Section 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 $i \geq 1$, $K_i := \inf\{k > K_{i-1} : S_k - S_{K_{i-1}} < 0\}$ be the successive decreasing ladder times of $(S_k)_{k\geq 0}$. Note that $K_1 = \sigma^-$.

We now consider the subsequence $(A_i)_{0\leq i\leq n}$ for a given length $n \in \mathbb{N} \setminus \{0\}$. Denote by $m(n) := \max\{i \geq 0 : K_i \leq n\}$ 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 ith excursion above 0.

Note that due to the negative drift we have $\mathrm{E}[K_1] < \infty$ (see Lemma 3.7) and $m(n) \to \infty$ a.s. when $n \to \infty$. With every excursion $i = 1, \ldots, m(n)$ we associate its *maximal segmental score* (also called *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] gave 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 $M_n = \max_{0\leq k\leq n} W_k$.

For every $\alpha, \beta \in \mathcal{A}$, we denote $q_{\alpha\beta} := \mathrm{P}_\alpha(A_{K_1} = \beta)$ 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 $q_{\alpha\beta} \geq p_{\alpha\beta} > 0$ for all $\alpha, \beta \in \mathcal{A}^-$. 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 $\mathrm{P}(A_{K_i} = \beta \mid A_{K_{i-1}} = \alpha) = q_{\alpha\beta}$ if $\alpha, \beta \in \mathcal{A}^-$ and 0 otherwise. Denote by $\tilde{z} > 0$ the stationary frequency vector of the irreducible stochastic matrix $\tilde{\mathbf{Q}}$, and let

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

$z := (z_\alpha)_{\alpha\in\mathcal{A}}$, 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 random variables $(Q_i)_{i\geq 1}$ 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 $F_{Q_1,\alpha,\beta}(y) := \mathrm{P}_\alpha(Q_1\leq y \mid A_{\sigma^-}=\beta)$ and $F_{Q_1,\alpha}(y) := \mathrm{P}_\alpha(Q_1\leq y)$. 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 ith excursion given that it starts in state α and ends in state β, i.e. $F_{Q_1,\alpha,\beta}(y) = \mathrm{P}(Q_i\leq y \mid A_{K_i}=\beta,\ A_{K_{i-1}}=\alpha)$, whereas $F_{Q_1,\alpha}$ represents the CDF of Q_i conditionally on the ith excursion starting in state α, i.e. $F_{Q_1,\alpha}(y) = \mathrm{P}(Q_i\leq y \mid A_{K_{i-1}}=\alpha)$. 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 $\sigma^+ := \inf\{k\geq 1 : S_k > 0\}$ with values in $\mathbb{N}\cup\{\infty\}$. Due to Hypothesis (1) we have $\mathrm{P}_\alpha(\sigma^+<\infty) < 1$ for all $\alpha\in\mathcal{A}$.

For every $\alpha,\beta\in\mathcal{A}$ and $\xi>0$, let $L_{\alpha\beta}(\xi) := \mathrm{P}_\alpha(S_{\sigma^+}\leq\xi,\ \sigma^+<\infty,\ A_{\sigma^+}=\beta)$. Note that $L_{\alpha\beta}(\xi)=0$ for $\beta\in\mathcal{A}\setminus\mathcal{A}^+$, and $L_{\alpha\beta}(\infty)\leq \mathrm{P}_\alpha(\sigma^+<\infty)<1$, therefore $\int_0^\infty \mathrm{d}L_{\alpha\beta}(\xi) = L_{\alpha\beta}(\infty) < 1$.

Let us also denote by $L_\alpha(\xi) := \sum_{\beta\in\mathcal{A}^+} L_{\alpha\beta}(\xi) = \mathrm{P}_\alpha(S_{\sigma^+}\leq\xi,\ \sigma^+<\infty)$ the conditional CDF of the first positive partial sum, when it exists, given that the Markov chain starts in state α, and $L_\alpha(\infty) := \lim_{\xi\to\infty} L_\alpha(\xi) = \mathrm{P}_\alpha(\sigma^+<\infty)$.

For any $\theta\in\mathbb{R}$ we introduce the matrix $\mathbf{\Phi}(\theta) := \big(p_{\alpha\beta}\cdot\exp(\theta f(\beta))\big)_{\alpha,\beta\in\mathcal{A}}$. 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 eigenvector $u(\theta) = (u_i(\theta))_{1\leq i\leq r}$ (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 $\rho'(0) = \mathrm{E}[f(A)]$, 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 us to recursively compute 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: $F_{S^+,\alpha}(\ell) := \mathrm{P}_\alpha(S^+\leq\ell)$ for all $\ell\in\mathbb{N}$, and for every $k\in\mathbb{N}\setminus\{0\}$ and $\beta\in\mathcal{A}$, $L_{\alpha\beta}^{(k)} := \mathrm{P}_\alpha(S_{\sigma^+}=k,\ \sigma^+<\infty,\ A_{\sigma^+}=\beta)$. 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 Section 4.2).

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

$$F_{S^+,\alpha}(0) = \mathrm{P}_\alpha(\sigma^+=\infty) = 1-L_\alpha(\infty),$$

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

The proof will be given in Section 3.

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

In Theorem 2.2 we obtain an asymptotic result for the tail behaviour 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 some more notation.

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

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

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; we denote by $\tilde{w} > 0$ 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 Section 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 $L^{(k)}_{\alpha\beta}$ terms different from zero, and therefore there are a finite number of non-null terms in the sum defining $G_{\alpha\beta}(\infty)$.

The following result is the analogue, in our setting, of Lemma 4.3 of [9].

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

$$\lim_{k\to+\infty} \frac{\mathrm{e}^{\theta^* k} \mathrm{P}_\alpha(S^+ > k)}{u_\alpha(\theta^*)} = \frac{1}{c} \cdot \sum_{\gamma\in\mathcal{A}^+} \frac{w_\gamma}{u_\gamma(\theta^*)} \sum_{\ell\geq 0} (L_\gamma(\infty) - L_\gamma(\ell))\mathrm{e}^{\theta^* \ell} := c(\infty), \tag{3}$$

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

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

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 $c(\infty)$:

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

Indeed, by the summation by parts formula

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

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

we obtain

$$\sum_{\ell=0}^{\infty}(L_\gamma(\infty)-L_\gamma(\ell))\mathrm{e}^{\theta^*\ell}=\frac{1}{\mathrm{e}^{\theta^*}-1}\sum_{\ell=0}^{\infty}(L_\gamma(\infty)-L_\gamma(\ell))(\mathrm{e}^{\theta^*(\ell+1)}-\mathrm{e}^{\theta^*\ell})$$

$$=\frac{1}{\mathrm{e}^{\theta^*}-1}$$

$$\times\left\{\lim_{k\to\infty}(L_\gamma(\infty)-L_\gamma(k))\mathrm{e}^{\theta^*k}-L_\gamma(\infty)-\sum_{\ell=0}^{\infty}(L_\gamma(\ell)-L_\gamma(\ell+1))\mathrm{e}^{\theta^*(\ell+1)}\right\}$$

$$=\frac{1}{\mathrm{e}^{\theta^*}-1}\left\{-L_\gamma(\infty)+\sum_{\ell=0}^{\infty}\mathrm{e}^{\theta^*(\ell+1)}\,\mathrm{P}_\gamma(S_{\sigma^+}=\ell+1,\ \sigma^+<\infty)\right\}$$

$$=\frac{1}{\mathrm{e}^{\theta^*}-1}\{\mathrm{E}_\gamma\big[\mathrm{e}^{\theta^*S_{\sigma^+}};\sigma^+<\infty\big]-L_\gamma(\infty)\}.$$

Before stating the next results, let us denote, for every integer $\ell<0$ and $\alpha,\beta\in\mathcal{A}$, $Q_{\alpha\beta}^{(\ell)}:=\mathrm{P}_\alpha(S_{\sigma^-}=\ell,\ A_{\sigma^-}=\beta)$. Note that $Q_{\alpha\beta}^{(\ell)}=0$ 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 $\dfrac{f_k}{g_k}\underset{k\to\infty}{\to}1$.

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*

$$\mathrm{P}_\alpha(Q_1>k)\underset{k\to\infty}{\sim}\mathrm{P}_\alpha(S^+>k)-\sum_{\ell<0}\sum_{\beta\in\mathcal{A}^-}\mathrm{P}_\beta\left(S^+>k-\ell\right)\cdot Q_{\alpha\beta}^{(\ell)}. \tag{4}$$

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):

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

Using 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:*

$$\mathrm{P}_\alpha\left(M_n\le\frac{\log(n)}{\theta^*}+x\right)\underset{n\to\infty}{\sim}\exp\left\{-\frac{n}{A^*}\sum_{\beta\in\mathcal{A}^-}z_\beta\mathrm{P}_\beta\left(S^+>\lfloor\log(n)/\theta^*+x\rfloor\right)\right\}$$

$$\times\exp\left\{\frac{n}{A^*}\sum_{k<0}\sum_{\gamma\in\mathcal{A}^-}\mathrm{P}_\gamma\left(S^+>\lfloor\log(n)/\theta^*+x\rfloor-k\right)\cdot\sum_{\beta\in\mathcal{A}^-}z_\beta Q_{\beta\gamma}^{(k)}\right\}, \tag{5}$$

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

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

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

Remark 2.6.

- Note that the asymptotic equivalent in (5) does not depend on the initial state α.
- We recall, for comparison, the asymptotic lower and upper bounds of [9] for the distribution of M_n:

$$\liminf_{n\to+\infty} \mathrm{P}_\alpha\left(M_n \leq \frac{\log(n)}{\theta*} + x\right) \geq \exp\left\{-K^+ \exp(-\theta^* x)\right\}, \tag{6}$$

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

  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 (5.1) and (5.2) of [9] for more details). A more explicit formula for K^* is given in Section 4.4 for an application in a simple case.
- Even if the expression of our asymptotic equivalent in (5) seems more cumbersome than the asymptotic bounds recalled in (6) and (7), the practical implementations are equivalent.

3. Proofs of the main results

3.1. Proof of Theorem 2.1

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

It then suffices to note that

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

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

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

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

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

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

$$\text{for all } \ell > 0, \text{ for all } \alpha \in \mathcal{A}, \quad b_\alpha(\ell) = a_\alpha(\ell) + \sum_\beta \sum_{k=0}^{\ell} b_\beta(\ell - k) G_{\alpha\beta}^{(k)}. \tag{8}$$

Since the restriction $\tilde{\mathbf{G}}(\infty)$ of $\mathbf{G}(\infty)$ to $\mathcal{A}^+$ is stochastic, its spectral radius equals 1 and the corresponding right eigenvector is the vector having all components equal to 1; the 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 [1] gives the formula in (3) for the limit $c(\infty)$ of $b_\alpha(\ell)$ when $\ell \to \infty$, which implies that $\lim_{k\to\infty} \mathrm{P}_\alpha(S^+ > k) = 0$.

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

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

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

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

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 $\mathrm{P}_\beta(S^+ > \ell - k) \to 0$ 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 Section 2.1, the sequence of random variables $(U_m(\theta))_{m\geq 0}$ defined by $U_0(\theta) := 1$ and*

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

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

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

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

By the Markov property we further have

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

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

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

We deduce that $\mathrm{E}[\exp(\theta f(A_{m+1}))u_{A_{m+1}}(\theta) \mid A_m] = u_{A_m}(\theta)\rho(\theta)$, and hence $\mathrm{E}[U_{m+1}(\theta) \mid \mathcal{F}_m] = U_m(\theta)$, which finishes the proof. □

Lemma 3.3. *With θ^* defined at the end of Section 2.1 we have, for all $\alpha \in \mathcal{A}$,*

$$\frac{1}{u_\alpha(\theta^*)} \sum_{\beta \in \mathcal{A}^+} \sum_{\ell=1}^{\infty} L_{\alpha\beta}^{(\ell)} \,\mathrm{e}^{\theta^* \ell}\, u_\beta(\theta^*) = 1.$$

Proof. The proof uses Lemma 3.1 and ideas inspired by [9] (Lemma 4.2). First note that the above equation is equivalent to $\mathrm{E}_\alpha[U_{\sigma^+}(\theta^*); \sigma^+ < \infty] = 1$, 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

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

We will show that $\mathrm{E}_\alpha[U_n(\theta^*); \sigma^+ > n] \to 0$ 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 $\lim_{n\to\infty} \mathrm{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n] = 0$.

For a fixed $a > 0$ we can write

$$\begin{aligned}\mathrm{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n] &= \mathrm{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, \text{there exists } k \leq n : S_k \leq -2a] \\ &\quad + \mathrm{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, -2a \leq S_k \leq 0, \ \text{for all } 0 \leq k \leq n]. \qquad (10)\end{aligned}$$

The first expectation on the right-hand side of (10) can be bounded further:

$$\begin{aligned}&\mathrm{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, \ \text{there exists } k \leq n : S_k \leq -2a] \\ &\quad \leq \mathrm{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n \leq -a] \\ &\qquad + \mathrm{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, S_n > -a, \ \text{there exists } k < n : S_k \leq -2a]. \qquad (11)\end{aligned}$$

We obviously have

$$\mathrm{E}_\alpha[\exp(\theta^* S_n); \sigma^+ > n, \ S_n \leq -a] \leq \exp(-\theta^* a). \qquad (12)$$

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

Let us further define the stopping time $T := \inf\{k \geq 1 : S_k \leq -2a\}$. Note that $T < \infty$ a.s., since $S_n \to -\infty$ a.s. when $n \to \infty$. Indeed, by the ergodic theorem, we have $S_n/n \to \mathrm{E}[f(A)] < 0$ a.s. when $n \to \infty$. Therefore we have

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

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

$$\begin{aligned}\limsup_{n\to\infty} \mathrm{E}_\alpha[\exp(\theta^* S_n);\, \sigma^+ > n,\ S_n > -a,\ \text{there exists } k < n : S_k \leq -2a]&\\ \leq \sum_{\beta\in\mathcal{A}^-} \mathrm{P}_\beta(S^+ > a).&\qquad (13)\end{aligned}$$

Considering the second expectation on the right-hand side of (10), we have

$$\lim_{n\to\infty} \mathrm{P}_\alpha(-2a \leq S_k \leq 0,\ \text{for all } 0 \leq k \leq n) = \mathrm{P}_\alpha(-2a \leq S_k \leq 0,\ \text{for all } k \geq 0) = 0, \qquad (14)$$

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

Equations (10)–(14) imply that, for every $a > 0$, we have

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

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

Proof of Theorem 2.2. For $\alpha \in \mathcal{A}^+$ the formula has already been 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 (9) implies the formula in (8).

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, for all $\beta \in \mathcal{A}^+$ and $k \geq 0$,

$$\frac{\mathrm{e}^{\theta^*(\ell-k)}\mathrm{P}_\beta(S^+ > \ell - k)}{u_\beta(\theta^*)} \underset{\ell\to\infty}{\longrightarrow} c(\infty).$$

We finally obtain

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

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

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

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

By applying the strong Markov property to the stopping time σ^- we can further decompose the last probability with respect to the values taken by S_{σ^-} and A_{σ^-}:

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

We thus obtain

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

By Theorem 2.2 we have $\mathrm{P}_\beta(S^+ > k) = O(\mathrm{e}^{-\theta^* k})$ as $k \to \infty$ for every $\beta \in \mathcal{A}^-$, from which we deduce that the left-hand side of the previous equation is $o(\mathrm{P}_\alpha(Q_1 > k))$ 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* $\mathrm{P}_\alpha(Q_1 > y \mid A_{\sigma^-} = \beta) \le C\mathrm{e}^{-\theta^* y}$.

Proof. The proof is partly inspired by [9]. Let $y > 0$ and denote by $\sigma(y)$ 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

$$\mathrm{E}_\alpha[U_{\sigma(y)}(\theta^*)] = \mathrm{E}_\alpha[U_0(\theta^*)] = 1. \tag{15}$$

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 $0 < U_{\min(\sigma(y),n)}(\theta^*) \le \tilde{C}$ a.s. Indeed, this follows from the fact that when $\sigma(y) > n$ we have $0 \le S_n \le y$, and when $\sigma(y) \le n$ either $S_{\sigma(y)} < 0$ or $y < S_{\sigma(y)} < y + \max\left\{f(\alpha) : \alpha \in \mathcal{A}^+\right\}$.

We deduce from (15) that, for some constant $K > 0$, we have

$$\begin{aligned}
1 = \mathrm{E}_\alpha\left[\mathrm{e}^{\theta^* S_{\sigma(y)}} \frac{u_{A_{\sigma(y)}}(\theta^*)}{u_{A_0}(\theta^*)}\right] &\ge K\mathrm{e}^{\theta^* y}\,\mathrm{E}_\alpha[\mathrm{e}^{\theta^*(S_{\sigma(y)} - y)} \mid S_{\sigma(y)} > y]\cdot \mathrm{P}_\alpha(S_{\sigma(y)} > y)\\
&\ge K\mathrm{e}^{\theta^* y}\,\mathrm{P}_\alpha(S_{\sigma(y)} > y) \ge K\mathrm{e}^{\theta^* y}\,\mathrm{P}_\alpha(S_{\sigma(y)} > y \mid A_{\sigma^-} = \beta) q_{\alpha\beta}.
\end{aligned}$$

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 $q_{\alpha\beta} = \mathrm{P}_\alpha(A_{\sigma^-} = \beta) \ge p_{\alpha\beta} \ge c$. In order to obtain the bound in the statement, it remains to note that $\mathrm{P}_\alpha(Q_1 > y \mid A_{\sigma^-} = \beta) = \mathrm{P}_\alpha(S_{\sigma(y)} > y \mid A_{\sigma^-} = \beta)$. □

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

Lemma 3.5. $\rho'(0)=\mathrm{E}[f(A)]<0.$

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

When differentiating the previous relation with respect to θ we obtain

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

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

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

On the other hand,

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

For $\theta=0$ we get

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

From (16) and (17) we deduce that $\frac{\rho'(0)}{r}+\sum_\alpha \pi_\alpha u'_\alpha(0)=\frac{1}{r}\mathrm{E}[f(A)]+\sum_\beta \pi_\beta u'_\beta(0)$, from which the stated result follows easily. □

Lemma 3.6. *There exists* $n_0\geq 0$ *such that for all* $n\geq n_0$ *and for all* $\alpha\in\mathcal{A}$ *we have*

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

Proof. By a large deviation principle for additive functionals of Markov chains (see Theorem 3.1.2 in [3]) we have $\limsup_{n\to+\infty}\frac{1}{n}\log\Big(\mathrm{P}_\alpha\Big(\frac{S_n}{n}\in\Gamma\Big)\Big)\leq-\mathcal{I}$, 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 $\rho'(0)=\mathrm{E}[f(A)]<0$ (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 $I(0)=\sup_{\theta\in\mathbb{R}}(-\log\rho(\theta))=-\inf_{\theta\in\mathbb{R}^+}\log(\rho(\theta))$. Again using 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 $\mathcal{I}=-\log(\inf_{\theta\in\mathbb{R}^+}\rho(\theta))>-\log\rho(0)=0$. □

Lemma 3.7. $\mathrm{E}_\alpha[K_1]<\infty$ *for every* $\alpha\in\mathcal{A}$.

Proof. Note that $\mathrm{P}_\alpha(K_1>n)\leq\mathrm{P}_\alpha(S_n\geq 0)$. With $n_0\in\mathbb{N}$ defined in Lemma 3.6, using a well-known alternative formula for the expectation we get

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

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

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* $A^* = \sum_\beta z_\beta \mathrm{E}_\beta[K_1]$ *a.s.*

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

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

First note that $\frac{K_1}{m} \to 0$ 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 $\mathrm{P}(K_i - K_{i-1} = \ell \mid A_{K_{i-1}} = \alpha) = \mathrm{P}_\alpha(K_1 = \ell)$. 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 $\mathrm{P}(Y_i = (\beta, \ell) \mid Y_{i-1} = (\alpha, k)) = q_{\alpha\beta}\mathrm{P}_\beta(K_1 = \ell)$. 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 $\sum_{\alpha,k}\pi(\alpha, k)\cdot q_{\alpha\beta}\mathrm{P}_\beta(K_1 = \ell) = \pi(\beta, \ell)$, and hence the Markov chain $(Y_i)_i$ is also irreducible, with invariant distribution $\pi(\alpha, k) := z_\alpha \mathrm{P}_\alpha(K_1 = k)$.

For fixed β, by 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 that

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

Taking the sum over β and using (18) gives the result in the statement. □

Proof of Theorem 2.4.

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

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

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 $y > 0$,

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

with $\psi_{\beta\gamma}(m) := \#\{i : 1 \leq i \leq m,\ A_{K_{i-1}} = \beta,\ A_{K_i} = \gamma\}/m$. Given that $A_0 = \alpha \in \mathcal{A}^-$, the $(A_{K_i})_{i\geq 0}$ 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

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

that $\psi_{\beta\gamma}(m) \to z_\beta q_{\beta\gamma}$ 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 $\psi_{\beta\gamma}(m) \to 0$ 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 $\psi_{\beta\gamma}(m) \to z_\beta q_{\beta\gamma}$ a.s. when $m \to \infty$ for every $\beta, \gamma \in \mathcal{A}$.

We will use the Taylor series expansion of the log function. Let us define

$$d_{\beta\gamma}(m) := m\left[1 - F_{\beta\gamma}\left(\frac{\log(m)}{\theta*} + x_m\right)\right] \quad \text{for every } m \geq 1.$$

Thanks to Lemma 3.4, the $d_{\beta\gamma}(m)$ are uniformly bounded in m, β, and γ. Since $0 \leq \psi_{\beta\gamma}(m) \leq 1$, we obtain

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

Since

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

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

But

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

and using Theorem 2.3 we get

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

This then leads to

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

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 the proof of this step in detail.

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

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

$$\mathrm{P}_\alpha \left(M_{K_{m(n)+1}} \leq \frac{\log (n)}{\theta *} + x \right) \leq \mathrm{P}_\alpha \left(M_n \leq \frac{\log (n)}{\theta *} + x \right) \leq \mathrm{P}_\alpha \left(M_{K_{m(n)}} \leq \frac{\log (n)}{\theta *} + x \right). \tag{19}$$

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

Now fix $\varepsilon > 0$. We have

$$\begin{aligned}
&\mathrm{P}_\alpha \left(M_{K_{m(n)}} \leq \frac{\log (n)}{\theta *} + x \right) \\
&\leq \mathrm{P}_\alpha \left(M_{K_{m(n)}} \leq \frac{\log (n)}{\theta *} + x, \left| \frac{m(n)}{n} - \frac{1}{A^*} \right| \leq \varepsilon \right) + \mathrm{P}_\alpha \left(\left| \frac{m(n)}{n} - \frac{1}{A^*} \right| > \varepsilon \right) \\
&\leq \mathrm{P}_\alpha \left(M_{K_{\lceil n/A^* - n\varepsilon \rceil}} \leq \frac{\log (n)}{\theta *} + x \right) + \mathrm{P}_\alpha \left(\left| \frac{m(n)}{n} - \frac{1}{A^*} \right| > \varepsilon \right).
\end{aligned} \tag{20}$$

Using the result of *Step 1*, we obtain

$$\frac{\mathrm{P}_\alpha \left(M_{K_{\lceil n/A^* - n\varepsilon \rceil}} \leq \frac{\log (n)}{\theta *} + x \right)}{E_n} \underset{n\to\infty}{\sim} R_n(\varepsilon), \tag{21}$$

where E_n is the asymptotic equivalent given in the statement

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

and

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

Using Theorem 2.2 we obtain

$$\limsup_{n\to\infty} R_n(\varepsilon) \leq \exp\{\varepsilon \cdot c(\infty) \mathrm{e}^{-\theta^* x} D^*\}, \tag{22}$$

with

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

Equations (19)–(22), together with the fact that $\frac{m(n)}{n} \to \frac{1}{A^*}$ a.s., imply that

$$\limsup_{n\to\infty} \frac{\mathrm{P}_\alpha \left(M_n \leq \frac{\log (n)}{\theta *} + x \right)}{E_n} \leq \exp\{\varepsilon \cdot c(\infty) \mathrm{e}^{-\theta^* x} D^*\}. \tag{23}$$

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

In a similar manner, we can show that

$$\liminf_{n\to\infty} \frac{\mathrm{P}_\alpha\left(M_n \le \frac{\log(n)}{\theta^*} + x\right)}{E_n} \ge \exp\{-\varepsilon \cdot c(\infty)\mathrm{e}^{-\theta^* x} G^*\}, \tag{24}$$

with

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

Taking the limit $\varepsilon \to 0$ in (23) and (24) gives

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

and hence $\mathrm{P}_\alpha\big(M_n \le \frac{\log(n)}{\theta^*} + x\big) \underset{n\to\infty}{\sim} E_n$, 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 $A^* = \sum_\alpha z_\alpha \mathrm{E}_\alpha(\sigma^-)$. 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 $\mathrm{E}_\alpha\left[U_{\min(\sigma^-,n)}(\theta)\right] = \mathrm{E}_\alpha\left[U_0(\theta)\right] = 1$. Furthermore, we have

$$1 = \mathrm{E}_\alpha\left[U_{\sigma^-}(\theta); \sigma^- \le n\right] + \mathrm{E}_\alpha\left[U_n(\theta); \sigma^- > n\right]. \tag{25}$$

We will show that $\mathrm{E}_\alpha\left[U_n(\theta); \sigma^- > n\right] \to 0$ when $n \to \infty$. It suffices to prove that

$$\mathrm{E}_\alpha\left[\frac{\mathrm{e}^{\theta S_n}}{\rho(\theta)^n}; \sigma^- > n\right] \to 0.$$

By the Cauchy–Schwartz inequality, we have

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

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

Since $\rho(\theta) \to 1$ when $\theta \to 0$, for sufficiently small θ we will have both $\theta < \frac{\theta^*}{2}$ and $\rho(\theta)^2 > \inf_{\theta\in\mathbb{R}^+} \rho(\theta)$, implying that

$$\mathrm{E}_\alpha\left[\frac{\mathrm{e}^{\theta S_n}}{\rho(\theta)^n}; \sigma^- > n\right] \to 0 \qquad \text{when } n \to \infty.$$

When passing to the limit as $n \to \infty$ in (25), we deduce that for θ sufficiently small we have $\mathrm{E}_\alpha\left[U_{\sigma^-}(\theta)\right] = \mathrm{E}_\alpha\left[U_0(\theta)\right] = 1$. Consequently,

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

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

We deduce that for θ sufficiently small we have

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

For θ sufficiently small, by differentiating the above relation we obtain

$$u'_\alpha(\theta) = \sum_\beta q_{\alpha\beta} u_\beta(\theta) \mathrm{E}_\alpha \left[\frac{S_{\sigma^-} \exp(\theta \cdot S_{\sigma^-}) \rho(\theta)^{\sigma^-} - \exp(\theta \cdot S_{\sigma^-}) \sigma^- \rho(\theta)^{\sigma^- - 1} \rho'(\theta)}{\rho(\theta)^{2\sigma^-}} \mid A_{\sigma^-} = \beta \right] + \sum_\beta q_{\alpha\beta} u'_\beta(\theta) \mathrm{E}_\alpha \left[\frac{\exp(\theta \cdot S_{\sigma^-})}{\rho(\theta)^{\sigma^-}} \mid A_{\sigma^-} = \beta \right].$$

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

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

Because $u(0) =^t (1/r, \ldots, 1/r)$, we further get

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

From the last relation we deduce that

$$\sum_\alpha z_\alpha u'_\alpha(0) = \frac{1}{r} \sum_\alpha z_\alpha \mathrm{E}_\alpha \left[S_{\sigma^-} \right] - \frac{\rho'(0)}{r} \sum_\alpha z_\alpha \mathrm{E}_\alpha[\sigma^-] + \sum_\alpha \sum_\beta z_\alpha q_{\alpha\beta} u'_\beta(0). \tag{26}$$

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

$$\sum_\alpha z_\alpha u'_\alpha(0) =^t z \cdot u'(0) =^t (z\mathbf{Q}) \cdot u'(0) = \sum_\beta {}^t(z\mathbf{Q})_\beta \cdot u'_\beta(0) = \sum_\beta \sum_\alpha z_\alpha q_{\alpha\beta} u'_\beta(0). \tag{27}$$

Equations (26) and (27) imply that $\sum_\alpha z_\alpha \mathrm{E}_\alpha \left[S_{\sigma^-} \right] = \rho'(0) \cdot \sum_\alpha z_\alpha \mathrm{E}_\alpha[\sigma^-]$, and thus $A^* = \sum_\alpha z_\alpha \mathrm{E}_\alpha[\sigma^-] = \frac{1}{\rho'(0)} \sum_\alpha z_\alpha \mathrm{E}_\alpha \left[S_{\sigma^-} \right]$. Using the fact that $\rho'(0) = \mathrm{E}[f(A)]$ (see Lemma 3.5) gives the stated expression for A^*.

4. Applications and computational methods

Let $-u, \ldots, 0, \ldots, v$ 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 $P^{(j)}_{\alpha\beta} := \mathrm{P}_\alpha(A_1 = \beta, f(A_1) = j)$ 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 by $\mathbf{Q}^{(\ell)}$ the matrix $(Q^{(\ell)}_{\alpha\beta})_{\alpha,\beta \in \mathcal{A}}$.

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

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

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

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

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,

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

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

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

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

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

hence

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

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

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

Example 1. 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

$$\mathbf{Q}^{(-1)} = \mathbf{P}^{(-1)} + \mathbf{P^{(0)}} \cdot \mathbf{Q}^{(-1)} + \mathbf{P^{(1)}}(\mathbf{Q}^{(-1)})^2 \quad \text{and } \mathbf{Q} = \mathbf{Q}^{(-1)}. \tag{28}$$

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

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

$$\mathbf{L}^{(\ell)} = \mathbf{P}^{(\ell)} + \mathbf{P}^{(\mathbf{0})}\mathbf{L}^{(\ell)} + \sum_{k=1}^{u} \mathbf{P}^{(-\mathbf{k})} \sum_{s} \sum_{t \in \mathcal{T}'_s} \prod_{i=1}^{s} \mathbf{L}^{(\mathbf{t_i})}. \tag{29}$$

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

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

Example 2. In the case where $u = v = 1$, (29) gives

$$\mathbf{L}^{(\mathbf{1})} = \mathbf{P}^{(\mathbf{1})} + \mathbf{P}^{(\mathbf{0})} \cdot \mathbf{L}^{(\mathbf{1})} + \mathbf{P}^{(-\mathbf{1})} \cdot (\mathbf{L}^{(\mathbf{1})})^2, \tag{30}$$

$$\mathbf{L}^{(\ell)} = 0 \text{ for } \ell > 1, \text{ thus } \mathbf{L}(\infty) = \mathbf{L}^{(\mathbf{1})} \cdot \mathbf{1}_r. \tag{31}$$

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 $F_{S^+,\alpha}(1) = 1 - L_\alpha(\infty) + \sum_{\beta \in \mathcal{A}} L_{\alpha\beta}^{(1)} \; F_{S^+,\beta}(0)$. With $\mathbf{L}(\infty) = (L_\alpha(\infty))_{\alpha \in \mathcal{A}}$, seen as a column vector, we can write

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

$$\mathbf{F}_{S^+,\cdot}(\ell) = 1 - \mathbf{L}(\infty) + \sum_{k=1}^{\ell} \mathbf{L}^{(\mathbf{k})}\mathbf{F}_{S^+,\cdot}(\ell - k), \quad \text{for all } \ell \geq 1.$$

See Section 4.2 for how to compute $\mathbf{L}^{(\mathbf{k})}$ 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, and 1, corresponding to the case $u = v = 1$. We will use the results in the previous subsections (see (28, 30, 31)) to derive the distribution of the maximal non-negative partial sum S^+. This distribution can be determined using the following matrix equalities:

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

with $\mathbf{L}^{(\mathbf{1})}$ given in (29) and

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

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

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.

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

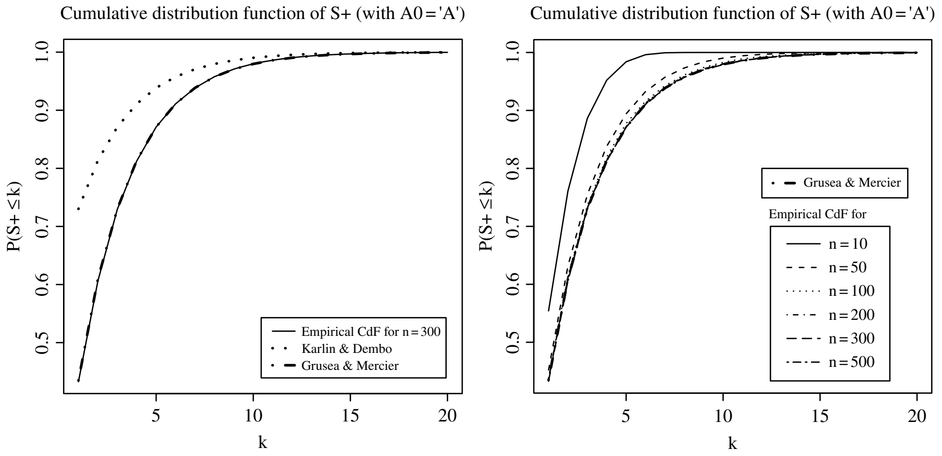

FIGURE 1: Cumulative distribution function of S^+ for the simple scoring scheme $(-1, 0, +1)$ and $A_0 = A$. Left panel: Comparison between the approximation of [9], a Monte Carlo estimation with sequences of length $n = 300$, and the exact formula proposed in Theorem 2.1. Right panel: Comparison, for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

We now present 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 (5.12) of [9] in order to compute $\mathbf{L}^{(1)}$ and $\mathbf{Q}^{(-1)}$, solutions of (28) and (30), from which we derive the approximate formulas proposed in 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 [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 the left panel of Figure 1 that for $n = 300$ the empirical CDF of S^+ and that obtained using Theorem 2.1 match perfectly. We can also visualize the fact that Theorem 2.1 improves the approximation of 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 us 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 satisfactorily in this example, even for the 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: Karlin and Dembo's [9] asymptotic bounds (the lower bound, depending on K^+ and recalled in (6), and the upper bound, depending on K^* and recalled in (7)), the approximation proposed in Theorem 2.4, and

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

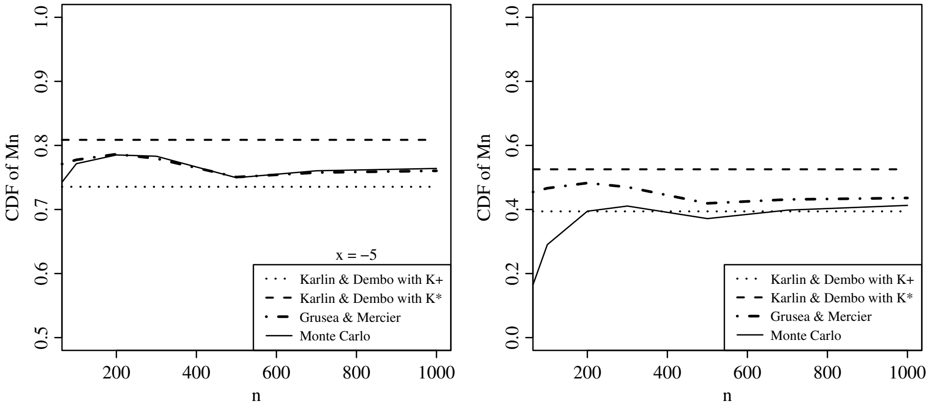

FIGURE 2: Comparison of the different approximations for $p(n, x) = \mathrm{P}\Big(M_n \leq \frac{\log(n)}{\theta^*} + x\Big)$ as a function of n, for fixed x and for the simple scoring scheme $(-1, 0, +1)$: the asymptotic lower and upper bounds of [9] (see (6) and (7)), the approximation proposed in Theorem 2.4, and a Monte Carlo estimation. Left panel: $p(n, x)$ for $x = -5$. Right panel: $p(n, x)$ for $x = -8$.

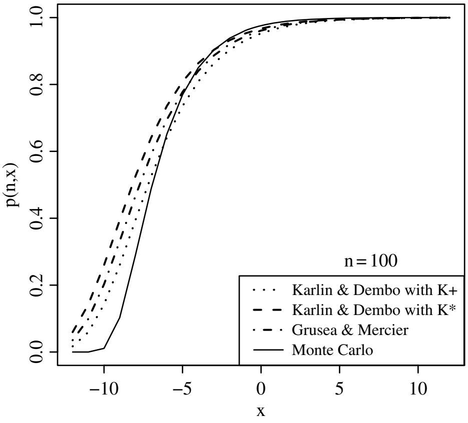

FIGURE 3: Comparison of the different approximations for $p(n, x) = \mathrm{P}\Big(M_n \leq \frac{\log(n)}{\theta^*} + x\Big)$ as a function of x, for fixed $n = 100$, and for the simple scoring scheme $(-1, 0, +1)$: the asymptotic lower and upper bounds of [9] (see (6) and (7)), the approximation proposed in Theorem 2.4, and a Monte Carlo estimation.

a Monte Carlo estimation. For the simple scoring scheme of this application, the parameter K^* appearing in the asymptotic bounds of [9] is given by their (5.6):

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

More precisely, in Figure 2 we plot the probability $p(n, x) := \mathrm{P}\big(M_n \leq \frac{\log(n)}{\theta^*} + x\big)$ as a function of n, for two fixed values $x = -5$ and -8. This illustrates the asymptotic behaviour of this probability with growing n. We can also observe the fact that Karlin and Dembo's asymptotic

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

bounds do not depend on n. In Figure 3, we compare the asymptotic bounds of [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 $\mathbb{P}$, for which we give each row P_x. For each example we also show the corresponding invariant probability π, which is in general close to the score frequencies, as expected. Biologists have warned us that since 1990 the sequences referenced in [8] may have changed a little bit due to the evolution of sequencing, which can explain some small differences in score frequencies between our sequences and those in [8]. Note that our Hypotheses (1) and (2) are both satisfied 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 [9] (p_{KDe} refers to the bound with K^* based on (7), and $p_{\mathrm{KDe}-K^+}$ refers to the bound with K^+ based on (6)), the approximation proposed 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 $p_{\mathrm{MC}} \leq p_{\mathrm{GMe}} \leq p_{\mathrm{KDe}} \leq p_{\mathrm{KDe}-K^+}$, except in Example (d)(ii), where we have $p_{\mathrm{GMe}} \leq p_{\mathrm{MC}} \leq p_{\mathrm{KDe}} \leq p_{\mathrm{KDe}-K^+}$. 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:

$$RE(\mathrm{KDe}) = 100 \cdot \frac{p_{\mathrm{KDe}} - p_{\mathrm{MC}}}{p_{\mathrm{MC}}}, \qquad RE(\mathrm{GMe}) = 100 \cdot \frac{p_{\mathrm{GMe}} - p_{\mathrm{MC}}}{p_{\mathrm{MC}}}. \tag{32}$$

The p-value given by [8] in the i.i.d. model ($p_{\mathrm{KDe\text{-}iid}}$) 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 [9], denoted KDe), and the empirical distribution obtained by Monte Carlo simulations, denoted MC:
the Kolmogorov–Smirnov distance:

$$d_{\mathrm{KS}}(\mathrm{GMe}) := \max_x \left(|P_{\mathrm{GMe}}(M_n \leq x) - P_{\mathrm{MC}}(M_n \leq x)|\right); \tag{33}$$

the Kullback–Leibler divergence:

$$d_{\mathrm{KL}}(\mathrm{GMe}) := \sum_x P_{\mathrm{MC}}(M_n = x) \cdot \log\left(\frac{P_{\mathrm{MC}}(M_n = x)}{P_{\mathrm{GMe}}(M_n = x)}\right). \tag{34}$$

Similarly, We define $d_{\mathrm{KS}}(\mathrm{KDe})$ and $d_{\mathrm{KL}}(\mathrm{KDe})$ using the asymptotic upper bound of [9] for the distribution of the local score (see (7)).

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

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

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

TABLE 1. Numerical comparison between our approximation for the local score distribution and the one from [9], using relative errors (see (32)) and two classical dissimilarity measures recalled in (33) and (34).

Example	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	$\approx 10^{-3}$	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%

Example (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. $f_{-1} = 82.2\%$; $f_2 = 17.8\%$. $P_{-1} = (0.784, 0.216)$; $P_2 = (0.821, 0.179)$. $p_{\text{KDe}} = 5.06 \cdot 10^{-3}$; $p_{\text{GMe}} = 1.51 \cdot 10^{-3}$; $p_{\text{MC}} = 1.41 \cdot 10^{-3}$. $\pi = [0.792; 0.208]$.

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

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

Zeste protein (UniProtKB-P09956): $n = 575$, $M_n = 11$, positions 194–209. $f_{-2} = 8.0\%$; $f_{-1} = 82.8\%$; $f_2 = 9.2\%$. $P_{-2} = (0.109, 0.696, 0.195)$; $P_{-1} = (0.078, 0.853, 0.069)$; $P_2 = (0.075, 0.717, 0.208)$. $p_{\text{KDe}} = 5.76 \cdot 10^{-1}$; $p_{\text{GMe}} = 5.21 \cdot 10^{-2}$; $p_{\text{MC}} = 5.04 \cdot 10^{-2}$; $p_{\text{KDe-iid}} = 3.7 \cdot 10^{-3}$. $\pi = [0.080; 0.828; 0.092]$.

Example (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. $f_{-2} = 9.2\%$; $f_{-1} = 79.7\%$; $f_2 = 11.1\%$. $P_{-2} = (0.208, 0.698, 0.094)$; $P_{-1} = (0.068, 0.827, 0.105)$; $P_2 = (0.172, 0.656, 0.172)$. $p_{\text{KDe}} = 13.9 \cdot 10^{-2}$; $p_{\text{GMe}} = 2.2 \cdot 10^{-2}$; $p_{\text{MC}} = 2.1 \cdot 10^{-2}$; $p_{\text{KDe-iid}} = 4 \cdot 10^{-3}$. $\pi = [0.093; 0.796; 0.111]$.

Example (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. $f_{-2} = 34.6\%$; $f_{-1} = 33.7\%$; $f_1 = 31.7\%$. $P_{-2} = (0.466, 0.230, 0.304)$; $P_{-1} = (0.254, 0.449, 0.297)$; $P_1 = (0.314, 0.337, 0.349)$. $p_{\text{KDe}} = 5.82 \cdot 10^{-4}$; $p_{\text{GMe}} = 7.31 \cdot 10^{-5}$; $p_{\text{MC}} = 6 \cdot 10^{-5}$; $p_{\text{KDe-iid}} = 1.8 \cdot 10^{-5}$. $\pi = [0.346; 0.338; 0.316]$.

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

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press

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

Acknowledgements

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.* **58**, 437–459.

[2] Cellier D., Charlot, F. and Mercier, S. (2003). An improved approximation for assessing the statistical 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. Prob.*, **19**, 1756–1767.

[4] Durbin, R., 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. *et al.* (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. *Molec. Ecol.* **26**, 3700–3714.

[6] Guedj, M. *et al.* (2006). Detecting local high-scoring segments: a first-stage approach for genome-wide association studies. *Statist. Appl. Genet. Mol. Biol.*, **5**, 22.

[7] Hassenforder, C. and Mercier, S. (2007). Exact distribution of the local score for Markovian sequences. *Ann. Inst. Statist. Math.*, **59**, 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. Comput. Biol.*, **8**, 373–380.

https://doi.org/10.1017/jpr.2019.75 Published online by Cambridge University Press