Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications 124 (2014) 268–288

stochastic processes and their applications

www.elsevier.com/locate/spa

Maximum likelihood estimator consistency for a ballistic random walk in a parametric random environment

Francis Comets[a], Mikael Falconnet[b], Oleg Loukianov[c,d], Dasha Loukianova[d], Catherine Matias[b,*]

[a] *Laboratoire Probabilités et Modèles Aléatoires, Université Paris Diderot, UMR CNRS 7599, France*
[b] *Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071, USC INRA, France*
[c] *Département Informatique, IUT de Fontainebleau, Université Paris Est, France*
[d] *Laboratoire Analyse et Probabilités, Université d'Évry Val d'Essonne, France*

Received 1 November 2012; received in revised form 23 April 2013; accepted 15 August 2013
Available online 26 August 2013

Abstract

We consider a one dimensional ballistic random walk evolving in an i.i.d. parametric random environment. We provide a maximum likelihood estimation procedure of the parameters based on a single observation of the path till the time it reaches a distant site, and prove that the estimator is consistent as the distant site tends to infinity. Our main tool consists in using the link between random walks and branching processes in random environments and explicitly characterising the limiting distribution of the process that arises. We also explore the numerical performance of our estimation procedure.

MSC: 62M05; 62F12; 60J25

Keywords: Ballistic regime; Branching process in random environment; Maximum likelihood estimation; Random walk in random environment

---

1. Introduction

Random walks in random environments (RWRE) have attracted much attention lately, mostly in the physics and probability theory literature. These processes were introduced originally by

---

[*] Corresponding author. Tel.: +33 164853552.

E-mail addresses: comets@math.univ-paris-diderot.fr (F. Comets), mikael.falconnet@genopole.cnrs.fr (M. Falconnet), oleg.loukianov@u-pec.fr (O. Loukianov), dasha.loukianova@univ-evry.fr (D. Loukianova), catherine.matias@genopole.cnrs.fr (C. Matias).

http://dx.doi.org/10.1016/j.spa.2013.08.002

Chernov [8] to model the replication of a DNA sequence. The idea underlying Chernov's model is that the protein that moves along the DNA strand during replication performs a random walk whose transition probabilities depend on the sequence letters, thus modelled as a random environment. Since then, RWRE have been developed far beyond this original motivation, resulting into a wealth of fine probabilistic results. Some recent surveys on the subject include [11,23].

Recently, these models have regained interest from biophysics, as they fit the description of some physical experiments that unzip the double strand of a DNA molecule. More precisely, some fifteen years ago, the first experiments on unzipping a DNA sequence have been conducted, relying on several different techniques (see [4,5], and the references therein). By that time, these experiments primarily took place in the quest for alternative (cheaper and/or faster) sequencing methods. When conducted in the presence of bounding proteins, such experiments also enabled the identification of specific locations at which proteins and enzymes bind to the DNA [15]. Nowadays, similar experiments are conducted in order to investigate molecular free energy landscapes with unprecedented accuracy [2,12]. Among other biophysical applications, one can mention the study of the formation of DNA or RNA hairpins [6].

Despite the emergence of data that is naturally modelled by RWRE, it appears that very few statistical issues on those processes have been studied so far. Very recently, Andreoletti and Diel [3] considered a problem inspired by an experiment on DNA unzipping [4,5,9], where the aim is to predict the sequence of bases relying on the observation of several unzipping of one finite length DNA sequence. Up to some approximations, the problem boils down to considering independent and identically distributed (i.i.d.) replicates of a one dimensional nearest neighbour path (*i.e.* the walk has ± 1 increments) in the same finite and two-sites dependent environment, up to the time each path reaches some value M (the sequence length). In this setup, the authors consider both a discrete time and a continuous time model. They provide estimates of the values of the environment at each site, which corresponds to estimating the sequence letters of the DNA molecule. Moreover, they obtain explicit formula for the probability to be wrong for a given estimator, thus evaluating the quality of the prediction.

In the present work, we study a different problem, also motivated by some DNA unzipping experiments: relying on an arbitrary long trajectory of a transient one-dimensional nearest neighbour path, we would like to estimate the parameters of the environment's distribution. Our motivation comes more precisely from the most recent experiments, that aim at characterising free binding energies between base pairs relying on the unzipping of a synthetic DNA sequence [17]. In this setup, the environment is still considered as random as those free energies are unknown and need to be estimated. While our asymptotic setup is still far from corresponding to the reality of those experiments, our work might give some insights on statistical properties of estimates of those binding free energies.

The parametric estimation of the environment distribution has already been studied in [1]. In their work, Adelman and Enriquez consider a very general RWRE and provide equations relating the distribution of some statistics of the trajectory to some moments of the environment distribution. In the specific case of a one-dimensional nearest neighbour path, those equations give moment estimators for the environment distribution parameters. It is worth mentioning that due to its great generality, the method is hard to understand at first, but it takes a simpler form when one considers the specific case of a one-dimensional nearest-neighbour path. Now, the method has two main drawbacks: first, it is not generic in the sense that it has to be designed differently for each parametric setup that is considered. Namely, the method relies on the choice of a one-to-one mapping between the parameters and some moments. In particular, when choosing a set of moment equations, injectivity of the induced mapping might even not be simple