Regenerative Block-Bootstrap Confidence Intervals for Tail and Extremal Indexes

Patrice Bertail, Stéphan Clémençon, Jessica Tressou

Regenerative Block-Bootstrap Confidence Intervals for Tail and Extremal Indexes

Patrice Bertail
patrice.bertail@u-paris10.fr
MODALX - Université Paris Ouest Nanterre
Stéphan Clémençon
stephan.clemencon@telecom-paristech.fr
LTCI - UMR Telecom ParisTech/CNRS No. 5141
Jessica Tressou*
jessica.tressou@agroparistech.fr
INRA Mét@risk - UR 1204 - 16 rue Claude Bernard
75231 Paris Cedex 5 - FRANCE

September 17, 2012

Abstract

A theoretically sound bootstrap procedure is proposed for building accurate confidence intervals of parameters describing the extremal behavior of instantaneous functionals $\{f(X_n)\}_{n\in\mathbb{N}}$ of a Harris Markov chain X, namely the extremal and tail indexes. Regenerative properties of the chain X (or of a Nummelin extension of the latter) are here exploited in order to construct consistent estimators of these parameters, following the approach developed in [10]. Their asymptotic normality is first established and the standardization problem is also tackled. It is then proved that, based on these estimators, the (approximate) regenerative block-bootstrap introduced in [7] yields asymptotically valid confidence intervals. In order to illustrate the performance of the methodology studied in this paper, simulation results are additionally displayed.

AMS classification: 60G70; 60J10; 60K20
Keywords: Regenerative Markov chain; Nummelin splitting technique; Extreme value statistics; Cycle submaximum; Hill estimator; Extremal index; Regenerative-block bootstrap

1 Introduction

As originally pointed out in [32], the extremal behavior of instantaneous functionals $f(X) = \{f(X_n)\}_{n\in\mathbb{N}}$ of a Harris recurrent Markov chain X may be described through the regenerative properties of the underlying chain, just like the asymptotic mean behavior. Following in the footsteps of this seminal contribution (see also [2]), the authors have recently investigated the performance of regeneration-based statistical procedures for estimating key parameters related to the extremal behavior analysis in the Markovian setup, see [10].

In particular, special attention has been paid to the problem of estimating the *extremal index* of the weakly dependent sequence $f(X)$, which measures to which extent extreme values tend to come in "small clusters", refer to [15], [11], [18] for an account of this notion. Various extremal index estimators have been recently proposed in the statistical literature [see 1, 23, 21, for instance], [29], [30] which generally rely on *blocking techniques*, where data segments of fixed (deterministic) length are considered in order to account for the dependence structure within the observations. Alternatively, an asymptotically valid methodology specifically tailored for (pseudo-) regenerative sequences has been proposed, based on data blocks of random length, corresponding to *cycles* in between successive regeneration times.

Proceeding in the same vein, it has been established in [10] that a regenerative version of the Hill estimator, computed from the set of *cycle submaxima*, namely maximum values observed in between consecutive renewal times, yields consistent estimation of the tail index of $f(X)$'s 1-d marginal distribution in the (supposedly existing) stationary regime, in the case when the latter belongs to the Fréchet maximum domain of attraction.

It is the purpose of this paper to continue this approach by investigating the problem of constructing confidence intervals for the extremal and tail indexes. We first prove the asymptotic normality of the regeneration-based estimators considered and then show how to studentize the latter in order to build asymptotic Gaussian confidence intervals. Next, we propose to extend the range of application of the *(approximate) regenerative block-bootstrap* (A-RBB in abbreviated form) originally introduced in [8] for bootstrapping Markovian sample means, to the present setting. Asymptotic validity of the ARBB procedure, when applied to the regeneration-based index estimates, is established and empirical simulations have been carried out, in order to evaluate empirically its performance when compared to Gaussian asymptotic intervals.

The article is structured as follows. Notations are first set out in Sec-

tion 2 and crucial notions related to the renewal properties of Harris Markov chains, that will be needed throughout the paper, are also briefly recalled. In Section 3, central limit theorems are stated for the regenerative versions of the "runs" and "blocks" estimators of the extremal index. Asymptotic normality of the regenerative Hill estimator is established and the studentization of these estimators is also investigated. Section 4 is devoted to the study of the (A)RBB methodology, when applied to the construction of confidence intervals based on the specific regeneration-based estimators considered. Finally, Section 5 displays preliminary simulation results, comparing the performance of bootstrap and Gaussian intervals. Technicalities are postponed to the Appendix.

2 Preliminaries

Throughout the article, we will denote by $X = \{X_n\}_{n\in\mathbb{N}}$ a time-homogeneous Harris recurrent Markov chain, valued in a measurable space $(E, \mathcal{E})$ with transition probability $\Pi(x, dy)$ and initial distribution ν [see 28, for an account of the Markov chain theory]. We also denote by $\mathbb{P}_\nu$ (respectively, by $\mathbb{P}_x$ with $x \in E$) the probability measure on the underlying space such that $X_0 \sim \nu$ (resp., $X_0 = x$) and by $\mathbb{E}_\nu[.]$ (resp., $\mathbb{E}_x[.]$) the corresponding expectation. We start off with recalling basic renewal properties of Harris Markov chains, while enhancing their connection with extremal behavior analysis.

2.1 Regenerative chains

Recall first that the chain X is said *regenerative* when it possesses a Harris recurrent atom, *i.e.*, a Harris set A such that: $\forall (x, y) \in A^2,\ \Pi(x, .) = \Pi(y, .)$. Set $\tau_A = \tau_A(1) = \inf\{n \geq 1,\ X_n \in A\}$ and $\tau_A(j) = \inf\{n > \tau_A(j-1),\ X_n \in A\}$ for $j \geq 2$. In the atomic case, by virtue of the strong Markov property, the sequence $\{\tau_A(k)\}_{k\geq 1}$ of successive return times to the atom forms a (possibly delayed) renewal process and more generally, the data segments, called *regeneration cycles*, determined by the times at which X forgets its past are i.i.d random variables valued in the torus $\mathbb{T} = \cup_{n=1}^{\infty} E^n$:

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

We denote by $\mathbb{P}_A$ the conditional probability measure given $X_0 \in A$ and by $\mathbb{E}_A[.]$ the $\mathbb{P}_A$-expectation.

In the regenerative setup, stochastic stability properties classically boil down to checking conditions related to the speed of return times to the

regenerative set. It is well-known for instance that X is *positive recurrent* if and only if $\alpha = \mathbb{E}_A[\tau_A] < \infty$ [see Theorem 10.2.2 in 25], and its (unique) invariant probability distribution μ is then the Pitman's occupation measure given by $\mu(B) = \alpha^{-1}\mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i \in B\}]$ for all $B \in \mathcal{E}$.

The following assumptions are involved in the subsequent analysis. Let $\kappa \geq 1$ and ν be any probability distribution on $(E, \mathcal{E})$.

$$\mathcal{H}(\kappa): \ \mathbb{E}_A[\tau_A^\kappa] < \infty \text{ and } \mathcal{H}(\nu, \kappa): \ \mathbb{E}_\nu[\tau_A^\kappa] < \infty.$$

Cycle submaxima. Let $f : (E, \mathcal{E}) \to \mathbb{R}$ be a measurable function. Consider the submaximum of the instantaneous functional $f(X) = \{f(X_n)\}_{n\in\mathbb{N}}$ over the j-th cycle, $j \geq 1$:

$$\zeta_j(f) = \max_{\tau_A(j)<k\leq\tau_A(j+1)} f(X_k).$$

It has been established in [32], see Theorem 3.1 therein, that, in the positive recurrent case, the distribution of the sampling maximum $M_n(f) = \max_{1\leq i\leq n} f(X_i)$ can be successfully approximated by the distribution of the maximum of $\lfloor n/\alpha \rfloor$ (roughly the mean number of cycles within a trajectory of length n) independent realizations of the cycle submaximum as $n \to \infty$, provided that the first (non regenerative) data segment plays no role in the extremal behavior, *i.e.* $\mathbb{P}_\nu(\max_{1\leq i\leq\tau_A} f(X_i) > \max_{1\leq j\leq l} \zeta_j(f)) \to 0$ as $l \to \infty$. More precisely, under these assumptions we have

$$\sup_{x\in\mathbb{R}} |\mathbb{P}_\nu(M_n(f) \leq x) - G_f(x)^{\lfloor n/\alpha \rfloor}| \to 0 \text{ as } n \to \infty, \tag{1}$$

where $G_f(x) = \mathbb{P}_A(\max_{1\leq i\leq\tau_A} f(X_i) \leq x)$ for all $x \in \mathbb{R}$. This shows that the tail behavior of the cycle submaximum's distribution $G_f(dx)$ rules the extremal behavior of the sequence $f(X)$.

2.2 Regenerative extensions of general Harris chains

Although the class of regenerative Markov chains includes all chains with countable state space as well as many Markov models used in Operations Research for modeling queuing/storage systems, the existence of Harris regenerative set is a very restrictive assumption in practice, that is not fulfilled by most Harris chains. Here we briefly recall a theoretical construction, termed the *splitting technique* and originally introduced in [26], extending in some sense the probabilistic structure of a general Harris chain, so as to artificially build a regeneration set, together with a practical method for approximating the regenerative extension. It is based on the notion of Harris

small set. Recall that a Harris set $S \in \mathcal{E}$ is *small* for the chain X if there exist $m \in \mathbb{N}^*$, a probability measure Φ supported by S, and $\delta > 0$ such that

$$\forall x \in S, \forall A \in \mathcal{E}, \quad \Pi^m(x, A) \geq \delta \Phi(A), \tag{2}$$

where Π^m denotes the m-th iterate of Π. Roughly speaking, the small sets are the ones on which an iterate of the transition probability is uniformly bounded below. When (2) holds, one says that X fulfills the *minorization condition* $\mathcal{M}(m, S, \delta, \Phi)$. We point out that small sets do exist for Harris chains, see [22]. Suppose now that condition (2) is satisfied. Rather than replacing the original chain by the chain $\{(X_{nm}, ..., X_{n(m+1)-1})\}_{n\in\mathbb{N}}$, we take $m = 1$. The regenerative Markov chain onto which X is embedded is constructed by expanding the sample space in order to define a specific sequence $(Y_n)_{n\in\mathbb{N}}$ of independent Bernoulli r.v.'s with parameter δ. The joint distribution is obtained by randomizing the transition Π each time the chain X hits S, which occurs with probability one (recall that the chain X is Harris). In order to obtain an insight into this construction, observe first that, when $X_n \in S$, the conditional distribution of X_{n+1} given X_n may be viewed as the following mixture

$$\Pi(X_n, dy) = (1 - \delta)\frac{\Pi(X_n, .) - \delta\Phi(dy)}{1 - \delta} + \delta\Phi(dy),$$

of which second component is independent from X_n. More precisely, the so-termed *split chain* $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is built the following way: suppose that $X_n \in S$, if $Y_n = 1$ (which occurs with probability $\delta \in]0, 1[$), X_{n+1} is drawn from Φ, otherwise (*i.e.* if $Y_n = 0$, which happens with probability $1 - \delta$), X_{n+1} is drawn from $(1 - \delta)^{-1}(\Pi(X_n, .) - \delta\Phi(.))$. Clearly, $S \times \{1\}$ is an atom for the split chain, the latter inheriting all the communication and stochastic stability properties from X. In particular the data segments in between consecutive visits to $S \times \{1\}$ are independent.

On approximating the regenerative extension. Unfortunately, the split chain is a theoretical construction and the Y_n's cannot be observed in practice. A "plug-in" approach has been nevertheless proposed in [8], in order to generate, conditionally to $X^{(n+1)} = (X_1, \ldots, X_{n+1})$, a random vector $(\widehat{Y}_1, \ldots, \widehat{Y}_n)$ from (supposedly known) parameters (S, δ, Φ) in a way that its conditional distribution approximates the distribution of $(Y_1, \ldots, Y_n)$ conditioned upon $X^{(n+1)}$ in a certain sense that will be specified below. Here we assume that the conditional distributions $\Pi(x, dy)$ with $x \in E$ are dominated by a σ-finite measure $\lambda(dy)$ of reference, in a way that $\Pi(x, dy) = \pi(x, y)\cdot\lambda(dy)$ for all $x \in E$. This clearly implies that $\Phi(dy)$ is also absolutely continuous

with respect to $\lambda(dy)$, and that

$$\forall x \in S, \;\; \pi(x,y) \geq \delta\phi(y), \; \lambda(dy) \text{ almost surely,} \tag{3}$$

where $\Phi(dy) = \phi(y) \cdot \lambda(dy)$. Given the sample path $X^{(n+1)}$, the Y_i's are independent random variables. Precisely, the conditional distribution of Y_i is the Bernoulli distribution with parameter

$$\frac{\delta\phi(X_{i+1})}{\pi(X_i, X_{i+1})} \cdot \mathbb{I}\{X_i \in S\} + \delta \cdot \mathbb{I}\{X_i \notin S\}. \tag{4}$$

A natural way of mimicking the Nummelin splitting construction consists of computing first an estimate $\widehat{\pi}_n(x,y)$ of the transition density over S^2 based on the available sample path and such that $\widehat{\pi}_n(x,y) \geq \delta\phi(y)$ a.s. for all $(x,y) \in S^2$, and then generating independent Bernoulli random variables $\widehat{Y}_1, \ldots, \widehat{Y}_n$ given $X^{(n+1)}$, the parameter of $\widehat{Y}_i$ being obtained by plugging $\widehat{\pi}_n(X_i, X_{i+1})$ into (4) in place of $\pi(X_i, X_{i+1})$. We point out that, from a practical viewpoint, it actually suffices to draw the $\widehat{Y}_i$'s only at times i when the chain hits the small set S, $\widehat{Y}_i$ indicating whether the trajectory should be cut at time point i or not. Let $\widehat{l}_n = \sum_{1 \leq k \leq n} \mathbb{I}\{X_k \in S, \, Y_k = 1\}$. Proceeding this way, one gets the sequence of *approximate regeneration times*, namely the successive time points $\widehat{\tau}_S(1), \ldots, \widehat{\tau}_S(\widehat{l}_n)$ at which $(X, \widehat{Y})$ visits the set $S \times \{1\}$. One may then form the *approximate regeneration blocks* $\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n - 1}$, as well as the *approximate cycle submaxima*:

$$\widehat{\zeta}_j(f) = \max_{1 + \widehat{\tau}_S(j) \leq i \leq \widehat{\tau}_S(j+1)} f(X_i) \text{ with } j = 1, \ldots, \widehat{l}_n - 1. \tag{5}$$

Knowledge of the parameters (S, δ, ϕ) of condition (3) is required for implementing this approximation method. A practical method for selecting those parameters in a fully data-driven manner is described at length in [9]. The question of accuracy of this approximation has been addressed in [8]. Under the following assumptions, a sharp bound for the deviation between the distribution of $((X_i, Y_i))_{1 \leq i \leq n}$ and the one of the $((X_i, \widehat{Y}_i))_{1 \leq i \leq n}$ in the sense of the Mallows or Wasserstein distance has been established, which essentially depends on the rate ρ_n of the uniform convergence of $\widehat{\pi}_n(x,y)$ to $\pi(x,y)$ over $S \times S$.

A1. The MSE of $\widehat{\pi}$ is of order ρ_n when error is measured by the sup norm over S^2:

$$\mathbb{E}_\nu \left[\sup_{(x,y) \in S^2} |\widehat{\pi}(x,y) - \pi(x,y)|^2 \right] = O(\rho_n) \text{ as } n \to +\infty,$$

where (ρ_n) denotes a sequence of nonnegative numbers decaying to zero at infinity.

A2. The parameters S and ϕ are chosen so that $\inf_{x \in S} \phi(x) > 0$.

A3. We have $\sup_{(x,y) \in S^2} \pi(x,y) < \infty$ and $\sup_{n \in \mathbb{N}} \sup_{(x,y) \in S^2} \hat{\pi}_n(x,y) < \infty$ $\mathbb{P}_\nu$-a.s. .

3 Regeneration-based Extreme Value Statistics

In this section, we recall how to construct estimators of the extremal and tail indexes based on the (approximate) cycle submaxima following in the footsteps of [10]. For each estimator considered, asymptotic normality is established and the standardization problem is tackled.

3.1 Asymptotically normal estimators of the extremal index

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 [15], [11] and [18] for an account of this notion. Precisely, for a positive recurrent Markov chain X with limiting probability distribution μ and any measurable function $f : (E, \mathcal{E}) \to \mathbb{R}$, there always exists $\theta = \theta(f) \in [0,1]$ such that

$$\mathbb{P}_\mu(\max_{1 \leq i \leq n} f(X_i)) \leq u_n) \sim F(u_n)^{n\theta} \text{ as } n \to \infty, \tag{6}$$

for any sequence of real numbers $\{u_n\}$ such that $n(1 - F(u_n)) \to \eta$ for some $\eta < \infty$, denoting by $F(x) = (\mathbb{E}_A[\tau_A])^{-1}\mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) \leq x\}]$ the cdf of $f(X_1)$ in steady-state, *i.e.* under $\mathbb{P}_\mu$. As already observed in [10], a positive recurrent chain is *a fortiori* strong mixing (*cf* Theorem A in [4]) and consequently satisfies Leadbetter's mixing condition $D(u_n)$, see [24].

In the remainder of this subsection, the function $f(x)$ is fixed and the index θ is assumed to be strictly positive. We point out that [31] have proved, under an extra technical assumption, that the extremal index of any geometrically ergodic Markov chain is strictly positive, refer to Theorem 4.1 therein.

3.1.1 The regenerative "blocks" estimator

As originally shown in [32], it follows from (1) and (6) that, for any sequence $\{u_n\}$ such that $n(1-F(u_n)) \to \eta$ for some $\eta < \infty$, $\theta = \lim_{n\to\infty} \theta(u_n)$, where

$$\theta(u) = \frac{\bar{G}_f(u)}{\Sigma_f(u)}, \tag{7}$$

with $\Sigma_f(u) = \alpha\bar{F}(u) = \mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) > u\}]$, denoting by $\bar{G}(x) = 1 - G(x)$ the survivor function of any cdf $G(x)$, and the convention that $0/0 = 0$.

In the regenerative case, from expression (7), which may be viewed as a regenerative version of the popular "blocks" estimator (see §8.1.2 in [15]), has been proposed in [10]:

$$\theta_n(u) = \frac{\bar{G}_{f,n}(u)}{\Sigma_{f,n}(u)}, \tag{8}$$

where, for all $u \in \mathbb{R}$,

$$G_{f,n}(u) = \frac{1}{l_n - 1} \sum_{j=1}^{l_n - 1} \mathbb{I}\{\zeta_j(f) \le u\} \text{ and } \Sigma_{f,n}(u) = \frac{1}{l_n - 1} \sum_{j=1}^{l_n - 1} S_j(u),$$

with $S_j(u) = \sum_{i=\tau_A(j)+1}^{\tau_A(j+1)} \mathbb{I}\{f(X_i) > u\}$, $l_n = \sum_{i=1}^{n} \mathbb{I}\{X_i \in A\}$, and the usual convention regarding empty summation and $\frac{0}{0} = 0$.

Expectedly, a counterpart of this quantity in the general Harris case is obtained by replacing the regeneration cycle submaxima by their approximate versions in (8):

$$\widehat{\theta}_n(u) = \frac{1 - \widehat{G}_{f,n}(u)}{\widehat{\Sigma}_{f,n}(u)}, \tag{9}$$

where, for all $u \in \mathbb{R}$, $\widehat{G}_{f,n}(u) = \frac{1}{\widehat{l}_n - 1} \sum_{j=1}^{\widehat{l}_n - 1} \mathbb{I}\{\widehat{\zeta}_j(f) \le u\}$ and $\widehat{\Sigma}_{f,n}(u) = \frac{1}{\widehat{l}_n - 1} \sum_{j=1}^{\widehat{l}_n - 1} \widehat{S}_j(u)$, with $\widehat{S}_j(u) = \sum_{i=\widehat{\tau}_S(j)+1}^{\widehat{\tau}_S(j+1)} \mathbb{I}\{f(X_i) > u\}$ for $1 \le j \le \widehat{l}_n - 1$.

These estimators have been proved consistent in [10] under mild moment assumptions, see Proposition 4 therein. For clarity's sake, we precisely recall the related result.

Proposition 1. *([10]) Suppose that $\theta > 0$. Let $(r_n)_{n\in\mathbb{N}}$ increase to infinity in a way that $r_n = o(\sqrt{n/\log\log n})$ as $n \to \infty$. Consider $(v_n)_{n\in\mathbb{N}}$ such that $r_n(1 - G_f(v_n))/\alpha \to \eta < \infty$ as $n \to \infty$.*

(i) In the regenerative case, suppose that $\mathcal{H}(\nu,1)$ and $\mathcal{H}(2)$ are fulfilled. Then,

$$\theta_n(\nu_n) \to \theta \ \mathbb{P}_\nu\text{-almost surely, as } n \to \infty. \tag{10}$$

(ii) In the general case, assume that moment assumptions $\mathcal{H}(\nu,1)$ and $\mathcal{H}(4)$ are fulfilled by the split chain and in addition that conditions $A_1 - A_3$ are satisfied. Then,

$$\widehat{\theta}_n(\nu_n) \to \theta \text{ in } \mathbb{P}_\nu\text{-probability, as } n \to \infty. \tag{11}$$

Remark 1. *(On moment assumptions for the split chain) We point out that, in the pseudo-regenerative setup described in §2.2, a sufficient condition for condition $\mathcal{H}(\kappa)$ (respectively, for condition $\mathcal{H}(\nu,\kappa)$) to hold is $\widehat{\mathcal{H}}(\kappa)$: $\sup_{x\in S}\mathbb{E}_x[\tau_S^\kappa] < \infty$ (resp., $\widehat{\mathcal{H}}(\kappa,\nu)$: $\mathbb{E}_\nu[\tau_S^\kappa] < \infty$). Practically, drift conditions of Foster-Lyapounov type are used for checking such moment conditions, refer to Chapter 11 in [25] for further details.*

Remark 2. *(On the empirical choice of the threshold sequence) In practice, the threshold sequence $\{\nu_n\}$ must be picked by the statistician. A natural choice, based on the available sample, consists of taking $\nu_n = G_{f,n}^{-1}(1-\eta/r_n)$ in the regenerative case (respectively, $\nu_n = \widehat{G}_{f,n}^{-1}(1-\eta/r_n)$ in the pseudo-regenerative case) and one may easily shows that assertion (i) (resp., assertion (ii)) of Proposition 1 remains valid.*

The next result reveals that, for a fixed threshold $u \in \mathbb{R}$, the asymptotic distribution of the quantity (8), respectively (9), is Gaussian. The technical proof is given in the Appendix section.

Theorem 2. *Let $u > 0$ be fixed.*

(i) In the regenerative case, under assumptions $\mathcal{H}(2)$ and $\mathcal{H}(\nu,1)$, there exists a constant $\sigma_f^2(u) < \infty$ such that

$$\sqrt{n}\,(\theta_n(u) - \theta(u)) \Rightarrow \mathcal{N}(0, \alpha\cdot\sigma_f^2(u)) \text{ as } n \to \infty, \tag{12}$$

where $\Rightarrow$ denotes the convergence in distribution.

(ii) In the pseudo-regenerative case, if the moment assumptions $\mathcal{H}(\nu,1)$ and $\mathcal{H}(4)$ are fulfilled by the split chain and if conditions $A_1 - A_3$ are in addition satisfied, then

$$\sqrt{n}\left(\widehat{\theta}_n(u) - \theta(u)\right) \Rightarrow \mathcal{N}(0, \alpha\cdot\sigma_f^2(u)) \text{ as } n \to \infty. \tag{13}$$

As shown in Theorem 2's proof, the asymptotic variance is given by

$$\sigma_f^2(u) = \left[\frac{\sigma_1^2(u)}{\Sigma_f(u)^2} - 2\frac{\sigma_{12}(u)\bar{G}_f(u)}{\Sigma_f(u)^3} + \frac{\bar{G}_f(u)^2\sigma_2^2(u)}{\Sigma_f(u)^4} \right], \tag{14}$$

where

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

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

These quantities may be straightforwardly estimated by computing their empirical counterparts based on the (approximate) regeneration cycles. However, the following result shows that, for a properly chosen threshold sequence $\{v_n\}$, increasing to infinity at a suitable rate, the second and third terms on the right hand side of (14) vanish, while the first one converges to $(\alpha\eta)^{-1}\theta^2$ as $n \to \infty$.

Proposition 3. *Let $(r_n)_{n\in\mathbb{N}}$ increase to infinity in a way that $r_n = o(\sqrt{n/\log\log n})$ as $n \to \infty$. Consider $(v_n)_{n\in\mathbb{N}}$ such that $r_n(1-G_f(v_n))/\alpha \to \eta < \infty$ as $n \to \infty$.*

(i) In the regenerative case, provided that assumptions $\mathcal{H}(2)$ and $\mathcal{H}(v,1)$ are fulfilled, the following convergence in distribution holds:

$$\sqrt{n/r_n}\,(\theta_n(v_n) - \theta(v_n)) \Rightarrow \mathcal{N}(0, \theta^2/\eta),\ \text{as } n \to \infty. \tag{15}$$

(ii) In the pseudo-regenerative case, if the split chain satisfies $\mathcal{H}(v,1)$ and $\mathcal{H}(4)$ and conditions $A_1 - A_3$ hold, we have the following convergence:

$$\sqrt{n/r_n}\left(\widehat{\theta}_n(v_n) - \theta(v_n)\right) \Rightarrow \mathcal{N}(0, \theta^2/\eta) \text{ as } n \to \infty. \tag{16}$$

We point out that, under maximum domain of attraction (MDA) assumption combined with additional technical conditions, the asymptotic bias may be proved to vanish. Indeed, recall that, under the assumption that $\theta > 0$, the probability distributions $G_f(dx)$ and $F(dx)$ necessarily belongs to the same MDA. Suppose for instance that they belong to the Fréchet MDA. There exists then $a > 0$ such that one may write $\bar{G}_f(x) = L_1(x)\cdot x^{-a}$ and $\bar{F}(x) = L_2(x)\cdot x^{-a}$, where $L_1(x)$ and $L_2(x)$ are slowly varying functions. In

this setup, the extremal index is thus proportional to the limiting ratio of these two functions:

$$\theta(u) = \frac{L_1(u)}{\alpha L_2(u)}.$$

Assume in addition that some second-order Hall-type conditions are fulfilled

$$L_i(x) = \lim_{y\to\infty} L_i(y) + C_i \cdot x^{-\beta_i} + o(x^{-\beta_i})$$

as $x \to \infty$ where $C_i < \infty$ and $\beta_i > 0$, $i = 1, 2$. Then, $\theta(v_n)$ converges to θ at the rate $v_n^{-\beta}$ with $\beta = \beta_1 \wedge \beta_2$ and $v_n \sim r_n^{1/\beta_1}$. Hence, as soon as (r_n) is picked such that $n/r_n^{1+2\beta/\beta_1} \to 0$, we have that $\sqrt{n/r_n}\,(\theta_n(v_n) - \theta) \Rightarrow \mathcal{N}(0, \theta^2/\eta)$ as $n \to \infty$ in the regenerative case, and a similar result holds true in the pseudo-regenerative case.

3.1.2 The regenerative "runs estimator"

Using the regenerative method, it has been proved in [32] that θ may be expressed as a limiting conditional probability: if $n(1-G_f(u_n))/\alpha \to \eta < \infty$, we have $\theta = \lim_{n\to\infty} \theta'(u_n)$ where: $\forall u \in \mathbb{R}$,

$$\theta'(u) = \mathbb{P}_A(\max_{2\le i\le \tau_A} f(X_i) \le u \mid X_1 > u). \tag{17}$$

Based on a path $X_1, \ldots, X_n$, the natural empirical counterpart of (17) in the regenerative setting is

$$\theta'_n(u) = \frac{\sum_{j=1}^{l_n-1} \mathbb{I}\{\max_{2+\tau_A(j)\le i\le \tau_A(j+1)} f(X_i) \le u < f(X_{1+\tau_A(j)})\}}{\sum_{j=1}^{l_n-1} \mathbb{I}\{f(X_{1+\tau_A(j)}) > u\}}. \tag{18}$$

Insofar as (17) measures the clustering tendency of high threshold exceedances within regeneration cycles only, it should be seen as a "regenerative version" of the *runs estimator*

$$\hat{\theta}_n^{(r)}(u) = \frac{\sum_{j=1}^{n-r} \mathbb{I}\{\max_{j+1\le i\le j+r} f(X_i) \le u < f(X_j)\}}{\sum_{j=1}^{n-r} \mathbb{I}\{f(X_j) > u\}}, \tag{19}$$

obtained by averaging over overlapping data segments of fixed length r.

In the pseudo-regenerative case, a practical estimate is built by means of the approximate regeneration times:

$$\widehat{\theta}'_n(u) = \frac{\sum_{j=1}^{\widehat{l}_n-1} \mathbb{I}\{\max_{2+\widehat{\tau}_S(j)\le i\le \widehat{\tau}_S(j+1)} f(X_i) \le u < f(X_{1+\widehat{\tau}_S(j)})\}}{\sum_{j=1}^{\widehat{l}_n-1} \mathbb{I}\{f(X_{1+\widehat{\tau}_S(j)}) > u\}} \tag{20}$$

Beyond its practical advantage (blocks are here entirely determined by the data), the estimator (18) may be proved *strongly consistent* as stated in the first part of the next theorem, while only weak consistency has been established for (19) but for a wider class of weakly dependent sequences, see [21].

Theorem 4. *Let* r_n *increase to infinity in a way that* $r_n = o(\sqrt{n/\log\log n})$ *as* $n \to \infty$.

(i) *Assume that* $\mathcal{H}(\nu, 1)$ *is fulfilled. Considering* $(\nu_n)_{n\in\mathbb{N}}$ *such that* $r_n(1 - F(\nu_n)) \to \eta < \infty$ *as* $n \to \infty$, *we then have*

$$\theta'_n(\nu_n) \to \theta, \quad \mathbb{P}_\nu\text{-almost surely, as } n \to \infty.$$

(i′) *Similarly, if the split chain fulfills moment conditions* $\mathcal{H}(\nu, 1)$ *and* $\mathcal{H}(4)$ *and conditions* $A_1 - A_3$ *hold, then weak consistency holds in the pseudo regenerative case:*

$$\widehat{\theta}'_n(\nu_n) \to \theta, \text{ in } \mathbb{P}_\nu\text{-probability, as } n \to \infty.$$

(ii) *In the regenerative case, provided that assumption* $\mathcal{H}(\nu, 1)$ *is fulfilled, the following convergence in distribution also holds:*

$$\sqrt{n/r_n}\left(\theta'_n(\nu_n) - \theta(\nu_n)\right) \Rightarrow \mathcal{N}(0, \theta^2(1-\theta)/\eta), \text{ as } n \to \infty. \tag{21}$$

(ii′) *In the pseudo-regenerative case, if the split chain satisfies* $\mathcal{H}(4)$ *and* $\mathcal{H}(\nu, 1)$ *and conditions* A_1–A_3 *hold, we have the following convergence:*

$$\sqrt{n/r_n}\left(\widehat{\theta}'_n(\nu_n) - \theta(\nu_n)\right) \Rightarrow \mathcal{N}(0, \theta^2(1-\theta)/\eta) \text{ as } n \to \infty. \tag{22}$$

The last statement of the preceding theorem and Proposition 3 *(i)* constitute the regenerative versions of Theorems 3 and 4 in [33], who first proved the CLT for the classical *runs* estimator (based on blocks of fixed length, cf. (19)). The proof of the preceding theorem follows the lines of those of Proposition 1, Theorem 2 and Proposition 3, as sketched in the appendix section.

3.2 Asymptotic normality of the regeneration-based Hill estimator

In the section, we assume that $\theta > 0$ and hence, as recalled in the previous section, the distributions $G_f(dx)$ and $F(dx)$ belong to the same MDA. We assume here they belong to the Fréchet MDA. In the regenerative setting, a natural way of estimating F's tail index, proposed in [10], thus consists in computing a Hill estimate of G_f's tail index from the observed cycle submaxima:

$$\xi_{n,k} = \left(k^{-1} \sum_{i=1}^{k} \log \frac{\zeta_{(i)}(f)}{\zeta_{(k+1)}(f)} \right)^{-1}, \tag{23}$$

with $1 \leq k \leq l_n - 1$ when $l_n > 1$, denoting by $\zeta_{(j)}(f)$ the j-th largest submaximum. As $l_n \to \infty$, $\mathbb{P}_\nu$- almost surely as $n \to \infty$, asymptotic results established in the case of i.i.d. observations extend straightforwardly to our setting, see part (i) of Theorem 5 below. We point out that in the i.i.d. setup one may take the whole state space as an atom, *i.e.* $A = E$, each cycle comprises then a single observation and (23) reduces to the standard Hill estimator.

In the general Harris case, one may naturally build an estimate by replacing the cycle submaxima by their approximate versions:

$$\widehat{\xi}_{n,\,k} = \left(k^{-1} \sum_{i=1}^{k} \log \frac{\widehat{\zeta}_{(i)}(f)}{\widehat{\zeta}_{(k+1)}(f)} \right)^{-1}, \tag{24}$$

with $1 \leq k \leq \hat{l}_n - 1$ when $\hat{l}_n > 1$ and denoting by $\widehat{\zeta}_{(j)}(f)$ the j-th largest approximate submaximum. It is shown in Proposition 5 of [10] that the approximation step does not compromise the consistency of the estimator, provided that the estimator of $\pi(x, y)$ over S^2 is accurate enough. In order to establish a rate of convergence, we will also consider the case where the transition estimate used in the approximation stage is computed from a trajectory of length $N >> n$ and will denote by $\widehat{H}^{(N)}_{k,n,}$ the corresponding estimator.

The consistency and the asymptotic normality of these estimators have been shown in [10] under the Von Mises condition recalled below, see Proposition 5 therein.

VM assumption. (Von Mises condition, [19]) Let $\rho \leq 0$. Suppose

$\bar{G}_f(x) = L(x)x^{-a}$,

$$\lim_{x\to\infty} \frac{\bar{G}_f(tx)/\bar{G}_f(x) - t^{-a}}{b(x)} = t^{-a}\frac{t^{\rho}-1}{\rho}, \quad t > 0$$

where $b(x)$ is a measurable function of constant sign, and with, by convention, $(t^{-\rho}-1)/\rho = \log t$ when $\rho = 0$. Equivalently, if $U_f(t) = G_f^{-1}(1-t^{-1})$,

$$\lim_{x\to\infty} \frac{U_f(tx)/U_f(x) - t^{-1/a}}{B(x)} = t^{1/\xi}\frac{t^{\rho/a}-1}{\rho/a},$$

where $B(x) = a^{-2}b(U_f(x))$.

Here, we formulate a central limit theorem in a more general fashion, revealing a bias-variance trade-off similarly to [13] in the i.i.d. setup. The proof is omitted as it follows by a straightforward modification of the proof of proposition 5 in [10] and the references therein.

Theorem 5. *Assume that* F *belongs to the Fréchet MDA and the VM assumption holds and consider an increasing sequence of integers* $\{k(n)\}$ *such that:* $k(n) < n$, $k(n) = o(n)$ *and* $\log\log n = o(k(n))$ *as* $n \to \infty$. *Assume further that*

$$\lim \sqrt{k}B(n/k) = \lambda \in \mathbb{R}, \tag{25}$$

(i) then, in the regenerative case, the following convergence in distribution holds

$$\sqrt{k(l_n)}\left(\xi_{n,\,k(l_n)} - \xi\right) \Rightarrow \mathcal{N}\left(\frac{\xi^3\lambda}{\rho-\xi},\ \xi^2\right) \text{ under } \mathbb{P}_\nu, \text{ as } n\to\infty. \tag{26}$$

(ii) in the pseudo-regenerative case, if conditions $A_1 - A_3$ *are in addition fulfilled, let* $(m_n)_{n\in\mathbb{N}}$ *be a sequence of integers increasing to infinity such that* $m_n\sqrt{\rho_n/k(m_n)} \to 0$ *as* $n\to\infty$, *then*

$$\sqrt{k(\widehat{l}_{m_n})}\left(\widehat{\xi}_{m_n,\,k(\widehat{l}_{m_n})} - \xi\right) \Rightarrow \mathcal{N}\left(\frac{\xi^3\lambda}{\rho-\xi},\ \xi^2\right) \text{ under } \mathbb{P}_\nu, \text{ as } n\to\infty. \tag{27}$$

4 Regenerative block-bootstrap confidence intervals

In this section, we recall the principle underlying the (approximate) regenerative block-bootstrap, originally introduced in [8] for bootstrapping Markovian sample means, and establish its asymptotic validity when applied to the estimators described in the preceding section.

4.1 The (A)RBB principle

Practically, the (A)RBB algorithm applies to any statistic $\widehat{T}_n = T(\mathcal{B}_1, \ldots, \mathcal{B}_{l_n - 1})$, based the (approximate) cycles with standardization $\widehat{\sigma}_n = \sigma\left[T(\mathcal{B}_1, \ldots, \mathcal{B}_{l_n - 1})\right]$. For notational simplicity, regeneration cycles and their approximate versions are here denoted the same manner. The resampling scheme consists of mimicking the underlying renewal structure by drawing data blocks with replacement until a trajectory of length n roughly is built. This way, the randomness in the number of renewals is reproduced during the procedure and, conditionally to the original data, the bootstrap series thus generated is regenerative.

Algorithm 1. *(A)RBB algorithm*

1. (Blocks.) *Identify the (pseudo-) blocks $\mathcal{B}_1, \ldots, \mathcal{B}_{l_n - 1}$ from the observed trajectory $X_0, \ldots, X_n$ as explained in Section 2.1 (resp. in Section 2.2 in the pseudo-regenerative case) and compute the statistic $\widehat{T}_n = T(\mathcal{B}_1, \ldots, \mathcal{B}_{l_n - 1})$, and its standard deviation $\widehat{\sigma}_n = \sigma(\mathcal{B}_1, \ldots, \mathcal{B}_{l_n - 1})$.*

2. (Sequential drawing.) *Draw sequentially and independently bootstrap data blocks $\mathcal{B}_1^*, \ldots, \mathcal{B}_k^*$ from the empirical distribution of the blocks defined at step 1. until the length of the bootstrap series $l^*(k) = \sum_{j=1}^{k} l(\mathcal{B}_j^*)$ is larger than n. Let $l_n^* = \inf\{k \geq 1, l^*(k) > n\}$. If one is just interested in asymptotic results, one may just draw $l_n - 1$ i.i.d blocks (conditionally to the trajectory so that l_n is fixed in the bootstrap procedure).*

3. (Bootstrap statistics.) *From the bootstrap data blocks generated at step 2, reconstruct a pseudo-trajectory by binding the blocks together, getting the reconstructed RBB sample path $X^{*(n)} = (\mathcal{B}_1^*, \ldots, \mathcal{B}_{l_n^* - 1}^*)$ of length $n^* = l^*(l_n^* - 1)$. Then compute the bootstrap version of the regenerative blocks estimator: $T_n^* = T(\mathcal{B}_1^*, \ldots, \mathcal{B}_{l_n^* - 1}^*)$ and its standard deviation $\widehat{\sigma}_n^* = \sigma(\mathcal{B}_1^*, \ldots, \mathcal{B}_{l_n^* - 1}^*)$.*

4. (Bootstrap CIs.) *Bootstrap confidence intervals (CI) at level $1-\alpha \in (1/2, 1)$ for the parameter of interest are 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 2-3):the basic percentile bootstrap CI is simply*

$$[q_{\alpha/2}^*,\ q_{1-\alpha/2}^*],$$

the Percentile bootstrap CI is defined as

$$\left[2\widehat{T}_n - q^*_{1-\alpha/2},\ 2\widehat{T}_n - q^*_{\alpha/2}\right]$$

and the t-Percentile bootstrap CI is given by

$$\left[\widehat{T}_n - t^*_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}},\ \widehat{T}_n - t^*_{\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}\right],$$

where t^*_p *is the* p^{th} *quantile of the studentized bootstrap root* $\frac{T^*_n - \widehat{T}_n}{\widehat{\sigma}^*_n/\sqrt{n}}$.

Remark 3. (Gaussian confidence intervals) *These bootstrap CI's can be compared to asymptotic CI's classically built from the statistic and its standardization*

$$\left[\widehat{T}_n - \Phi^{-1}_{1-\alpha/2}.\widehat{\sigma}_n/\sqrt{n},\ \widehat{T}_n - \Phi^{-1}_{\alpha/2}\widehat{\sigma}_n/\sqrt{n}\right],$$

where Φ^{-1}_p *is the* p^{th} *quantile of the standard normal distribution, or replacing* $\widehat{\sigma}_n/\sqrt{n}$ *with a new standardization estimator defined as the empirical standard deviation of* T^*_n *given by* $\widetilde{\sigma}^{*2} = \sum_b (T^*_n - \overline{T}^*_n)^2/n$.

4.2 Asymptotic validity of (A)RBB distribution estimates

The results stated below show that the bootstrap procedure described in the previous subsection is asymptotically valid. Let $\mathbb{P}^*(.)$ be the conditional probability given the observed trajectory. The following assertions hold true.

Theorem 6. *1.* ("Blocks" estimator) *Suppose that the assumptions of Theorem 2 are fulfilled. Let* $\widetilde{\theta}_n(u)$ *denote the estimator* $\theta_n(u)$ *in the regenerative case,* $\widehat{\theta}_n(u)$ *in the pseudo-regenerative case, and let* $\widetilde{\theta}^*_n(u)$ *be its (A)RBB version. Then, we have, as* $n \to \infty$*:*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}^*\left(\sqrt{n}\left(\widetilde{\theta}^*_n(u) - \widetilde{\theta}_n(u)\right) \le x\right) - \mathbb{P}_\nu\left(\sqrt{n}\left(\widetilde{\theta}_n(u) - \theta(u)\right) \le x\right)\right| \to 0.$$

2. ("Runs" estimator) *Suppose that the hypotheses of Theorem 4 are satisfied. Denote by* $\widetilde{\theta}'_n(u)$ *the estimator* $\theta'_n(u)$ *in the regenerative case,* $\widehat{\theta}'_n(u)$ *in the pseudo-regenerative case, and let* $\widetilde{\theta}'^*_n(u)$ *be its (A)RBB version. Then, we have, as* $n \to \infty$*:*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}^*\left(\sqrt{n}\left(\widetilde{\theta}'^*_n(u) - \widetilde{\theta}'_n(u)\right) \le x\right) - \mathbb{P}_\nu\left(\sqrt{n}\left(\widetilde{\theta}'_n(u) - \theta(u)\right) \le x\right)\right| \to 0.$$

Such results may also be used to estimate the mean-square error of $\sqrt{n}(\theta_n(u) - \theta(u))$ and to calibrate the level u by minimizing the MSE, in the same spirit as [20] or [12], and as illustrated in the simulation section.

4.3 Markov subsampling and the Hill estimator

As claimed by the following proposition, the (A)RBB algorithm can also be successfully applied to tail index estimation provided that the sequential drawing (step 2 in the previous algorithm) is replaced with a subsampling drawing without replacement (see [14]). Proving that the procedure is still valid in absence of subsampling deserves a much thorougher analysis, far beyond the scope of this paper. We thus introduce the following subsampling variant of Algorithm 1.

Algorithm 2. *RBB subsampling*

1. (Blocks.) *As described in step 1 of Algorithm 1.*
2. (Subsampling drawing.) *Choose a subsampling size m_n large enough but small compare to n and compute l_{m_n} as the observed number of blocks in a stretch of length m_n : typically, l_{m_n} is of order $[\frac{m_n}{\mathbb{E}_A \tau_A}]$ and is thus asymptotically equivalent to $\widetilde{l}_{m_n} = \left[l_n \frac{m_n}{n}\right]$, where $[x]$ is the integer part of x. Draw $\widetilde{l}_{m_n}$ bootstrap data blocks $\mathcal{B}_1^*, \ldots, \mathcal{B}_k^*$ by sampling without replacement into the blocks $\mathcal{B}_1, \ldots, \mathcal{B}_{l_n - 1}$.*
3. (Subsampling statistics.) *Apply step 3. and 4. of Algorithm 1 to the reconstructed RBB sample path $X^{*(n)} = (\mathcal{B}_1^*, \ldots, \mathcal{B}^*_{\widetilde{l}_{m_n}-1})$.*

Theorem 7. *Suppose that assumptions of Theorem 5 are fulfilled. Denote by $\widetilde{\xi}_{n,k}$ the estimator $\xi_{n,k}$ in the regenerative case, $\widehat{\xi}_{n,k}$ in the pseudo-regenerative case, and let $\widetilde{\xi}^*_{n,k}$ be its subsampling counterpart.*

Let $m_n > 1$ such that $m_n \to +\infty$ and $m_n/n \to 0$ as $n \to +\infty$. If we assume in addition that $k(\widetilde{l}_{m_n})/k(l_n) \to 0$, we then have, as $n \to +\infty$,

$$\sup_{x \in \mathbb{R}} \left| \widetilde{H}^*_n(x) - \widetilde{H}_n(x) - \right| \to 0,$$

where $\widetilde{H}^*_n(x) = \mathbb{P}^* \left(\sqrt{k(\widetilde{l}_{m_n})} \left(\widetilde{\xi}^*_{m_n,\, k(l_{m_n})} - \widetilde{\xi}_{n,\, k(l_n)} \right) \le x \right)$ *and* $\widetilde{H}_n(x) = \mathbb{P}_\nu \left(\sqrt{k(l_n)} \left(\widetilde{\xi}_{n,\, k(l_n)} - \xi \right) \le x \right)$.

In the subsampling context, higher order accuracy can not be established. It is thus sufficient to consider a simple form of the standardization in order to prove the asymptotic validity. The issue of choosing the subsampling size m_n and the tuning parameter k is discussed in next section.

5 Simulation results

In this section, we present illustrative simulation results to provide empirical evidence of the nice behavior of the estimators and confidence intervals proposed in this paper. Whenever possible, a comparison with other estimators and confidence intervals is conducted.

5.1 Regenerative examples

Considering waiting times of certain queuing processes, we compute and discuss the regeneration-based "blocks" and "runs" estimators of the extremal index and the regeneration-based Hill estimator of the tail parameter.

Regeneration based extremal index estimators. We first consider the waiting times of a M/M/1 process (*cf* [3]) with parameters $\lambda = 0.2, \mu = 0.8$ and sample path length n. As underlined in [10], there exists a closed analytical form for the extremal index in this case, it is equal to $\theta = (1-\lambda/\mu)^2 = 0.5625$ and all the required assumptions are satisfied. The estimators $\theta_n(u)$ and $\theta'_n(u)$ of the extremal index proposed in this paper are both defined based on a threshold u, supposed to be large.

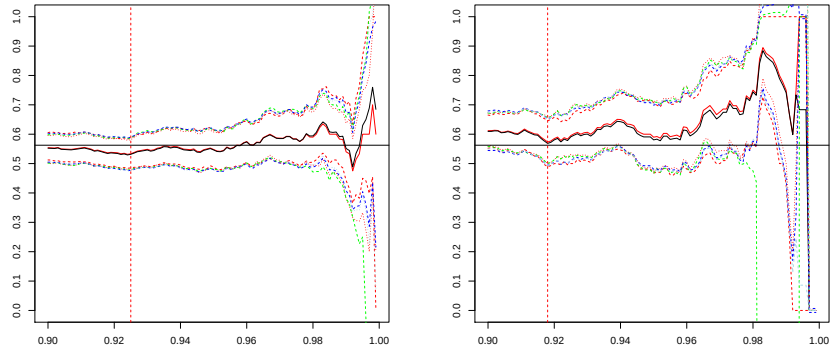
RBB confidence intervals. Figures 1(a) and 1(b) show the asymptotic and bootstrap confidence intervals of the regenerative "blocks" estimator and the regenerative "runs" estimator, respectively. These CI's are quite similar except for the largest values of u. In the sequel, when a bootstrap CI is computed, it will be the basic percentile bootstrap confidence interval. The coverage probabilities of the basic bootstrap percentile CI for the M/M/1 waiting process is estimated over $M = 300$ trajectories, as shown in Figure 2.

Choosing the threshold. As mentioned after Theorem 6, the threshold u can be chosen by minimizing an estimation of the mean-square error of $\sqrt{n}(\theta_n(u)-\theta(u))$, so that the optimal threshold value u^* can be determined as

$$u^* = \arg\min_{u>0} \widehat{MSE}(u),$$

with $\widehat{MSE}(u) = \sigma_f^2(u_n + (\theta_n(u) - \bar{\theta^*}_n(u))^2$, where $\bar{\theta^*}_n(u)$ is the mean of the bootstrap statistics. The same process can be applied to the regenerative "runs" estimator. Applying this to the M/M/1 queue yields $\theta^*\theta_n(u^*) = 0.5263$ with CI $(0.4431\ .6610)$ (which includes the targeted extremal index 0.5625).

Another possibility (which does not require any bootstrap) arises from the fact that the ratio of the asymptotic variances of our 2 regenerative



(a) Regenerative "Blocks" estimator (b) Regenerative "Runs" estimator

Figure 1: Extremal index estimation for waiting times of the M/M/1 queue with $\lambda = 0.2, \mu = 0.8, \theta = 0.56$ (the x-axis gives the percentiles of the simulated (W_n), $n = 1000$, $B = 199$ bootstrap samples, solid red for the regenerative estimator, solid black for the mean bootstrap estimator, dashed red for the basic percentile bootstrap CI, dotted red for the percentile bootstrap CI, dashed green for the t-percentile bootstrap CI, dashed blue for the asymptotic CI based on the regenerative standardization, dashed light blue for the asymptotic CI based on the bootstrap standardization, horizontal black line is θ, vertical dashed red line is the optimal u value as determined by minimizing (28).

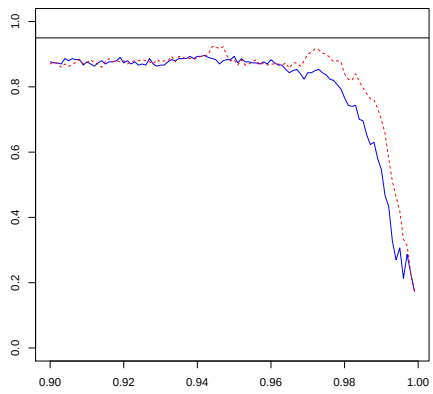


Figure 2: Coverage probabilities of the basic percentile bootstrap CI for the regenerative "blocks" estimator and the regenerative "runs" estimators. M/M/1 queue with $\lambda = 0.2, \mu = 0.8, \theta = 0.56$ (the x-axis gives the percentiles of the simulated (X_n), $n = 1000$, $1-\alpha = 95\%$-CI, $B = 199$ bootstrap samples, $M = 300$, the solid blue curve is that of the "blocks" estimator, the dashed red curve is that of the "runs" estimator).

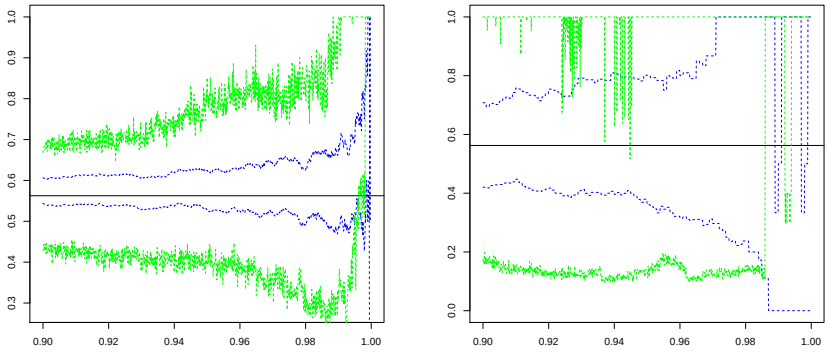
estimators is asymptotically constant $(\sigma'_f(u_n)^2/\sigma_f^2(u_n) \to 1-\theta$ for a properly chosen sequence of thresholds u_n, see Theorem 3, assertion (i) and theorem 4, assertion (ii)[1].) Hence, one may define an optimal threshold value u^*, and hence a unique estimator of the extremal index, by minimizing in u the function

$$\left(\sigma'^2_f(u)/\sigma_f^2(u) - (1-\theta_n(u))\right)^2, \tag{28}$$

and defining $\theta^* = \theta_n(u^*)$. Applying this process to the MM1 queue yields $\theta^* = 0.5179$ with CI $(0.4947\ .6370)$ (which covers the targeted extremal index 0.5625).

Alternative estimators. In [10], the regenerative blocks estimator was compared to the intervals estimator proposed by [16] and to various fixed lengths block estimators and runs estimators (see Fig. 2 therein). Its mean squared error was generally lower than those of the alternative estimators. As far as CI's are concerned, the authors of [16] also proposed a bootstrap procedure based on an automatic declustering of the process relying on the estimation of the extremal index (see section 4 therein). Figure 3 illustrates

[1] The asymptotic variance of the regenerative "runs" estimator for a fixed threshold u is $\sigma'_f(u)^2 = s_f^2(u)/\alpha$, with $s_f^2(u)$ given in eq. (30) in the appendix section.



(a) $n = 10000$ (b) $n = 1000$

Figure 3: Comparison of our bootstrap basic Percentile CI to that proposed in [16]. ($B = 199$, dashed blue for ours and dashed green for theirs, solid black is the true θ).

that our bootstrap CI is much sharper than theirs on this example.

Regeneration based Hill Estimator. We consider now the waiting times of a M/G/1 process with Pareto service times, with parameters $\lambda = 0.2$ and $a = 3$. The subsampling size was fixed to $m_n = \lfloor n/\log(n) \rfloor$. For each of the M trajectories, for each of the B bootstrap samples, the regenerative Hill estimator is first computed for various values of k, from $k = 10$ to the number of blocks $k = l_{m_n}$. The optimal k is then determined by computing a bias corrected Hill estimator (as in [6, 17]) and choosing the value k^* that minimizes the estimated MSE

$$\widehat{MSE}(k) = \widehat{H}^2_{k,\,n}/k + (H_{k,\,n} - \widehat{H}_{k,\,n})^2,$$

where $\widehat{H}_{k,\,n}$ is a bias corrected version of the Hill estimator. The regenerative standardization is then computed as $H_{k^*,n}/\sqrt{k^*}$. Results of this simulation are presented in Table 1. Note that the basic percentile CI and asymptotic CI with bootstrap variance have the best coverage probabilities and are also very easy to compute (it does better than the asymptotic CI which has however the advantage of not requiring the bootstrap resampling). Regarding the choice of the subsampling size m_n, various values were tested and larger values do keep a nice coverage probability with reduced mean length. The application of Algorithm 1 yields particularly nice results questioning the validity of such procedure for the regenrative Hill estimator and hence

CI name	Lower b.	Upper b.	Coverage	Mean length	MSE
Basic Percentile CI	0.248	0.655	100.0%	0.449	0.0093
Percentile CI	-0.077	0.330	57.3%	0.449	0.0093
Asymptotic CI	0.196	0.382	64.7%	0.161	0.0093
Asymptotic CI [a]	0.075	0.503	99.3%	0.471	0.0222
t-Percentile CI	0.095	0.314	49.7%	0.207	0.0093
Standard bootstrap CI [b]	0.156	0.201	0.0%	0.048	0.0547

Table 1: Confidence intervals around tail index estimators: M/G/1 queue with Pareto service times $\lambda = 0.5$, $1/a = 1/3$, sample path of length $n = 10,000$, $m_n = [n/\log(n)] = 1,085$, $B = 199$ bootstrap samples, $M = 300$ Monte-Carlo replications to compute the coverage probabilities, mean lengths of the CI's and mean squared error of the estimator (MSE) - [a] Based on the bootstrap variance - [b] the last line refers to the Standard Hill estimator while the rest of the table refers to the Regenerative Hill estimator.

the validity of the bootstrap of the Hill estimator in the i.i.d. case as well (theoretical work in progress).

Alternative estimator. In [10], the regenerative Hill estimator is compared to the standard Hill estimator computed directly from the largest waiting times, as proposed by [27]. The same bias correction method was applied to the standard Hill estimator in order to determine the optimal k value. In their paper [27], the authors do not propose any confidence interval for their estimator but one could compute a bootstrap CI as proposed in [5] in the iid case (the principle is to resample directly the log differences that are iid exponential rather than the upper statistics). This approach results in very small CI's that fail to compensate for the fact that the Standard Hill estimator is quite bad on this example and hence have a null coverage probability, see the last line of Table 1.

5.2 Pseudo-regenerative examples

We now turn to examples for which a regenerative extension must be approximated and show that this additional step does not damage the accuracy of the method.

Approximate regeneration based extremal index estimator. For the pseudo regenerative case, we consider a first order autoregressive model with Cauchy noise, with parameters $\rho = 0.8$ and $\sigma = 1$, yielding an extremal

index θ equal to $1-\rho$, see [10] for details, namely section 5.2 therein for a precise description of the construction of the pseudo-blocks. The bootstrap CI's and their coverage probabilities are shown in Figure 4. Note that the percentiles of X used for the "runs" estimator are a lot lower than those used for the "blocks" estimator. The CI's for the "blocks" estimator is better than that of the "runs" estimator in terms of coverage probability.

Approximate regeneration based Hill estimator. With the AR(1)-Cauchy example again, we investigated the estimation of the tail index equal to 1 here, see [10] for details. The regenerative Hill estimator was computed for $M = 100$ trajectories of length $n = 10,000$, using a subsampling size $m_n = n/\log(n) = 1,085$ and $B = 199$ bootstrap replications in each case: we obtained $\widehat{H}_{n,\,k^*} = 1.14$ for $k^* = 104$ ($sd = 0.111$) with a basic percentile bootstrap CI of $(0.592 - 1.945)$; a coverage probability of 94% and a mean length of 1.457. Again, when the subsampling size is increased, the coverage probability remains aroung the desired 95% while the mean length of the CI is drastically reduced, which puts questions to the validity of the full regenerative bootstrap for the regenerative Hill estimator as proposed in Algorithm 1 and used for the regenerative extremal index estimators.

A Technical Proofs

A.1 Proof of Theorem 2

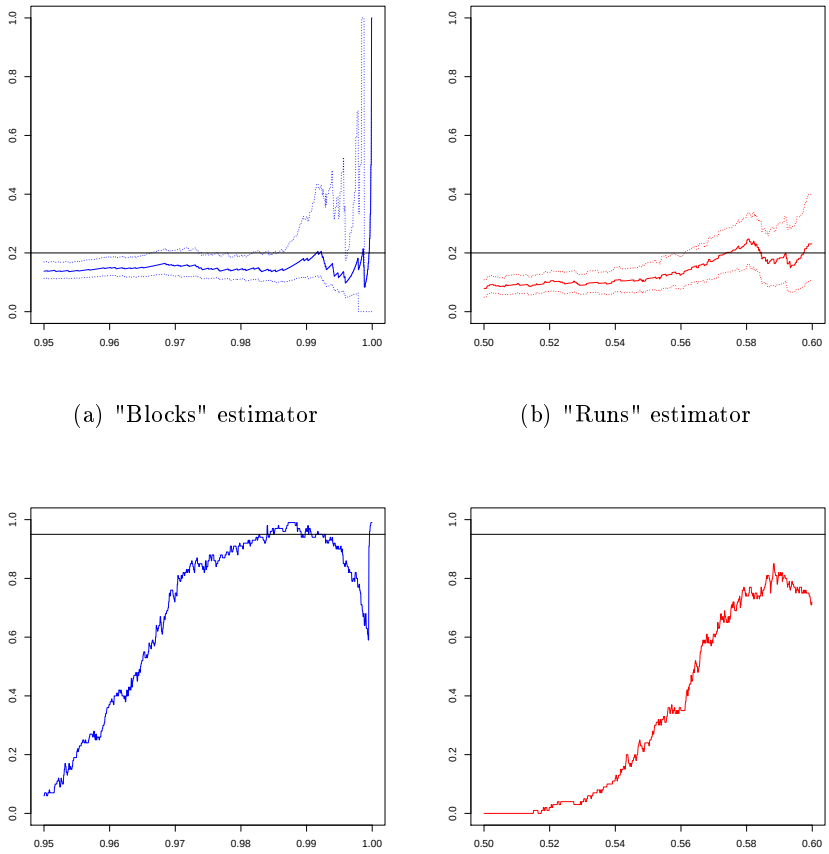
For assertion **(i)**, observe that $\theta_n(u)$ is simply the ratio of the components of the bivariate vector $(\bar{G}_{f,n}(u), \Sigma_{f,n}(u))'$, which is asymptotically normal under the specified moment conditions (see the proof of the CLT stated in Theorem 17.2.2 of [25]) for the atomic case:

$$\sqrt{n}\left[\begin{array}{c}\bar{G}_{f,n}(u)\\ \Sigma_{f,n}(u)\end{array} - \begin{array}{c}\bar{G}_f(u)\\ \Sigma_f(u)\end{array}\right] \Rightarrow \mathcal{N}\left(0, \alpha\cdot\begin{pmatrix}\sigma_1^2(u) & \sigma_{12}(u)\\ \sigma_{12}(u) & \sigma_2^2(u)\end{pmatrix}\right),$$

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

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

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



(a) "Blocks" estimator

(b) "Runs" estimator

(c) Coverage "blocks" estimator

(d) Coverage "runs" estimator

Figure 4: Extremal index estimation for waiting times of the AR1 Cauchy process with $\rho = 0.8$ and $\sigma = 1, \theta = 0.2$ (the x-axis gives the percentiles of the simulated (X_t), $n = 10000$, $B = 199$ bootstrap samples, $M = 100$ Monte Carlo replications

Application of the Delta method finally yields (12), with

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

The demonstration of assertion **(ii)** relies on similar arguments regarding the asymptotic normal behavior of the bivariate vector $(1-\widehat{G}_{f,n}(u), \widehat{\Sigma}_{f,n}(u))'$ obtained from the CLT stated in Theorem 17.3.6 of [25].

A.2 Proof of Proposition 3

Consider $(v_n)_{n\in\mathbb{N}}$ such that $r_n\bar{G}_f(v_n)/\alpha \to \eta < \infty$ as $n \to \infty$, with $r_n = o(\sqrt{(}n/\log\log n)$. A preliminary step consists of studying the behavior of the various components of $\sigma_f^2(v_n)$ as $n \to \infty$, as stated in the next lemma.

Lemma 8. *We have*

$$r_n\Sigma_{f,n}(v_n)/\alpha \to \eta/\theta, \quad r_nG_{f,n}(v_n)/\alpha \to \eta$$
$$r_n\sigma_1^2(v_n)/\alpha \to \eta, \quad \sigma_2^2(v_n) = O(r_n^{-2}), \quad \sigma_{12}(v_n) = O(r_n^{-3/2}).$$

Proof. Consider the solution of the Poisson equation:

$$\widehat{g}(x,u) = \mathbb{E}_x\left[\sum_{i=1}^{\tau_A} \mathbb{I}\{f(X_i) > u\} - \Sigma_f(u)\right]$$

Observe that

$$\begin{aligned}\sigma_2^2(v_n) &\leq \mathbb{E}_\mu[\widehat{g}(X_1,v_n)^2] \leq \mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}\mathbb{I}\{f(X_i) > v_n\}\right)^2\right] + \Sigma_f(v_n)^2 \\ &\leq \mathbb{E}_A[\tau_A^2] + \Sigma_f(v_n)^2.\end{aligned}$$

By Cauchy Schwarz, we yield the last order of magnitude: $\sigma_{12}(v_n) \leq \sigma_1(v_n)\sigma_2(v_n)$ □

Now, define $X_n = \sum_{i=1}^{l_n-1} X_{j,n} = \sum_{i=1}^{l_n-1} \frac{\mathbb{I}\{\zeta_j(f)>v_n\}-\bar{G}_f(v_n)}{\sqrt{l_n-1}\sigma_1(v_n)}$, and observe that

$$\begin{aligned}\sqrt{n/r_n}(\theta_n(v_n) - \theta(v_n)) &= \frac{\sqrt{n}}{\sqrt{l_n-1}}\frac{\sqrt{r_n}\sigma_1(v_n)}{r_n\Sigma_{f,n}(v_n)}X_n \\ &- \frac{r_n\sigma_2(v_n)}{r_n\Sigma_{f,n}(v_n)}\frac{\theta(v_n)}{\sqrt{r_n}}\sqrt{n}\frac{\Sigma_{f,n}(v_n)-\Sigma_f(v_n)}{\sigma_2(v_n)}.\end{aligned}$$

Given the rates in the preliminary lemma, the second term on the RHS is of order $O_{\mathbb{P}}(r_n^{-1/2})$ so that the asymptotic behavior of $\sqrt{n/r_n}(\theta_n(\nu_n) - \theta(\nu_n))$ is determined by that of the first term. A direct application of the Lindeberg Feller theorem is not possible here since $l_n - 1$ is not a stopping time for (X_n). However it can be shown through the application of a slight modification of the arguments given page 425 of [25] that if we denote by $\widetilde{X}_n$ the vector X_n in which l_n is replaced with $\lfloor n/\alpha \rfloor$, then $n^{-1/2}|X_n - \widetilde{X}_n|$ converges to zero in probability. Then the asymptotic standard normality of $\widetilde{X}_n$, and hence that of X_n, results from the Lindeberg Feller theorem under the assumed moment conditions. We then conclude from the above lemma that gives the limit as $\to \infty$ of the term in front of X_n, namely $\theta/\sqrt{\eta}$. Let us now check the conditions of the Lindeberg Feller theorem on $\widetilde{X}_n$.

H1 The sequence $\left(\widetilde{X}_{j,n}\right)$ is asymptotically negligible since

$$\sum_{j=1}^{\lfloor n/\alpha \rfloor - 1} \mathbb{P}_A(|X_{j,n}| \geq \varepsilon) \leq \frac{(\lfloor n/\alpha \rfloor - 1)\mathbb{E}_A[X_{j,n}^4]}{\varepsilon^4} \leq \frac{(\lfloor n/\alpha \rfloor - 1)^{-1}}{\varepsilon^4 \sigma_1^4(\nu_n)} \xrightarrow{P} 0$$

since $\sigma_1^4(\nu_n) = O(r_n^{-2}) = O(\log\log n/n)$ and $\mathbb{E}_A[(\mathbb{I}\{\zeta_j(f) > \nu_n\} - G_f(\nu_n))^4] \leq 1$.

H2 The second condition $\sum_{j=1}^{\lfloor n/\alpha \rfloor - 1} \mathbb{E}_A[X_{j,n}] = 0$ also holds since $\mathbb{E}_A[\mathbb{I}\{\zeta_j(f) > \nu_n\}] = \bar{G}_f(\nu_n)$.

H3 Finally, the condition on the variance is also satisfied since:

$$\Gamma_n = \sum_{k=1}^{\lfloor n/\alpha \rfloor - 1} \mathbb{V}_A(X_{j,n}) = (\lfloor n/\alpha \rfloor - 1)\mathbb{V}_A(X_{j,n}) = \frac{(\lfloor n/\alpha \rfloor - 1)\sigma_1^2(\nu_n)}{(\lfloor n/\alpha \rfloor - 1)\sigma_1^2(\nu_n)} \xrightarrow{P} 1,$$

denoting by $\mathbb{V}_A(.)$ the variance under $\mathbb{P}_A(.)$.

Note that the assumptions $\mathcal{H}(2)$ and $\mathcal{H}(\nu, 1)$ are only needed to ensure that $\sigma_2(u)$ is defined for all u, and the length of the first non regenerative block has a finite first order moment so that our regeneration based extremal index estimator does not differ much of that with denominator $(l_n - 1)/n \sum_{i=1}^n \mathbb{I}\{f(X_i) > u\}$.

A.3 Proof of Theorem 4

Writing (17) as the ratio of $G_f^1(u) = \mathbb{P}_A(\{\max_{2 \leq i \leq \tau_A} f(X_i) \leq u\} \cap \{f(X_1) > u\})$ and $\bar{F}^1(u) = \mathbb{P}_A(f(X_1) > u)$, then the regenerative "runs" estimator given

in (18) is simply the ratio of the empirical counterparts of the probabilities, which are denoted $G^1_{f,n}(u)$ and $\bar{F}^1_n(u)$ in the sequel. Hence, the proof of (i) and (ii) exactly follows that of the strong consistency of the *regenerative blocks* estimator, (8), provided in [10], and that of its asymptotic normality given above, provided that the next lemma holds true under the stated assumptions.

Lemma 9. *$\bar{G}_f(u) \sim G^1_f(u)$, $\bar{F}(u) \sim \bar{F}^1(u)$ so that we can state a LIL for $G^1_f(u)$, a LIL for $\bar{F}^1(u)$ and get asymptotic equivalences similar to those stated in the previous lemma. Let $r_n \uparrow \infty$ in a way that $r_n = o(\sqrt{n/\log\log n})$ as $n \to \infty$, considering $(\nu_n)_{n\in\mathbb{N}}$ such that $r_n(1 - F(\nu_n)) \to \eta < \infty$ as $n \to \infty$, we have $r_n\bar{F}^1(\nu_n) \to \eta/\theta$.*

More precisely, we can state the asymptotic normality of the bivariate vector $(G^1_{f,n}(u), \bar{F}^1_n(u))'$, for all fixed u,

$$\sqrt{n}\begin{bmatrix} G^1_{f,n}(u) \\ \bar{F}^1_n(u) \end{bmatrix} - \begin{bmatrix} G^1_f(u) \\ \bar{F}^1(u) \end{bmatrix} \Rightarrow \mathcal{N}\left(0, \alpha \cdot \mathbb{V}\left((G^1_f(u), \bar{F}^1(u))'\right)\right),$$

with $\mathbb{V}\left((G^1_f(u), \bar{F}^1(u))'\right) = \begin{pmatrix} G^1_f(u)(1 - G^1_f(u)) & G^1_f(u)(1 - \bar{F}^1(u)) \\ G^1_f(u)(1 - \bar{F}^1(u) & \bar{F}^1(u)(1 - \bar{F}^1(u)) \end{pmatrix}$.

Note that because of the specific "probability" form of the covariance terms here (they are all lower or equal to one), we don't need extra moment assumptions as we did in the case of the regenerative blocks estimator.

An application of the Delta method finally yields the following asymptotic variance for all fixed u

$$s^2_f(u) = \alpha \times \begin{bmatrix} \frac{G^1_f(u)(1-G^1_f(u))}{\bar{F}^1(u)^2} - 2\frac{G^1_f(u)^2(1-\bar{F}^1(u))}{\bar{F}^1(u)^3} \\ + \frac{G^1_f(u)^2\bar{F}^1(u)(1-\bar{F}^1(u))}{\bar{F}^1(u)^4} \end{bmatrix}. \tag{29}$$

$$= \alpha \times \left[\theta'(u)\frac{(1 - G^1_f(u))}{\bar{F}^1(u)} - \theta'(u)^2\frac{1 - \bar{F}^1(u)}{\bar{F}^1(u)}\right] \tag{30}$$

Since $r_n\bar{F}^1(\nu_n) \to \eta/\theta$, we can identify s^2 in (21) and (22) as $\alpha\theta^2(1-\theta)$. A formal application of the Lindeberg-Feller theorem similarly to the regenerative blocks estimator proof (the X_n's are bivariate) leads to the same result and it is easily seen that, because only indicator functions are involved, no moment assumption on τ_A is needed.

For the pseudo-regenerative version (i') and (ii'), it is sufficient to observe that under the stated assumptions, we can prove similarly to theorem 2 in [10], that $\sup_{x\in\mathbb{R}} |\hat{G}^1_{f,n}(x) - G^1_{f,n}(x)| = O_{\mathbb{P}_\nu}(\mathcal{R}_n(\hat{\pi}_n,\ \pi)^{1/2})$, as $n \to \infty$,

and similarly to Lemma 6.2 in [8], that we have $N(n) \times \sup_{u\in\mathbb{R}} |\hat{F}^1_{N(n)}(u) - F^1_{N(n)}(u)| = O_{\mathbb{P}_\nu}(\mathcal{R}_{N(n)}(\hat{\pi}_{N(n)}, \pi)^{1/2})$ as $N \to \infty$.

A.4 Proof of Proposition 4.2

The bootstrap version of the "Blocks" estimator of the extremal index is given by the ratio of the bootstrap version bivariate vector $(\bar{G}_{f,n}(u), \Sigma_{f,n}(u))'$. Since by [8], the (A)RBB (both in its regenerative and pseudo-regenerative versions) is asymptotically valid, it follows immediately that, for fixed u, $\sqrt{n}(\theta^*_n(u) - \theta_n(u))$ has the same limiting distribution as $\sqrt{n}(\theta_n(u) - \theta(u))$. The same result remains valid even if l_n is fixed in the bootstrap procedure. The same arguments may be used for the "runs" estimator.

A.5 Proof of Proposition 7

When using $\widetilde{l}_{m_n}$, the proposed procedure boils down to a subsampling procedure in an i.i.d framework. Using continuity and standard U-statistics arguments (see [14]), by mimicking the proof of [14] (see p. 44 therein), we obtain that

$$\begin{aligned}
&\mathbb{P}^*\left(\sqrt{k(\widetilde{l}_{m_n})}\left(\widetilde{\xi}^*_{m_n,\,k(l_{m_n})} - \widetilde{\xi}_{n,\,k(l_n)}\right) \le x\right) \\
&= \mathbb{P}^*\left(\sqrt{k(\widetilde{l}_{m_n})}\left(\widetilde{\xi}^*_{m_n,\,k(\widetilde{l}_{m_n})} - a\right) \le x\right) + o(1) \\
&= \mathbb{P}\left(\sqrt{k(\widetilde{l}_{m_n})}\left(\widetilde{\xi}_{m_n,\,k(\widetilde{l}_{m_n})} - a\right) \le x\right) + o(1)
\end{aligned}$$

The first equality is a straightforward consequence of the continuity of the limiting distribution of the Hill estimator and of the assumption stating that $k(l_n \frac{m_n}{n})/k(l_n) \to 0$.

Now using the fact that $m_n \to \infty$ and the fact that the Hill estimator so normalized has a nondegenerate distribution, we get the result.

References

[1] M.A. Ancona-Navarette and J.A. Tawn. A comparison of methods for estimating the extremal index. *Extremes*, 3(1):5–38, 2000.

[2] S. Asmussen. Extreme value theory for queues via cycle maxima. *Extremes*, 1(2):137–168, 1998.

[3] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, 2003.

[4] K. B. Athreya and S. G. Pantula. Mixing properties of Harris chains and autoregressive processes. *J. Appl. Probab.*, 23(4):880–892, 1986.

[5] J.N. Bacro and M. Brito. A tail bootstrap procedure for estimating the tail Pareto-index. *J. Statist. Planning Inference*, 71(1–2):245–260, 1998.

[6] J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys. Tail index estimation and an exponential regression model. *Extremes*, 2(2):177–200, 1999.

[7] P. Bertail and S. Clémençon. Regeneration-based statistics for Harris recurrent Markov chains. In P. Bertail, P. Soulier, and P. Doukhan, editors, *Dependence in Probability and Statistics*, volume 187 of *Lecture Notes in Statistics*, pages 3–54, 2006.

[8] P. Bertail and S. Clémençon. Regenerative-block bootstrap for Markov chains. *Bernoulli*, 12(4):689–712, 2006.

[9] P. Bertail and S. Clémençon. Approximate regenerative block-bootstrap for Markov chains. *Computational Statistics and Data Analysis*, 52(5):2739–2756, 2007.

[10] P. Bertail, S. Clémençon, and J. Tressou. Extreme value statistics for Markov chains via the (pseudo-) regenerative method. *Extremes*, 12:327–360, 2009.

[11] S. Coles. *An introduction to statistical modelling of Extreme Values*. Springer series in Statistics. Springer, 2001.

[12] J. Danielsson, L. de Haan, L. Peng, and C.G. de Vries. Using a Bootstrap method to choose the sample fraction in tail index estimation. *J. Multivariate Analysis*, 76(2):226–248, 2001.

[13] L. de Haan and L. Peng. Comparison of tail index estimators. *Statist. Neerlandica*, 52:60–70, 1998.

[14] D.N.Politis, J.P.Romano, and M.Wolf. On the asymptotic theory of subsampling. *Statistica Sinica*, 114(4):1105–1124, 2001.

[15] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Applications of Mathematics. Springer-Verlag, 1997.

[16] C.A.T. Ferro and J. Segers. Inference for clusters of extreme values. *J. R. Statist. Soc.*, 65(2):545–556, 2003.

[17] A. Feuerverger and P. Hall. Estimating a tail exponent by modelling departure from a Pareto Distribution. *Ann. Statist.*, 27:760–781, 1999.

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

[19] C. M. Goldie and R. L. Smith. Slow variation with remainder: theory and applications. *Quart. J. Math. Oxford*, 38(1):45–71, 1987.

[20] P. Hall. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivariate Analysis*, 32:177–203, 1990.

[21] T. Hsing. Extremal index estimation for a weakly dependent stationary sequence. *Ann. Statist.*, 21(4):2043–2071, 1993.

[22] J. Jain and B. Jamison. Contributions to Doeblin's theory of Markov processes. *Z. Wahrsch. Verw. Geb.*, 8:19–40, 1967.

[23] F. Laurini and J.A. Tawn. New estimators for the extremal index and other cluster characteristics. *Extremes*, 6(3):189–211, 2003.

[24] M.R. Leadbetter. Extremes and local dependence in stationary sequences. *Z. Wahrscheinlichkeitsch*, 65:291–306, 1983.

[25] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1996.

[26] E. Nummelin. A splitting technique for Harris recurrent chains. *Z. Wahrsch. Verw. Gebiete*, 43:309–318, 1978.

[27] S. Resnick and C. Stărică. Tail index estimation for dependent data. *Ann. Appl. Probab.*, 8:1156–1183, 1998.

[28] D. Revuz. *Markov Chains*. 2nd edition, North-Holland, 1984.

[29] C. Y. Robert. Inference for the limiting cluster size distribution of extreme values. *Ann.Statist.*, 37(1):271–310, 2009.

[30] C.Y. Robert, J. Segers, and C.A.T. Ferro. A sliding blocks estimator for the extremal index. *Electronic Journal of Statistics*, 3:993–1020, 2009.

[31] G.O. Roberts, J.S. Rosenthal, J. Segers, and B. Sousa. Extremal indices, geometric ergodicity of Markov chains, and MCMC. *Extremes*, 9:213–229, 2006.

[32] H. Rootzén. Maxima and exceedances of stationary Markov chains. *Adv. Appl. Probab.*, 20:371–390, 1988.

[33] I. Weissman and S. Y. Novak. On blocks and runs estimators of the extremal index. *J. Statist. Planning Inference*, 66:281–288, 1998.