Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications 124 (2014) 4202–4223

stochastic processes and their applications

www.elsevier.com/locate/spa

Elements related to the largest complete excursion of a reflected BM stopped at a fixed time. Application to local score

Claudie Chabriac[a], Agnès Lagnoux[a,*], Sabine Mercier[a], Pierre Vallois[b]

[a] *Institut de Mathématiques de Toulouse, UMR 5219, Université Toulouse 2, 5 Allées Antonio Machado, 31058 Toulouse, France*

[b] *Institut Elie Cartan, Université de Lorraine, CNRS UMR 7502, INRIA, BIGS, Campus Sciences, BP 70239, Vandoeuvre-lès-Nancy Cedex, 54506, France*

Received 26 June 2013; received in revised form 16 June 2014; accepted 1 July 2014
Available online 18 July 2014

Abstract

We calculate the density function of $\left(U^*(t), \theta^*(t)\right)$, where $U^*(t)$ is the maximum over $[0, g(t)]$ of a reflected Brownian motion U, where $g(t)$ stands for the last zero of U before t, $\theta^*(t) = f^*(t) - g^*(t)$, $f^*(t)$ is the hitting time of the level $U^*(t)$, and $g^*(t)$ is the left-hand point of the interval straddling $f^*(t)$. We also calculate explicitly the marginal density functions of $U^*(t)$ and $\theta^*(t)$. Let U_n^* and θ_n^* be the analogs of $U^*(t)$ and $\theta^*(t)$ respectively where the underlying process (U_n) is the Lindley process, i.e. the difference between a centered real random walk and its minimum. We prove that $\left(\frac{U_n^*}{\sqrt{n}}, \frac{\theta_n^*}{n}\right)$ converges weakly to $\left(U^*(1), \theta^*(1)\right)$ as $n \to \infty$.

MSC: 60F17; 60G17; 60G40; 60G44; 60G50; 60G52; 60J55; 60J65

Keywords: Lindley process; Local score; Donsker invariance theorem; Reflected Brownian motion; Inverse of the local time; Brownian excursions

* Corresponding author. Tel.: +33 561504611.
E-mail address: lagnoux@univ-tlse2.fr (A. Lagnoux).
URL: http://www.lsp.ups-tlse.fr/Fp/Lagnoux (A. Lagnoux).

http://dx.doi.org/10.1016/j.spa.2014.07.003

1. Introduction

1.1. The local score is a probabilistic tool which is often used by molecular biologists to study sequences of either amino-acids or nucleotides as DNA. In particular its statistical properties allow to determine the most significant segment in a given sequence, see for instance [11,17]. Any position i in the sequence is allocated a random value ϵ_i. For example, ϵ_i can measure either physical or chemical property of the ith amino acid or nucleotide of the sequence. It can also code the similarity between two components of two sequences. It is assumed that $(\epsilon_i)_{i\geq 1}$ is a sequence of independent and identically distributed random variables. Rather than considering $(\epsilon_i)_{i\geq 1}$, it is more useful to deal with:

$$S_n = \epsilon_1 + \cdots + \epsilon_n \quad \text{for } n \geq 1;\ S_0 = 0. \tag{1.1}$$

Obviously, (S_n) is the random walk starting at 0, with independent increments $(\epsilon_i)_{i\geq 1}$. Let us introduce:

$$\underline{S}_n = \min_{0\leq i\leq n} S_i, \quad n \geq 0. \tag{1.2}$$

The two following processes (U_n) and $(\overline{U}_n)$ play an important role in the study of biological sequences. The first one is called the Lindley process and is defined as:

$$U_n = S_n - \underline{S}_n = S_n - \min_{i\leq n} S_i, \quad n \geq 0. \tag{1.3}$$

The process (U_n) is non negative and further properties can be found either in (Chapter III of [1]) (or Chapter I [6]). The local score $\overline{U}_n$ is the supremum of the Lindley process up to time n.

Molecular biologists are interested in "unexpected" large values of (U_n), see [17].

The exact distribution of $\overline{U}_n$ has been determined in [12], using the exponentiation of a suitable matrix and classical tools related to Markov chains theory. Although the given formula in [12] is efficient whatever the sign of $\mathbb{E}(\epsilon_i)$, in practice, it can be only applied to short sequences. However, we are sometimes faced with long sequences and in these situations it is often assumed that they have a negative trend, i.e. $\mathbb{E}(\epsilon_i) < 0$. Then, the local score $\overline{U}_n$ grows as $\ln(n)$ (see [18]) and an asymptotic approximation of the distribution of $\overline{U}_n$ as n is large has been given in [11,9], using the renewal theory. When $\mathbb{E}(\epsilon_n) = 0$, the asymptotic behavior of the tail distribution of $\overline{U}_n$ has been determined in [7] and the rate of convergence is given in [10].

Although the study of biological sequences is the starting point of this paper, the remainder will only consider the probabilistic model.

Here we consider that the $(\epsilon_i)_{i\geqslant 1}$ are centered with unit variance.

It is clear that the trajectory of (U_n) can be composed of a succession of 0 and excursions above 0. However, we only deal with *complete* excursions up to a fixed time. This leads us to introduce the maximum U_n^* of the heights of all the complete excursions up to time n. The second variable which will play an important role is θ_n^*, the time necessary to reach its maximum height U_n^*. See Section 3 for more information and detailed definitions of the previous RVs.

We believe that the knowledge of the joint distribution of the pair (U_n^*, θ_n^*) should permit the associated bi-dimensional statistical tests to be more powerful than the usual ones based on the first component. This program should be developed in a forthcoming paper.

1.2. Unfortunately, it is difficult to determine explicitly the law of $(U_n^*,\ \theta_n^*)$ for a fixed n. Bearing in mind applications with long biological sequences, it is relevant to study the distribution of $(U_n^*,\ \theta_n^*)$ where n is large. The functional convergence theorem of Donsker tells us that the