Available online at www.sciencedirect.com
ScienceDirect
Stochastic Processes and their Applications ■ (■■■■) ■■■–■■■
stochastic processes and their applications
www.elsevier.com/locate/spa

Non parametric estimation for random walks in random environment

Roland Diel*, Matthieu Lerasle

Université Côte d'Azur, CNRS, LJAD, France

Received 20 September 2016; received in revised form 3 April 2017; accepted 22 April 2017
Available online xxxx

Abstract

We investigate the problem of estimating the cumulative distribution function (c.d.f.) F of a distribution ν from the observation of one trajectory of the random walk in i.i.d. random environment with distribution ν on $\mathbb{Z}$. We first estimate the moments of ν, then combine these moment estimators to obtain a collection of estimators $(\widehat{F}_n^M)_{M\geq 1}$ of F, our final estimator is chosen among this collection by Goldenshluger–Lepski's method. This estimator is easily computable. We derive convergence rates for this estimator depending on the Hölder regularity of F and on the divergence rate of the walk. Our rate is minimal when the chain realizes a trade-off between a fast exploration of the sites, allowing to get more information and a larger number of visits of each site, allowing a better recovery of the environment itself.

MSC: primary 62G05; secondary 62E17; 60K3
Keywords: Random walk in random environment; Non-parametric estimation; Oracle inequalities; Adaptive estimation

1. Introduction

Since its introduction by Chernov [6] to model DNA replication, random walks in random environment (RWRE) on $\mathbb{Z}^d$ have been widely studied in the probabilistic literature. This model is now well understood in the case $d = 1$, the case $d > 1$ is more complex and only partial results have been obtained. A recent overview can be found for example in [25].

In this paper, we are interested in estimating the distribution ν from the observation of one trajectory of a random walk in random environment ν on $\mathbb{Z}$. The topic of statistical inference for

* Corresponding author.
E-mail addresses: roland.diel@unice.fr (R. Diel), mlerasle@unice.fr (M. Lerasle).

http://dx.doi.org/10.1016/j.spa.2017.04.011

RWRE is quite new, it emerged with the appearance of statistical data fitting this model such as data related to the DNA-unzipping experiment or DNA-polymerase phenomenon. The problem considered may also be useful in applications where one wants to recover the environment itself. A purely frequentist estimator of the environment has poor performances on sites where the walk only spent a short time, it is likely to be outperformed by a Bayesian estimator, provided that the prior distribution is reasonably chosen. An estimator of ν could provide such a reasonable prior.

In the literature, the problem was originally considered in [1] who introduced an estimator of the moments of the distribution. The state space of the walk in [1] is more general than $\mathbb{Z}$ but their estimators have a huge variance, they are therefore unstable and cannot really be used in practice. More recently, [7,8,11,12] considered the random walk on $\mathbb{Z}$ and investigated the problem in a parametric framework. They proved consistency and asymptotic normality of the maximum likelihood estimator in the ballistic and sub-ballistic regimes and its efficiency in the ballistic regime. Some consistency results have been extended to the recurrent regime in [7] for a slightly different estimator. The case of Markovian environment has also been investigated in [2]. Although very interesting, this approach suffers several drawbacks both for practical applications and from a statistical perspective. First, the results are stated in a purely asymptotic framework where the number n of sites visited by the walk tends to infinity. Next, the quality of the estimator strongly relies on the assumption that the unknown distribution lies in a parametric model. Both assumptions impose severe restrictions for applications. The robustness of the procedure to a misspecified model for the unknown distribution, or the dependence of the performances of the maximum likelihood estimator with respect to an increasing number of parameters to recover are not considered. Moreover, the maximum likelihood estimator can be evaluated only after solving a maximization problem that is computationally intractable in general. Finally, the estimators of [7,8,11,12] are not exactly the same depending on the regime of the walk (recurrent or transient). This is an important problem from a statistical perspective since the regime depends on the unknown distribution of the observations, see Section 2 for details.

In this paper, we propose by contrast a non-asymptotic and non-parametric approach to tackle the estimation of the unknown cumulative distribution function (c.d.f.) of the environment from one observation of the walk. All our concentration results are valid in any regime, the only difference between the regimes lies in the convergence rate of the c.d.f. estimator. Our approach is based on the estimation of the moments of the unknown distribution, these estimations can always be performed in linear time. Those primary estimators are then combined to build a collection of estimators with non-increasing bias and non-decreasing variance and the final estimator is chosen among them according to Goldenshluger–Lepski's method [14]. The resulting estimator is therefore easily computable and provides at least a starting point to an optimization algorithm computing the maximum likelihood. It satisfies an oracle type inequality, meaning that it performs as well as the best estimator of the original collection. The oracle type inequality is used to obtain rates of convergence under regularity assumptions on the unknown c.d.f. More precisely, the rate of convergence of our estimator, stated in Theorem 1 in terms of the number n of visited sites, is given in the recurrent case by $\frac{\log n}{\sqrt{n}}$ and in the transient case by $\left(\frac{\log n}{n}\right)^{\gamma/(2\gamma+4\kappa)}$ where γ is the Hölder regularity of the unknown c.d.f. and $\kappa > 0$ is a parameter related to the rate at which the chain derives to infinity, see Section 2 for details. This rate can be compared with the one we would achieve if we observed the environment $(\omega_x)_{1\leq x\leq n}$. Actually, the empirical c.d.f. is known to converge at rate $1/\sqrt{n}$, without assumptions on the regularity of F by the Kolmogorov–Smirnov theorem and Dvoretzky–Kiefer–Wolfowitz inequality [10,19] gives a precise non-asymptotic concentration inequality. Our result is therefore much weaker, which is