# Regenerative Block-Bootstrap Confidence Intervals for the Extremal Index

Patrice Bertail, Stéphan Clémençon, Jessica Tressou

# Regenerative Block-Bootstrap Confidence Intervals for the Extremal Index

Patrice Bertail[1], Stéphan Clémençon[2], and Jessica Tressou[3]

[1] MODAL'X - Université Paris X Nanterre
(e-mail: pbertail@u-paris10.fr)

[2] LTCI UMR Institut Telecom/CNRS 5141- Telecom Paristech
(e-mail: stephan.clemencon@telecom-paristech.fr)

[3] Unité Met@risk - INRA
(email: jessica.tressou@agroparistech.fr)

**Abstract.** This paper is devoted to show how the *regenerative block-bootstrap* methodology (RBB), proved asymptotically valid in the case of sample means and $U$-statistics based on data drawn from a regenerative Markov chain $X = \{X_n\}_{n\in\mathbb{N}}$ in Bertail and Clémençon (2006a), Bertail and Clémençon (2006b), may be successfully extended for constructing confidence intervals for the extremal index of *instantaneous functions* $\{f(X_n)\}_{n\in\mathbb{N}}$. Precisely, this boils down to applying the RBB procedure to the *regenerative blocks estimator* proposed in Bertail et al. (2007b) for measuring the clustering tendency of high threshold exceedances. Asymptotic normality of this estimator is established, together with the asymptotic validity of the bootstrap distribution estimate, under mild stochastic stability assumptions. Eventually, a preliminary numerical experiment is presented, with the aim to empirically evaluate the capacity of the RBB for building accurate confidence intervals for the extremal index of the waiting process related to a M/M/1 queue.
**Keywords.** Atomic Markov chain, extremal index, regenerative block-bootstrap.

## 1 Introduction

The *regenerative block-bootstrap* (RBB) is a resampling procedure specifically tailored for regenerative processes, and thus, in particular, for atomic Markov chains. It is designed for estimating the distribution of specific statistics of interest, namely functionals of *regenerative blocks* (*i.e.* data segments corresponding to observations in between consecutive *regeneration times*) observed along a sample path. The RBB theoretical properties have been thoroughly investigated in the case of sample mean statistics in Bertail and Clémençon (2006b) (refer also to Bertail and Clémençon (2007a) for simulation studies), and preliminary encouraging results have been established for $U$-statistics in Bertail and Clémençon (2006a) (see also Bertail et al. (2008) for further developments). The principle underlying the RBB approach consists of resampling a random number of regenerative blocks and binding them together until the reconstructed trajectory is approximately the same size as the original one.

This way, the resampling method fully mimics the regenerative structure of the chain. Besides, the latter methodology can be extended to general Harris Markov chains, even if it entails to approximately simulate a regenerative Nummelin extension of the original chain (see §3.2 in Bertail and Clémençon (2006b)). Continuating these works, the present paper intends to apply the RBB to the (strongly consistent) extremal index estimator, based on regenerative data blocks (and termed *regenerative blocks estimator* for this reason), introduced in Bertail et al. (2007b) for instantaneous functions $\{f(X_n)\}_{n\in\mathbb{N}}$ of a general Harris chain $X$. Owing to space limitations, here we solely deal with the atomic/regenerative case, extensions to the general Harris setup being straightforward.

The outline of the paper is as follows. In section 2, notations are set out and the theoretical notions on the extremal index $\theta$ of an instantaneous function of a regenerative chain needed throughout the paper are briefly recalled. The RBB method for producing confidence intervals for $\theta$ from the regenerative blocks estimator $\hat{\theta}$ is presented in an algorithmic fashion in section 3. Asymptotic normality of $\hat{\theta}$ is proved under mild assumptions, as well as the asymptotic validity of the RBB in this setup. A short simulation study is displayed in section 4. Technical details are postponed to the Appendix.

## 2 Background

In the following, $X = \{X_n\}_{n\in\mathbb{N}}$ is a Harris positive recurrent Markov chain, valued in a measurable space $(E, \mathcal{E})$ with transition probability $\Pi(x, dy)$, invariant probability distribution $\mu$ and initial distribution $\nu$ (see Revuz (1984) for basics of the Markov chain theory).

Recall that $X$ is termed *regenerative* or *atomic* when it possesses an accessible atom, *i.e.*, a Harris measurable set $A$ such that $\Pi(x, .) = \Pi(y, .)$ for all $x$, $y$ in $A$. Denoting then by $\tau_A = \tau_A(1) = \inf\{n \geq 1,\ X_n \in A\}$ the hitting time on $A$ and by $\tau_A(j) = \inf\{n > \tau_A(j-1),\ X_n \in A\}$ for $j \geq 2$ the successive return times to $A$, it follows from the *strong Markov property* that the data blocks determined by the latter (namely, the *regeneration cycles*)

$$\mathcal{B}_1 = (X_{\tau_A(1)+1}, ...,\ X_{\tau_A(2)}), ...,\ \mathcal{B}_j = (X_{\tau_A(j)+1}, ...,\ X_{\tau_A(j+1)}), ...,$$

are i.i.d. random variables, valued in the torus $\mathbb{T} = \cup_{n=1}^{\infty} E^n$.

We also denote by $\mathbb{P}_\nu$ (respectively $\mathbb{P}_A$) the probability measure on the underlying space such that $X_0 \sim \nu$ (resp. $X_0 \in A$) and by $\mathbb{E}_\nu[.]$ (resp. $\mathbb{E}_A[.]$) the corresponding expectations. Let $f : E \to \mathbb{R}$ be a measurable function.

A key parameter in the extremal behavior analysis of an instantaneous function $\{f(X_n)\}_{n\in\mathbb{N}}$ of the chain $X$ is the *extremal index* $\theta \in (0, 1)$, measuring to which extent extreme values tend to come in "small clusters" (refer to Embrechts et al. (1997), Coles (2001), Finkenstadt and Rootzén (2003) for an account of this notion). Precisely, $\mathbb{P}_\mu(\max_{1\leq i\leq n} f(X_i)) \leq u_n) \underset{n\to\infty}{\sim}$

$F(u_n)^{n\theta}$, for any sequence $u_n$ such that $n(1-F(u_n)) \to \eta < \infty$, denoting by $F(x) = (\mathbb{E}_A[\tau_A])^{-1}\mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) \le x\}]$ the cdf of $f(X_1)$ in steady-state (*i.e.* under $\mathbb{P}_\mu$). It has been shown in Rootzén (1988) that $\theta = \lim_{n\to\infty} \theta(u_n)$ where

$$\theta(u) = \frac{\mathbb{P}_A(\max_{1\le i\le \tau_A} f(X_i) > u)}{\mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) > u\}]}. \tag{1}$$

From expression (1), the following estimator based on the observation of a trajectory of length $n$ has been proposed in Bertail et al. (2007b)

$$\widehat{\theta}_n(u) = \frac{\sum_{j=1}^{l_n-1} \mathbb{I}\{\zeta_j(f) > u\}}{\sum_{j=1}^{l_n-1} S_j(u)}, \tag{2}$$

with $\zeta_j(f) = \max_{1+\tau_A(j)\le i\le \tau_A(j+1)} f(X_i)$, $S_j(u) = \sum_{i=\tau_A(j)+1}^{\tau_A(j+1)} \mathbb{I}\{f(X_i) > u\}$, $l_n = \sum_{i=1}^{n} \mathbb{I}\{X_i \in A\}$ and the usual convention regarding empty summation (as well as $0/0 = 0$). The estimate (2) being a functional of the regenerative blocks, $\widehat{\theta}_n(u) = H_u(\mathcal{B}_1, \ldots, \mathcal{B}_{l_n-1})$, it may naturally be considered as an appropriate target for the RBB algorithm. The technical assumptions below are required in the analysis.
**Assumptions.** Let $\kappa \ge 1$. $\mathcal{H}(\kappa)\ :\ \mathbb{E}_A[\tau_A^\kappa] < \infty$ and $\mathcal{H}(\nu,\kappa)\ :\ \mathbb{E}_\nu[\tau_A^\kappa] < \infty$.

## 3 RBB confidence intervals for the extremal index

Given a trajectory $X^{(n)} = (X_0, X_1, \ldots, X_n)$ and the related regenerative blocks $\mathcal{B}_1, \ldots, \mathcal{B}_{l_n-1}$, in order to produce confidence intervals for $\theta(u)$, the RBB algorithm is then implemented in three steps as follows.

**Algorithm 1**

1. *Draw sequentially bootstrap data blocks $\mathcal{B}_1^*, \ldots, \mathcal{B}_k^*$ independently from the empirical distribution of the blocks $\mathcal{B}_1, \ldots, \mathcal{B}_{l_n-1}$ until the length of the bootstrap series $l^*(k) = \sum_{j=1}^{k} l(\mathcal{B}_j^*)$ is larger than $n$. Let $l_n^* = \inf\{k \ge 1, l^*(k) > n\}$.*
2. *From the bootstrap data blocks generated at step 1, reconstruct a pseudo-trajectory by binding the blocks together, getting the reconstructed RBB sample path $X^{*(n)} = (\mathcal{B}_1^*, \ldots, \mathcal{B}_{l_n-1}^*)$. Then compute the bootstrap version of the regenerative blocks estimator: $\theta_n^*(u) = H_u(\mathcal{B}_1^*, \ldots, \mathcal{B}_{l_n-1}^*)$.*
3. *A bootstrap confidence interval at level $1-\alpha \in (1/2, 1)$ for the parameter $\theta_{(u)}$ is obtained by computing the bootstrap root's quantiles $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$, of orders $\alpha/2$ and $1-\alpha/2$ respectively (in practice, the latter are approximated in a Monte-Carlo fashion by iterating steps 1-2):*

$$\mathcal{I}_{1-\alpha}^* = [\widehat{\theta}_n(u) + q_{\alpha/2}^*, \widehat{\theta}_n(u) + q_{1-\alpha/2}^*].$$

The next theorem establishes the asymptotic normality of (2) as an estimator of $\theta(u)$ as well as the asymptotic validity of the RBB distribution estimate, when applied to the latter estimate.

**Theorem 1.** *Let $u > 0$ be fixed. Under assumptions $\mathcal{H}(2)$ and $\mathcal{H}(\nu, 1)$, there exists a constant $\sigma_f^2(u) < \infty$ such that, as $n \to \infty$,*

$$\sqrt{n}(\widehat{\theta}_n(u) - \theta(u)) \to \mathcal{N}(0, \sigma_f^2(u)).$$

*Moreover, the RBB distribution estimate is asymptotically valid: under $\mathbb{P}_\nu$,*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}^* \left( \sqrt{n^*} \left( \theta_n^*(u) - \widehat{\theta}_n(u) \right) \leq x \right) - \mathbb{P}_\nu \left( \sqrt{n} \left( \widehat{\theta}_n(u) - \theta(u) \right) \leq x \right) \right| \to 0,$$

*where $\mathbb{P}^*(.)$ denotes the conditional probability given the original data $X^{(n)}$.*

A sketch of the proof is given in the Appendix.

## 4 Numerical experiment

A depiction of the numerical results produced by the RBB method applied in purpose of setting confidence limits for the extremal index of the waiting time process $X$ related to a M/M/1 queue is given in Fig. 1: $X_{n+1} = \max\{X_n + U_n - \Delta T_{n+1},\ 0\}$ where inter-arrival times and service times $(\Delta T_n)_{n\geq 1}$ and $(U_n)_{n\geq 1}$ are assumed independent from each other and i.i.d. with exponential distributions of respective intensities $\lambda$ and $\mu$. Assuming the *load condition* "$\lambda/\mu < 1$" is fulfilled, $X$ is classically positive recurrent with the empty file $\{0\}$ as atom and its extremal index is $\theta = (1 - \lambda/\mu)^2$ (see Hooghiemstra and Meester (1995)). Using threshold levels $u$ corresponding to high percentiles of the $X_k$'s with $n = 10000$ (represented along the $x$-axis), $\widehat{\theta}_n(u)$ is represented together with the 95%-RBB confidence interval obtained from $B = 200$ replications of the bootstrap sample ($y$-axis). This preliminary empirical result is encouraging, although, as may be proved by lengthy computations (see Bertail et al. (2008)), the length of the confidence interval goes to infinity as $u \uparrow \infty$.

## A Appendix - Proof of Theorem 1

Consider the empirical counterparts of $\bar{G}_f(u) = \mathbb{P}_A(\max_{1\leq i\leq \tau_A} f(X_i) > u)$ and $\Sigma_f(u) = \mathbb{E}_A(\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) > u\})$:

$$\bar{G}_n(u) = \frac{1}{l_n - 1} \sum_{j=1}^{l_n - 1} \mathbb{I}\{\zeta_j(f) > u\} \text{ and } \Sigma_{f,n}(u) = \frac{1}{l_n - 1} \sum_{j=1}^{l_n - 1} S_j(u).$$

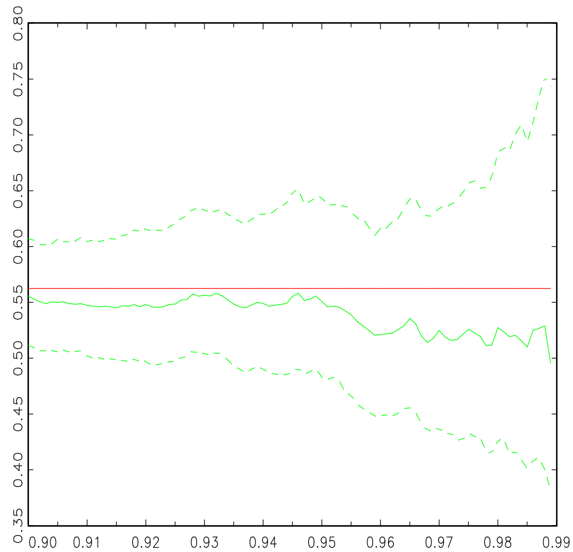


**Fig. 1.** Estimation of the extremal index in the M/M/1 queue with parameters $\lambda = 0.2$, $\mu = 0.8$, $\theta = 0.56$ and 95% RBB confidence interval.

Equipped with this notation, $\widehat{\theta}_n(u)$ is simply the ratio of the components of the bivariate vector $(\bar{G}_{f,n}(u), \Sigma_{f,n}(u))'$, which is asymptotically normal under the specified moment conditions (see the proof of the CLT stated in Theorem 17.3.6 of Meyn and Tweedie (1996)):

$$\sqrt{n}\begin{bmatrix} \bar{G}_{f,n}(u) \\ S_{f,n}(u) \end{bmatrix} - \begin{matrix} \bar{G}_f(u) \\ S_f(u) \end{matrix}\Bigg] \to \mathcal{N}\left(0, \mathbb{E}_A[\tau_A] \cdot \begin{pmatrix} \sigma_1^2(u) & \sigma_{12}(u) \\ \sigma_{12}(u) & \sigma_2^2(u) \end{pmatrix}\right),$$

since $(l_n - 1)/n \to \mathbb{E}_A[\tau_A]$, $\mathbb{P}_\nu$-a.s. as $n \to \infty$, and with

$$\sigma_1^2(u) = \bar{G}_f(u)(1 - \bar{G}_f(u)),\ \sigma_2^2(u) = \mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) > u\} - \Sigma_f(u)\right)^2\right],$$

$$\sigma_{12}(u) = \mathbb{E}_A\left[\left(\mathbb{I}\{\max_{1\leq i\leq\tau_A} f(X_i) > u\} - \bar{G}_f(u)\right)\left(\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) > u\} - \Sigma_f(u)\right)\right].$$

The first part of Theorem 1 is then classically proved by using the Delta Method, yielding

$$\sigma_f^2(u) = \mathbb{E}_A[\tau_A]\left[\frac{\sigma_1^2(u)}{\Sigma_f(u)^2} - 2\frac{\sigma_{12}(u)\bar{G}_f(u)}{\Sigma_f(u)^3} + \frac{\bar{G}_f(u)^2\sigma_2^2(u)}{\Sigma_f(u)^4}\right],$$

and the second part immediately follows. Details are omitted owing to space limitations.

## References


Ancona-Navarette, M. A. and Tawn, J. A. (2000): A comparison of methods for estimating the extremal index. *Extremes*, *3*, 5–38.

Bertail, P. and Clémençon, S. (2006a): *Regeneration-based statistics for Harris recurrent Markov chains*, pages 1–54. Number 187 in Lecture notes in Statistics. Springer.

Bertail, P. and Clémençon, S. (2006b): Regenerative-block bootstrap for Markov chains. *Bernoulli*, *12*, 689–712.

Bertail, P. and Clémençon, S. (2007a): Approximate regenerative block-bootstrap for markov chains. *To appear in Computational Statistics and Data Analysis*.

Bertail, P., Clémençon, S. and Tressou, J. (2007b): Extreme value statistics for Markov chains via the (pseudo-)regenerative method. Available at `http://hal.archives-ouvertes.fr/hal-00165652`.

Bertail, P., Clémençon, S. and Tressou, J. (2008): Regeneration-based $u$-statistics for markovian data. Working paper.

Coles, S. (2001): *An introduction to statistical modelling of Extreme Values*. Springer series in Statistics. Springer.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997): *Modelling Extremal Events for Insurance and Finance*. Applications of Mathematics. Springer-Verlag.

Finkenstadt, B. and Rootzén, H. (2003): *Extreme values in Finance, Telecommunications and the Environment*. volume 99 of *Monograph on Statistics and Applied Probability*. Chapman & Hall.

Hooghiemstra, G. and Meester, L. E. (1995): Computing the extremal index of special Markov chains and queues. *Stoch. Proc. Appl.*, *65*, 171–185.

Meyn, S. P. and Tweedie, R. L. (1996): *Markov Chains and Stochastic Stability*. Springer-Verlag.

Revuz, D. (1984): *Markov Chains*. 2nd edition, North-Holland.

Rootzén; H. (1988): Maxima and exceedances of stationary Markov chains. *Adv. Appl. Probab.*, *20*, 371–390.