1

Regeneration-based statistics for Harris recurrent Markov chains

Patrice Bertail[1] and Stéphan Clémençon[2]

[1] CREST-LS, 3, ave Pierre Larousse, 94205 Malakoff, France
Patrice.Bertail@ensae.fr
[2] MODAL'X - Université Paris X Nanterre
LPMA - UMR CNRS 7599 - Universités Paris VI et Paris VII
sclemenc@u-paris10.fr

Abstract : Harris Markov chains make their appearance in many areas of statistical modeling, in particular in time series analysis. Recent years have seen a rapid growth of statistical techniques adapted to data exhibiting this particular pattern of dependence. In this paper an attempt is made to present how renewal properties of Harris recurrent Markov chains or of specific extensions of the latter may be practically used for statistical inference in various settings. When the study of probabilistic properties of general Harris Markov chains may be classically carried out by using the regenerative method (cf [82]), via the theoretical construction of regenerative extensions (see [62]), statistical methodologies may also be based on regeneration for general Harris chains. In the regenerative case, such procedures are implemented from data blocks corresponding to consecutive observed regeneration times for the chain. And the main idea for extending the application of these statistical techniques to general Harris chains X consists in generating first a sequence of approximate renewal times for a regenerative extension of X from data $X_1, ..., X_n$ and the parameters of a minorization condition satisfied by its transition probability kernel, and then applying the latter techniques to the data blocks determined by these pseudo-regeneration times as if they were exact regeneration blocks. Numerous applications of this estimation principle may be considered in both the stationary and nonstationary (including the null recurrent case) frameworks. This article deals with some important procedures based on (approximate) regeneration data blocks, from both practical and theoretical viewpoints, for the following topics: mean and variance estimation, confidence intervals, U-statistics, Bootstrap, robust estimation and statistical study of extreme values.

1.1 Introduction

1.1.1 On describing Markov chains via Renewal processes

Renewal theory plays a key role in the analysis of the asymptotic structure of many kinds of stochastic processes, and especially in the development of asymptotic properties of general irreducible Markov chains. The underlying ground consists in the fact that limit theorems proved for sums of independent random vectors may be easily extended to regenerative random processes, that is to say random processes that may be decomposed at random times, called *regeneration times*, into a sequence of mutually independent blocks of observations, namely *regeneration cycles* (see [82]). The method based on this principle is traditionally called the *regenerative method.* Harris chains that possess an atom, i.e. a Harris set on which the transition probability kernel is constant, are special cases of regenerative processes and so directly fall into the range of application of the regenerative method (Markov chains with discrete state space as well as many markovian models widely used in operational research for modeling storage or queuing systems are remarkable examples of atomic chains). The theory developed in [62] (and in parallel the closely related concepts introduced in [6]) showed that general Markov chains could all be considered as regenerative in a broader sense (*i.e.* in the sense of the existence of a theoretical regenerative extension for the chain, see § 2.3), as soon as the Harris recurrence property is satisfied. Hence this theory made the regenerative method applicable to the whole class of Harris Markov chains and allowed to carry over many limit theorems to Harris chains such as LLN, CLT, LIL or Edgeworth expansions.

In many cases, parameters of interest for a Harris Markov chain may be thus expressed in terms of regeneration cycles. While, for atomic Markov chains, statistical inference procedures may be then based on a random number of observed regeneration data blocks, in the general Harris recurrent case the regeneration times are theoretical and their occurrence cannot be determined by examination of the data only. Although the *Nummelin splitting technique* for constructing regeneration times has been introduced as a theoretical tool for proving probabilistic results such as limit theorems or probability and moment inequalities in the markovian framework, this article aims to show that it is nevertheless possible to make a practical use of the latter for extending regeneration-based statistical tools. Our proposal consists in an empirical method for building approximatively a realization drawn from a Nummelin extension of the chain with a regeneration set and then recovering "approximate regeneration data blocks". As will be shown further, though the implementation of the latter method requires some prior knowledge about the behaviour of the chain and crucially relies on the computation of a consistent estimate of its transition kernel, this methodology allows for numerous statistical applications.

We finally point out that, alternatively to regeneration-based statistical methods, inference techniques based on data (moving) blocks of fixed length may also be used in our markovian framework. But as will be shown throughout the article, such blocking techniques, introduced for dealing with general time series (in the weakly dependent setting) are less powerful, when applied to Harris Markov chains, than the methods we promote here, which are specifically tailored for (pseudo) regenerative processes.

1.1.2 Outline

The outline of the paper is as follows. In section 2, notations are set out and key concepts of the Markov chain theory as well as some basic notions about the regenerative method and the Nummelin splitting technique are recalled. Section 3 presents and discusses how to practically construct (approximate) regeneration data blocks, on which statistical procedures we investigate further are based. Sections 4 and 5 mainly survey results established at length in [8], [9], [10], [11]. More precisely, the problem of estima-ting additive functionals of the stationary distribution in the Harris positive recurrent case is considered in section 4. Estimators based on the (pseudo) regenerative blocks, as well as estimates of their asymptotic variance are exhibited, and limit theorems describing the asymptotic behaviour of their bias and their sampling distribution are also displayed. Section 5 is devoted to the study of a specific resampling procedure, which crucially relies on the (approximate) regeneration data blocks. Results proving the asymptotic validity of this particular bootstrap procedure (and its optimality regarding to second order properties in the atomic case) are stated. Section 6 shows how to extend some of the results of sections 4 and 5 to V and U-statistics. A specific notion of robustness for statistics based on the (approximate) regenerative blocks is introduced and investigated in section 7. And asymptotic properties of some regeneration-based statistics related to the extremal behaviour of Markov chains are studied in section 8 in the regenerative case only. Finally, some concluding remarks are collected in section 9 and further lines of research are sketched.

1.2 Theoretical background

1.2.1 Notation and definitions

We now set out the notations and recall a few definitions concerning the communication structure and the stochastic stability of Markov chains (for further detail, refer to [72] or [60]). Let $X = (X_n)_{n\in\mathbb{N}}$ be an aperiodic irreducible Markov chain on a countably generated state space $(E,\ \mathcal{E})$, with transition probability Π, and initial probability distribution ν. For any $B \in \mathcal{E}$ and any $n \in \mathbb{N}$, we thus have

$$X_0 \sim \nu \text{ and } \mathbb{P}(X_{n+1} \in B \mid X_0, ...,\ X_n) = \Pi(X_n,\ B) \text{ a.s. .}$$

In what follows, $\mathbb{P}_\nu$ (respectively $\mathbb{P}_x$ for x in E) will denote the probability measure on the underlying probability space such that $X_0 \sim \nu$ (resp. $X_0 = x$), $\mathbb{E}_\nu\,(.)$ the $\mathbb{P}_\nu$-expectation (resp. $\mathbb{E}_x\,(.)$ the $\mathbb{P}_x$-expectation), $\mathbb{I}\{\mathcal{A}\}$ will denote the indicator function of the event $\mathcal{A}$ and $\Rightarrow$ the convergence in distribution.

A measurable set B is *Harris recurrent* for the chain if for any $x \in B$, $\mathbb{P}_x(\sum_{n=1}^{\infty} \mathbb{I}\{X_n \in B\} = \infty) = 1$. The chain is said *Harris recurrent* if it is ψ-irreducible and every measurable set B such that $\psi(B) > 0$ is Harris recurrent. When the chain is Harris recurrent, we have the property that $\mathbb{P}_x(\sum_{n=1}^{\infty} \mathbb{I}\{X_n \in B\} = \infty) = 1$ for any $x \in E$ and any $B \in \mathcal{E}$ such that $\psi(B) > 0$.

A probability measure μ on E is said invariant for the chain when $\mu\Pi = \mu$, where $\mu\Pi(dy) = \int_{x\in E} \mu(dx)\Pi\,(x, dy)$. An irreducible chain is said *positive recurrent* when it admits an invariant probability (it is then unique).

Now we recall some basics concerning the regenerative method and its application to the analysis of the behaviour of general Harris chains via the Nummelin splitting technique (refer to [63] for further detail).

1.2.2 Markov chains with an atom

Assume that the chain is ψ-irreducible and possesses an accessible atom, that is to say a measurable set A such that $\psi(A) > 0$ and $\Pi(x, .) = \Pi(y, .)$ for all x, y in A. Denote by $\tau_A = \tau_A(1) = \inf\{n \geq 1,\ X_n \in A\}$ the hitting time on A, by $\tau_A(j) = \inf\{n > \tau_A(j-1),\ X_n \in A\}$ for $j \geq 2$ the successive return times to A and by $\mathbb{E}_A\,(.)$ the expectation conditioned on $X_0 \in A$. Assume further that the chain is Harris recurrent, the probability of returning infinitely often to the atom A is thus equal to one, no matter what the starting point. Then, it follows from the *strong Markov property* that, for any initial distribution ν, the sample paths of the chain may be divided into i.i.d. blocks of random length corresponding to consecutive visits to A:

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

taking their values in the torus $\mathbb{T} = \cup_{n=1}^{\infty} E^n$. The sequence $(\tau_A(j))_{j\geqslant 1}$ defines successive times at which the chain forgets its past, called *regeneration times*. We point out that the class of atomic Markov chains contains not only chains with a countable state space (for the latter, any recurrent state is an accessible atom), but also many specific Markov models arising from the field of operational research (see [2] for regenerative models involved in queuing theory, as well as the examples given in § 4.3). When an accessible atom exists, the *stochastic stability* properties of the chain amount to properties concerning the speed of return time to the atom only. For instance, in this framework, the following result, known as Kac's theorem, holds (*cf* Theorem 10.2.2 in [60]).

Theorem 1. *The chain X is positive recurrent iff $\mathbb{E}_A(\tau_A) < \infty$. The (unique) invariant probability distribution μ is then the Pitman's occupation measure given by*

$$\mu(B) = \mathbb{E}_A(\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i \in B\})/\mathbb{E}_A(\tau_A),\ \textit{for all } B \in \mathcal{E}.$$

For atomic chains, limit theorems can be derived from the application of the corresponding results to the i.i.d. blocks $(\mathcal{B}_n)_{n\geqslant 1}$. One may refer for example to [60] for the LLN, CLT, LIL, [16] for the Berry-Esseen theorem, [56], [57], [58] and [8] for other refinements of the CLT. The same technique can also be applied to establish moment and probability inequalities, which are not asymptotic results (see [26]). As mentioned above, these results are established from hypotheses related to the distribution of the $\mathcal{B}_n$'s. The following assumptions shall be involved throughout the article. Let $\kappa > 0$, $f : E \to \mathbb{R}$ be a measurable function and ν be a probability distribution on $(E, \mathcal{E})$.

Regularity conditions:

$$\mathcal{H}_0(\kappa) :\ \mathbb{E}_A(\tau_A^\kappa) < \infty,$$
$$\mathcal{H}_0(\kappa,\ \nu) :\ \mathbb{E}_\nu(\tau_A^\kappa) < \infty.$$

Block-moment conditions:

$$\mathcal{H}_1(\kappa,\ f) :\ \mathbb{E}_A((\sum_{i=1}^{\tau_A} |f(X_i)|)^\kappa) < \infty,$$
$$\mathcal{H}_1(\kappa,\ \nu,\ f) :\ \mathbb{E}_\nu((\sum_{i=1}^{\tau_A} |f(X_i)|)^\kappa) < \infty.$$

Remark 1. We point out that conditions $\mathcal{H}_0(\kappa)$ and $\mathcal{H}_1(\kappa,\ f)$ do not depend on the accessible atom chosen : if they hold for a given accessible atom A, they are also fulfilled for any other accessible atom (see Chapter 11 in [60]). Besides, the relationship between the "block moment" conditions and the rate of decay of mixing coefficients has been investigated in [17]: for instance, $\mathcal{H}_0(\kappa)$ (as well as $\mathcal{H}_1(\kappa,\ f)$ when f is bounded) is typically fulfilled as soon as the strong mixing coefficients sequence decreases at an arithmetic rate $n^{-\rho}$, for some $\rho > \kappa - 1$.

1.2.3 General Harris recurrent chains

The Nummelin splitting technique

We now recall the *splitting technique* introduced in [62] for extending the probabilistic structure of the chain in order to construct an artificial regeneration set in the general Harris recurrent case. It relies on the crucial notion of *small set*. Recall that, for a Markov chain valued in a state space $(E, \mathcal{E})$ with transition probability Π, a set $S \in \mathcal{E}$ is said to be *small* if there exist $m \in \mathbb{N}^*$, $\delta > 0$ and a probability measure Γ supported by S such that, for all $x \in S$, $B \in \mathcal{E}$,

$$\Pi^m(x,B) \geq \delta\Gamma(B), \tag{1.1}$$

denoting by Π^m the m-th iterate of Π. When this holds, we say that the chain satisfies the *minorization condition* $\mathcal{M}(m,S,\delta,\Gamma)$. We emphasize that accessible small sets always exist for ψ-irreducible chains: any set $B\in\mathcal{E}$ such that $\psi(B)>0$ actually contains such a set (*cf* [48]). Now let us precise how to construct the atomic chain onto which the initial chain X is embedded, from a set on which an iterate Π^m of the transition probability is uniformly bounded below. Suppose that X satisfies $\mathcal{M}=\mathcal{M}(m,S,\delta,\Gamma)$ for $S\in\mathcal{E}$ such that $\psi(S)>0$. Even if it entails replacing the chain $(X_n)_{n\in\mathbb{N}}$ by the chain $\big((X_{nm},...,X_{n(m+1)-1})\big)_{n\in\mathbb{N}}$, we suppose $m=1$. The sample space is expanded so as to define a sequence $(Y_n)_{n\in\mathbb{N}}$ of independent Bernoulli r.v.'s with parameter δ by defining the joint distribution $\mathbb{P}_{\nu,\mathcal{M}}$ whose construction relies on the following randomization of the transition probability Π each time the chain hits S (note that it happens a.s. since the chain is Harris recurrent and $\psi(S)>0$). If $X_n\in S$ and

- if $Y_n=1$ (which happens with probability $\delta\in\,]0,1[$), then X_{n+1} is distributed according to Γ,
- if $Y_n=0$, (which happens with probability $1-\delta$), then X_{n+1} is drawn from $(1-\delta)^{-1}(\Pi(X_{n+1},.)-\delta\Gamma(.))$.

Set $Ber_\delta(\beta)=\delta\beta+(1-\delta)(1-\beta)$ for $\beta\in\{0,1\}$. We now have constructed a bivariate chain $X^{\mathcal{M}}=((X_n,Y_n))_{n\in\mathbb{N}}$, called the *split chain*, taking its values in $E\times\{0,1\}$ with transition kernel $\Pi_{\mathcal{M}}$ defined by

- for any $x\notin S$, $B\in\mathcal{E}$, β and β' in $\{0,1\}$,

$$\Pi_{\mathcal{M}}\left((x,\beta),B\times\{\beta'\}\right)=Ber_\delta(\beta')\times\Pi\left(x,B\right),$$

- for any $x\in S$, $B\in\mathcal{E}$, β' in $\{0,1\}$,

$$\Pi_{\mathcal{M}}\left((x,1),B\times\{\beta'\}\right)=Ber_\delta(\beta')\times\Gamma(B),$$
$$\Pi_{\mathcal{M}}\left((x,0),A\times\{\beta'\}\right)=Ber_\delta(\beta')\times(1-\delta)^{-1}(\Pi\left(x,B\right)-\delta\Gamma(B)).$$

Basic assumptions

The whole point of the construction consists in the fact that $S\times\{1\}$ is an atom for the split chain $X^{\mathcal{M}}$, which inherits all the communication and stochastic stability properties from X (irreducibility, Harris recurrence,...), in particular (for the case $m=1$ here) the blocks constructed for the split chain are independent. Hence the splitting method enables to extend the regenerative method, and so to establish all of the results known for atomic chains, to general Harris chains. It should be noticed that if the chain X satisfies $\mathcal{M}(m,S,\delta,\Gamma)$ for $m>1$, the resulting blocks are not independent anymore but 1-dependent, a form of dependence which may be also easily handled. For simplicity 's sake, we suppose in what follows that condition $\mathcal{M}$ is fulfilled with $m = 1$, we shall also omit the subscript $\mathcal{M}$ and abusively denote by $\mathbb{P}_\nu$ the extensions of the underlying probability we consider. The following assumptions, involving the speed of return to the small set S shall be used throughout the article. Let $\kappa > 0$, $f : E \to \mathbb{R}$ be a measurable function and ν be a probability measure on $(E, \mathcal{E})$.

Regularity conditions:

$$\mathcal{H}_0'(\kappa) : \sup_{x\in S} \mathbb{E}_x(\tau_S^\kappa) < \infty,$$
$$\mathcal{H}_0'(\kappa,\ \nu) : \mathbb{E}_\nu(\tau_S^\kappa) < \infty.$$

Block-moment conditions:

$$\mathcal{H}_1'(\kappa,\ f) : \sup_{x\in S} \mathbb{E}_x((\sum_{i=1}^{\tau_S} |f(X_i)|)^\kappa) < \infty,$$
$$\mathcal{H}_1'(\kappa,\ f,\ \nu) :\ \mathbb{E}_\nu((\sum_{i=1}^{\tau_S} |f(X_i)|)^\kappa) < \infty.$$

Remark 2. It is noteworthy that assumptions $\mathcal{H}_0'(\kappa)$ and $\mathcal{H}_1'(\kappa,\ f)$ do not depend on the choice of the small set S (if they are checked for some accessible small set S, they are fulfilled for all accessible small sets *cf* § 11.1 in [60]). Note also that in the case when $\mathcal{H}_0'(\kappa)$ (resp. $\mathcal{H}_0'(\kappa,\ \nu)$) is satisfied, $\mathcal{H}_1'(\kappa,\ f)$ (resp., $\mathcal{H}_1'(\kappa, f, \nu)$) is fulfilled for any bounded f. Moreover, recall that positive recurrence, conditions $\mathcal{H}_1'(\kappa)$ and $\mathcal{H}_1'(\kappa, f)$ may be practically checked by using test functions methods (*cf* [49], [83]). In particular, it is well known that such block moment assumptions may be replaced by drift criteria of Lyapounov's type (refer to Chapter 11 in [60] for further details on such conditions and many illustrating examples, see also [29]).

We recall finally that such assumptions on the initial chain classically imply the desired conditions for the split chain: as soon as X fulfills $\mathcal{H}_0'(\kappa)$ (resp., $\mathcal{H}_0'(\kappa,\ \nu)$, $\mathcal{H}_1'(\kappa,\ f)$, $\mathcal{H}_1'(\kappa,\ f,\ \nu)$), $X^{\mathcal{M}}$ satisfies $\mathcal{H}_0(\kappa)$ (resp., $\mathcal{H}_0(\kappa,\ \nu)$, $\mathcal{H}_1(\kappa,\ f)$, $\mathcal{H}_1(\kappa,\ f,\ \nu)$).

The distribution of $(Y_1, ..., Y_n)$ conditioned on $(X_1, ..., X_{n+1})$.

As will be shown in the next section, the statistical methodology for Harris chains we propose is based on approximating the conditional distribution of the binary sequence $(Y_1, ..., Y_n)$ given $X^{(n+1)} = (X_1, ..., X_{n+1})$. We thus precise the latter. Let us assume further that the family of the conditional distributions $\{\Pi(x, dy)\}_{x\in E}$ and the initial distribution ν are dominated by a σ-finite measure λ of reference, so that $\nu(dy) = f(y)\lambda(dy)$ and $\Pi(x, dy) = p(x, y)\lambda(dy)$, for all $x \in E$. Notice that the minorization condition entails that Γ is absolutely continuous with respect to λ too, and that

$$p(x,y) \geq \delta\gamma(y),\ \lambda(dy) \text{ a.s.} \tag{1.2}$$

for any $x \in S$, with $\Gamma(dy) = \gamma(y)dy$. The distribution of $Y^{(n)} = (Y_1, ..., Y_n)$ conditionally to $X^{(n+1)} = (x_1, ..., x_{n+1})$ is then the tensor product of Bernoulli distributions given by: for all $\beta^{(n)} = (\beta_1, ..., \beta_n) \in \{0,1\}^n$, $x^{(n+1)} = (x_1, ..., x_{n+1}) \in E^{n+1}$,

$$\mathbb{P}_\nu\left(Y^{(n)} = \beta^{(n)} \mid X^{(n+1)} = x^{(n+1)}\right) = \prod_{i=1}^{n} \mathbb{P}_\nu(Y_i = \beta_i \mid X_i = x_i,\ X_{i+1} = x_{i+1}),$$

with, for $1 \leqslant i \leqslant n$,

$$\mathbb{P}_\nu(Y_i = 1 \mid X_i = x_i,\ X_{i+1} = x_{i+1}) = \delta, \text{ if } x_i \notin S,$$
$$\mathbb{P}_\nu(Y_i = 1 \mid X_i = x_i,\ X_{i+1} = x_{i+1}) = \frac{\delta\gamma(x_{i+1})}{p(x_i, x_{i+1})}, \text{ if } x_i \in S.$$

Roughly speaking, conditioned on $X^{(n+1)}$, from $i = 1$ to n, Y_i is drawn from the Bernoulli distribution with parameter δ, unless X has hit the small set S at time i: in this case Y_i is drawn from the Bernoulli distribution with parameter $\delta\gamma(X_{i+1})/p(X_i, X_{i+1})$. We denote by $\mathcal{L}^{(n)}(p, S, \delta, \gamma, x^{(n+1)})$ this probability distribution.

1.3 Dividing the sample path into (approximate) regeneration cycles

In the preceding section, we recalled the Nummelin approach for the theoretical construction of regeneration times in the Harris framework. Here we now consider the problem of approximating these random times from data sets in practice and propose a basic preprocessing technique, on which estimation methods we shall discuss further are based.

1.3.1 Regenerative case

Let us suppose we observed a trajectory $X_1, ..., X_n$ of length n drawn from the chain X. In the regenerative case, when an atom A for the chain is *a priori* known, regeneration blocks are naturally obtained by simply examining the data, as follows.

Algorithm 1 *(Regeneration blocks construction)*

1. *Count the number of visits* $l_n = \sum_{i=1}^{n} \mathbb{I}\{X_i \in A\}$ *to* A *up to time* n.
2. *Divide the observed trajectory* $X^{(n)} = (X_1, ..., X_n)$ *into* $l_n + 1$ *blocks corresponding to the pieces of the sample path between consecutive visits to the atom* A*,*

$$\mathcal{B}_0 = (X_1, ..., \ X_{\tau_A(1)}), \ \mathcal{B}_1 = (X_{\tau_A(1)+1}, ..., \ X_{\tau_A(2)}), ...,$$
$$\mathcal{B}_{l_n-1} = (X_{\tau_A(l_n-1)+1}, ..., \ X_{\tau_A(l_n)}), \ \mathcal{B}_{l_n}^{(n)} = (X_{\tau_A(l_n)+1}, ..., \ X_n),$$

with the convention $\mathcal{B}_{l_n}^{(n)} = \emptyset$ *when* $\tau_A(l_n) = n$*.*

3. *Drop the first block* $\mathcal{B}_0$*, as well as the last one* $\mathcal{B}_{l_n}^{(n)}$*, when non-regenerative (i.e. when* $\tau_A(l_n) < n$*).*

The regeneration blocks construction is illustrated by Fig. 1 in the case of a random walk on the half line $\mathbb{R}^+$ with $\{0\}$ as an atom.

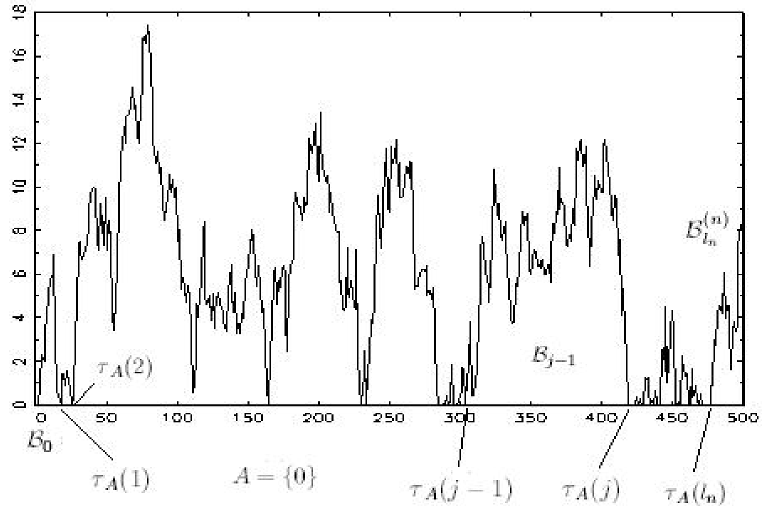

Figure 1 : Dividing the trajectory of a random walk on the half line into regeneration data blocks corresponding to successive visits to $A = 0$

1.3.2 General Harris case

The principle

Suppose now that observations $X_1, ..., \ X_{n+1}$ are drawn from a Harris chain X satisfying the assumptions of § 2.3.3 (refer to the latter paragraph for the notations). If we were able to generate binary data $Y_1, ..., \ Y_n$, so that

$X^{\mathcal{M}\,(n)} = ((X_1, Y_1), ..., (X_n, Y_n))$ be a realization of the split chain $X^{\mathcal{M}}$ described in § 2.3, then we could apply the *regeneration blocks construction* procedure to the sample path $X^{\mathcal{M}\,(n)}$. Unfortunately, knowledge of the transition density $p(x, y)$ for $(x, y) \in S^2$ is required to draw practically the Y_i's this way. We propose a method relying on a preliminary estimation of the "nuisance parameter" $p(x, y)$. More precisely, it consists in approximating the splitting construction by computing an estimator $p_n(x, y)$ of $p(x, y)$ using data $X_1, ..., X_{n+1}$, and to generate a random vector $(\widehat{Y}_1, ..., \widehat{Y}_n)$ conditionally to $X^{(n+1)} = (X_1, ..., X_{n+1})$, from distribution $\mathcal{L}^{(n)}(p_n, S, \delta, \gamma, X^{(n+1)})$, which approximates in some sense the conditional distribution $\mathcal{L}^{(n)}(p, S, \delta, \gamma, X^{(n+1)})$ of $(Y_1, ..., Y_n)$ for given $X^{(n+1)}$. Our method, which we call *approximate regeneration blocks construction (ARB construction* in abbreviated form) amounts then to apply the *regeneration blocks construction* procedure to the data $((X_1, \widehat{Y}_1), ..., (X_n, \widehat{Y}_n))$ as if they were drawn from the atomic chain $X^{\mathcal{M}}$. In spite of the necessary consistent transition density estimation step, we shall show in the sequel that many statistical procedures, that would be consistent in the ideal case when they would be based on the regeneration blocks, remain asymptotically valid when implemented from the approximate data blocks. For given parameters $(\delta,\ S,\ \gamma)$ (see § 3.2.2 for a data driven choice of these parameters), the approximate regeneration blocks are constructed as follows.

Algorithm 2 *(Approximate regeneration blocks construction)*

1. *From the data* $X^{(n+1)} = (X_1, ..., X_{n+1})$*, compute an estimate* $p_n(x, y)$ *of the transition density such that* $p_n(x, y) \geq \delta\gamma(y)$*,* $\lambda(dy)$ *a.s., and* $p_n(X_i, X_{i+1}) > 0$*,* $1 \leqslant i \leqslant n$*.*
2. *Conditioned on* $X^{(n+1)}$*, draw a binary vector* $(\widehat{Y}_1, ..., \widehat{Y}_n)$ *from the distribution estimate* $\mathcal{L}^{(n)}(p_n, S, \delta, \gamma, X^{(n+1)})$*. It is sufficient in practice to draw the* $\widehat{Y}_i$*'s at time points* i *when the chain visits the set* S *(i.e. when* $X_i \in S$*), since at these times and at these times only the split chain may regenerate. At such a time point* i*, draw* $\widehat{Y}_i$ *according to the Bernoulli distribution with parameter* $\delta\gamma(X_{i+1})/p_n(X_i, X_{i+1})$*).*
3. *Count the number of visits* $\widehat{l}_n = \sum_{i=1}^{n} \mathbb{I}\{X_i \in S, \widehat{Y}_i = 1)$ *to the set* $A_{\mathcal{M}} = S \times \{1\}$ *up to time* n *and divide the trajectory* $X^{(n+1)}$ *into* $\widehat{l}_n + 1$ *approximate regeneration blocks corresponding to the successive visits of* $(X,\ \widehat{Y})$ *to* $A_{\mathcal{M}}$*,*

$$\widehat{\mathcal{B}}_0 = (X_1, ...,\ X_{\widehat{\tau}_{A_{\mathcal{M}}}(1)}),\ \widehat{\mathcal{B}}_1 = (X_{\widehat{\tau}_{A_{\mathcal{M}}}(1)+1}, ...,\ X_{\widehat{\tau}_{A_{\mathcal{M}}}(2)}), ...,$$

$$\widehat{\mathcal{B}}_{\widehat{l}_n - 1} = (X_{\widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n - 1)+1}, ...,\ X_{\widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n)}),\ \widehat{\mathcal{B}}_{\widehat{l}_n}^{(n)} = (X_{\widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n)+1}, ...,\ X_{n+1}),$$

   where $\widehat{\tau}_{A_{\mathcal{M}}}(1) = \inf\{n \geqslant 1,\ X_n \in S, \widehat{Y}_n = 1\}$ *and* $\widehat{\tau}_{A_{\mathcal{M}}}(j+1) = \inf\{n > \widehat{\tau}_{A_{\mathcal{M}}}(j),\ X_n \in S, \widehat{Y}_n = 1\}$ *for* $j \geqslant 1$*.*
4. *Drop the first block* $\widehat{\mathcal{B}}_0$ *and the last one* $\widehat{\mathcal{B}}_{\widehat{l}_n}^{(n)}$ *when* $\widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n) < n$*.*

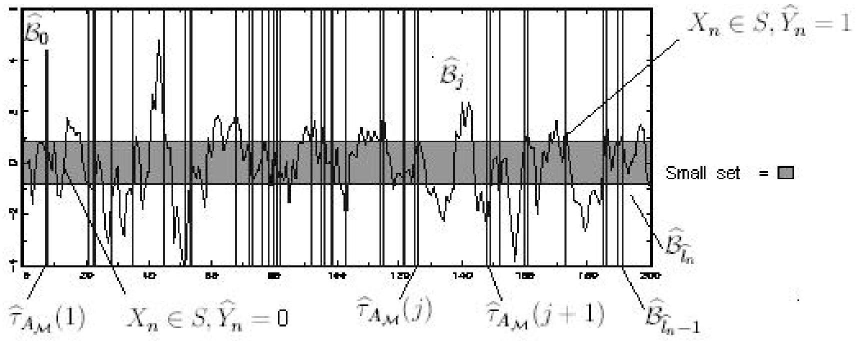

Figure 2: ARB construction for an AR(1) simulated time-series

Such a division of the sample path is illustrated by Fig. 2 : from a practical viewpoint the trajectory may only be cut when hitting the small set. At such a point, drawing a Bernoulli r.v. with the estimated parameter indicates whether one should cut here the time series trajectory or not.

Practical choice of the minorization condition parameters

Because the construction above is highly dependent on the minorization condition parameters chosen, we now discuss how to select the latter with a data-driven technique so as to construct enough blocks for computing meaningful statistics. As a matter of fact, the rates of convergence of the statistics we shall study in the sequel increase as the mean number of regenerative (or pseudo-regenerative) blocks, which depends on the size of the small set chosen (or more exactly, on how often the chain visits the latter in a trajectory of finite length) and how sharp is the lower bound in the minorization condition: the larger the size of the small set is, the smaller the uniform lower bound for the transition density. This leads us to the following trade-off. Roughly speaking, for a given realization of the trajectory, as one increases the size of the small set S used for the data blocks construction, one naturally increases the number of points of the trajectory that are candidates for determining a block (*i.e.* a cut in the trajectory), but one also decreases the probability of cutting the trajectory (since the uniform lower bound for $\{p(x,y)\}_{(x,y)\in S^2}$ then decreases). This gives an insight into the fact that better numerical results for statistical procedures based on the ARB construction may be obtained in practice for some specific choices of the small set, likely for choices corresponding to a maximum expected number of data blocks given the trajectory, that is

$$N_n(S) = \mathbb{E}_\nu(\sum_{i=1}^{n} \mathbb{I}\{X_i \in S, Y_i = 1\} \,|X^{(n+1)}).$$

Hence, when no prior information about the structure of the chain is available, here is a practical data-driven method for selecting the minorization condition parameters in the case when the chain takes real values. Consider a collection $\mathcal{S}$ of borelian sets S (typically compact intervals) and denote by $\mathcal{U}_S(dy) = \gamma_S(y).\lambda(dy)$ the uniform distribution on S, where $\gamma_S(y) = \mathbb{I}\{y \in S\}/\lambda(S)$ and λ is the Lebesgue measure on $\mathbb{R}$. Now, for any $S \in \mathcal{S}$, set $\delta(S) = \lambda(S). \inf_{(x,y)\in S^2} p(x,y)$. We have for any x, y in S, $p(x,y) \geq \delta(S)\gamma_S(y)$. In the case when $\delta(S) > 0$, the ideal criterion to optimize may be then expressed as

$$N_n(S) = \frac{\delta(S)}{\lambda(S)} \sum_{i=1}^{n} \frac{\mathbb{I}\{(X_i, X_{i+1}) \in S^2\}}{p(X_i, X_{i+1})}. \qquad (1.3)$$

However, as the transition kernel $p(x,y)$ and its minimum over S^2 are unknown, a practical empirical criterion is obtained by replacing $p(x,y)$ by an estimate $p_n(x,y)$ and $\delta(S)$ by a lower bound $\delta_n(S)$ for $\lambda(S).p_n(x,y)$ over S^2 in expression (1.3). Once $p_n(x,y)$ is computed, calculate $\delta_n(S) = \lambda(S). \inf_{(x,y)\in S^2} p_n(x,y)$ and maximize thus the empirical criterion over $S \in \mathcal{S}$

$$\widehat{N}_n(S) = \frac{\delta_n(S)}{\lambda(S)} \sum_{i=1}^{n} \frac{\mathbb{I}\{(X_i, X_{i+1}) \in S^2\}}{p_n(X_i, X_{i+1})}. \qquad (1.4)$$

More specifically, one may easily check at hand on many examples of real valued chains (see § 4.3 for instance), that any compact interval $V_{x_0}(\varepsilon) = [x_0 - \varepsilon, x_0 + \varepsilon]$ for some well chosen $x_0 \in \mathbb{R}$ and $\varepsilon > 0$ small enough, is a small set, choosing γ as the density of the uniform distribution on $V_{x_0}(\varepsilon)$. For practical purpose, one may fix x_0 and perform the optimization over $\varepsilon > 0$ only (see [10]) but both x_0 and ε may be considered as tuning parameters. A possible numerically feasible selection rule could rely then on searching for (x_0, ε) on a given pre-selected grid $\mathcal{G} = \{(x_0(k), \varepsilon(l)),\ 1 \leqslant k \leqslant K,\ 1 \leqslant l \leqslant L\}$ such that $\inf_{(x,y)\in V_{x_0}(\varepsilon)^2} p_n(x,y) > 0$ for any $(x_0, \varepsilon) \in \mathcal{G}$.

Algorithm 3 *(ARB construction with empirical choice of the small set)*

1. *Compute an estimator $p_n(x,y)$ of $p(x,y)$.*
2. *For any $(x_0, \varepsilon) \in \mathcal{G}$, compute the estimated expected number of pseudo-regenerations:*

$$\widehat{N}_n(x_0, \varepsilon) = \frac{\delta_n(x_0, \varepsilon)}{2\varepsilon} \sum_{i=1}^{n} \frac{\mathbb{I}\{(X_i, X_{i+1}) \in V_{x_0}(\varepsilon)^2\}}{p_n(X_i, X_{i+1})},$$

with $\delta_n(x_0, \varepsilon) = 2\varepsilon. \inf_{(x,y)\in V_{x_0}(\varepsilon)^2} p_n(x,y)$.

3. *Pick* (x_0^*, ε^*) *in* $\mathcal{G}$ *maximizing* $\widehat{N}_n(x_0, \varepsilon)$ *over* $\mathcal{G}$, *corresponding to the set* $S^* = [x_0^* - \varepsilon^*,\ x_0^* + \varepsilon^*]$ *and the minorization constant* $\delta_n^* = \delta_n(x_0^*, \varepsilon^*)$.
4. *Apply Algorithm 2 for ARB construction using* S^*, δ_n^* *and* p_n.

Remark 3. Numerous consistent estimators of the transition density of Harris chains have been proposed in the literature. Refer to [76], [77], [78], [74], [15], [31], [67], [4] or [25] for instance in positive recurrent cases, [50] in specific null recurrent cases.

This method is illustrated by Fig. 3 in the case of an $AR(1)$ model: $X_{i+1} = \alpha X_i + \varepsilon_{i+1}$, $i \in \mathbb{N}$, with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $\alpha = 0.95$ and $X_0 = 0$, for a trajectory of length $n = 200$. Taking $x_0 = 0$ and letting ε grow, the expected number regeneration blocks is maximum for ε^* close to 0.9. The true minimum value of $p(x, y)$ over the corresponding square is actually $\delta = 0.118$. The first graphic in this panel shows the *Nadaraya-Watson estimator*

$$p_n(x, y) = \frac{\sum_{i=1}^n K(h^{-1}(x - X_i))K(h^{-1}(y - X_{i+1}))}{\sum_{i=1}^n K(h^{-1}(x - X_i))},$$

computed from the gaussian kernel $K(x) = (2\pi)^{-1}\exp(-x^2/2)$ with an optimal bandwidth h of order $n^{-1/5}$. The second one plots $\widehat{N}_n(\varepsilon)$ as a function of ε. The next one indicates the set S^* corresponding to our empirical selection rule, while the last one displays the "optimal" ARB construction.

Note finally that other approaches may be considered for determining practically small sets and establishing accurate minorization conditions, which conditions do not necessarily involve uniform distributions besides. Refer for instance to [73] for Markov diffusion processes.

A two-split version of the ARB construction

When carrying out the theoretical study of statistical methods based on the ARB construction, one must deal with difficult problems arising from the dependence structure in the set of the resulting data blocks, due to the preliminary estimation step. Such difficulties are somehow similar as the ones that one traditionally faces in a semiparametric framework, even in the i.i.d. setting. The first step of semiparametric methodologies usually consists in a preliminary estimation of some infinite dimensional nuisance parameter (typically a density function or a nonparametric curve), on which the remaining (parametric) steps of the procedure are based. For handling theoretical difficulties related to this dependence problem, a well known method, called the *splitting trick*, amounts to split the data set into two parts, the first subset being used for estimating the nuisance parameter, while the parameter of interest is then estimated from the other subset (using the preliminary estimate). An analogous principle may be implemented in our framework using an additional split of the data in the "middle of the trajectory", for ensuring that

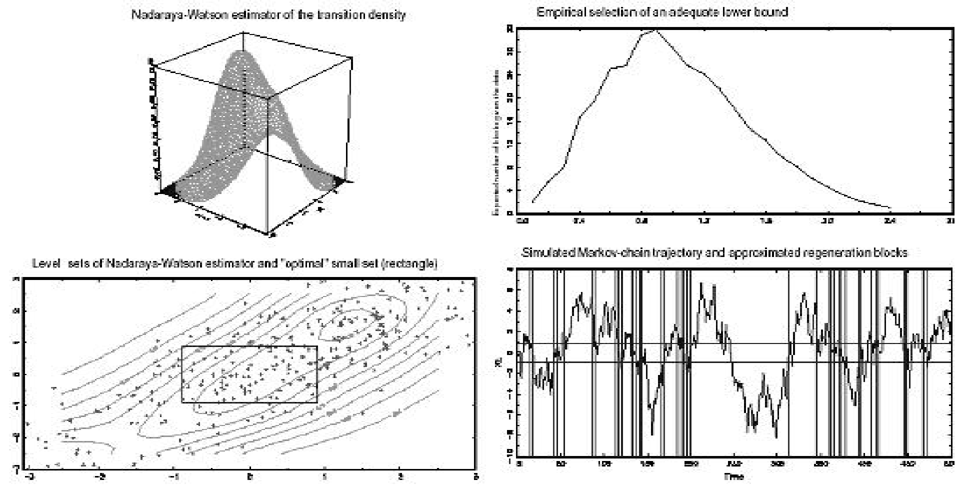

Figure 3 : Illustration of Algorithm 3 : ARB construction with empirical choice of the small set.

a regeneration at least occurs in between with an overwhelming probability (so as to get two independent data subsets, see step 2 in the algorithm below). For this reason, we consider the following variant of the ARB construction. Let $1 < m < n$, $1 \leqslant p < n - m$.

Algorithm 4 *(two-split ARB construction)*

1. *From the data* $X^{(n+1)} = (X_1, ..., X_{n+1})$, *keep only the first* m *observations* $X^{(m)}$ *for computing an estimate* $p_m(x, y)$ *of* $p(x, y)$ *such that* $p_m(x, y) \geq \delta\gamma(y)$, $\lambda(dy)$ *a.s. and* $p_m(X_i, X_{i+1}) > 0$, $1 \leqslant i \leqslant n - 1$.
2. *Drop the observations between time* $m + 1$ *and time* $m^* = m + p$ *(under standard assumptions, the split chain regenerates once at least between these times with large probability).*
3. *From remaining observations* $X^{(m^*, n)} = (X_{m^*+1}, ..., X_n)$ *and estimate* p_m, *apply steps 2-4 of* ***Algorithm 2*** *(respectively of* ***Algorithm 3****).*

This procedure is similar to the *2-split method* proposed in [79], except that here the number of deleted observations is arbitrary and easier to interpret in term of regeneration. Of course, the more often the split chain regenerates, the smaller p may be chosen. And the main problem consists in picking $m = m_n$ so that $m_n \to \infty$ as $n \to \infty$ for the estimate of the transition kernel to be accurate enough, while keeping enough observation $n - m^*$ for the block construction step: one typically chooses $m = o(n)$ as $n \to \infty$. Further assumptions are required for investigating precisely how to select m. In [11], a choice based on the rate of convergence α_m of the estimator $p_m(x,y)$ (for the MSE when error is measured by the sup-norm over $S \times S$, see assumption $\mathcal{H}_2$ in § 4.2) is proposed: when considering smooth markovian models for instance, estimators with rate $\alpha_m = m^{-1}\log(m)$ may be exhibited and one shows that $m = n^{2/3}$ is then an optimal choice (up to a log(n)). However, one may argue, as in the semiparametric case, that this methodology is motivated by our limitations in the analysis of asymptotic properties of the estimators only, whereas from a practical viewpoint it may deteriorate the finite sample performance of the initial algorithm. To our own experience, it is actually better to construct the estimate $p(x,y)$ from the whole trajectory and the interest of **Algorithm 4** is mainly theoretical.

1.4 Mean and variance estimation

In this section, we suppose that the chain X is positive recurrent with unknown stationary probability μ and consider the problem of estimating an additive functional of type $\mu(f) = \int f(x)\mu(dx) = \mathbb{E}_\mu(f(X_1))$, where f is a μ-integrable real valued function defined on the state space $(E, \mathcal{E})$. Estimation of additive functionals of type $\mathbb{E}_\mu(F(X_1, ..., X_k))$, for fixed $k \geqslant 1$, may be investigated in a similar fashion. We set $\overline{f}(x) = f(x) - \mu(f)$.

1.4.1 Regenerative case

Here we assume further that X admits an *a priori* known accessible atom A. As in the i.i.d. setting, a natural estimator of $\mu(f)$ is the sample mean statistic,

$$\mu_n'(f) = n^{-1}\sum_{i=1}^{n} f(X_i). \tag{1.5}$$

When the chain is stationary (*i.e.* when $\nu = \mu$), the estimator $\mu_n'(f)$ has zero-bias. However, its bias is significant in all other cases, mainly because of the presence of the first and last (non-regenerative) data blocks $\mathcal{B}_0$ and $\mathcal{B}_{l_n}^{(n)}$ (see Proposition 4.1 below). Besides, by virtue of Theorem 2.1, $\mu(f)$ may be expressed as the mean of the $f(X_i)$'s over a regeneration cycle (renormalized by the mean length of a regeneration cycle)

$$\mu(f) = \mathbb{E}_A(\tau_A)^{-1}\mathbb{E}_A(\sum_{i=1}^{\tau_A} f(X_i)).$$

This suggests to introduce the following estimators of the mean $\mu(f)$. Define the sample mean based on the observations (eventually) collected after the first regeneration time only by $\widetilde{\mu}_n(f) = (n-\tau_A)^{-1}\sum_{i=1+\tau_A}^{n} f(X_i)$ with the convention $\widetilde{\mu}_n(f) = 0$, when $\tau_A > n$, as well as the sample mean based on the observations collected between the first and last regeneration times before n by $\overline{\mu}_n(f) = (\tau_A(l_n) - \tau_A)^{-1} \sum_{i=1+\tau_A}^{\tau_A(l_n)} f(X_i)$ with $l_n = \sum_{i=1}^{n} \mathbb{I}\{X_i \in A\}$ and the convention $\overline{\mu}_n(f) = 0$, when $l_n \leqslant 1$ (observe that, by Markov's inequality, $\mathbb{P}_\nu(l_n \leqslant 1) = O(n^{-1})$ as $n \to \infty$, as soon as $\mathcal{H}_0(1, \nu)$ and $\mathcal{H}_0(2)$ are fulfilled).

Let us introduce some additional notation for the block sums (resp. the block lengths), that shall be used here and throughout. For $j \geqslant 1$, $n \geqslant 1$, set

$$l(\mathcal{B}_0) = \tau_A,\ l(\mathcal{B}_j) = \tau_A(j+1) - \tau_A(j),\ l(\mathcal{B}_{l_n}^{(n)}) = n - \tau_A(l_n)$$

for the length of the blocks and

$$f(\mathcal{B}_0) = \sum_{i=1}^{\tau_A} f(X_i),\ f(\mathcal{B}_j) = \sum_{i=1+\tau_A(j)}^{\tau_A(j+1)} f(X_i),\ f(\mathcal{B}_{l_n}^{(n)}) = \sum_{i=1+\tau_A(l_n)}^{n} f(X_i)$$

for the value of the functional on the blocks. With these notations, the estimators above may be rewritten as

$$\mu_n'(f) = \frac{f(\mathcal{B}_0) + \sum_{j=1}^{l_n-1} f(\mathcal{B}_j) + f(\mathcal{B}_{l_n}^{(n)})}{l(\mathcal{B}_0) + \sum_{j=1}^{l_n-1} l(\mathcal{B}_j) + l(\mathcal{B}_{l_n}^{(n)})},$$

$$\widetilde{\mu}_n(f) = \frac{\sum_{j=1}^{l_n-1} f(\mathcal{B}_j) + f(\mathcal{B}_{l_n}^{(n)})}{\sum_{j=1}^{l_n-1} l(\mathcal{B}_j) + l(\mathcal{B}_{l_n}^{(n)})},\ \overline{\mu}_n(f) = \frac{\sum_{j=1}^{l_n-1} f(\mathcal{B}_j)}{\sum_{j=1}^{l_n-1} l(\mathcal{B}_j)}.$$

Let $\mu_n(f)$ designs any of the three estimators $\mu_n'(f)$, $\widetilde{\mu}_n(f)$ or $\overline{\mu}_n(f)$. If X fulfills conditions $\mathcal{H}_0(2)$, $\mathcal{H}_0(2, \nu)$, $\mathcal{H}_1(f, 2, A)$, $\mathcal{H}_1(f, 2, \nu)$ then the following CLT holds under $\mathbb{P}_\nu$ (*cf* Theorem 17.2.2 in [60])

$$n^{1/2}\sigma^{-1}(f)(\mu_n(f) - \mu(f)) \Rightarrow \mathcal{N}(0, 1)\,, \text{ as } n \to \infty,$$

with a normalizing constant

$$\sigma^2(f) = \mu\,(A)\,\mathbb{E}_A((\sum_{i=1}^{\tau_A} f(X_i) - \mu(f)\tau_A)^2), \tag{1.6}$$

From this expression we propose the following estimator of the asymptotic variance, adopting the usual convention regarding to empty summation,

$$\sigma_n^2(f) = n^{-1} \sum_{j=1}^{l_n-1} (f(\mathcal{B}_j) - \overline{\mu}_n(f) l(\mathcal{B}_j))^2. \tag{1.7}$$

Notice that the first and last data blocks are not involved in its construction. We could have proposed estimators involving different estimates of $\mu(f)$, but as will be seen later, it is preferable to consider an estimator based on regeneration blocks only. The following quantities shall be involved in the statistical analysis below. Define

$$\alpha=\mathbb{E}_A(\tau_A),\ \beta=\mathbb{E}_A(\tau_A\sum_{i=1}^{\tau_A}\overline{f}(X_i))=Cov_A(\tau_A,\sum_{i=1}^{\tau_A}\overline{f}(X_i)),$$

$$\varphi_\nu=\mathbb{E}_\nu(\sum_{i=1}^{\tau_A}\overline{f}(X_i)),\ \gamma=\alpha^{-1}\mathbb{E}_A(\sum_{i=1}^{\tau_A}(\tau_A-i)\overline{f}(X_i)).$$

We also introduce the following technical conditions.

(C1) (*Cramer condition*)

$$\overline{\lim_{t\to\infty}}\mid\mathbb{E}_A(\exp(it\sum_{i=1}^{\tau_A}\overline{f}(X_i)))\mid<1.$$

(C2) (*Cramer condition*)

$$\overline{\lim_{t\to\infty}}\mid\mathbb{E}_A(\exp(it(\sum_{i=1}^{\tau_A}\overline{f}(X_i))^2))\mid<1.$$

(C3) *There exists $N\geqslant 1$ such that the N-fold convoluted density g^{*N} is bounded, denoting by g the density of the $(\sum_{i=1+\tau_A(j)}^{\tau_A(j+1)}\overline{f}(X_i)\ -\alpha^{-1}\beta)^2$'s.*

(C4) *There exists $N\geqslant 1$ such that the N-fold convoluted density G^{*N} is bounded, denoting by G the density of the $(\sum_{i=1+\tau_A(j)}^{\tau_A(j+1)}\overline{f}(X_i))^2$'s.*

These two conditions are automatically satisfied if $\sum_{i=1+\tau_A(1)}^{\tau_A(2)}\overline{f}(X_i)$ has a bounded density.

The result below is a straightforward extension of Theorem 1 in [56](see also Proposition 3.1 in [8]).

Proposition 1. *Suppose that $\mathcal{H}_0(4)$, $\mathcal{H}_0(2,\ \nu)$, $\mathcal{H}_1(4,\ f)$, $\mathcal{H}_1(2,\ \nu,\ f)$ and Cramer condition (C1) are satisfied by the chain. Then, as $n\to\infty$, we have*

$$\mathbb{E}_\nu(\mu_n'(f))=\mu(f)+(\varphi_\nu+\gamma-\beta/\alpha)n^{-1}+O(n^{-3/2}),\tag{1.8}$$

$$\mathbb{E}_\nu(\widetilde{\mu}_n(f))=\mu(f)+(\gamma-\beta/\alpha)n^{-1}+O(n^{-3/2}),\tag{1.9}$$

$$\mathbb{E}_\nu(\overline{\mu}_n(f))=\mu(f)-(\beta/\alpha)n^{-1}+O(n^{-3/2}).\tag{1.10}$$

If the Cramer condition (C2) is also fulfilled, then

$$\mathbb{E}_\nu(\sigma_n^2(f))=\sigma^2(f)+O(n^{-1}),\ \textit{as } n\to\infty,\tag{1.11}$$

and we have the following CLT under $\mathbb{P}_\nu$,

$$n^{1/2}(\sigma_n^2(f)-\sigma^2(f))\Rightarrow\mathcal{N}(0,\xi^2(f)),\ \textit{as } n\to\infty,\tag{1.12}$$

with $\xi^2(f)=\mu(A)Var_A((\sum_{i=1}^{\tau_A}\overline{f}(X_i))^2-2\alpha^{-1}\beta\sum_{i=1}^{\tau_A}\overline{f}(X_i))$.

Proof. The proof of (1.8)-(1.11) is given in [8] and the linearization of $\sigma_n^2(f)$ below follows from their Lemma 6.3

$$\sigma_n^2(f) = n^{-1}\sum_{j=1}^{l_n-1} g(\mathcal{B}_j) + r_n, \tag{1.13}$$

with $g(\mathcal{B}_j) = \overline{f}(\mathcal{B}_j)^2 - 2\alpha^{-1}\beta\overline{f}(\mathcal{B}_j)$, for $j \geqslant 1$, and for some $\eta_1 > 0$, $\mathbb{P}_\nu(nr_n > \eta_1 \log(n)) = O(n^{-1})$, as $n \to \infty$. We thus have, as $n \to \infty$,

$$n^{1/2}(\sigma_n^2(f) - \sigma^2(f)) = (l_n/n)^{1/2} l_n^{-1/2} \sum_{j=1}^{l_n-1} (g(\mathcal{B}_j) - \mathbb{E}(g(\mathcal{B}_j)) + o_{\mathbb{P}_\nu}(1),$$

and (13) is established with the same argument as for Theorem 17.3.6 in [60], as soon as $Var(g(\mathcal{B}_j)) < \infty$, that is ensured by assumption $\mathcal{H}_1(4,\, f)$.

Remark 4. We emphasize that in a non i.i.d. setting, it is generally difficult to construct an accurate (positive) estimator of the asymptotic variance. When no structural assumption, except stationarity and square integrability, is made on the underlying process X, a possible method, currently used in practice, is based on so-called *blocking techniques*. Indeed under some appropriate mixing conditions (which ensure that the following series converge), it can be shown that the variance of $n^{-1/2}\mu_n'(f)$ may be written $Var(n^{-1/2}\mu_n'(f)) = \Gamma(0) + 2\sum_{t=1}^n (1 - t/n)\Gamma(t)$ and converges to $\sigma^2(f) = \sum_{t=\infty}^{\infty} \Gamma(t) = 2\pi g(0)$, where $g(w) = (2\pi)^{-1}\sum_{t=-\infty}^{\infty} \Gamma(t)\cos(wt)$ and $(\Gamma(t))_{t\geqslant 0}$ denote respectively the spectral density and the autocovariance sequence of the discrete-time stationary process X. Most of the estimators of $\sigma^2(f)$ that have been proposed in the literature (such as the Bartlett spectral density estimator, the moving-block jackknife/subsampling variance estimator, the overlapping or non-overlapping batch means estimator) may be seen as variants of the basic *moving-block bootstrap estimator* (see [52], [55])

$$\hat{\sigma}_{M,n}^2 = \frac{M}{Q}\sum_{i=1}^{Q} (\overline{\mu}_{i,M,L} - \mu_n(f))^2, \tag{1.14}$$

where $\overline{\mu}_{i,M,L} = M^{-1}\sum_{t=L(i-1)+1}^{L(i-1)+M} f(X_t)$ is the mean of f on the i-th data block $(X_{L(i-1)+1}, \dots,\, X_{L(i-1)+M})$. Here, the size M of the blocks and the amount L of 'lag' or overlap between each block are deterministic (eventually depending on n) and $Q = [\frac{n-M}{L}] + 1$, denoting by $[\cdot]$ the integer part, is the number of blocks that may be constructed from the sample $X_1, ..., X_n$. In the case when $L = M$, there is no overlap between block i and block $i+1$ (as the original solution considered by [42], [24]), whereas the case $L = 1$ corresponds to maximum overlap (see [64], [66] for a survey). Under suitable regularity conditions (mixing and moments conditions), it can be shown that if $M \to \infty$ with $M/n \to 0$ and $L/M \to a \in [0,1]$ as $n \to \infty$, then we have

$$\mathbb{E}(\hat{\sigma}_{M,n}^2) - \sigma^2(f) = O(1/M) + O(\sqrt{M/n}), \tag{1.15}$$

$$Var(\hat{\sigma}_{M,n}^2) = 2c\frac{M}{n}\sigma^4(f) + o(M/n), \tag{1.16}$$

as $n \to \infty$, where c is a constant depending on a, taking its smallest value (namely $c = 2/3$) for $a = 0$. This result shows that the bias of such estimators may be very large. Indeed, by optimizing in M we find the optimal choice $M \sim n^{1/3}$, for which we have $\mathbb{E}(\hat{\sigma}^2_{M,n}) - \sigma^2(f) = O(n^{-1/3})$. Various extrapolation and jackknife techniques or kernel smoothing methods have been suggested to get rid of this large bias (refer to [64], [40], [7] and [12]). The latter somehow amount to make use of Rosenblatt smoothing kernels of order higher than two (taking some negative values) for estimating the spectral density at 0. However, the main drawback in using these estimators is that they take negative values for some n, and lead consequently to face problems, when dealing with studentized statistics. In our specific Markovian framework, the estimate $\sigma^2_n(f)$ in the atomic case (or latter $\widehat{\sigma}^2_n(f)$ in the general case) is much more natural and allows to avoid these problems. This is particularly important when the matter is to establish Edgeworth expansions at orders higher than two in such a non i.i.d. setting. As a matter of fact, the bias of the variance may completely cancel the accuracy provided by higher order Edgeworth expansions (but also the one of its Bootstrap approximation) in the studentized case, given its explicit role in such expansions (see [40]).

From Proposition 4.1, we immediately derive that

$$t_n = n^{1/2} \sigma_n^{-1}(f)(\mu_n(f) - \mu(f)) \Rightarrow \mathcal{N}(0,1)\,, \text{ as } n \to \infty,$$

so that asymptotic confidence intervals for $\mu(f)$ are immediately available in the atomic case. This result also shows that using estimators $\widetilde{\mu}_n(f)$ or $\overline{\mu}_n(f)$ instead of $\mu'_n(f)$ allows to eliminate the only quantity depending on the initial distribution ν in the first order term of the bias, which may be interesting for estimation purpose and is crucial when the matter is to deal with an estimator of which variance or sampling distribution may be approximated by a resampling procedure in a nonstationary setting (given the impossibility to approximate the distribution of the "first block sum" $\sum_{i=1}^{\tau_A} f(X_i)$ from one single realization of X starting from ν). For these estimators, it is actually possible to implement specific Bootstrap methodologies, for constructing second order correct confidence intervals for instance (see [9], [10] and section 5). Regarding to this, it should be noticed that Edgeworth expansions (E.E. in abbreviated form) may be obtained using the regenerative method by partitioning the state space according to all possible values for the number l_n regeneration times before n and for the sizes of the first and last block as in [57]. [8] proved the validity of an E.E. in the studentized case, of which form is recalled below. Notice that actually (C3) corresponding to their v) in Proposition 3.1 in [8] is not needed in the unstudentized case. Let $\Phi(x)$ denote the distribution function of the standard normal distribution and set $\phi(x) = d\Phi(x)/dx$.

Theorem 2. *Let* $b(f) = \lim_{n\to\infty} n(\mu_n(f) - \mu(f))$ *be the asymptotic bias of* $\mu_n(f)$*. Under conditions* $\mathcal{H}_0(4)$, $\mathcal{H}_0(2,\,\nu)$, $\mathcal{H}_1(4,\,f)$, $\mathcal{H}_1(2,\,\nu,\,f)$, *(C1), we have the following E.E.,*

$$\sup_{x\in\mathbb{R}} |\mathbb{P}_\nu\left(n^{1/2}\sigma(f)^{-1}(\mu_n(f)-\mu(f))\leq x\right) - E_n^{(2)}(x)| = O(n^{-1}),\ \textit{as } n\to\infty,$$

with

$$E_n^{(2)}(x) = \Phi(x) - n^{-1/2}\frac{k_3(f)}{6}(x^2-1)\phi(x) - n^{-1/2}b(f)\phi(x), \tag{1.17}$$

$$k_3(f) = \alpha^{-1}(M_{3,A} - \frac{3\beta}{\sigma(f)}),\ M_{3,A} = \frac{\mathbb{E}_A((\sum_{i=1}^{\tau_A}\overline{f}(X_i))^3)}{\sigma(f)^3}. \tag{1.18}$$

A similar limit result holds for the studentized statistic under the further hypothesis that (C2), (C3), $\mathcal{H}_0(s)$ and $\mathcal{H}_1(s,\ f)$ are fulfilled with $s = 8+\varepsilon$ for some $\varepsilon > 0$:

$$\sup_{x\in\mathbb{R}} |\mathbb{P}_\nu(n^{1/2}\sigma_n^{-1}(f)(\mu_n(f)-\mu(f))\leq x) - F_n^{(2)}(x)| = O(n^{-1}\log(n)), \tag{1.19}$$

as $n\to\infty$, with $F_n^{(2)}(x) = \Phi(x) + n^{-1/2}\frac{1}{6}k_3(f)(2x^2+1)\phi(x) - n^{-1/2}b(f)\phi(x)$. When $\mu_n(f) = \overline{\mu}_n(f)$, under (C4) , $O(n^{-1}\log(n))$ may be replaced by $O(n^{-1})$.

This theorem may serve for building accurate confidence intervals for $\mu(f)$ (by E.E. inversion as in [1] or [41]). It also paves the way for studying precisely specific bootstrap methods, as in [10]. It should be noted that the skewness $k_3(f)$ is the sum of two terms: the third moment of the recentered block sums and a correlation term between the block sums and the block lengths. The coefficients involved in the E.E. may be directly estimated from the regenerative blocks. Once again by straightforward CLT arguments, we have the following result.

Proposition 2. *For $s \geqslant 1$, under $\mathcal{H}_1(f,\ 2s)$, $\mathcal{H}_1(f,\ 2,\ \nu)$, $\mathcal{H}_0(2s)$ and $\mathcal{H}_0(2,\ \nu)$, $M_{s,A} = \mathbb{E}_A((\sum_{i=1}^{\tau_A}\overline{f}(X_i))^s)$ is well-defined and we have*

$$\widehat{\mu}_{s,n} = n^{-1}\sum_{i=1}^{l_n-1}(f(\mathcal{B}_j) - \overline{\mu}_n(f)l(\mathcal{B}_j))^s = \alpha^{-1}M_{s,A} + O_{\mathbb{P}_\nu}(n^{-1/2}),\ \textit{as } n\to\infty.$$

1.4.2 Positive recurrent case

We now turn to the general positive recurrent case (refer to § 2.3 for assumptions and notation). It is noteworthy that, though they may be expressed using the parameters of the minorization condition $\mathcal{M}$, the constants involved in the CLT are independent from these latter. In particular the mean and the asymptotic variance may be written as

$$\mu(f) = \mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-1}\mathbb{E}_{A_{\mathcal{M}}}(\sum_{i=1}^{\tau_{A_{\mathcal{M}}}} f(X_i)),$$

$$\sigma^2(f) = \mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-1}\mathbb{E}_{A_{\mathcal{M}}}((\sum_{i=1}^{\tau_{A_{\mathcal{M}}}} \overline{f}(X_i))^2),$$

where $\tau_{A_{\mathcal{M}}} = \inf\{n \geqslant 1,\ (X_n,\ Y_n) \in S \times \{1\}\}$ and $\mathbb{E}_{A_{\mathcal{M}}}(.)$ denotes the expectation conditionally to $(X_0, Y_0) \in A_{\mathcal{M}} = S \times \{1\}$. However, one cannot use the estimators of $\mu(f)$ and $\sigma^2(f)$ defined in the atomic setting, applied to the split chain, since the times when the latter regenerates are unobserved. We thus consider the following estimators based on the *approximate regeneration times* (*i.e.* times i when $(X_i, \widehat{Y}_i) \in S \times \{1\}$), as constructed in § 3.2,

$$\widehat{\mu}_n(f) = \widehat{n}_{A_{\mathcal{M}}}^{-1} \sum_{j=1}^{\widehat{l}_n - 1} f(\widehat{\mathcal{B}}_j) \text{ and } \widehat{\sigma}_n^2(f) = \widehat{n}_{A_{\mathcal{M}}}^{-1} \sum_{j=1}^{\widehat{l}_n - 1} \{f(\widehat{\mathcal{B}}_j) - \widehat{\mu}_n(f) l(\widehat{\mathcal{B}}_j)\}^2,$$

with, for $j \geqslant 1$,

$$f(\widehat{\mathcal{B}}_j) = \sum_{i=1+\widehat{\tau}_{A_{\mathcal{M}}}(j)}^{\widehat{\tau}_{A_{\mathcal{M}}}(j+1)} f(X_i),\ l(\widehat{\mathcal{B}}_j) = \widehat{\tau}_{A_{\mathcal{M}}}(j+1) - \widehat{\tau}_{A_{\mathcal{M}}}(j),$$

$$\widehat{n}_{A_{\mathcal{M}}} = \widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n) - \widehat{\tau}_{A_{\mathcal{M}}}(1) = \sum_{j=1}^{\widehat{l}_n - 1} l(\widehat{\mathcal{B}}_j).$$

By convention, $\widehat{\mu}_n(f) = 0$ and $\widehat{\sigma}_n^2(f) = 0$ (resp. $\widehat{n}_{A_{\mathcal{M}}} = 0$), when $\widehat{l}_n \leqslant 1$ (resp., when $\widehat{l}_n = 0$). Since the ARB construction involves the use of an estimate $p_n(x, y)$ of the transition kernel $p(x, y)$, we consider conditions on the rate of convergence of this estimator. For a sequence of nonnegative real numbers $(\alpha_n)_{n\in\mathbb{N}}$ converging to 0 as $n \to \infty$,

$\mathcal{H}_2$: *$p(x,\ y)$ is estimated by $p_n(x,\ y)$ at the rate α_n for the MSE when error is measured by the L^∞ loss over $S \times S$:*

$$\mathbb{E}_\nu(\sup_{(x,y)\in S\times S} |p_n(x, y) - p(x, y)|^2) = O(\alpha_n),\ \text{as } n \to \infty.$$

See Remark 3.1 for references concerning the construction and the study of transition density estimators for positive recurrent chains, estimation rates are usually established under various smoothness assumptions on the density of the joint distribution $\mu(dx)\Pi(x, dy)$ and the one of $\mu(dx)$. For instance, under classical Hölder constraints of order s, the typical rate for the risk in this setup is $\alpha_n \sim (\ln n/n)^{s/(s+1)}$ (refer to [25]).

$\mathcal{H}_3$: *The "minorizing" density γ is such that* $\inf_{x\in S} \gamma(x) > 0$.

$\mathcal{H}_4$: *The transition density $p(x, y)$ and its estimate $p_n(x, y)$ are bounded by a constant $R < \infty$ over S^2.*

Some asymptotic properties of these statistics based on the approximate regeneration data blocks are stated in the following theorem (their proof is omitted since it immediately follows from the argument of Theorem 3.2 and Lemma 5.3 in [10]),

Theorem 3. *If assumptions $\mathcal{H}_0'(2,\,\nu)$, $\mathcal{H}_0'(8)$, $\mathcal{H}_1'(f,\,2,\,\nu)$, $\mathcal{H}_1'(f,\,8)$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$ are satisfied by X, as well as conditions (C1) and (C2) by the split chain, we have, as $n \to \infty$,*

$$\mathbb{E}_\nu(\widehat{\mu}_n(f)) = \mu(f) - \beta/\alpha\; n^{-1} + O(n^{-1}\alpha_n^{1/2}),$$
$$\mathbb{E}_\nu(\widehat{\sigma}_n^2(f)) = \sigma^2(f) + O(\alpha_n \vee n^{-1}),$$

and if $\alpha_n = o(n^{-1/2})$,then

$$n^{1/2}(\widehat{\sigma}_n^2(f) - \sigma^2(f)) \Rightarrow \mathcal{N}(0, \xi^2(f))$$

where α, β and $\xi^2(f)$ are the quantities related to the split chain defined in Proposition 4.1 .

Remark 5. The condition $\alpha_n = o(n^{-1/2})$ as $n \to \infty$ may be ensured by smoothness conditions satisfied by the transition kernel $p(x, y)$: under Hölder constraints of order s such rates are achieved as soon as $s > 1$, that is a rather weak assumption.

We also define the *pseudo-regeneration based standardized* (resp., *studentized*) *sample mean by*

$$\widehat{\varsigma}_n = n^{1/2}\sigma^{-1}(f)(\widehat{\mu}_n(f) - \mu(f)),$$
$$\widehat{t}_n = \widehat{n}_{A_\mathcal{M}}^{1/2}\,\widehat{\sigma}_n(f)^{-1}(\widehat{\mu}_n(f) - \mu(f)).$$

The following theorem straightforwardly results from Theorem 3.

Theorem 4. *Under the assumptions of Theorem 3, we have as $n \to \infty$*

$$\widehat{\varsigma}_n \Rightarrow \mathcal{N}(0,1) \;\; \textit{and}\; \widehat{t}_n \Rightarrow \mathcal{N}(0,1).$$

This shows that from pseudo-regeneration blocks one may easily construct a consistent estimator of the asymptotic variance $\sigma^2(f)$ and asymptotic confidence intervals for $\mu(f)$ in the general positive recurrent case (see Section 5 for more accurate confidence intervals based on a regenerative bootstrap method). In [8], an E.E. is proved for the studentized statistic $\widehat{t}_n$. The main problem consists in handling computational difficulties induced by the dependence structure, that results from the preliminary estimation of the transition density. For partly solving this problem, one may use **Algorithm 4**, involving the *2-split trick*. Under smoothness assumptions for the transition kernel (which are often fulfilled in practice), [11] established the validity of the E.E. up to $O(n^{-5/6}\log(n))$, stated in the result below.

For this we need to introduce the following Cramer condition which is somehow easier to check than the Cramer condition C1 for the split-chain

($C1\prime$) Assume that, for the chosen small set S,

$$\overline{lim}_{|t|\to\infty}\sup_{x\in S}|E_x(\exp(it(\sum_{i=1}^{\tau_S}\{f(X_i)-\mu(f)\}))|<1$$

Theorem 5. *Suppose that* ($C1\prime$), $\mathcal{H}_0'(\kappa,\ \nu)$, $\mathcal{H}_1'(\kappa,\ f,\ \nu)$, $\mathcal{H}_0'(\kappa)$, $\mathcal{H}_1'(\kappa,\ f)$ *with* $\kappa>6$, $\mathcal{H}_2$, $\mathcal{H}_3$ *and* $\mathcal{H}_4$ *are fulfilled. Let* m_n *and* p_n *be integer sequences tending to* ∞ *as* $n\to\infty$, *such that* $n^{1/\gamma}\le p_n\le m_n$ *and* $m_n=o(n)$ *as* $n\to\infty$. *Then, the following limit result holds for the pseudo-regeneration based standardized sample mean obtained via Algorithm 4*

$$\sup_{x\in\mathbb{R}}|\mathbb{P}_\nu\left(\widehat{\varsigma}_n\le x\right)-E_n^{(2)}(x)|=O(n^{-1/2}\alpha_{m_n}^{1/2}\vee n^{-3/2}m_n),\ as\ n\to\infty,$$

and if in addition ($C4$) *holds for the split chain and the preceding assumptions with* $\kappa>8$ *and are satisfied, we also have*

$$\sup_{x\in\mathbb{R}}|\mathbb{P}_\nu(\widehat{t}_n\le x)-F_n^{(2)}(x)|=O(n^{-1/2}\alpha_{m_n}^{1/2}\vee n^{-3/2}m_n),\ \ as\ n\to\infty,$$

where $E_n^{(2)}(x)$ *and* $F_n^{(2)}(x)$ *are the expansions defined in Theorem 4.2 related to the split chain. In particular, if* $\alpha_{m_n}=m_n\log(m_n)$, *by picking* $m_n=n^{2/3}$, *these E.E. hold up to* $O(n^{-5/6}\log(n))$.

The conditions are satisfied for a wide range of Markov chains, including nonstationary cases and chains with polynomial decay of $\alpha-$mixing coefficients (*cf* remark 2.1) that do not fall into the validity framework of the Moving Block Bootstrap methodology. In particular it is worth noticing that these conditions are weaker than [39]'s conditions (in a strong mixing setting).

As stated in the following proposition, the coefficients involved in the E.E.'s above may be estimated from the approximate regeneration blocks.

Proposition 3. *Under* $\mathcal{H}_0'(2s,\ \nu)$, $\mathcal{H}_1'(2s,\ \nu,\ f)$, $\mathcal{H}_0'(2s\vee 8)$, $\mathcal{H}_1'(2s\vee 8,\ f)$ *with* $s\ge 2$, $\mathcal{H}_2$, $\mathcal{H}_3$ *and* $\mathcal{H}_4$, *the expectation* $M_{s,A_\mathcal{M}}=\mathbb{E}_{A_\mathcal{M}}((\sum_{i=1}^{\tau_{A_\mathcal{M}}}\overline{f}(X_i))^s)$ *is well-defined and we have, as* $n\to\infty$,

$$\widehat{\mu}_{s,n}=n^{-1}\sum_{i=1}^{l_n-1}(f(\widehat{\mathcal{B}}_j)-\widehat{\mu}_n(f)l(\widehat{\mathcal{B}}_j))^s=\mathbb{E}_{A_\mathcal{M}}(\tau_{A_\mathcal{M}})^{-1}M_{s,A_\mathcal{M}}+O_{\mathbb{P}_\nu}(\alpha_{m_n}^{1/2}).$$

1.4.3 Some illustrative examples

Here we give some examples with the aim to illustrate the wide range of applications of the results previously stated.

Example 1 : countable Markov chains.

Let X be a general irreducible chain with a countable state space E. For such a chain, any recurrent state $a \in E$ is naturally an accessible atom and conditions involved in the limit results presented in § 4.1 may be easily checked at hand. Consider for instance Cramer condition (C1), denote by Π the transition matrix and set $A = \{a\}$. We have, for any $k \in \mathbb{N}^*$:

$$\begin{aligned}\left|\mathbb{E}_A(e^{it\sum_{j=1}^{\tau_A} f(X_j)})\right| &= \left|\sum_{l=1}^{\infty} \mathbb{E}_A(e^{it\sum_{j=1}^{l} f(X_j)}|\tau_A = l)\mathbb{P}_A(\tau_A = l)\right| \\ &\leqslant \left|\mathbb{E}_A(e^{it\sum_{j=1}^{k} f(X_j)}|\tau_A = k)\mathbb{P}_A(\tau_A = k) + 1 - \mathbb{P}_A(\tau_A = k)\right|.\end{aligned}$$

and

$$\begin{aligned}\left|\mathbb{E}_A(e^{it\sum_{j=1}^{k} f(X_j)}|\tau_A = k)\right| &= \left|\sum_{x_1\neq a,\ldots,x_{k-1}\neq a} e^{it\sum_{j=1}^{k} f(x_j)}\Pi(a,x_1)...\Pi(x_{k-1},a)\right| \\ &\leqslant \sum_{x_1\neq a,\ldots,x_{k-1}\neq a} \Pi(a,x_1)...\Pi(x_{k-1},a) = \mathbb{P}_A(\tau_A = k).\end{aligned}$$

We thus have $|E_A(e^{itS_A(f)})| \leq \mathbb{P}_A(\tau_A = k)^2 + 1 - \mathbb{P}_A(\tau_A = k)$. Hence, as soon as there exists $k_0 \geqslant 1$ such that the probability that the chain returns to state a in k_0 steps is strictly positive and strictly less than 1, (C1) is fulfilled. Notice that the only case for which such condition does not hold corresponds to the case when the return time to the atom is deterministic (observe that this includes the discrete i.i.d. case, that corresponds to the case when the whole state space is a Harris atom).

Example 2 : modulated random walk on $\mathbb{R}_+$.

Consider the model

$$X_0 = 0 \text{ and } X_{n+1} = (X_n + W_n)_+ \text{ for } n \in \mathbb{N}, \tag{1.20}$$

where $x_+ = \max(x, 0)$, (X_n) and (W_n) are sequences of r.v.'s such that, for all $n \in \mathbb{N}$, the distribution of W_n conditionally to $X_0, ..., X_n$ is given by $U(X_n, .)$ where $U(x, w)$ is a transition kernel from $\mathbb{R}_+$ to $\mathbb{R}$. Then, X_n is a Markov chain on $\mathbb{R}_+$ with transition probability kernel $\Pi(x,\ dy)$ given by

$$\begin{aligned}\Pi(x,\ \{0\}) &= U(x,\ ]-\infty,\ -x]), \\ \Pi(x,\ ]y,\ \infty[) &= U(x,\ ]y-x,\ \infty[),\end{aligned}$$

for all $x \geqslant 0$. Observe that the chain Π is δ_0-irreducible when $U(x, .)$ has infinite left tail for all $x \geqslant 0$ and that $\{0\}$ is then an accessible atom for X. The chain is shown to be positive recurrent iff there exists $b > 0$ and a test function $V : \mathbb{R}_+ \to [0, \infty]$ such that $V(0) < \infty$ and the drift condition below holds for all $x \geqslant 0$

$$\int \Pi(x, dy)V(y) - V(x) \leqslant -1 + b\mathbb{I}\{x = 0\},$$

(see in [60]). The times at which X reaches the value 0 are thus regeneration times, and allow to define regeneration blocks dividing the sample path, as shown in Fig. 1. Such a modulated random walk (for which, at each step n, the increasing W_n depends on the actual state $X_n = x$), provides a model for various systems, such as the popular *content-dependent storage process* studied in [44] (see also [19]) or the *work-modulated single server queue* in the context of queuing systems (*cf* [20]). For such atomic chains with continuous state space (refer to [60], [33], [34] and [2] for other examples of such chains), one may easily check conditions used in § 3.1 in many cases. One may show for instance that (C1) is fulfilled as soon as there exists $k \geqslant 1$ such that $0 < \mathbb{P}_A(\tau_A = k) < 1$ and the distribution of $\sum_{i=1}^{k} f(X_i)$ conditioned on $X_0 \in A$ and $\tau_A = k$ is absolutely continuous. For the regenerative model described above, this sufficient condition is fulfilled with $k = 2$, $f(x) = x$ and $A = \{0\}$, when it is assumed for instance that $U(x, dy)$ is absolutely continuous for all $x \geqslant 0$ and $\emptyset \neq \text{supp}U(0, dy) \cap \mathbb{R}_+^* \neq \mathbb{R}_+^*$.

Example 3: nonlinear time series.

Consider the heteroskedastic autoregressive model

$$X_{n+1} = m(X_n) + \sigma(X_n)\varepsilon_{n+1},\ n \in \mathbb{N},$$

where $m : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}_+^*$ are measurable functions, $(\varepsilon_n)_{n\in\mathbb{N}}$ is a i.i.d. sequence of r.v.'s drawn from $g(x)dx$ such that, for all $n \in \mathbb{N}$, ε_{n+1} is independent from the X_k's, $k \leqslant n$ with $\mathbb{E}(\varepsilon_{n+1}) = 0$ and $\mathbb{E}(\varepsilon_{n+1}^2) = 1$. The transition kernel density of the chain is given by $p(x, y) = g((y - m(x))/\sigma(x))$, $(x,\ y) \in \mathbb{R}^2$. Assume further that g, m and σ are continuous functions and there exists $x_0 \in \mathbb{R}$ such that $p(x_0, x_0) > 0$. Then, the transition density is uniformly bounded from below over some neighborhood $V_{x_0}(\varepsilon)^2 = [x_0 - \varepsilon, x_0 + \varepsilon]^2$ of $(x_0,\ x_0)$ in $\mathbb{R}^2$: there exists $\delta = \delta(\varepsilon) \in]0,\ 1[$ such that,

$$\inf_{(x,y)\in V_{x_0}^2} p(x, y) \geqslant \delta(2\varepsilon)^{-1}. \tag{1.21}$$

We thus showed that the chain X satisfies the minorization condition $\mathcal{M}(1, V_{x_0}(\varepsilon), \delta, \mathcal{U}_{V_{x_0}(\varepsilon)})$. Furthermore, block-moment conditions for such time series model may be checked via the practical conditions developed in [29] (see their example 3).

1.5 Regenerative block-bootstrap

[5] and [27] proposed a specific bootstrap methodology for atomic Harris positive recurrent Markov chains, which exploits the renewal properties of the

latter. The main idea underlying this method consists in resampling a deterministic number of data blocks corresponding to regeneration cycles. However, because of some inadequate standardization, the *regeneration-based bootstrap* method proposed in [27] is not second order correct when applied to the sample mean problem (its rate is $O_{\mathbb{P}}(n^{-1/2})$ in the stationary case). Prolongating this work, [9] have shown how to modify suitably this resampling procedure to make it second order correct up to $O_{\mathbb{P}}(n^{-1}\log(n))$ in the unstudentized case (*i.e.* when the variance is known) when the chain is stationary. However this Bootstrap method remains of limited interest from a practical viewpoint, given the necessary modifications (standardization and recentering) and the restrictive stationary framework required to obtain the second order accuracy: it fails to be second order correct in the nonstationary case, as a careful examination of the second order properties of the sample mean statistic of a positive recurrent chain based on its E.E. shows (*cf* [57], [8]). A powerful alternative, namely the *Regenerative Block-Bootstrap* (*RBB),* have been thus proposed and studied in [10], that consists in imitating further the renewal structure of the chain by resampling regeneration data blocks, until the length of the reconstructed Bootstrap series is larger than the length n of the original data series, so as to approximate the distribution of the (random) number of regeneration blocks in a series of length n and remove some bias terms (see section 4). Here we survey the asymptotic validity of the RBB for the studentized mean by an adequate estimator of the asymptotic variance. This is the useful version for confidence intervals but also for practical use of the Bootstrap (*cf* [43]) and for a broad class of Markov chains (including chains with strong mixing coefficients decreasing at a polynomial rate), the accuracy reached by the RBB is proved to be of order $O_{\mathbb{P}}(n^{-1})$ both for the standardized and the studentized sample mean. The rate obtained is thus comparable to the optimal rate of the Bootstrap distribution in the i.i.d. case, contrary to the *Moving Block Bootstrap* (*cf* [40], [53]). The proof relies on the E.E. for the studentized sample mean stated in § 4.1 (see Theorems 4.2, 4.6). In [10] a straightforward extension of the RBB procedure to general Harris chains based on the ARB construction (see § 3.1) is also proposed (it is called *Approximate Regenerative Block-Bootstrap*, *ARBB* in abbreviated form). Although it is based on the approximate regenerative blocks, it is shown to be still second order correct when the estimate p_n used in the ARB algorithm is consistent. We also emphasize that the principles underlying the (A)RBB may be applied to any (eventually continuous time) regenerative process (and not necessarily markovian) or with a regenerative extension that may be approximated (see [84]).

1.5.1 The (approximate) regenerative block-bootstrap algorithm.

Once true or approximate regeneration blocks $\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n-1}$ are obtained (by implementing *Algorithm 1, 2, 3* or *4*), the (*approximate) regenerative* *block-bootstrap* algorithm for computing an estimate of the sample distribution of some statistic $T_n = T(\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n-1})$ with standardization $S_n = S(\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n-1})$ is performed in 3 steps as follows.

Algorithm 5 *(Approximate) Regenerative Block-Bootstrap*

1. Draw sequentially bootstrap data blocks $\mathcal{B}_1^*, ..., \mathcal{B}_k^*$ (with length denoted by $l(\mathcal{B}_j^*)$, $j = 1, ..., k$) independently from the empirical distribution $\widehat{\mathcal{L}}_n = (\widehat{l}_n - 1)^{-1} \sum_{j=1}^{\widehat{l}_n-1} \delta_{\widehat{\mathcal{B}}_j}$ of the initial blocks $\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n-1}$, until the length of the bootstrap data series $l^*(k) = \sum_{j=1}^{k} l(\mathcal{B}_j^*)$ is larger than n. Let $l_n^* = \inf\{k \geqslant 1,\, l^*(k) > n\}$.
2. From the bootstrap data blocks generated at step 1, reconstruct a pseudo-trajectory by binding the blocks together, getting the reconstructed *(A)RBB sample path*
$$X^{*(n)} = (\mathcal{B}_1^*, ..., \mathcal{B}_{l_n^*-1}^*).$$
Then compute the *(A)RBB statistic* and its *(A)RBB standardization*
$$T_n^* = T(X^{*(n)}) \text{ and } S_n^* = S(X^{*(n)}).$$
3. The *(A)RBB distribution* is then given by
$$H_{(A)RBB}(x) = \mathbb{P}^*(S_n^{*-1}(T_n^* - T_n) \leqslant x),$$
where $\mathbb{P}^*$ denotes the conditional probability given the original data.

Remark 6. A Monte-Carlo approximation to $H_{(A)RBB}(x)$ may be straightforwardly computed by repeating independently N times this algorithm.

1.5.2 Atomic case: second order accuracy of the RBB

In the case of the sample mean, the bootstrap counterparts of the estimators $\overline{\mu}_n(f)$ and $\sigma_n^2(f)$ considered in § 4.1 (using the notation therein) are

$$\mu_n^*(f) = n_A^{*-1} \sum_{j=1}^{l_n^*-1} f(\mathcal{B}_j^*) \text{ and } \sigma_n^{*2}(f) = n_A^{*-1} \sum_{j=1}^{l_n^*-1} \left\{ f(\mathcal{B}_j^*) - \mu_n^*(f) l(\mathcal{B}_j^*) \right\}^2, \tag{1.22}$$

with $n_A^* = \sum_{j=1}^{l_n^*-1} l(\mathcal{B}_j^*)$. Let us consider the RBB distribution estimates of the unstandardized and studentized sample means

$$H_{RBB}^U(x) = \mathbb{P}^*(n_A^{1/2} \sigma_n(f)^{-1} \{\mu_n^*(f) - \overline{\mu}_n(f)\} \leq x),$$
$$H_{RBB}^S(x) = \mathbb{P}^*(n_A^{*-1/2} \sigma_n^{*-1}(f) \{\mu_n^*(f) - \overline{\mu}_n(f)\} \leq x).$$

The following theorem established in [9] shows that the RBB is asymptotically valid for the sample mean. Moreover it ensures that the RBB attains the optimal rate of the i.i.d. Bootstrap. The proof of this result crucially relies on the E.E. given in [57] in the standardized case and its extension to the studentized case proved in [8].

Theorem 6. *Suppose that (C1) is satisfied. Under $\mathcal{H}_0'(2,\ \nu)$, $\mathcal{H}_1'(2,\ f,\ \nu)$, $\mathcal{H}_0'(\kappa)$ and $\mathcal{H}_1(\kappa,\ f)$ with $\kappa > 6$, the RBB distribution estimate for the unstandardized sample mean is second order accurate in the sense that*

$$\Delta_n^U = \sup_{x\in\mathbb{R}} |H_{RBB}^U(x) - H_\nu^U(x)| = O_{\mathbb{P}_\nu}(n^{-1}),\ \text{as } n\to\infty,$$

with $H_\nu^U(x) = \mathbb{P}_\nu(n_A^{1/2}\sigma_f^{-1}\{\overline{\mu}_n(f) - \mu(f)\} \leq x)$. And if in addition (C4), $\mathcal{H}_0'(\kappa)$ and $\mathcal{H}_1(\kappa,\ f)$ are checked with $\kappa > 8$, the RBB distribution estimate for the standardized sample mean is also 2nd order correct

$$\Delta_n^S = \sup_{x\in\mathbb{R}} |H_{RBB}^S(x) - H_\nu^S(x)| = O_{\mathbb{P}_\nu}(n^{-1}),\ \text{as } n\to\infty,$$

with $H_\nu^S(x) = \mathbb{P}_\nu(n_A^{1/2}\sigma_n^{-1}(f)\{\overline{\mu}_n(f) - \mu(f)\} \leq x)$.

1.5.3 Asymptotic validity of the ARBB for general chains

The ARBB counterparts of the statistics $\widehat{\mu}_n(f)$ and $\widehat{\sigma}_n^2(f)$ considered in § 4.2 (using the notation therein) may be expressed as

$$\mu_n^*(f) = n_{A_\mathcal{M}}^{*-1}\sum_{j=1}^{l_n^*-1} f(\mathcal{B}_j^*) \text{ and } \sigma_n^{*2}(f) = n_{A_\mathcal{M}}^{*-1}\sum_{j=1}^{l_n^*-1}\left\{f(\mathcal{B}_j^*) - \mu_n^*(f)l(\mathcal{B}_j^*)\right\}^2,$$

denoting by $n_{A_\mathcal{M}}^* = \sum_{j=1}^{l_n^*-1} l(\mathcal{B}_j^*)$ the length of the ARBB data series. Define the ARBB versions of the pseudo-regeneration based unstudentized and studentized sample means (*cf* § 4.2) by

$$\widehat{\varsigma}_n^* = n_{A_\mathcal{M}}^{1/2}\frac{\mu_n^*(f) - \widehat{\mu}_n(f)}{\widehat{\sigma}_n(f)} \text{ and } \widehat{t}_n^* = n_{A_\mathcal{M}}^{*1/2}\frac{\mu_n^*(f) - \widehat{\mu}_n(f)}{\sigma_n^*(f)}.$$

The unstandardized and studentized version of the ARBB distribution estimates are then given by

$$H_{ARBB}^U(x) = \mathbb{P}^*(\widehat{\varsigma}_n^* \leq x \mid X^{(n+1)}) \text{ and } H_{ARBB}^S(x) = \mathbb{P}^*(\widehat{t}_n^* \leq x \mid X^{(n+1)}).$$

This is the same construction as in the atomic case, except that one uses the approximate regeneration blocks instead of the exact regenerative ones (*cf* Theorem 3.3 in [10]).

Theorem 7. *Under the hypotheses of Theorem 4.2, we have the following convergence results in distribution under $\mathbb{P}_\nu$*

$$\Delta_n^U = \sup_{x\in\mathbb{R}} |H_{ARBB}^U(x) - H_\nu^U(x)| \to 0,\ \text{as } n\to\infty,$$
$$\Delta_n^S = \sup_{x\in\mathbb{R}} |H_{ARBB}^S(x) - H_\nu^S(x)| \to 0,\ \text{as } n\to\infty.$$

Second order properties of the ARBB using the 2-split trick

To bypass the technical difficulties related to the dependence problem induced by the preliminary step estimation, assume now that the pseudo regenerative blocks are constructed according to Algorithm 4 (possibly including the selection rule for the small set of Algorithm 3 when using only the m_n first observations). It is then easier (at the price of a small loss in the 2nd order term) to get second order results both in the case of standardized and studentized statistics, as stated below (refer to [10] for the technical proof).

Theorem 8. *Under assumptions $(C1')$, $\mathcal{H}_0'(\kappa,\ \nu)$, $\mathcal{H}_1'(\kappa,\ f,\ \nu)$, $\mathcal{H}_0'(f,\ \kappa)$, $\mathcal{H}_1'(f,\ \kappa)$ with $\kappa > 6$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$, we have the second order validity of the ARBB distribution both in the standardized and unstandardized case up to order*

$$\Delta_n^U = O_{\mathbb{P}_\nu}(n^{-1/2}\alpha_{m_n}^{1/2} \vee n^{-1/2}n^{-1}m_n\}),\ as\ n \to \infty.$$

And if in addition, $(C4)$ holds for the split chains and the preceding assumptions hold with $\kappa > 8$, we have

$$\Delta_n^S = O_{\mathbb{P}_\nu}(n^{-1/2}\alpha_{m_n}^{1/2} \vee n^{-1/2}n^{-1}m_n),\ as\ n \to \infty$$

In particular if $\alpha_m = m\log(m)$, by choosing $m_n = n^{2/3}$, the ARBB is second order correct up to $O(n^{-5/6}\log(n))$.

It is worth noticing that the rate that can be attained by the 2-split trick variant of the ARBB for such chains is faster than the optimal rate the MBB may achieve, which is typically of order $O(n^{-3/4})$ under very strong assumptions (see [40], [53]). Other variants of the bootstrap (sieve bootstrap) for time-series (see [21]) may also yield (at least pratically) very accurate approximation (see [22]). When some specific non-linear structure is assumed for the chain (see our example 3), nonparametric method estimation and residual based resampling methods may also be used : see for instance [35]. However to our knowledge, there is no explicit rate of convergence available for these kinds of bootstrap techniques. An empirical comparison of all these recent methods is under the scope of this paper but would be certainly of great help.

1.6 Some extensions to U-statistics

We now turn to extend some of the asymptotic results stated in sections 4 and 5 for sample mean statistics to a wider class of functionals and shall consider statistics of the form $\sum_{1\leqslant i\neq j\leqslant n} U(X_i, X_j)$. For the sake of simplicity, we confined the study to U-statistics of degree 2, in the real case only. As will be shown below, asymptotic validity of inference procedures based on such statistics does not straightforwardly follow from results established in the previous sections, even for atomic chains. Furthermore, whereas asymptotic validity of the (approximate) regenerative block-bootstrap for these functionals may be easily obtained, establishing its second order validity and give precise rate is much more difficult from a technical viewpoint and is left to a further study. Besides, arguments presented in the sequel may be easily adapted to V-statistics $\sum_{1\leqslant i,\ j\leqslant n} U(X_i, X_j)$.

1.6.1 Regenerative case

Given a trajectory $X^{(n)} = (X_1, ..., X_n)$ of a Harris positive atomic Markov chain with stationary probability law μ (refer to § 2.2 for assumptions and notation), we shall consider in the following U-statistics of the form

$$T_n = \frac{1}{n(n-1)} \sum_{1\leqslant i\neq j\leqslant n} U(X_i,\ X_j), \tag{1.23}$$

where $U : E^2 \to \mathbb{R}$ is a kernel of degree 2. Even if it entails introducing the symmetrized version of T_n, it is assumed throughout the section that the kernel $U(x, y)$ is symmetric. Although such statistics have been mainly used and studied in the case of i.i.d. observations, in dependent settings such as ours, these statistics are also of interest, as shown by the following examples.

• In the case when the chain takes real values and is positive recurrent with stationary distribution μ, the variance of the stationary distribution $s^2 = \mathbb{E}_\mu((X - \mathbb{E}_\mu(X))^2)$, if well defined (note that it differs in general from the asymptotic variance of the mean statistic studied in § 4.1), may be consistently estimated under adequate block moment conditions by

$$\widehat{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_n)^2 = \frac{1}{n(n-1)} \sum_{1\leqslant i\neq j\leqslant n} (X_i - X_j)^2/2,$$

where $\mu_n = n^{-1}\sum_{i=1}^{n} X_i$, which is a U-statistic of degree 2 with symmetric kernel $U(x, y) = (x - y)^2/2$.

• In the case when the chain takes its values in the multidimensional space $\mathbb{R}^p$, endowed with some norm $||.\ ||$, many statistics of interest may be written as a U-statistic of the form

$$U_n = \frac{1}{n(n-1)} \sum_{1\leqslant i\neq j\leqslant n} H(||X_i - X_j||),$$

where $H : \mathbb{R} \to \mathbb{R}$ is some measurable function. And in the particular case when $p = 2$, for some fixed t in $\mathbb{R}^2$ and some smooth function h, statistics of type

$$U_n = \frac{1}{n(n-1)} \sum_{1\leqslant i\neq j\leqslant n} h(t,\ X_i,\ X_j)$$

arise in the study of the *correlation dimension* for dynamic systems (see [18]). *Depth statistical functions* for spatial data are also particular examples of such statistics (*cf* [81]).

In what follows, the parameter of interest is

$$\mu(U)=\int_{(x,y)\in E^2}U(x,y)\mu(dx)\mu(dy), \tag{1.24}$$

which quantity we assume to be finite. As in the case of i.i.d. observations, a natural estimator of $\mu(U)$ in our markovian setting is T_n. We shall now study its consistency properties and exhibit an adequate sequence of renormalizing constants for the latter, by using the *regeneration blocks construction* once again. For later use, define $\omega_U:\mathbb{T}^2\to\mathbb{R}$ by

$$\omega_U(x^{(k)},y^{(l)})=\sum_{i=1}^{k}\sum_{j=1}^{l}U(x_i,y_j),$$

for any $x^{(k)}=(x_1,...,x_k)$, $y^{(l)}=(y_1,...,y_l)$ in the torus $\mathbb{T}=\cup_{n=1}^{\infty}E^n$ and observe that ω_U is symmetric, as U.

"Regeneration-based Hoeffding's decomposition"

By the representation of μ as a Pitman's occupation measure (*cf* Theorem 2.1), we have

$$\begin{aligned}\mu(U)&=\alpha^{-2}\mathbb{E}_A(\sum_{i=1}^{\tau_A(1)}\sum_{l=\tau_A(1)+1}^{\tau_A(2)}U(X_i,X_j))\\&=\alpha^{-2}\mathbb{E}(\omega_U(\mathcal{B}_l,\mathcal{B}_k)),\end{aligned}$$

for any integers k, l such that $k\neq l$. In the case of U-statistics based on dependent data, the classical (orthogonal) Hoeffding decomposition (*cf* [80]) does not hold anymore. Nevertheless, we may apply the underlying projection principle for establishing the asymptotic normality of T_n by approximatively rewriting it as a U-statistic of degree 2 computed on the regenerative blocks only, in a fashion very similar to the *Bernstein blocks technique* for strongly mixing random fields (*cf* [30], [7]), as follows. As a matter of fact, the estimator T_n may be decomposed as

$$T_n=\frac{(l_n-1)(l_n-2)}{n(n-1)}U_{l_n-1}+T_n^{(0)}+T_n^{(n)}+\Delta_n, \tag{1.25}$$

where,

$$U_L=\frac{2}{L(L-1)}\sum_{1\leqslant k<l\leqslant L}\omega_U(\mathcal{B}_k,\mathcal{B}_l),$$

$$T_n^{(0)}=\frac{2}{n(n-1)}\sum_{1\leqslant k\leqslant l_n-1}\omega_U(\mathcal{B}_k,\mathcal{B}_0),\ T_n^{(n)}=\frac{2}{n(n-1)}\sum_{0\leqslant k\leqslant l_n-1}\omega_U(\mathcal{B}_k,\mathcal{B}_{l_n}^{(n)}),$$

$$\Delta_n=\frac{1}{n(n-1)}\{\sum_{k=0}^{l_n-1}\omega_U(\mathcal{B}_k,\mathcal{B}_k)+\omega_U(\mathcal{B}_{l_n}^{(n)},\mathcal{B}_{l_n}^{(n)})-\sum_{i=1}^{n}U(X_i,X_i)\}.$$

Observe that the "block diagonal part" of T_n, namely Δ_n, may be straightforwardly shown to converge $\mathbb{P}_\nu$- a.s. to 0 as $n \to \infty$, as well as $T_n^{(0)}$ and $T_n^{(1)}$ by using the same arguments as the ones used in § 4.1 for dealing with sample means, under obvious block moment conditions (see conditions *(ii)-(iii)* below). And, since $l_n/n \to \alpha^{-1}$ $\mathbb{P}_\nu$- a.s. as $n \to \infty$, asymptotic properties of T_n may be derived from the ones of U_{l_n-1}, which statistic depends on the regeneration blocks only. The key point relies in the fact that the theory of U-statistics based on i.i.d. data may be straightforwardly adapted to functionals of the i.i.d. regeneration blocks of the form $\sum_{k<l}\omega_U(\mathcal{B}_k,\mathcal{B}_l)$. Hence, the asymptotic behaviour of the U-statistic U_L as $L \to \infty$ essentially depends on the properties of the linear and quadratic terms appearing in the following variant of *Hoeffding's decomposition.* For k, $l \geqslant 1$, define

$$\widetilde{\omega}_U(\mathcal{B}_k,\mathcal{B}_l) = \sum_{i=\tau_A(k)+1}^{\tau_A(k+1)} \sum_{j=\tau_A(l)+1}^{\tau_A(l+1)} \{U(X_i,X_j)-\mu(U)\}.$$

(notice that $\mathbb{E}(\widetilde{\omega}_U(\mathcal{B}_k,\mathcal{B}_l))=0$ when $k \neq l$) and for $L \geqslant 1$ write the expansion

$$U_L - \mu(U) = \frac{2}{L}\sum_{k=1}^{L}\omega_U^{(1)}(\mathcal{B}_k) + \frac{2}{L(L-1)}\sum_{1\leqslant k<l\leqslant L}\omega_U^{(2)}(\mathcal{B}_k,\mathcal{B}_l), \tag{1.26}$$

where, for any $b_1=(x_1,...,x_l)\in\mathbb{T}$,

$$\omega_U^{(1)}(b_1) = \mathbb{E}(\widetilde{\omega}_U(\mathcal{B}_1,\mathcal{B}_2)|\mathcal{B}_1=b_1) = \mathbb{E}_A(\sum_{i=1}^{l}\sum_{j=1}^{\tau_A}\widetilde{\omega}_U(x_i,X_j))$$

is the linear term (see also our definition of the *influence function* of the parameter $\mathbb{E}(\omega(\mathcal{B}_1,\mathcal{B}_2))$ in section 7) and for all b_1, b_2 in $\mathbb{T}$,

$$\omega_U^{(2)}(b_1,b_2) = \widetilde{\omega}_U(b_1,b_2) - \widetilde{\omega}_U^{(1)}(b_1) - \widetilde{\omega}_U^{(1)}(b_2)$$

is the quadratic degenerate term (gradient of order 2). Notice that by using the Pitman's occupation measure representation of μ, we have as well, for any $b_1=(x_1,...,x_l)\in\mathbb{T}$,

$$(E_A\tau_A)^{-1}\omega_U^{(1)}(b_1) = \sum_{i=1}^{l}\mathbb{E}_\mu(\widetilde{\omega}_U(x_i,X_1)).$$

For resampling purposes, we also introduce the U-statistic based on the data between the first regeneration time and the last one only:

$$\widetilde{T}_n = \frac{2}{\widetilde{n}(\widetilde{n}-1)}\sum_{1+\tau_A\leqslant i<j\leqslant\tau_A(l_n)} U(X_i,X_j),$$

with $\widetilde{n}=\tau_A(l_n)-\tau_A$ and $\widetilde{T}_n=0$ when $l_n\leqslant 1$ by convention.

Asymptotic normality and asymptotic validity of the RBB

Now suppose that the following conditions, which are involved in the next result, are fulfilled by the chain.

(i) (Non degeneracy of the U-statistic)

$$0 < \sigma_U^2 = \mathbb{E}(\omega_U^{(1)}(\mathcal{B}_1)^2) < \infty.$$

(ii) (Block-moment conditions: linear part) For some $s \geqslant 2$,

$$\mathbb{E}(\omega_{|U|}^{(1)}(\mathcal{B}_1)^s) < \infty \text{ and } \mathbb{E}_\nu(\omega_{|U|}^{(1)}(\mathcal{B}_0)^2) < \infty.$$

(iii) (Block-moment conditions: quadratic part) For some $s \geqslant 2$,

$$\mathbb{E}|\omega_{|U|}(\mathcal{B}_1, \mathcal{B}_2)|^s < \infty \text{ and } \mathbb{E}|\omega_{|U|}(\mathcal{B}_1, \mathcal{B}_1)|^s < \infty,$$
$$\mathbb{E}_\nu|\omega_{|U|}(\mathcal{B}_0, \mathcal{B}_1)|^2 < \infty \text{ and } \mathbb{E}_\nu|\omega_{|U|}(\mathcal{B}_0, \mathcal{B}_0)|^2 < \infty.$$

By construction, under *(ii)-(iii)* we have the crucial orthogonality property:

$$Cov(\omega_U^{(1)}(\mathcal{B}_1),\ \omega_U^{(2)}(\mathcal{B}_1, \mathcal{B}_2)) = 0. \tag{1.27}$$

Now a slight modification of the argument given in [47] allows to prove straightforwardly that $\sqrt{L}(U_L - \mu(U))$ is asymptotically normal with zero mean and variance $4\sigma_U^2$. Furthermore, by adapting the classical CLT argument for sample means of Markov chains (refer to [60] for instance) and using (1.27) and $l_n/n \to \alpha^{-1}$ $\mathbb{P}_\nu$-a.s. as $n \to \infty$, one deduces that $\sqrt{n}(T_n - \mu(U)) \Rightarrow \mathcal{N}(0, \varSigma^2)$ as $n \to \infty$ under $\mathbb{P}_\nu$, with $\varSigma^2 = 4\alpha^{-3}\sigma_U^2$.

Besides, estimating the normalizing constant is important (for constructing confidence intervals or bootstrap counterparts for instance). So we define the natural estimator $\sigma_{U,\ l_n-1}^2$ of σ_U^2 based on the (asymptotically i.i.d.) $l_n - 1$ regeneration data blocks by

$$\sigma_{U,\ L}^2 = (L-1)(L-2)^{-2}\sum_{k=1}^{L}[(L-1)^{-1}\sum_{l=1, k\neq l}^{L}\omega_U(\mathcal{B}_k, \mathcal{B}_l) - U_L]^2,$$

for $L \geqslant 1$. The estimate $\sigma_{U,\ L}^2$ is a simple transposition of the *jackknife estimator* considered in [23] to our setting and may be easily shown to be strongly consistent (by adapting the SLLN for U-statistics to this specific functional of the i.i.d regeneration blocks). Furthermore, we derive that $\varSigma_n^2 \to \varSigma^2$ $\mathbb{P}_\nu$-a.s., as $n \to \infty$, where

$$\varSigma_n^2 = 4(l_n/n)^3\sigma_{U,\ l_n-1}^2.$$

We also consider the regenerative block-bootstrap counterparts T_n^* and $\varSigma_n^{*2}$ of $\widetilde{T}_n$ and $\varSigma_n^2$ respectively, constructed via *Algorithm 5*:

$$T_n^* = \frac{2}{n^*(n^*-1)} \sum_{1 \leqslant i < j \leqslant n^*} U(X_i^*,\ X_j^*),$$

$$\Sigma_n^{*2} = 4(l_n^*/n^*)^3 \sigma^{*2}_{U,\ l_n^*-1},$$

where n^* denotes the length of the RBB data series $X^{*(n)} = (X_1, ..., X_{n^*})$ constructed from the $l_n^* - 1$ bootstrap data blocks, and

$$\sigma^{*2}_{U,\ l_n^*-1} = (l_n^*-2)(l_n^*-3)^{-2} \sum_{k=1}^{l_n^*-1} [(l_n^*-2)^{-1} \sum_{l=1, k \neq l}^{l_n^*-1} \omega_U(\mathcal{B}_k^*, \mathcal{B}_l^*) - U^*_{l_n^*-1}]^2, \tag{1.28}$$

$$U^*_{l_n^*-1} = \frac{2}{(l_n^*-1)(l_n^*-2)} \sum_{1 \leqslant k < l \leqslant l_n^*-1} \omega_U(\mathcal{B}_k^*, \mathcal{B}_l^*).$$

We may then state the following result.

Theorem 9. *If conditions (i)-(iii) are fulfilled with* $s = 4$, *then we have the CLT under* $\mathbb{P}_\nu$

$$\sqrt{n}(T_n - \mu(U))/\Sigma_n \Rightarrow \mathcal{N}(0,1),\ as\ n \to \infty.$$

This limit result also holds for $\widetilde{T}_n$, *as well as the asymptotic validity of the RBB distribution: as* $n \to \infty$,

$$\sup_{x \in \mathbb{R}} |\mathbb{P}^*(\sqrt{n^*}(T_n^* - \widetilde{T}_n))/\Sigma_n^* \leq x) - \mathbb{P}_\nu(\sqrt{n}(\widetilde{T}_n - \mu(U))/\Sigma_n \leq x)| \overset{\mathbb{P}_\nu}{\to} 0.$$

Whereas proving the asymptotic validity of the RBB for U-statistics under these assumptions is straightforward (its second order accuracy up to $o(n^{-1/2})$ seems also quite easy to prove by simply adapting the argument used by [46] under appropriate Cramer condition on $\omega_U^{(1)}(\mathcal{B}_1)$ and block-moment assumptions), establishing an exact rate, $O(n^{-1})$ for instance as in the case of sample mean statistics, is much more difficult. Even if one try to reproduce the argument in [8] consisting in partitioning the underlying probability space according to every possible realization of the regeneration times sequence between 0 and n, the problem boils down to control the asymptotic behaviour of the distribution $\mathbb{P}(\sum_{1 \leqslant i \neq j \leqslant m} \omega_U^{(2)}(\mathcal{B}_i, \mathcal{B}_j)/\sigma^2_{U,\ m} \leqslant y,\ \sum_{j=1}^m l(\mathcal{B}_j) = l)$ as $m \to \infty$, which is a highly difficult technical task (due to the need of a simultaneous control of the lattice component and of the degenerate part of the U-statistics).

Remark 7. We point out that the approach developed here to deal with the statistic U_L naturally applies to more general functionals of the regeneration blocks $\sum_{k<l} \omega(\mathcal{B}_k, \mathcal{B}_l)$, with $\omega : \mathbb{T}^2 \to \mathbb{R}$ being some measurable function. For instance, the estimator of the asymptotic variance $\widehat{\sigma}_n^2(f)$ proposed in § 4.1 could be derived from such a functional, that may be seen as a U-statistic based on observation blocks with kernel $\omega(\mathcal{B}_k, \mathcal{B}_l) = (f(\mathcal{B}_k) - f(\mathcal{B}_l))^2/2$.

1.6.2 General case

Suppose now that the observed trajectory $X^{(n+1)} = (X_1, ..., X_{n+1})$ is drawn from a general Harris positive chain with stationary probability μ (see § 2.2 for assumptions and notation). Using the split chain, we have the representation of the parameter $\mu(U)$:

$$\mu(U) = \mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-2}\mathbb{E}_{A_{\mathcal{M}}}(\omega_U(\mathcal{B}_1, \mathcal{B}_2)).$$

Using the pseudo-blocks $\widehat{\mathcal{B}}_l$, $1 \leqslant l \leqslant \widehat{l}_n - 1$, as constructed in § 3.2, we consider the sequence of renormalizing constants for T_n :

$$\widehat{\Sigma}_n^2 = 4(\widehat{l}_n/n)^3\widehat{\sigma}^2_{U,\,\widehat{l}_n-1}, \tag{1.29}$$

with

$$\widehat{\sigma}^2_{U,\,\widehat{l}_n-1} = (\widehat{l}_n-2)(\widehat{l}_n-3)^{-2}\sum_{k=1}^{\widehat{l}_n-1}[(\widehat{l}_n-2)^{-1}\sum_{l=1,k\neq l}^{\widehat{l}_n-1}\omega_U(\widehat{\mathcal{B}}_k,\widehat{\mathcal{B}}_l)-\widehat{U}_{\widehat{l}_n-1}]^2,$$

$$\widehat{U}_{\widehat{l}_n-1} = \frac{2}{(\widehat{l}_n-1)(\widehat{l}_n-2)}\sum_{1\leqslant k<l\leqslant \widehat{l}_n-1}\omega_U(\widehat{\mathcal{B}}_k,\widehat{\mathcal{B}}_l).$$

We also introduce the U-statistic computed from the first approximate regeneration time and the last one:

$$\widehat{T}_n = \frac{2}{\widehat{n}(\widehat{n}-1)}\sum_{1+\widehat{\tau}_A(1)\leqslant i<j\leqslant\widehat{\tau}_A(l_n)}U(X_i,X_j),$$

with $\widehat{n} = \widehat{\tau}_A(\widehat{l}_n) - \widehat{\tau}_A(1)$. Let us define the bootstrap counterparts T_n^* and Σ_n^* of $\widehat{T}_n$ and $\widehat{\Sigma}_n^2$ constructed from the pseudo-blocks via *Algorithm 5*. Although approximate blocks are used here instead of the (unknown) regenerative ones $\mathcal{B}_l$, $1 \leqslant l \leqslant l_n - 1$, asymptotic normality still holds under appropriate assumptions, as shown by the theorem below, which we state in the only case when the kernel U is bounded (with the aim to make the proof simpler).

Theorem 10. *Suppose that the kernel $U(x,\,y)$ is bounded and that $\mathcal{H}_2$, $\mathcal{H}_3$, $\mathcal{H}_4$ are fulfilled, as well as (i)-(iii) for $s = 4$. Then we have as $n \to \infty$,*

$$\widehat{\Sigma}_n^2 \to \Sigma^2 = 4\mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-3}\mathbb{E}_{A_{\mathcal{M}}}(\omega_U^{(1)}(\mathcal{B}_1)^2),\ \textit{in } \mathbb{P}_\nu\textit{-pr.}$$

Moreover as $n \to \infty$, under $\mathbb{P}_\nu$ we have the convergence in distribution

$$n^{1/2}\widehat{\Sigma}_n^{-1}(T_n - \mu(U)) \Rightarrow \mathcal{N}(0,1),$$

as well as the asymptotic validity of the ARBB counterpart

$$\sup_{x\in\mathbb{R}}|\mathbb{P}^*(\sqrt{n^*}(T_n^* - \widehat{T}_n))/\Sigma_n^* \leq x) - \mathbb{P}_\nu(\sqrt{n}(\widehat{T}_n - \mu(U))/\widehat{\Sigma}_n \leq x)| \underset{n\to\infty}{\overset{\mathbb{P}_\nu}{\to}} 0.$$

Proof. By applying the results of § 6.1 to the split chain, we get that the variance of the limiting (normal) distribution of $\sqrt{n}(T_n - \mu(U))$ is $\Sigma^2 = 4\mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-3}\mathbb{E}_{A_{\mathcal{M}}}(\omega_U^{(1)}(\mathcal{B}_1)^2)$. The key point of the proof consists in considering an appropriate coupling between $(X_i,\ Y_i)_{1\leqslant i\leqslant n}$ and $(X_i,\ \widehat{Y}_i)_{1\leqslant i\leqslant n}$ (or equivalently between the sequence of the "true" regeneration times between 0 and n and the sequence of approximate ones), so as to control the deviation between functionals constructed from the regeneration blocks and their counterparts based on the approximate ones. The coupling considered here is the same as the one used in the proof of Theorem 3.1 in [10] (refer to the latter article for a detailed construction). We shall now evaluate how $\widehat{\sigma}^2_{U,\ \widehat{l}_n-1}$ differs from $\sigma^2_{U,\ l_n-1}$, its counterpart based on the "true" regeneration blocks. Observe first that

$$T_n = \frac{\widehat{n}(\widehat{n}-1)}{n(n-1)}\widehat{T}_n + \widehat{T}_n^{(0)} + \widehat{T}_n^{(n)} + \widehat{\Delta}_n,$$

where

$$\widehat{T}_n^{(0)} = \frac{2}{n(n-1)}\sum_{1\leqslant k\leqslant \widehat{l}_n-1}\omega_U(\widehat{\mathcal{B}}_k,\widehat{\mathcal{B}}_0),\ \widehat{T}_n^{(n)} = \frac{2}{n(n-1)}\sum_{0\leqslant k\leqslant l_n-1}\omega_U(\widehat{\mathcal{B}}_k,\widehat{\mathcal{B}}^{(n)}_{\widehat{l}_n}),$$

$$\widehat{\Delta}_n = \frac{1}{n(n-1)}\{\sum_{k=0}^{\widehat{l}_n-1}\omega_U(\widehat{\mathcal{B}}_k,\widehat{\mathcal{B}}_k) + \omega_U(\widehat{\mathcal{B}}^{(n)}_{\widehat{l}_n},\widehat{\mathcal{B}}^{(n)}_{\widehat{l}_n}) - \sum_{i=1}^{n}U(X_i,X_i)\}.$$

Now following line by line the proof of lemma 5.2 in [10], we obtain that, as $n\to\infty$, $\widehat{n}/n-1 = O_{\mathbb{P}_\nu}(1)$, $\widehat{\Delta}_n - \Delta_n$, $\widehat{T}_n^{(0)} - \widehat{T}_n^{(0)}$ and $\widehat{T}_n^{(n)} - \widehat{T}_n^{(n)}$ are $O_{\mathbb{P}_\nu}(n^{-1})$. It follows thus that $\widehat{T}_n = T_n + o_{\mathbb{P}_\nu}(n^{-1/2})$ as $n\to\infty$, and $\sqrt{n}(\widehat{T}_n - \mu(U))$ is asymptotically normal with variance Σ^2. The same limit results is straightforwardly available then for the Bootstrap version by standard regenerative arguments. Furthermore, by Lemma 5.3 in [10] we have $\mid \widehat{l}_n/n - l_n/n \mid = O_{\mathbb{P}_\nu}(\alpha_n^{1/2})$ as $n\to\infty$, and thus $\widehat{l}_n/n \to \mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-1}$ in $\mathbb{P}_\nu$-*pr.* as $n\to\infty$. It then follows by simple (especially when U is bounded) but tedious calculations that $\widehat{\Sigma}_n^2 - \Sigma_n^2 = D_n + o_{\mathbb{P}_\nu}(1)$ as $n\to\infty$, with

$$\begin{aligned}D_n = 4(l_n/n)^3[\widehat{l}_n^{-1}&\sum_{i=1}^{\widehat{l}_n-1}\{\frac{1}{\widehat{l}_n-2}\sum_{j=1,j\neq i}^{\widehat{l}_n-1}\omega_U(\widehat{\mathcal{B}}_i,\widehat{\mathcal{B}}_j)\}^2\\ &- l_n^{-1}\sum_{i=1}^{l_n-1}\{\frac{1}{l_n-2}\sum_{j=1,j\neq i}^{l_n-1}\omega_U(\mathcal{B}_i,\mathcal{B}_j)\}^2].\end{aligned}$$

Now set $\widehat{g}_n(\widehat{\mathcal{B}}_i) = (\widehat{l}_n-2)^{-1}\sum_{j=1,j\neq i}^{\widehat{l}_n-1}\omega_U(\widehat{\mathcal{B}}_i,\widehat{\mathcal{B}}_j)$ for $i\in\{1,...,\widehat{l}_n-1\}$ and $g_n(\mathcal{B}_i) = (l_n-2)^{-1}\sum_{j=1,j\neq i}^{l_n-1}\omega_U(\mathcal{B}_i,\mathcal{B}_j)$ for $i\in\{1,...,\widehat{l}_n-1\}$. By standard arguments on U-statistics (see for instance [46] and the references therein) and using once again lemma 5.1 and 5.2 in [9], we have uniformly in $i\in$

$\{1,...,\widehat{l}_n-1\}$ (resp. in $i\in\{1,...,\widehat{l}_n-1\}$), $\widehat{g}_n(\widehat{\mathcal{B}}_i)=\omega_U^{(1)}(\widehat{\mathcal{B}}_i)+o_{\mathbb{P}_\nu}(1)$ (resp. $g_n(\mathcal{B}_i)=\omega_U^{(1)}(\mathcal{B}_i)+o_{\mathbb{P}_\nu}(1)$) as $n\to\infty$. Such uniform bounds are facilitated by the boundedness assumption on U, but one may expect that with refined computations the same results could be established for unbounded kernels.

It follows that as $n\to\infty$,

$$\varDelta_n=4(l_n/n)^3[\widehat{l}_n^{-1}\sum_{i=1}^{\widehat{l}_n-1}\{\omega_U^{(1)}(\widehat{\mathcal{B}}_i)\}^2-l_n^{-1}\sum_{i=1}^{l_n-1}\{\omega_U^{(1)}(\mathcal{B}_i)\}^2]+o_{\mathbb{P}_\nu}(1).$$

The first term in the right hand side is also $o_{\mathbb{P}_\nu}(1)$ by lemma 5.2 in [10]. The proof of the asymptotic validity of the Bootstrap version is established by following the preceding lines: it may be easily checked by first reducing the problem to a sum and following the proof of Theorem 3.3 in [10]. As in the i.i.d case, this asymptotic result essentially boils down then to check that the empirical moments converge to the theoretical ones. This can be done by adapting standard SLLN arguments for U-statistics.

1.7 Robust functional parameter estimation

Extending the notion of *influence function* and/or *robustness* to the framework of general time series is a difficult task (see [51] or [59]). Such concepts are important not only to detect *"outliers"* among the data or influential observations but also to generalize the important notion of *efficient estimation* in semiparametric frameworks (see the recent discussion in [13] for instance). In the markovian setting, a recent proposal based on martingale approximation has been made by [61]. Here we propose an alternative definition of the influence function based on the (approximate) regeneration blocks construction, which is easier to manipulate and immediately leads to central limit and convolution theorems.

1.7.1 Defining the influence function on the torus

The leitmotiv of this paper is that most parameters of interest related to Harris chains are functionals of the distribution $\mathcal{L}$ of the regenerative blocks (observe that $\mathcal{L}$ is a distribution on the torus $\mathbb{T}=\cup_{n\geqslant1}E^n$), namely the distribution of $(X_1,....,X_{\tau_A})$ conditioned on $X_0\in A$ when the chain possesses an atom A, or the distribution of $(X_1,....,X_{\tau_{A_\mathcal{M}}})$ conditioned on $(X_0,Y_0)\in A_\mathcal{M}$ in the general case when one considers the split chain (refer to section 2 for assumptions and notation, here we shall omit the subscript A and $\mathcal{M}$ in what follows to make the notation simpler). In view of Theorem 2.1, this is obviously true in the positive recurrent case for any functional of the stationary law μ. But, more generally, the probability distribution $\mathbb{P}_\nu$ of the Markov chain X starting from ν may be decomposed as follows: $\mathbb{P}_\nu((X_n)_{n\geqslant1})=\mathcal{L}_\nu((X_1,....,X_{\tau_{A(1)}}))\prod_{k=1}^\infty\mathcal{L}((X_{1+\tau_A(k)},....,X_{\tau_A(k+1)}))$, denoting by $\mathcal{L}_\nu$ the distribution of $(X_1, ..., X_{\tau_A})$ conditioned on $X_0 \sim \nu$. Thus any functional of the law of $(X_n)_{n \geqslant 1}$ may be seen as a functional of $(\mathcal{L}_\nu, \mathcal{L})$. However, pointing out that the distribution of $\mathcal{L}_\nu$ cannot be estimated in most cases encountered in practice, only functionals of $\mathcal{L}$ are of practical interest. The object of this subsection is to propose the following definition of the influence function for such functionals. Let $\mathcal{P}_{\mathbb{T}}$ denote the set of all probability measures on the torus $\mathbb{T}$ and for any $b \in \mathbb{T}$, set $l(b) = k$ if $b \in E^k$, $k \geqslant 1$. We then have the following natural definition, that straightforwardly extends the classical notion of influence function in the i.i.d. case, with the important novelty that distributions on the torus are considered here.

Definition 1. *Let $T : \mathcal{P}_{\mathbb{T}} \to \mathbb{R}$ be a functional on $\mathcal{P}_{\mathbb{T}}$. If for $\mathcal{L}$ in $\mathcal{P}_{\mathbb{T}}$, $t^{-1}(T((1-t)\mathcal{L} + t\delta_b) - T(\mathcal{L}))$ has a finite limit as $t \to 0$ for any $b \in \mathbb{T}$, then the influence function $T^{(1)}$ of the functional T is well defined, and by definition one has for all b in $\mathbb{T}$,*

$$T^{(1)}(b,\ \mathcal{L}) = \lim_{t \to 0} \frac{T((1-t)\mathcal{L} + t\delta_b) - T(\mathcal{L})}{t}. \tag{1.30}$$

1.7.2 Some examples

The relevance of this definition is illustrated through the following examples, which aim to show how easy it is to adapt known calculations of influence function on $\mathbb{R}$ to this framework.

a) Suppose that X is positive recurrent with stationary distribution μ. Let $f : E \to \mathbb{R}$ be μ-integrable and consider the parameter $\mu_0(f) = \mathbb{E}_\mu(f(X))$. Denote by $\mathcal{B}$ a r.v. valued in $\mathbb{T}$ with distribution $\mathcal{L}$ and observe that $\mu_0(f) = \mathbb{E}_{\mathcal{L}}\left(f(\mathcal{B})\right)/\mathbb{E}_{\mathcal{L}}\left(l(\mathcal{B})\right) = T(\mathcal{L})$ (recall the notation $f(b) = \sum_{i=1}^{l(b)} f(b_i)$ for any $b \in \mathbb{T}$). A classical calculation for the influence function of ratios yields then

$$T^{(1)}(b, \mathcal{L}) = \frac{d}{dt}(T((1-t)\mathcal{L} + tb)|_{t=0} = \frac{f(b) - \mu(f)l(b)}{\mathbb{E}_{\mathcal{L}}\left(l(\mathcal{B})\right)}$$

Notice that $\mathbb{E}_{\mathcal{L}}(T^{(1)}(\mathcal{B}, \mathcal{L})) = 0$.

b) Let θ be the unique solution of the equation: $\mathbb{E}_\mu(\psi(X,\ \theta)) = 0$, where $\psi :\ \mathbb{R}^2 \to \mathbb{R}$ is $\mathcal{C}^2$. Observing that it may be rewritten as $\mathbb{E}_{\mathcal{L}}(\psi(\mathcal{B},\ \theta)) = 0$, a similar calculation to the one used in the i.i.d. setting (if differentiating inside the expectation is authorized) gives in this case

$$T_\psi^{(1)}(b, \mathcal{L}) = -\frac{\psi(b,\ \theta)}{\mathbb{E}_A(\sum_{i=1}^{\tau_A} \frac{\partial \psi(X_i, \theta)}{\partial \theta})}.$$

By definition of θ, we naturally have $\mathbb{E}_{\mathcal{L}}(T_\psi^{(1)}(B, \mathcal{L})) = 0$.

c) Assuming that the chain takes real values and its stationary law μ has zero mean and finite variance, let ρ be the correlation coefficient between consecutive observations under the stationary distribution:

$$\rho = \frac{\mathbb{E}_\mu(X_n X_{n+1})}{\mathbb{E}_\mu(X_n^2)} = \frac{\mathbb{E}_A(\sum_{n=1}^{\tau_A} X_n X_{n+1})}{\mathbb{E}_A(\sum_{n=1}^{\tau_A} X_n^2)}.$$

For all b in $\mathbb{T}$, the influence function is

$$T_\rho^{(1)}(b, \mathcal{L}) = \frac{\sum_{i=1}^{l(b)} b_i(b_{i+1} - \rho b_i)}{\mathbb{E}_A(\sum_{t=1}^{\tau_A} X_t^2)},$$

and one may check that $\mathbb{E}_\mathcal{L}(T_\rho^{(1)}(\mathcal{B}, \mathcal{L})) = 0$.

d) It is now possible to reinterpret the results obtained for U-statistics in section 6. With the notation above, the parameter of interest may be rewritten

$$\mu(U) = \mathbb{E}_\mathcal{L}\,(l(\mathcal{B}))^{-2}\mathbb{E}_{\mathcal{L}\times\mathcal{L}}(U(\mathcal{B}_1, \mathcal{B}_2)),$$

yielding the influence function: $\forall b \in \mathbb{T}$,

$$\mu^{(1)}(b, \mathcal{L}) = 2\mathbb{E}_\mathcal{L}\,(l(\mathcal{B}))^{-2}\mathbb{E}_\mathcal{L}(\widetilde{\omega}_U(\mathcal{B}_1, \mathcal{B}_2)|\mathcal{B}_1 = b).$$

1.7.3 Main results

In order to lighten the notation, the study is restricted to the case when X takes real values, *i.e.* $E \subset \mathbb{R}$, but straightforwardly extends to a more general framework. Given an observed trajectory of length n, natural empirical estimates of parameters $T(\mathcal{L})$ are of course the *plug-in estimators* $T(\mathcal{L}_n)$ based on the empirical distribution of the observed regeneration blocks $\mathcal{L}_n = (l_n - 1)^{-1} \sum_{j=1}^{l_n - 1} \delta_{\mathcal{B}_j} \in \mathcal{P}_\mathbb{T}$ in the atomic case, which is defined as soon as $l_n \geqslant 2$ (notice that $\mathbb{P}_\nu(l_n \leqslant 1) = O(n^{-1})$ as $n \to \infty$, if $\mathcal{H}_0(1, \nu)$ and $\mathcal{H}_0(2)$ are satisfied). For measuring the closeness between $\mathcal{L}_n$ and $\mathcal{L}$, consider the bounded Lipschitz type metric on $\mathcal{P}_\mathbb{T}$

$$d_{BL}(\mathcal{L}, \mathcal{L}') = \sup_{f \in Lip_\mathbb{T}^1} \{\int f(b)\mathcal{L}(db) - \int f(b)\mathcal{L}'(db)\}, \tag{1.31}$$

for any $\mathcal{L}$, $\mathcal{L}'$ in $\mathcal{P}_\mathbb{T}$, denoting by $Lip_\mathbb{T}^1$ the set of functions $F : \mathbb{T} \to \mathbb{R}$ of type $F(b) = \sum_{i=1}^{l(b)} f(b_i)$, $b \in \mathbb{T}$, where $f : E \to \mathbb{R}$ is such that $\sup_{x\in E} |f(x)| \leqslant 1$ and is 1-Lipschitz. Other metrics (of Zolotarev type for instance, *cf* [68]) may be considered. In the general Harris case (refer to § 3.2 for notation), the influence function based on the atom of the split chain, as well as the empirical distribution of the (unobserved) regeneration blocks have to be approximated to be of practical interest. Once again, we shall use the approximate regeneration blocks $\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n - 1}$ (using *Algorithm 2, 3*) in the general case and consider

$$\widehat{\mathcal{L}}_n = (\widehat{l}_n - 1) \sum_{j=1}^{\widehat{l}_n - 1} \delta_{\widehat{\mathcal{B}}_j},$$

when $\widehat{l}_n \geqslant 2$. The following theorem provides an asymptotic bound for the error committed by replacing the empirical distribution $\mathcal{L}_n$ of the "true" regeneration blocks by $\widehat{\mathcal{L}}_n$, when measured by d_{BL}.

Theorem 11. *Under $\mathcal{H}_0^{'}(4), \mathcal{H}_0^{'}(4,\nu), \mathcal{H}_2,\ \mathcal{H}_3$ and $\mathcal{H}_4$, we have*

$$d_{BL}(\mathcal{L}_n, \widehat{\mathcal{L}}_n) = O(\alpha_n^{1/2}),\ as\ n \to \infty.$$

And if in addition $d_{BL}(\mathcal{L}_n, \mathcal{L}) = O(n^{-1/2})$ as $n \to \infty$, then

$$d_{BL}(\mathcal{L}_n, \widehat{\mathcal{L}}_n) = O(\alpha_n^{1/2} n^{-1/2}),\ as\ n \to \infty.$$

Proof. With no loss of generality, we assume the X_i's centered. From lemma 5.3 in [10], we have $l_n/\widehat{l}_n - 1 = O_{\mathbb{P}_\nu}(\alpha_n^{1/2})$ as $n \to \infty$. Besides, writing

$$d_{BL}(\mathcal{L}_n, \widehat{\mathcal{L}}_n) \leq (\frac{l_n - 1}{\widehat{l}_n - 1} - 1) \sup_{f \in Lip_{\mathbb{T}}^1} |\frac{1}{l_n - 1} \sum_{j=1}^{l_n - 1} f(\mathcal{B}_j)|$$
$$+ \frac{n}{\widehat{l}_n - 1} \sup_{f \in Lip_{\mathbb{T}}^1} |n^{-1} \sum_{j=1}^{l_n - 1} f(\mathcal{B}_j) - n^{-1} \sum_{j=1}^{\widehat{l}_n - 1} f(\widehat{\mathcal{B}}_j)|, \tag{1.32}$$

and observing that $\sup_{f \in Lip_{\mathbb{T}}^1} |(l_n - 1)^{-1} \sum_{j=1}^{l_n - 1} f(\mathcal{B}_j)| \leqslant 1$, we get that the first term in the right hand side is $O_{\mathbb{P}_\nu}(\alpha_n^{1/2})$ as $n \to \infty$. Now as $\sup_{x \in E} |f(x)| \leqslant 1$, we have

$$|n^{-1}(\sum_{j=1}^{l_n} f(\mathcal{B}_j) - \sum_{j=1}^{\widehat{l}_n} f(\widehat{\mathcal{B}}_j))| \leq n^{-1}(|\widehat{\tau}_{A_\mathcal{M}}(1) - \tau_{A_\mathcal{M}}(1)| + |\widehat{\tau}_{A_\mathcal{M}}(l_n) - \widehat{\tau}_{A_\mathcal{M}}(l_n)|),$$

and from lemma 5.1 in by [9], the term in the right hand side is $o_{\mathbb{P}_\nu}(n^{-1})$ as $n \to \infty$. We thus get

$$d_{BL}(\mathcal{L}_n, \widehat{\mathcal{L}}_n) \leq \alpha_n^{1/2} d_{BL}(\mathcal{L}_n, \mathcal{L}) + o_{\mathbb{P}_\nu}(n^{-1}), \text{ as } n \to \infty.$$

And this completes the proof.

Given the metric on $\mathcal{P}_{\mathbb{T}}$ defined by d_{BL}, we consider now the *Fréchet differentiability* for functionals $T : \mathcal{P}_{\mathbb{T}} \to \mathbb{R}$.

Definition 2. *We say that T is Fréchet-differentiable at $\mathcal{L}_0 \in \mathcal{P}_{\mathbb{T}}$, if there exists a linear operator $DT_{\mathcal{L}_0}^{(1)}$ and a function $\epsilon^{(1)}(., \mathcal{L}_0)$: $\mathbb{R} \to \mathbb{R}$, continuous at 0 with $\epsilon^{(1)}(0, \mathcal{L}_0) = 0$, such that:*

$$\forall \mathcal{L} \in \mathcal{P}_{\mathbb{T}},\ T(\mathcal{L}) - T(\mathcal{L}_0) = D^{(1)} T_{\mathcal{L}_0}(\mathcal{L} - \mathcal{L}_0) + R^{(1)}(\mathcal{L}, \mathcal{L}_0),$$

with $R^{(1)}(\mathcal{L}, \mathcal{L}_0) = d_{BL}(\mathcal{L}, \mathcal{L}_0)\epsilon^{(1)}(d_{BL}(\mathcal{L}, \mathcal{L}_0), \mathcal{L}_0)$. Moreover, T is said to have a canonical gradient (or influence function) $T^{(1)}(., \mathcal{L}_0)$, if one has the following representation for $DT_{\mathcal{L}_0}^{(1)}$:

$$\forall \mathcal{L} \in \mathcal{P}_{\mathbb{T}},\ DT_{\mathcal{L}_0}^{(1)}(\mathcal{L} - \mathcal{L}_0) = \int_{\mathbb{T}} T^{(1)}(b, \mathcal{L}_0)\mathcal{L}(db).$$

Now it is easy to see that from this notion of differentiability on the torus one may directly derive CLT's, provided the distance $d(\mathcal{L}_n, \mathcal{L})$ may be controlled.

Theorem 12. *In the regenerative case, if* $T : \mathcal{P}_{\mathbb{T}} \to \mathbb{R}$ *is Fréchet differentiable at* $\mathcal{L}$ *and* $d_{BL}(\mathcal{L}_n, \mathcal{L}) = O_{\mathbb{P}_\nu}(n^{-1/2})$ *(or* $R^{(1)}(\mathcal{L}_n, \mathcal{L}) = o_{\mathbb{P}_\nu}(n^{-1/2})$*) as* $n \to \infty$*, and if* $\mathbb{E}_A(\tau_A) < \infty$ *and* $0 < Var_A(T^{(1)}(\mathcal{B}_1, \mathcal{L})) < \infty$ *then under* $\mathbb{P}_\nu$*,*

$$n^{1/2}(T(\mathcal{L}_n) - T(\mathcal{L})) \Rightarrow \mathcal{N}(0, \mathbb{E}_A(\tau_A) Var_A(T^{(1)}(\mathcal{B}_1, \mathcal{L})), \text{ as } n \to \infty.$$

In the general Harris case, if the split chain satisfies the assumptions above (with A *replaced by* $A_{\mathcal{M}}$*), under the assumptions of Theorem 11, as* $n \to \infty$ *we have under* $\mathbb{P}_\nu$*,*

$$n^{1/2}(T(\widehat{\mathcal{L}}_n) - T(\mathcal{L})) \Rightarrow \mathcal{N}(0, \mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}}) Var_{A_{\mathcal{M}}}(T^{(1)}(\mathcal{B}_1, \mathcal{L})).$$

The proof is straightforward and left to the reader. Observe that if one renormalizes by $l_n^{1/2}$ instead of renormalizing by $n^{1/2}$ in the atomic case (resp., by $\widehat{l}_n^{1/2}$ in the general case), the asymptotic distribution would be simply $\mathcal{N}(0, Var_A(T^{(1)}(\mathcal{B}_1, \mathcal{L}))$ (resp., $Var_{A_{\mathcal{M}}}(T^{(1)}(\mathcal{B}_1, \mathcal{L}))$, which depends on the atom chosen (resp. on the parameters of condition $\mathcal{M}$).

Then going back to the preceding examples, we straightforwardly deduce the following results.

a) Noticing that $n^{1/2}/l_n^{1/2} \to \mathbb{E}_A(\tau_A)^{1/2}$ $\mathbb{P}_\nu$- a.s. as $n \to \infty$, we immediately get that under $\mathbb{P}_\nu$, as $n \to \infty$,

$$n^{1/2}(\mu_n(f) - \mu(f)) \Rightarrow \mathcal{N}(0, \mathbb{E}_A(\tau_A)^{-1} Var_A(\sum_{i=1}^{\tau_A}(f(X_i) - \mu(f)).$$

b) In a similar fashion, under smoothness assumptions ensuring Fréchet differentiability, the M-estimator $\widehat{\theta}_n$ being the (unique) solution of the block-estimating equation

$$\sum_{i=\tau_A+1}^{\tau_A(l_n)} \psi(X_i, \theta) = \sum_{j=1}^{l_n} \sum_{i=\tau_A(j)+1}^{\tau_A(j+1)} \psi(X_i, \theta) = 0,$$

we formally obtain that, if $\mathbb{E}_A(\sum_{i=1}^{\tau_A} \frac{\partial \psi(X_i,\theta)}{\partial \theta}) \neq 0$ and θ is the true value of the parameter, then under $\mathbb{P}_\nu$, as $n \to \infty$,

$$n^{1/2}(\widehat{\theta}_n - \theta) \Rightarrow \mathcal{N}(0,\ [\frac{\mathbb{E}_A(\sum_{i=1}^{\tau_A} \frac{\partial \psi(X_i,\theta)}{\partial \theta})}{\mathbb{E}_A(\tau_A)}]^{-2} \frac{Var_A(\sum_{i=1}^{\tau_A} \psi(X_i, \theta))}{\mathbb{E}_A(\tau_A)}).$$

Observe that both factors in the variance are independent from the atom A chosen. It is worth noticing that, by writing the asymptotic variance in this way, as a function of the distribution of the blocks, a consistent estimator for the latter is readily available, from the (approximate) regeneration blocks. Examples c) and d) may be treated similarly.

Remark 8. The concepts developed here may also serve as a tool for robustness purpose, for deciding whether a specific data block has an important influence on the value of some given estimate or not, and/or whether it may be considered as "outlier". The concept of robustness we introduce is related to blocks of observations, instead of individual observations. Heuristically, one may consider that, given the regenerative dependency structure of the process, a single suspiciously outlying value at some time point n may have a strong impact on the trajectory, until the (split) chain regenerates again, so that not only this particular observation but the whole "contaminated" segment of observations should be eventually removed. Roughly stated, it turns out that examining (approximate) regeneration blocks as we propose before, allows to identify more accurately outlying data in the sample path, as well as their nature (in the time series context, different type of outliers may occur, such as additive or innovative outliers). By comparing the data blocks (their length, as well as the values of the functional of interest on these blocks) this way, one may detect the ones to remove eventually from further computations.

1.8 Some extreme values statistics

We now turn to statistics related to the extremal behaviour of functionals of type $f(X_n)$ in the atomic positive Harris recurrent case, where $f:(E,\mathcal{E})\to\mathbb{R}$ is a given measurable function. More precisely, we shall focus on the limiting distribution of the maximum $M_n(f)=\max_{1\leqslant i\leqslant n} f(X_i)$ over a trajectory of length n, in the case when the chain X possesses an accessible atom A (see [3] and the references therein for various examples of such processes X in the area of queuing systems and a theoretical study of the tail properties of $M_n(f)$ in this setting).

Submaxima over regeneration blocks

For $j\geqslant 1$, we define the "submaximum" over the j-th cycle of the sample path:

$$\zeta_j(f)=\max_{1+\tau_A(j)\leqslant i\leqslant \tau_A(j+1)} f(X_i).$$

The $\zeta_j(f)$'s are i.i.d. r.v.'s with common d.f. $G_f(x)=\mathbb{P}(\zeta_1(f)\leqslant x)$. The following result established by [75] shows that the limiting distribution of the sample maximum of $f(X)$ is entirely determined by the tail behaviour of the df G_f and relies on the crucial observation that the maximum value $M_n(f)=\max_{1\leqslant i\leqslant n} f(X_i)$ over a trajectory of length n, may be expressed in terms of "submaxima" over regeneration blocks as follows

$$M_n(f)=\max(\zeta_0(f),\ \max_{1\leqslant j\leqslant l_n-1}\zeta_j(f),\ \zeta_{l_n}^{(n)}(f)),$$

where $\zeta_0(f)=\max_{1\leqslant i\leqslant \tau_A} f(X_i)$ and $\zeta_{l_n}^{(n)}(f)=\max_{1+\tau_A(l_n)\leqslant i\leqslant n} f(X_i)$ denote the maxima over the non regenerative data blocks, and with the usual convention that the maximum over an empty set equals $-\infty$.

Proposition 4. *(see [75]). Let $\alpha = \mathbb{E}_A(\tau_A)$ be the mean return time to the atom A. Under the assumption (A1) that the first (non-regenerative) block does not affect the extremal behaviour, i.e. $\mathbb{P}_\nu(\zeta_0(f) > \max_{1\leqslant k\leqslant l}\zeta_k(f)) \to 0$ as $l \to \infty$, we have*

$$\sup_{x\in\mathbb{R}} \mid \mathbb{P}_\nu(M_n(f) \leqslant x) - G_f(x)^{n/\alpha} \mid \to 0, \text{ as } n \to \infty. \tag{1.33}$$

Hence, as soon as condition (A1) is fulfilled, the asymptotic behaviour of the sample maximum may be deduced from the tail properties of G_f. In particular, the limiting distribution of $M_n(f)$ (for a suitable normalization) is the extreme df $H_\xi(x)$ of shape parameter $\xi \in \mathbb{R}$ (with $H_\xi(x) = \exp(-x^{-1/\xi})\mathbb{I}\{x > 0\}$ when $\xi > 0$, $H_0(x) = \exp(-\exp(-x))$ and $H_\xi(x) = \exp(-(-x)^{-1/\xi})\mathbb{I}\{x < 0\}$ if $\xi < 0$) iff G_f belongs to the maximum domain of attraction $MDA(H_\xi)$ of the latter df (refer to [69] for basics in extreme value theory). Thus, when $G_f \in MDA(H_\xi)$, there are sequences of norming constants a_n and b_n such that $G_f(a_n x + b_n)^n \to H_\xi(x)$ as $n \to \infty$, we then have $\mathbb{P}_\nu(M_n(f) \leqslant a'_n x + b_n) \to H_\xi(x)$ as $n \to \infty$, with $a'_n = a_n/\alpha^\xi$.

Tail estimation based on submaxima over regeneration blocks

In the case when assumption (A1) holds, one may straightforwardly derive from (1.33) estimates of $H_{f,\,n}(x) = \mathbb{P}_\nu(M_n(f) \leqslant x)$ as $n \to \infty$ based on the observation of a random number of submaxima $\zeta_j(f)$ over a sample path, as proposed in [36]:

$$\widehat{H}_{f,\,n,\,l}(x) = (\widehat{G}_{f,\,n}(x))^l,$$

with $1 \leqslant l \leqslant l_n$ and denoting by $\widehat{G}_{f,\,n}(x) = \frac{1}{l_n-1}\sum_{i=1}^{l_n-1}\mathbb{I}\{\zeta_j(f) \leqslant x\}$ the empirical df of the $\zeta_j(f)$'s (with $\widehat{G}_{f,\,n}(x) = 0$ by convention when $l_n \leqslant 1$). We have the following limit result (see also Proposition 3.6 in [36] for a different formulation, stipulating the observation of a deterministic number of regeneration cycles).

Proposition 5. *Let (u_n) be such that $n(1 - G_f(u_n))/\alpha \to \eta < \infty$ as $n \to \infty$. Suppose that assumptions $\mathcal{H}_0(1,\nu)$ and (A1) holds, then $H_{f,\,n}(u_n) \to \exp(-\eta)$ as $\eta \to \infty$. And let $N_n \in \mathbb{N}$ such that $N_n/n^2 \to 0$ as $n \to \infty$, then we have*

$$\widehat{H}_{f,\,N_n,\,l_n}(u_n)/H_{f,\,n}(u_n) \to 1 \textit{ in } \mathbb{P}_\nu\textit{- probability, as } n \to \infty. \tag{1.34}$$

Moreover if $N_n/n^{2+\rho} \to \infty$ as $n \to \infty$ for some $\rho > 0$, this limit result also holds $\mathbb{P}_\nu$- a.s. .

Proof. First, the convergence $H_{f,\,n}(u_n) \to \exp(-\eta)$ as $\eta \to \infty$ straightforwardly follows from Proposition 8.1. Now we shall show that $l_n(1 - \widehat{G}_{f,\,N_n}(u_n)) \to \eta$ in $\mathbb{P}_\nu$- *pr.* as $n \to \infty$. As $l_n/n \to \alpha^{-1}$ $\mathbb{P}_\nu$- a.s. as $n \to \infty$ by the SLLN, it thus suffices to prove that

$$n(G_f(u_n) - \widehat{G}_{f,\,N_n}(u_n)) \to 0 \text{ in } \mathbb{P}_\nu - pr. \text{ as } n \to \infty. \tag{1.35}$$

Write

$$n(G_f(u_n) - \widehat{G}_{f,\ N_n}(u_n)) = \frac{N_n}{l_{N_n} - 1} \frac{n}{N_n} \sum_{j=1}^{l_{N_n}-1} \{\mathbb{I}\{\zeta_j(f) \leqslant u_n\} - G_f(u_n)\},$$

and observe that $N_n/(l_{N_n} - 1) \to \alpha$, $\mathbb{P}_\nu$- a.s. as $n \to \infty$ by the SLLN again. Besides, from the argument of Theorem 15 in [26], we easily derive that there exist constants C_1 and C_2 such that for all $\varepsilon > 0$, $n \in \mathbb{N}$

$$\mathbb{P}_\nu\left(\left|\sum_{j=1}^{l_{N_n}-1} \{\mathbb{I}\{\zeta_j(f) \leqslant u_n\} - G_f(u_n)\}\right| \geqslant \varepsilon\right) \leqslant C_1 \exp(-C_2\varepsilon^2/N_n) \\ + \mathbb{P}_\nu\left(\tau_A \geqslant N_n\right).$$

From this bound, one immediately establishes (1.35). And in the case when $N_n = n^{2+\rho}$ for some $\rho > 0$, Borell-Cantelli's lemma, combined with the latter bound shows that the convergence also takes place $\mathbb{P}_\nu$-almost surely.

This result indicates that observation of a trajectory of length N_n, with $n^2 = o(N_n)$ as $n \to \infty$, is required for estimating consistently the extremal behaviour of the chain over a trajectory of length n. As shall be shown below, it is nevertheless possible to estimate the tail of the sample maximum $M_n(f)$ from the observation of a sample path of length n only, when assuming some type of behaviour for the latter, namely under maximum domain of attraction hypotheses. As a matter of fact, if one assume that $G_f \in MDA(H_\xi)$ for some $\xi \in \mathbb{R}$, of which sign is *a priori* known, one may implement classical inference procedures (refer to § 6.4 in [32] for instance) from the observed submaxima $\zeta_1(f), ..., \zeta_{l_n-1}(f)$ for estimating the shape parameter ξ of the extremal distribution, as well as the norming constants a_n and b_n. We now illustrate this point in the Fréchet case (*i.e.* when $\xi > 0$), through the example of the Hill inference method.

Heavy-tailed stationary distribution

As shown in [75], when the chain takes real values, assumption (A1) is checked for $f(x) = x$ (for this specific choice, we write $M_n(f) = M_n$, $G_f = G$, and $\zeta_j(f) = \zeta_j$ in what follows) in the particular case when the chain is stationary, i.e. when $\nu = \mu$. Moreover, it is known that when the chain is positive recurrent there exists some index θ, namely the *extremal index* of the sequence $X = (X_n)_{n\in\mathbb{N}}$ (see Leadbetter & [75] for instance), such that

$$\mathbb{P}_\mu(M_n \leqslant x) \underset{n\to\infty}{\sim} F_\mu(x)^{n\theta}, \tag{1.36}$$

denoting by $F_\mu(x) = \mu(]-\infty,\ x]) = \alpha^{-1}\mathbb{E}_A(\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i \leqslant x\})$ the stationary df. In this case, as remarked in [75], if (u_n) is such that $n(1 - G(u_n))/\alpha \to \eta < \infty$, we deduce from Proposition 8.1 and (1.36) that

$$\theta = \lim_{n\to\infty} \frac{\mathbb{P}_A(\max_{1\leqslant i\leqslant \tau_A} X_i > u_n)}{\mathbb{E}_A(\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i > u_n\})}.$$

We may then propose a natural estimate of the extremal index θ based on the observation of a trajectory of length N,

$$\widehat{\theta}_N = \frac{\sum_{j=1}^{l_N-1} \mathbb{I}\{\zeta_j > u_n\}}{\sum_{i=1}^{N} \mathbb{I}\{X_i > u_n\}},$$

which may be shown to be consistent (resp., strongly consistent) under $\mathbb{P}_\mu$ when $N = N_n$ is such that $N_n/n^2 \to \infty$ (resp. $N_n/n^{2+\rho} \to \infty$ for some $\rho > 0$) as $n \to \infty$ and $\mathcal{H}_0(2)$ is fulfilled by reproducing the argument of Proposition 9.2. And Proposition 8.1 combined with (1.36) also entails that for all ξ in $\mathbb{R}$,

$$G \in MDA(H_\xi) \ \Leftrightarrow F_\mu \ \in MDA(H_\xi).$$

Regeneration-based Hill estimator

This crucial equivalence holds in particular in the Fréchet case, *i.e.* for $\xi > 0$. Recall that assuming that a df F belongs to $MDA(H_\xi)$ classically amounts then to suppose that it satisfies the tail regularity condition

$$1 - F(x) = L(x)x^{-a},$$

where $a = \xi^{-1}$ and L is a slowly varying function, *i.e.* a function L such that $L(tx)/L(x) \to 1$ as $x \to \infty$ for any $t > 0$ (*cf* Theorem 8.13.2 in [14]). Since the seminal contribution of [45], numerous papers have been devoted to the development and the study of statistical methods in the i.i.d. setting for estimating the tail index $a > 0$ of a regularly varying df. Various inference methods, mainly based on an increasing sequence of upper order statistics, have been proposed for dealing with this estimation problem, among which the popular *Hill estimator*, relying on a conditional maximum likelihood approach. More precisely, based on i.i.d. observations $X_1, ..., X_n$ drawn from F, the Hill estimator is given by

$$H^X_{k,\ n} = (k^{-1} \sum_{i=1}^{k} \ln \frac{X_{(i)}}{X_{(k+1)}})^{-1}, \tag{1.37}$$

where $X_{(i)}$ denotes the i-th largest order statistic of the sample $X^{(n)} = (X_1, ..., X_n)$, $1 \leqslant i \leqslant n$, $1 \leqslant k < n$. Strong consistency (*cf* [28]) of this estimate has been established when $k = k_n \to \infty$ at a suitable rate, namely for $k_n = o(n)$ and $\ln \ln n = o(k_n)$ as $n \to \infty$, as well as asymptotic normality (see [38]) under further conditions on F and k_n, $\sqrt{k_n}(H^X_{k_n,n} - a) \Rightarrow \mathcal{N}(0,\, a^2)$, as $n \to \infty$. Now let us define the *regeneration-based Hill estimator* from the observation of the $l_n - 1$ submaxima $\zeta_1, ..., \zeta_{l_n-1}$, denoting by $\xi_{(j)}$ the j-th largest submaximum,

$$\widehat{a}_{n,\ k} = H^{\zeta}_{k,\ l_n-1} = (k^{-1}\sum_{i=1}^{k}\ln\frac{\zeta_{(i)}}{\zeta_{(k+1)}})^{-1}.$$

Given that $l_n \to \infty$, $\mathbb{P}_\nu$- a.s. as $n \to \infty$, results established in the case of i.i.d. observations straightforwardly extend to our setting (for comparison purpose, see [71] for properties of the classical Hill estimate in dependent settings).

Proposition 6. *Suppose that $F_\mu \in MDA(H_{a^{-1}})$ with $a > 0$. Let (k_n) be an increasing sequence of integers such that $k_n \leqslant n$ for all n, $k_n = o(n)$ and $\ln\ln n = o(k_n)$ as $n \to \infty$. Then the regeneration-based Hill estimator is strongly consistent*

$$\widehat{a}_{n,\ k_{l_n-1}} \to a,\ \mathbb{P}_\nu\text{- a.s., as } n \to \infty.$$

Under the further assumption that F_μ satisfies the Von Mises condition and that k_n is chosen accordingly (cf [38]), it is moreover asymptotically normal in the sense that

$$\sqrt{k_{l_n-1}}(\widehat{a}_{n,\ k_{l_n-1}} - a) \Rightarrow \mathcal{N}(0,\ a^2)\ \text{under}\ \mathbb{P}_\nu,\ \text{as}\ n \to \infty.$$

1.9 Concluding remarks

Although we are far from having covered the unifying theme of statistics based on (pseudo-) regeneration for Harris Markov chains, an exhaustive treatment of the possible applications of this methodology being naturally beyond the scope of the present survey article, we endeavoured to present here enough material to illustrate the power of this method. Most of the results reviewed in this paper are very recent (or new) and this line of research is still in development. Now we conclude by making a few remarks raising several open questions among the topics we focussed on, and emphasizing the potential gain that the regeneration-based statistical method could provide in further applications.

- We point out that establishing sharper rates for the 2nd order accuracy of the ARBB when applied to sample mean statistics in the general Harris case presents considerable technical difficulties (at least to us). However, one might expect that this problem could be successfully addressed by refining some of the (rather loose) bounds put forward in the proof. Furthermore, as previously indicated, extending the argument to U-statistics requires to prove preliminar nonuniform limit theorems for U-statistics of random vectors with a lattice component.
- In numerous applications it is relevant to consider null recurrent (eventually regenerative) chains: such chains frequently arise in queuing/network systems, related to teletraffic data for instance (see [70] or [37] for example), with heavy-tailed cycle lengths. Hence, exploring the theoretical properties of the (A)RBB for these specific time series provides thus another subject

of further research: as shown by [50], consistent estimates of the transition kernel, as well as rates of convergence for the latter, may still be exhibited for β-recurrent null chains (*i.e.* chains for which the return time to an atom is in the domain of attraction of a stable law with $\beta \in]0,1[$ being the stable index), so that extending the asymptotic validity of the (A)RBB distribution in this case seems conceivable.

• Turning to the statistical study of extremes now (which matters in insurance and finance applications for instance), a thorough investigation of the asymptotic behaviour of extreme value statistics based on the approximate regeneration blocks remains to be carried out in the general Harris case.

We finally mention ongoing work on empirical likelihood estimation in the markovian setting, for which methods based on (pseudo-) regeneration blocks are expected to provide significant results.

References

1. Abramovitz L., Singh K.(1985). Edgeworth Corrected Pivotal Statistics and the Bootstrap, *Ann. Stat.*, **13** ,116-132.
2. Asmussen, S. (1987). *Applied Probabilities and Queues.* Wiley.
3. Asmussen, S. (1998). Extremal Value Theory for Queues Via Cycle Maxima. *Extremes*, **1**, No 2, 137-168.
4. Athreya, K.B., Atuncar, G.S. (1998). Kernel estimation for real-valued Markov chains. *Sankhya*, **60**, series A, No 1, 1-17.
5. Athreya, K.B., Fuh, C.D. (1989). Bootstrapping Markov chains: countable case. *Tech. Rep.* B-89-7, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC.
6. Athreya, K.B., Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.,* **245**, 493-501.
7. Bertail, P. (1997). Second order properties of an extrapolated bootstrap without replacement: the i.i.d. and the strong mixing cases, *Bernoulli*, **3**, 149-179.
8. Bertail, P., Clémençon, S. (2004a). Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains. *Prob. Th. Rel. Fields*, **130**, 388–414 .
9. Bertail, P., Clémençon, S. (2004b). Note on the regeneration-based bootstrap for atomic Markov chains. *To appear in Test.*
10. Bertail, P. , Clémençon, S. (2004c). Regenerative Block Bootstrap for Markov Chains. *Submitted.*
11. Bertail, P. , Clémençon, S. (2004d). Approximate Regenerative Block-Bootstrap for Markov Chains: second-order properties. In *Compstat 2004 Proc.* Physica Verlag.
12. Bertail, P., Politis, D. (2001). Extrapolation of subsampling distribution estimators in the i.i.d. and strong-mixing cases, *Can. J. Stat.*, **29**, 667-680.
13. Bickel, P.J., Kwon, J. (2001). Inference for Semiparametric Models: Some Current Frontiers. *Stat. Sin.*, **11**, No. 4, 863-960.
14. Bingham N.H., Goldie G.M., Teugels J.L. (1989): *Regular Variation*, Cambridge University Press.

15. Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahr. verw. Gebiete,* **65**, 181-237.
16. Bolthausen, E. (1980). The Berry-Esseen Theorem for strongly mixing Harris recurrent Markov Chains. *Z. Wahr. Verw. Gebiete*, **54**, 59-73.
17. Bolthausen, E. (1982). The Berry-Esseen Theorem for strongly mixing Harris recurrent Markov Chains. *Z. Wahr. Verw. Gebiete*, **60**, 283-289.
18. Borovkova,S., Burton R., Dehling H. (1999). Consistency of the Takens estimator for the correlation dimension. Ann. Appl. Prob., **9**, No. 2, 376-390.
19. Brockwell, P.J., Resnick, S.J., Tweedie, R.L. (1982). Storage processes with general release rules and additive inputs. *Adv. Appl. Probab.,* **14**, 392-433.
20. Browne, S., Sigman, K. (1992). Work-modulated queues with applications to storage processes. *J. Appl. Probab.,* **29**, 699-712.
21. Bühlmann, P. (1997). Sieve Bootstrap for time series. *Bernoulli,* **3**, 123-148.
22. Bühlmann, P. (2002). Bootstrap for time series. *Stat. Sci.*, **17**, 52-72.
23. Callaert, H., Veraverbeke, N. (1981). The order of the normal approximation for a Studentized statistic. *Ann. Stat.,* **9**, 194-200.
24. Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.*, **14**, 1171-1179.
25. Clémençon, S. (2000). Adaptive estimation of the transition density of a regular Markov chain. *Math. Meth. Stat.,* **9**, No. 4, 323-357.
26. Clémençon, S. (2001). Moment and probability inequalities for sums of bounded additive functionals of regular Markov chains via the Nummelin splitting technique. *Stat. Prob. Letters,* **55**, 227-238.
27. Datta, S., McCormick W.P. (1993). Regeneration-based bootstrap for Markov chains. *Can. J. Statist.,* **21**, No.2, 181-193.
28. Deheuvels, P. Häusler, E., Mason, D.M. (1988). Almost sure convergence of the Hill estimator. *Math. Proc. Camb. Philos. Soc.,* **104**, 371-381.
29. Douc, R., Fort, G., Moulines, E., Soulier, P. (2004). Practical drift conditions for subgeometric rates of convergence. *Ann. Appl. Prob.,* **14**, No 3, 1353-1377.
30. Doukhan, P. (1994). *Mixing: Properties and Examples.* Lecture Notes in Statist., 85. Springer, New York.
31. Doukhan, P., Ghindès, M. (1983). Estimation de la transition de probabilité d'une chaîne de Markov Doeblin récurrente. *Stochastic Process. Appl.,* **15**, 271-293.
32. Embrechts, P., Klüppelberg, C., Mikosch, T. (2001). *Modelling Extremal Events.* Springer-Verlag.
33. Feller, W. (1968). *An Introduction to Probability Theory and its Applications: vol. I.* John Wiley & Sons, NY, 2nd edition.
34. Feller, W. (1971). *An Introduction to Probability Theory and its Applications: vol. II.* John Wiley & Sons, NY, 3rd edition
35. Franke, J. , Kreiss, J. P., Mammen, E. (2002). Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli*, **8**, 1–37.
36. Glynn, W.P., Zeevi, A. (2000). Estimating Tail Probabilities in Queues via Extremal Statistics. In *Analysis of Communication Networks: Call Centres, Traffic, and Performance* [D.R. McDonald and S.R. Turner, eds.] AMS, Providence, Rhode Island, 135-158.
37. Glynn, W.P., Whitt, W. (1995). Heavy-Traffic Extreme-Value Limits for Queues. Op. Res. Lett. **18**, 107-111.
38. Goldie, C.M. (1991). Implicit renewal theory and tails of solutions of random equations. *Ann. Appl. Prob.,* **1**, 126-166.

39. Götze, F., Hipp, C. (1983). Asymptotic expansions for sums of weakly dependent random vectors. *Zeit. Wahrschein. verw. Geb.*, **64**, 211-239.
40. Götze, F., Künsch, H.R. (1996). Second order correctness of the blockwise bootstrap for stationary observations. *Ann. Statist.*, **24**, 1914-1933.
41. Hall P. (1983). Inverting an Edgeworth Expansion. *Ann. Statist.*, **11**, 569-576.
42. Hall, P. (1985). Resampling a coverage pattern. *Stoch. Process. Applic.*, **20**, 231-246.
43. Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer.
44. Harrison, J.M., Resnick, S.J. (1976). The stationary distribution and first exit probabilities of a storage process with general release rule. *Math. Oper. Res.,* **1**, 347-358.
45. Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3, 1163-1174
46. Helmers, R (1991). On the Edgeworth expansion and the bootstrap approximation for a studentized statistics. *Ann. Statist.* ,**19**, 470-484.
47. Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Stat.*, **19**, 293–325.
48. Jain, J., Jamison, B. (1967). Contributions to Doeblin's theory of Markov processes. *Z. Wahrsch. Verw. Geb.,* **8**, 19-40.
49. Kalashnikov, V.V. (1978). *The Qualitative Analysis of the Behavior of Complex Systems by the Method of Test Functions.* Nauka, Moscow.
50. Karlsen, H.A., Tjøstheim, D. (2001). Nonparametric estimation in null recurrent time series. *Ann. Statist.,* **29** (2), 372-416.
51. Künsch, H.R. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.,* **12**, 843-863.
52. Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.,* **17**, 1217-1241.
53. Lahiri, S.N. (2003). *Resampling methods for dependent Data*, Springer.
54. Leadbetter, M.R., Rootzén, H. (1988). Extremal Theory for Stochastic Processes. *Ann. Prob.*, **16**, No. 2, 431-478.
55. Liu R., Singh K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring The Limits of The Bootstrap*. Ed. Le Page R. and Billard L., John Wiley, NY.
56. Malinovskii, V. K. (1985). On some asymptotic relations and identities for Harris recurrent Markov Chains. *Statistics and Control of Stochastic Processes*, 317-336.
57. Malinovskii, V. K. (1987). Limit theorems for Harris Markov chains I. *Theory Prob. Appl.*, **31**, 269-285.
58. Malinovskii, V. K. (1989). Limit theorems for Harris Markov chains II. *Theory Prob. Appl.*, **34**, 252-265.
59. Martin, R.D., Yohai, V.J. (1986). Influence functionals for time series. *Ann. Stat.,* **14**, 781-818.
60. Meyn, S.P., Tweedie, R.L., (1996). *Markov chains and stochastic stability.* Springer.
61. Müller, U.U., Schick, A.,Wefelmeyer, W., (2001). Improved estimators for constrained Markov chain models. *Stat. Prob. Lett.,* **54**, 427-435.
62. Nummelin, E. (1978). A splitting technique for Harris recurrent chains. *Z. Wahrsch. Verw. Gebiete,* **43**, 309-318.
63. Nummelin, E. (1984). *General irreducible Markov chains and non negative operators.* Cambridge University Press, Cambridge.

64. Politis, D.N. , Romano, J.P. (1992). A General Resampling Scheme for Triangular Arrays of alpha-mixing Random Variables with Application to the Problem of Spectral Density Estimation, *Ann. Statist.*, **20**, 1985-2007.
65. Politis, D.N., Romano, J.P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.,* **22**, 2031-2050.
66. Politis, D.N., Romano, J.P., Wolf, T. (2000). *Subsampling.* Springer Series in Statistics, Springer, NY
67. Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation.* Academic Press, NY.
68. Rachev, S. T., Rüschendorf, L. (1998). *Mass Transportation Problems***.** *Vol. I* and *II.* Springer.
69. Resnick, S. (1987). *Extreme Values, Regular Variation and Point Processes.* Springer, NY.
70. Resnick, S. (1997). Heavy Tail Modeling And Teletraffic Data. *Ann. Stat., **25**, 1805-1869.*
71. Resnick, S., Starica, C. (1995). Consistency of Hill estimator for dependent data. *J. Appl. Prob.*, **32**, 139-167.
72. Revuz, D (1984). *Markov chains.* North-Holland, 2nd edition.
73. Roberts, G.O., Rosenthal, J.S. (1996). Quantitative bounds for convergence rates of continuous time Markov processes. *Electr. Journ. Prob.*, No 9, 1-21.
74. Rosenblatt, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference,* Ed. M. Puri, 199-210.
75. Rootzén, H. (1988). Maxima and exceedances of stationary Markov chains. *Adv. Appl. Prob.*, **20**, 371-390.
76. Roussas, G. (1969). Nonparametric Estimation in Markov Processes. *Ann. Inst. Stat. Math.*, 73-87.
77. Roussas, G. (1991a). Estimation of transition distribution function and its quantiles in Markov Processes. In *Nonparametric Functional Estimation and Related Topics*, Ed. G. Roussas, 443-462.
78. Roussas, G. (1991b). Recursive estimation of the transition distribution function of a Markov Process. *Stat. Probab. Letters*, **11**, 435-447.
79. Schick, A (2001). Sample splitting with Markov Chains, *Bernoulli*, 7(1), 33-61.
80. Serfling J. (1981). *Approximation Theorems of Mathematical Statistics*, Wiley, NY.
81. Serfling, R., Zuo, Y., (2000). General Notions of Statistical Depth Function (in Data Depth). *Ann. Stat.*, **28**, No. 2., 461-482.
82. Smith, W. L. (1955). Regenerative stochastic processes. *Proc. Royal Stat. Soc., A,* **232**, 6-31.
83. Tjøstheim, D. (1990). Non Linear Time series, *Adv. Appl. Prob.*, 22, 587-611.
84. Thorisson, H. (2000). *Coupling, Stationarity and Regeneration.* Springer.