Available online at www.sciencedirect.com

SciVerse ScienceDirect

Stochastic Processes and their Applications 121 (2011) 2488–2506

stochastic processes and their applications

www.elsevier.com/locate/spa

Context tree selection: A unifying view

A. Garivier[a,*], F. Leonardi[b]

[a] LTCI, CNRS, Telecom ParisTech, 75634 Paris Cedex 13, France
[b] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

Received 10 November 2010; received in revised form 27 June 2011; accepted 30 June 2011
Available online 14 July 2011

Abstract

Context tree models have been introduced by Rissanen in [25] as a parsimonious generalization of Markov models. Since then, they have been widely used in applied probability and statistics. The present paper investigates non-asymptotic properties of two popular procedures of context tree estimation: Rissanen's algorithm Context and penalized maximum likelihood. First showing how they are related, we prove finite horizon bounds for the probability of over- and under-estimation. Concerning over-estimation, no boundedness or loss-of-memory conditions are required: the proof relies on new deviation inequalities for empirical probabilities of independent interest. The under-estimation properties rely on classical hypotheses for processes of infinite memory. These results improve on and generalize the bounds obtained in Duarte et al. (2006) [12], Galves et al. (2008) [18], Galves and Leonardi (2008) [17], Leonardi (2010) [22], refining asymptotic results of Bühlmann and Wyner (1999) [4] and Csiszár and Talata (2006) [9].

Keywords: Algorithm Context; Penalized maximum likelihood; Model selection; Variable length Markov chains; Bayesian information criterion; Deviation inequalities

1. Introduction

Context tree models (CTM), first introduced by Jorma Rissanen in [25] as efficient tools in Information Theory, have been successfully studied and used since then in many fields of Probability and Statistics, including Bioinformatics [2,5], Universal Coding [27], Mathematical

* Corresponding author.
E-mail addresses: aurelien.garivier@telecom-paristech.fr, garivier@telecom-paristech.fr (A. Garivier), florencia@usp.br (F. Leonardi).

0304-4149/$ - see front matter
doi:10.1016/j.spa.2011.06.012

Statistics [4] or Linguistics [15]. Sometimes also called Variable Length Markov Chain (VLMC), a context tree process is informally defined as a Markov chain whose memory length depends on past symbols. This property makes it possible to represent the set of memory sequences as a tree, called the *context tree* of the process.

A remarkable tradeoff between expressivity and simplicity explains this success: no more difficult to handle than Markov chains, they appear to be much more flexible and parsimonious, including memory only where necessary. Not only do they provide more efficient models for fitting the data: it appears also that, in many applications, the shape of the context tree has a natural and informative interpretation. In Bioinformatics, the contexts trees of a sample have been useful to test the relevance of protein families databases [5] and in Linguistics, tree estimation highlights structural discrepancies between Brazilian and European Portuguese [15].

Of course, practical use of CTM requires the possibility of constructing efficient estimators of the model T_0 generating the data. It could be feared that, as a counterpart of the model multiplicity, increased difficulty would be encountered in model selection. Actually, this is not the case, and soon several procedures have been proposed and proved to be consistent. Roughly speaking, two families of context tree estimators are available. The first family, derived from the so-called algorithm *Context* introduced by Rissanen in [25], is based on the idea of *tree pruning*. They are somewhat reminiscent of the CART [3] pruning procedures: a measure of discrepancy between a node's children determines whether they have to be removed from the tree or not. The second family of estimators are based on a classical approach of mathematical statistics: *Penalized Maximum Likelihood* (PML). For each possible model, a criterion is computed which balances the quality of fit and the complexity of the model. In the framework of Information Theory, these procedures can be interpreted as derivations of the *Minimum Description Length* principle [1].

In the case of bounded memory processes, the problem of consistent estimation is clear: an estimator $\hat{T}$ is strongly consistent if it is equal to T_0 eventually almost surely as the sample size grows to infinity. As soon as 1983, Rissanen proved consistency results for the algorithm Context in this case. But later, the possibility of handling infinite memory processes was also addressed. In [9], an estimator $\hat{T}$ is called *strongly consistent* if for every positive integer K, its truncation $\hat{T}_{|K}$ at level K is equal to the truncation $T_0|_K$ of T_0 eventually almost surely. With this definition, PML estimators are shown to be strongly consistent if the penalties are appropriately chosen and if the maximization is restricted to a proper set of models. This last restriction was proven to be unnecessary in the finite memory case [19].

More recently, the problem of deriving *non-asymptotic* bounds for the probability of incorrect estimation was considered. In [18], non-universal inequalities were derived for a version of the algorithm Context in the case of finite context trees. These results were generalized to the case of infinite trees in [17], and to PML estimators in [22]. Using recent advances in weak dependence theory, all these results strongly rely on mixing hypotheses of the process.

For all these results, a distinction has to be made between two potential errors: under- and over-estimation. A context of T_0 is said to be *under-estimated* if one of its proper suffixes appears in the estimated tree $\hat{T}$, whereas it is called *over-estimated* if it appears as an internal node of $\hat{T}$. Over- and under-estimation appear to be of different natures: while under-estimation is eventually avoided by the existence of a strictly positive distance between a process and all processes with strictly smaller context trees, controlling over-estimation requires bounds on the fluctuations of empirical processes.

In this article, we present a unified analysis of the two families of context tree estimators. We contribute to a completely non-asymptotic analysis: we show that for appropriate parameters