A Renewal Approach to Markovian U-statistics

Patrice Bertail, Stéphan Clémençon

A Renewal Approach to Markovian U-statistics

Patrice Bertail
CREST - INSEE & MODAL'X - Université Paris X

Stéphan Clémençon*
Institut Telecom - LTCI UMR Telecom ParisTech/CNRS No. 5141

July 8, 2010

Abstract

In this paper we describe a novel approach to the study of U-statistics in the markovian setup, based on the (pseudo-) regenerative properties of Harris Markov chains. Exploiting the fact that any sample path $X_1, \ldots, X_n$ of a general Harris chain X may be divided into asymptotically i.i.d. data blocks $\mathcal{B}_1, \ldots, \mathcal{B}_N$ of random length corresponding to successive (pseudo-) regeneration times, we introduce the notion of *regenerative U-statistic* $\Omega_N = \sum_{k \neq l} \omega_h(\mathcal{B}_k, \mathcal{B}_l)/(N(N-1))$ related to a U-statistic $U_n = \sum_{i \neq j} h(X_i, X_j)/(n(n-1))$. We show that, under mild conditions, these two statistics are asymptotically equivalent up to the order $O_{\mathbb{P}}(n^{-1})$. This result serves as a basis for establishing limit theorems related to statistics of the same form as U_n. Beyond its use as a technical tool for proving results of a theoretical nature, the regenerative method is also employed here in a constructive fashion for estimating the limiting variance or the sampling distribution of certain U-statistics through resampling. The proof of the asymptotic validity of this statistical methodology is provided, together with an illustrative simulation result.

Keywords and phrases: Markov chain, regenerative process, Nummelin splitting technique, U-statistics, Hoeffding decomposition, limit theorems, Berry-Esseen inequality, regenerative block-bootstrap.
AMS 2000 Mathematics Subject Classification: 62M05, 62F10, 62F12.

1 Introduction

Whereas the asymptotic properties of U-statistics based on independent and identically distributed data are well understood since the sixties (see Chapter 5 in [39] and the references therein), the study of this specific class of statistics, generalizing sample means, for dependent data has recently received special attention in the statistical literature, see

*Corresponding author: Telecom ParisTech, 46 rue Barrault 75634 Paris Cedex 13, France, e-mail: stephan.clemencon@telecom-paristech.fr, tel: +33 1 45 81 78 07, fax: +33 1 45 81 71 58.

[21, 4, 14, 15, 20] for instance. Indeed, this class includes numerous statistics widely used in practice such as the sample variance, and many other statistics may be well approximated by a U-statistic. The problem of extending classical limit theorems for U-statistics in the i.i.d. setup to the weakly dependent framework is generally tackled by means of *coupling techniques*. Through this approach, under adequate mixing assumptions, the limiting behavior of a U-statistic $\sum_{i \neq j} h(X_i, X_j)$ computed from a stationary weakly dependent sequence $X_1, \ldots, X_n$ may be deduced from that of a certain counterpart $\sum_{i \neq j} h(X'_i, X'_j)$ based on an i.i.d. sequence $X'_1, \ldots, X'_n$ with the same one-dimensional marginal μ. Hence, coupling is the main tool used until now for deriving asymptotic results of dependent U-statistics. Precisely, the Law of Large Numbers (LLN) has been established in [1] in the case when the kernel $h(x, y)$ is bounded and, in addition, either the stochastic process is β-mixing (or *absolutely regular* in other terms) or else $h(x, y)$ is continuous $\mu \otimes \mu$-almost surely, while the general situation of unbounded kernels is handled in [14, 15]. The Central Limit Theorem (CLT) for U-statistics based on data drawn from a β-mixing stationary ergodic process has been established in [45]; see also a refinement in [21]. An extension to the case of U-statistics of Lipschitz functionals of a β-mixing process has been subsequently considered in [22], and in [16] with a weakened continuity assumption.

The purpose of this paper is to develop an alternative to the coupling methodology, specifically tailored for regenerative processes or stochastic processes for which a regenerative extension may be built, namely *pseudo-regenerative processes*, see [43, 28]. This includes the important case of general Harris Markov chains on which the present study focuses. Indeed, sample paths of a Harris chain may be classically divided into i.i.d. *regeneration blocks*, namely data segments between random times at which the chain forgets its past, termed *regeneration times*. Hence, many results established in the i.i.d. setup may be extended to the markovian framework by applying the latter to (functionals of) the regeneration blocks. Refer to [33] for the Strong Law of Large Numbers and the Central Limit Theorem, as well as [13, 31, 32, 6] for refinements of the CLT; see also [19, 11] for moment and deviation inequalities. This approach to the study of the behavior of Markov chains based on renewal theory is known as the *regenerative method*, see [42]. In the present article, we develop further this view, in order to accurately investigate the asymptotic properties of U-statistics of positive Harris chains. Our approach crucially relies on the notion of *regenerative U-statistic approximant* of a markovian U-statistic. As the approximant is itself a standard U-statistic based on regeneration blocks, classical theorems apply to the latter and consequently yield the corresponding results for the original statistic. This way, a Strong Law of Large Numbers, a Central Limit Theorem and a Berry-Esseen bound (where the constant involved can be possibly explicitly bounded) are established for markovian U-statistics under weak hypotheses. We also examine the question of studentizing markovian U-statistics in connection with the construction of confidence intervals. Following in the footsteps of [7, 9], regeneration data blocks or approximants of the latter are used here in a practical fashion for computing a consistent estimate of the limiting variance. Beyond gaussian asymptotic confidence intervals, we propose to bootstrap certain markovian U statistics, using the specific resampling proce-

dure introduced in [7] producing bootstrap data series with a renewal structure mimicking that of the original chain. The asymptotic validity of the (approximate) regenerative block-bootstrap of U-statistics is rigorously established. For illustration purpose, some simulation results are displayed.

The rest of the paper is organized as follows. In Section 2, notation and the main assumptions are first set out. The conceptual background related to the renewal properties of Harris chains and the regenerative method is briefly exposed, together with some important examples of U-statistics in the markovian setup. The adaptation of the *projection method* to our framework, leading to the notion of *regenerative Hoeffding decomposition*, is tackled in Section 3 as a preliminary step to proving consistency and asymptotic normality of markovian U-statistics in the general positive recurrent case. Further limit results are also given, together with some hints to establish moment and probability inequalities. Section 4 is dedicated to the studentization of markovian U-statistics for the purpose of building confidence intervals and to the extension of the (approximate) regenerative block-bootstrap in the U-statistics setup. Finally, our methodology is illustrated on a simulation example in Section 5. Technical proofs are deferred to the Appendix.

2 Theoretical background

We start off with setting out the notations needed in the sequel and then briefly recall the concepts related to the Markov chain theory that shall be used in the subsequent analysis.

2.1 Notation and primary assumptions

Throughout the article, $X = (X_n)_{n\in\mathbb{N}}$ denotes a ψ-irreducible[1] time-homogeneous Markov chain, valued in a measurable space $(E, \mathcal{E})$ with transition probability $\Pi(x, dy)$ and initial distribution ν (refer to [36] for an account of the Markov chain theory). In addition, $\mathbb{P}_\nu$ denotes the probability measure on the underlying space such that $X_0 \sim \nu$, we write $\mathbb{P}_x$ when considering the Dirac mass at $x \in E$. The expectations under $\mathbb{P}_\nu$ and $\mathbb{P}_x$ will be denoted by $\mathbb{E}_\nu[.]$ and $\mathbb{E}_x[.]$ respectively, the indicator function of any event $\mathcal{A}$ by $\mathbb{I}_\mathcal{A}$. We assume further that the chain X is Harris recurrent, meaning that the chain visits an infinite number of times any subset $B \in \mathcal{E}$ such that $\psi(B) > 0$ with probability one whatever the initial state, ψ being a maximal irreducibility measure, *i.e.* $\mathbb{P}_x(\sum_{n\geq 1} \mathbb{I}_{\{X_n\in B\}} = \infty) = 1$, for all $x \in E$.

[1] Let ψ be a positive measure on a countably generated measurable space $(E, \mathcal{E})$. Recall that a Markov chain X with state space E is said ψ-irreducible if and only if, for all $B \in \mathcal{E}$, the chain visits the subset B with strictly positive probability as soon as $\psi(B) > 0$ whatever the initial state, *i.e.* $\forall x \in E$, $\sum_{n\geq 1} \mathbb{P}_x(X_n \in B) > 0$. Moreover, an irreducibility measure is said *maximal* if it dominates any other irreducibility measure. Such a measure always exists for an irreducible chain, see Theorem 4.0.1 in [33].

2.2 (Pseudo-) Regenerative Markov chains

Within this framework, a Markov chain is said to be *regenerative* when it possesses an accessible atom, *i.e.*, a measurable set A such that $\psi(A) > 0$ and $\Pi(x, .) = \Pi(y, .)$ for all $(x, y) \in A^2$. Denote then 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, by $\mathbb{P}_A[.]$ the probability measure on the underlying space such that $X_0 \in A$ and by $\mathbb{E}_A[.]$ the $\mathbb{P}_A$-expectation.

In the atomic case, it follows from the *strong Markov property* that the blocks of observations in between consecutive visits to the atom

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

form a collection of i.i.d. random variables, valued in the torus $\mathbb{T} = \cup_{n=1}^{\infty} E^n$, and the sequence $\{\tau_A(j)\}_{j\geq 1}$, corresponding to successive times at which the chain forgets its past, is a (possibly delayed) renewal process. We point out that the class of atomic chains is not as restrictive as it seems at first glance. It contains all chains with a countable state space (any recurrent state is an accessible atom), as well as many specific Markov models arising from the field of operations research, see [2] for instance.

In the regenerative setting, all stochastic stability properties may be expressed in terms of speed of return to the atom. For instance, the chain is *positive recurrent*[2] if and only if the expected return time to the atom is finite, *i.e.* $\mathbb{E}_A[\tau_A] < \infty$, see Theorem 10.2.2 in [33]. Its invariant probability distribution μ is then the occupation measure given by

$$\mu(B) = \frac{1}{\mathbb{E}_A[\tau_A]} \mathbb{E}_A\left[\sum_{i=1}^{\tau_A} \mathbb{I}_{\{X_i \in B\}}\right],\ \text{for all } B \in \mathcal{E}. \tag{2}$$

There loosely exists no such accessible atom in the general Harris case. However, it is always possible to go back to the regenerative setup. It is indeed possible to construct an artificial regeneration set through the Nummelin *splitting technique*, see [34]. This relies on the following condition referred to as *minorization condition*:

$$\forall (x, B) \in E \times \mathcal{E},\ \ \Pi^m(x, B) \geq s(x) \cdot \Phi(B), \tag{3}$$

where $m \in \mathbb{N}^*$, $s : E \to [0, 1[$ is a μ-integrable function such that $\mu(s) = \int_{x\in E} s(x)\mu(dx)$ and $\Phi(dx)$ is a probability measure on $(E, \mathcal{E})$, denoting by Π^m the m-th iterate of the transition kernel Π. Recall that, as soon as μ is positive recurrent and $\mathcal{E}$ is countably generated, it is always possible to find (m, s, Φ) such that condition (3) holds , see [34]. For simplicity, we assume that $m = 1$ here and throughout (notice that this framework

[2]Recall that an irreducible chain X with transition probability $\Pi(x, dy)$ is said positive recurrent when there exists a probability distribution μ that is invariant for the latter, *i.e.* $\mu(dy) = \int_{x\in E} \mu(dx)\Pi(x, dy)$. If it exists, such a probability is unique and is called the *stationary distribution*, see Theorem 10.0.1 in [33].

includes most examples encountered in practice, see [10]). Writing then the one-step transition probability as a conditional mixture of two distributions

$$\Pi(x,dy) = s(x)\Phi(dy) + (1-s(x))\frac{\Pi(x,dy) - s(x)\Phi(dy)}{1-s(x)}, \text{ for all } x \in E,$$

the Nummelin technique then consists in building a sequence $Y = (Y_n)_{n\in\mathbb{N}}$ of independent Bernoulli r.v.'s such that (X,Y) is a bivariate Markov chain, referred to as the *split chain*, with state space $E \times \{0,1\}$, the marginal Y indicating whether the chain (X,Y), and consequently the original one, regenerates. Precisely, given $X_n = x$, Y_n is distributed according to the Bernoulli distribution of parameter $s(x)$: if $Y_n = +1$, which thus happens with probability $s(x)$, X_{n+1} is then drawn from $\Phi(dy)$, otherwise it is distributed according to $(\Pi(x,dy) - s(x)\Phi(dy))/(1-s(x))$. This way, $A_s = E \times \{1\}$ is an accessible atom for the split chain and the latter inherits all communication and stochastic stability properties from the original chain. Notice finally that, when $m > 1$, the blocks determined by the successive visits to $E \times \{0,1\}$ are not independent any more, but 1-dependent (see §17.3.1 in [33]), a form of dependence that can also be handled, see [5].

2.3 The regenerative method

The regenerative method originates from [42] and has been thoroughly developed since then in the purpose of investigating the properties of Markov chains. One may refer to Chapter 17 in [33], for a systematic use of this approach with the aim to study averages of instantaneous functionals of positive Harris chains, $S_n(f) = n^{-1}\sum_{i=1}^n f(X_i)$ where $f : E \to \mathbb{R}$ is a measurable mapping. Roughly speaking, the task is to exploit the decomposition

$$S_n(f) = \frac{1}{n}\sum_{i=1}^{\tau_A} f(X_i) + \frac{1}{n}\sum_{j=1}^{l_n-1} f(\mathcal{B}_j) + \frac{1}{n}\sum_{i=1+\tau_A(l_n)+1}^{n} f(X_i), \tag{4}$$

where $f(\mathcal{B}_j) = \sum_{i=1+\tau_A(j)}^{\tau_A(j+1)} f(X_i)$ for $j \geq 1$ and $l_n = \sum_{i=1}^n \mathbb{I}_{\{X_i \in A\}}$ is the number of regenerations up to n, with the convention that empty summation equals to zero. This may be viewed as the natural counterpart of the parameter $\mu(f) = \int_{x\in E} f(x)\mu(dx)$, which we assume well-defined. As the $f(\mathcal{B}_j)$'s are i.i.d. and $l_n/n \to 1/\mathbb{E}_A[\tau_A]$ almost-surely as $n \to \infty$ from basic renewal theory, standard limit theorems such as the SLLN, CLT or LIL may be straightforwardly derived using classical results available in the i.i.d. setup.

We point out that, when dealing with higher order limit theorems or non asymptotic results, applying the regenerative method involves in addition the use of a specific *partitioning technique* due to the fact that, for fixed n, the blocks $\mathcal{B}_1, \ldots, \mathcal{B}_{l_n-1}$ are not independent, the sum of their lengths is indeed less than n. The partition actually corresponds to all the possible ways for the chain of regenerating up to time n. Refinements of the CLT for $S_n(f)$ have been established this way, refer to [13] for a local Berry-Esseen theorem and [32, 6] for Edgeworth expansions. In [11], it is also shown how the regenerative method yields sharp tail bounds for $S_n(f)$.

This approach has also been employed from a statistical perspective. In [3, 18], it is used for investigating the consistency properties of certain delta estimators of the stationary and transition densities. A practical use of the decomposition into regeneration blocks or approximate of the latter (see section 4) is at the center of a general methodology for constructing bootstrap confidence intervals of $\mu(f)$ and extreme-value statistics, see [7, 9]. It is the major goal of this paper to extend the range of applicability of the regenerative method to the study of the consistency properties of markovian U-statistics such as the ones mentioned below. Beyond the asymptotic study, we shall also tackle here the question of constructing gaussian and bootstrap confidence intervals based on such statistics.

We also point out that the use of the regenerative method is naturally not restricted to the markovian setup, the latter applies to any stochastic process with a regenerative extension.

2.4 U-statistics in the Markov setup - Examples

From now on, the chain X is assumed positive recurrent with limiting probability distribution μ. We focus here on parameters of type

$$\mu(h) = \int_{x_1 \in E} \dots \int_{x_k \in E} h(x_1, \dots, x_k) \mu(dx_1) \dots \mu(dx_k), \tag{5}$$

where $k \geq 2$ and $h : E^k \to \mathbb{R}^l$ is a measurable function, $l \geq 1$, such that the quantity (5) is well-defined. For simplicity's sake, we shall restrict ourselves to the case where $k = 2$ and the kernel $h(x, y)$ is symmetric, *i.e.* $\forall (x, y) \in E^2$, $h(x, y) = h(y, x)$. All results of this paper straightforwardly extend to the general case. As in the i.i.d. setting, a natural counterpart of (5) based on a sample path $X_1, \dots, X_n$ is given by

$$U_n(h) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(X_i, X_j). \tag{6}$$

Many statistics widely used in practice fall within this class, among which popular dispersion estimators such as the following ones.

- The sampling variance. Suppose that $E \subset \mathbb{R}$. Provided that $\int_{x \in E} x^2 \mu(dx) < \infty$, a natural estimate of μ's variance σ_μ^2 is

$$\hat{\sigma}_\mu^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2,$$

  where $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ is the sample mean. It may be indeed rewritten as (6) with $h(x, y) = (x - y)^2/2$.

- THE GINI MEAN DIFFERENCE. The Gini index provides another popular measure of dispersion. It corresponds to the case where $E \subset \mathbb{R}$ and $h(x,y) = |x-y|$:

$$G_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} |X_i - X_j|.$$

Other important examples arise from the field of nonparametric testing. For instance, the next markovian U-statistic may be considered when testing symmetry around zero for the stationary distribution, following the example of the i.i.d. situation.

- THE WILCOXON STATISTIC. Suppose that $E \subset \mathbb{R}$ is symmetric around zero. As an estimate of the quantity $\int_{(x,y)\in E^2} \{2\mathbb{I}_{\{x+y>0\}} - 1\}\mu(dx)\mu(dy)$, it is pertinent to consider the statistic

$$W_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \{2 \cdot \mathbb{I}_{\{X_i + X_j > 0\}} - 1\},$$

which is relevant for testing whether μ is located at zero or not.

Regarding analysis of extreme values in the multidimensional setup, it is noteworthy that particular markovian U-statistics arise in the study of *depth statistical functions* for dynamic systems.

- THE TAKENS ESTIMATOR. Suppose that $E \subset \mathbb{R}^d$, $d \geq 1$. Denote by $||.||$ the usual euclidian norm on $\mathbb{R}^d$. In [14], the following estimate of the *correlation integral*, $C_\mu(r) = \int_{(x,x')} \mathbb{I}_{\{||x-x'||\leq r\}} \mu(dx)\mu(dx')$ with $r > 0$, is considered:

$$C_n(r) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \mathbb{I}_{\{||X_i - X_j|| \leq r\}}.$$

In the case where a scaling law holds for the correlation integral, *i.e.* when there exists $(\alpha, r_0, c) \in \mathbb{R}_+^{*3}$ such that $C_\mu(r) = c \cdot r^{-\alpha}$ for $0 < r \leq r_0$, the U-statistic

$$T_n = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \log\left(\frac{||X_i - X_j||}{r_0}\right)$$

is used in order to build the Takens estimator $\hat{\alpha}_n = -T_n^{-1}$ of the *correlation dimension* α.

Many other functionals that are relevant in the field of statistical analysis of markovian data may also be approximated by U-statistics. It is the case of ratio statistics for instance, such as the estimator of the extreme value index introduced in [9]. Some of the results established in this paper are particularly useful for investigating their asymptotic properties, see [12].

The reason for investigating asymptotic properties of markovian U-statistics naturally arises from the ubiquity of the Markov assumption in time-series analysis. Additionally, we point out that, in many cases, where no explicit closed analytical form for a distribution μ of interest, allowing for direct computation or crude Monte-Carlo estimation of the parameter (5), is available, the popular *MCMC* approach consists in considering the target distribution μ as the limiting probability measure of a positive recurrent Markov chain, where simulation is computationally feasible. Statistical inference of μ's features is then based on a sample path of the chain with long runlength, which corresponds to the asymptotic framework of this article. For instance, refer to [25] or [38] for recent accounts of the *Markov Chain Monte-Carlo* methodology.

3 Asymptotic theory of markovian U-statistics

Throughout this section, we suppose that the chain X is regenerative, with an atom A and denote by $\mathcal{B}_j$ the corresponding regeneration blocks of observations. All results carry over to the general positive Harris case, using the Nummelin technique, see §2.2. As previously mentioned, we confine the present the study to statistics of the form

$$U_n(h) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(X_i, X_j),$$

where $h : \mathbb{R}^2 \to \mathbb{R}$ is a symmetric kernel. In the subsequent analysis, we investigate the asymptotic properties of this statistic, as an estimator of the parameter

$$\mu(h) = \int_{(x,y) \in E^2} h(x,y) \mu(dx) \mu(dy), \tag{7}$$

which we assume to be well defined, *i.e.* $\mu(|h|) < \infty$. In the regenerative setup, the parameter of interest may be rewritten as follows, *cf* Eq. (2):

$$\mu(h) = \frac{1}{\alpha^2} \cdot \mathbb{E}_A \left[\sum_{i=1}^{\tau_A(1)} \sum_{j=1+\tau_A(1)}^{\tau_A(2)} h(X_i, X_j) \right], \tag{8}$$

where $\alpha = \mathbb{E}_A[\tau_A]$ denotes the mean cycle length.

Extensions to more general settings are straightforward, including non symmetric kernels, U-statistics of higher order $k > 2$, V-statistics, as well as generalized U and V-statistics. Owing to space limitations, details are omitted here.

3.1 Regenerative U-statistics

In the i.i.d. setup, a major tool in establishing the asymptotic theory of U-statistics is the *projection method*, introduced in [27] and popularized in more general situations under

the name of *Hajek's projection*, see [26]. It consists in approximating the recentered U-statistic by its orthogonal projection onto the space of averages of zero-mean i.i.d. random variables, viewed as a subspace of the Hilbert space of centered square integrable r.v.'s, under suitable moment assumptions. This way, limit results established for sums of i.i.d. r.v.'s carry over to U-statistics, which enables the asymptotic theory of more general functionals to be derived.

In the markovian context, the theory of sums of i.i.d. variables also yields the limit distribution theory of sums $\sum_{i\leq n} f(X_i)$ via the regenerative method. Indeed, the original average is approximated by an average of (asymptotically) i.i.d. *block sums* $f(\mathcal{B}_j)$, see §2.3. In order to successively exploit these two approximation methods for investigating the limit behavior of markovian U-statistics, we introduce the following notion.

Definition 1 (Regenerative kernel) *Let $h : E^2 \to \mathbb{R}$ be a kernel. The regenerative kernel related to h is the kernel $\omega_h : \mathbb{T}^2 \to \mathbb{R}$ given by*

$$\omega_h\left((x_1,\ldots,x_n),\ (y_1,\ldots,y_m)\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} h(x_i,y_j),$$

for all $x^{(n)} = (x_1,\ldots,x_n)$ and $y^{(m)} = (y_1,\ldots,y_m)$ in the torus $\mathbb{T} = \cup_{n\geq 1} E^n$.

It is noteworthy that the kernel ω_h is symmetric, as soon as h is. It will be useful in the following to consider U-statistics based on regeneration data segments solely.

Definition 2 (Regenerative U-statistic) *Let $h : E^2 \to \mathbb{R}$ be a symmetric kernel such that $\mu(|h|) < \infty$ and set $\tilde{h} = h - \mu(h)$. A regenerative U-statistic related to the kernel h is a U-statistic with kernel $\omega_{\tilde{h}}$:*

$$R_L(h) = \frac{2}{L(L-1)} \sum_{1\leq k<l\leq L} \omega_{\tilde{h}}(\mathcal{B}_k,\mathcal{B}_l),$$

where $L \geq 1$ and $\mathcal{B}_1,\ \ldots,\ \mathcal{B}_L$ are regeneration blocks of the chain X.

We point out that $R_L(h)$ is a standard U-statistic with mean zero. Hence, we may consider its *Hoeffding decomposition*: $R_L(h) = 2S_L(h) + D_L(h)$, where

$$S_L(h) = \frac{1}{L}\sum_{k=1}^{L} h_1(\mathcal{B}_k) \text{ and } D_L(h) = \frac{2}{L(L-1)} \sum_{1\leq k<l\leq L} h_2(\mathcal{B}_k,\mathcal{B}_l),$$

with $\forall (b_1,b_2) \in \mathbb{T}^2$,

$$h_1(b_1) = \mathbb{E}[\omega_{\tilde{h}}(b_1,\mathcal{B}_1)] \text{ and } h_2(b_1,b_2) = \omega_{\tilde{h}}(b_1,b_2) - h_1(b_1) - h_1(b_2).$$

The U-statistic $D_L(h)$ is *degenerate*, *i.e.* its kernel satisfies: $\forall b_1 \in \mathbb{T}$, $\mathbb{E}[h_2(b_1,\mathcal{B}_1)] = 0$. Assuming that $\mathbb{E}[\omega_h^2(\mathcal{B}_1,\mathcal{B}_2)]/\alpha^2 = \int_{(x,y)\in E^2} h^2(x,y)\mu(dx)\mu(dy) < \infty$, its variance is of

order $1/L^2$ and $cov(h_1(\mathcal{B}_1), h_2(\mathcal{B}_1, \mathcal{B}_2)) = 0$. The leading term in this orthogonal decomposition is thus the average of i.i.d. random variables $2S_L(h)$, when the conditional expectation h_1 is non zero. As $L \to \infty$, $L^{-1/2}R_L(h)$ then converges in distribution to the normal $\mathcal{N}(0, 4s^2(h))$, with $s^2(h) = \mathbb{E}[h_1^2(\mathcal{B}_1)]$. One may refer to Chapter 5 in [39] for a detailed account of the theory of U-statistics in the i.i.d. setting.

The following technical assumptions are involved in the subsequent analysis.

$\mathbf{A_0}$ (Block-length: moment assumption.) Let $q \geq 1$, we have $\mathbb{E}_A\left[\tau_A^q\right] < \infty$.

$\mathbf{A_1}$ (Block-sums: moment assumptions.) Let $k \geq 1$, we have

$$\mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}\sum_{j=1}^{\tau_A}|h(X_i, X_j)|\right)^k\right] < \infty \text{ and } \mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}\sum_{j=1+\tau_A}^{\tau_A(2)}|h(X_i, X_j)|\right)^k\right] < \infty.$$

$\mathbf{A_2}$ (Non-regenerative block.) Let $l \geq 1$, we have $\mathbb{E}_\nu\left[\tau_A^l\right] < \infty$ as well as

$$\mathbb{E}_\nu\left[\left(\sum_{i=1}^{\tau_A}\sum_{j=1}^{\tau_A}|h(X_i, X_j)|\right)^l\right] < \infty \text{ and } \mathbb{E}_\nu\left[\left(\sum_{i=1}^{\tau_A}\sum_{j=1+\tau_A}^{\tau_A(2)}|h(X_i, X_j)|\right)^l\right] < \infty.$$

$\mathbf{A_3}$ (Linear term: moment assumptions.) Let $m \geq 0$, we have

$$\mathbb{E}_\nu\left[\left(\sum_{i=1}^{\tau_A}h_1(X_i)\right)^m\right] < \infty \text{ and } \mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}h_1(X_i)\right)^{m+2}\right] < \infty.$$

$\mathbf{A_4}$ (Uniform moment assumptions.) Let $p \geq 0$, we have

$$\sup_{x\in E}\mathbb{E}_\nu\left[\left(\sum_{j=1}^{\tau_A}\bar{h}(x, X_j)\right)^p\right] < \infty \text{ and } \sup_{x\in E}\mathbb{E}_A\left[\left(\sum_{j=1}^{\tau_A}\bar{h}(x, X_j)\right)^{p+2}\right] < \infty,$$

where $\forall (x, y) \in E^2$, $\bar{h}(x, y) = h(x, y) - \int_{z\in E} h(x, z)\mu(dz)$.

$\mathbf{A_5}$ (Non-degeneracy.) We have

$$\mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}h_1(X_i)\right)^2\right] > 0.$$

Notice that assumption $\mathbf{A_1}$ combined with $\mathbf{A_2}$ implies assumption $\mathbf{A_3}$ when $k = m \geq 1$ by Jensen's inequality. In a similar fashion, using in addition Fubini's theorem, assumption $\mathbf{A_4}$ implies $\mathbf{A_3}$ when $p = m$, since $h_1(x)/\alpha = \int_{x \in E} h(x,y)\mu(dy) - \mu(h)$ for all $x \in E$. In the case where the kernel h is bounded, conditions $\mathbf{A_1} - \mathbf{A_4}$ can be reduced to moment assumptions related to the return time τ_A solely.

Remark 3 *(On moment conditions) From a practical perspective, we recall that block-moment assumptions can be generally checked for the split Markov chain by establishing drift conditions of Lyapounov's type for the original chain, see Chapter 11 in [33] and [23], as well as the references therein. We also refer to [8] for an explicit checking of such conditions on several important examples and to §4.1.2 for sufficient conditions formulated in terms of uniform speed of return to small sets.*

3.2 Consistency and asymptotic normality

We start off by stating the key result in deriving the asymptotic theory of markovian U-statistics by means of the regenerative method. It reveals that the U-statistic $U_n(h)$ can be approximated by the corresponding regenerative U-statistic based on the (random number of) observed regeneration blocks up to a multiplicative factor.

Proposition 4 (Regenerative approximation) *Suppose that assumptions $\mathbf{A_0} - \mathbf{A_2}$ are fulfilled with $q = k = l = 2$. Set $W_n(h) = U_n(h) - \mu(h) - (l_n - 1)(l_n - 2)R_{l_{n-1}}(h)/(n(n-1))$. Then, as $n \to \infty$, the following stochastic convergences hold:*

(i) $W_n(h) \to 0$, $\mathbb{P}_\nu$*-almost surely,*

(ii) $\mathbb{E}_\nu \left[(W_n(h))^2 \right] = O(n^{-2})$.

This result yields the strong consistency of markovian U-statistics. The next theorem is immediate, since the SLLN holds for the U-statistic $R_L(h)$.

Theorem 5 (Strong Law of Large Numbers) *Suppose that assumptions $\mathbf{A_0} - \mathbf{A_2}$ are fulfilled with $q = k = l = 2$. Then, as $n \to \infty$, we have:*

$$U_n(h) \to \mu(h), \ \ \mathbb{P}_\nu\text{-almost surely.}$$

In a similar fashion, using the approximation result stated in Proposition 4, the CLT applied to the supposedly non degenerate U-statistic $R_L(h)$ yields the analogous result for markovian U-statistics.

Theorem 6 (Central Limit Theorem) *Suppose that assumptions $\mathbf{A_0} - \mathbf{A_2}$ with $q = k = l = 2$ and $\mathbf{A_5}$ are fulfilled. Then, we have the convergence in distribution under $\mathbb{P}_\nu$:*

$$\sqrt{n}\,(U_n(h) - \mu(h)) \Rightarrow \mathcal{N}(0, \sigma^2(h)), \ \text{as } n \to \infty,$$

where $\sigma^2(h) = 4\mathbb{E}_A \left[\left(\sum_{i=1}^{\tau_A} h_1(X_i) \right)^2 \right] / \alpha^3$.

Remark 7 (Stochastic stability assumptions) *We point out that, in contrast to results established by means of coupling techniques, stationarity is not required here and moment assumptions involved in the theorem above are weaker than those stipulated in [21] (whose results hold true however for a more general class of weakly dependent processes). Indeed, though expressed in terms of β-mixing coefficients decay rate, the latter boil down to conditions $\mathbf{A_0} - \mathbf{A_2}$ with q, k and l all strictly larger than 2, refer to Chapter 9 in [37] for a precise study of relationships between mixing conditions and block-moment assumptions.*

Remark 8 (On the asymptotic variance) *Observe that, for all $x \in E$, one may write $h_1(x)/\alpha = \int_{y\in E} h(x,y)\mu(dy) - \mu(h) \stackrel{def}{=} h_0(x)$ by virtue of Kac's formula (2). In addition, it follows from the analysis carried out in §17.4.3 of [33] that the asymptotic variance of $U_n(h)$ can be expressed as follows*

$$\sigma^2(h) = 4\int_{x\in E} \{2\hat{h}_0(x)h_0(x) - h_0^2(x)\}\mu(dx),$$

where $\hat{h}_0(x) = \mathbb{E}_x[\sum_{n=0}^{\infty} h_0(X_i)]$ is the solution of the Poisson equation:

$$\hat{h}_0(x) - \int_{y\in E} \hat{h}_0(x)\Pi(x,dy) = h_0(x),\ x \in E.$$

In the regenerative case, one may also take $\hat{h}_0(x) = \mathbb{E}_x[\sum_{i=1}^{\tau_A} h_0(X_i)]$.

The next result concludes the subsection by showing that the distance of the markovian U-statistic $U_n(h)$ to its asymptotic mean $\mu(h)$ is maximally of the order $\sqrt{\log\log n/n}$ as $n \to \infty$.

Theorem 9 (Law of Iterated Logarithm) *Suppose that assumptions $\mathbf{A_0} - \mathbf{A_2}$ with $q = k = l = 2$ and $\mathbf{A_5}$ are fulfilled. Then,*

$$\limsup_{n\to\infty} \frac{\sqrt{n}\,(U_n(h) - \mu(h))}{\sqrt{2\sigma^2(h)\log\log n}} = +1 \quad \mathbb{P}_\nu\text{-almost surely}\,.$$

3.3 Further asymptotic results

Here, we consider some refinements of the limit theorems stated in the previous subsection. The first result shows that, as an estimator of $\mu(h)$, the bias of the U-statistic (6) is of order $O(n^{-1})$. In order to state it precisely, we consider the following technical conditions.

$\mathbf{A_6}$ (Cramer condition - linear term.) We have:

$$\limsup_{t\to\infty} \left| \mathbb{E}_A \left[\exp\left\{ it \sum_{i=1}^{\tau_A} h_1(x) \right\} \right] \right| < 1.$$

$\mathbf{A_7}$ (Uniform Cramer condition.) We have:

$$\sup_{x\in E}\limsup_{t\to\infty}\left|\mathbb{E}_A\left[\exp\left\{it\sum_{i=1}^{\tau_A}\bar{h}(x,X_i)\right\}\right]\right|<1,$$

where $\forall (x,y)\in E^2$, $\bar{h}(x,y)=h(x,y)-\int_{z\in E}h(x,z)\mu(dz)$.

We point out that condition $\mathbf{A_6}$ (respectively, condition $\mathbf{A_6}$) is fulfilled as soon as there exists no regular grid (*i.e.* grid of the form $\{a+b\cdot h:\ h\in\mathbb{Z}\}$ for $(a,b)\in\mathbb{R}^2$) that contains the set $\{h_1(x):\ x\in E\}$ (respectively, that contains the set $\{\bar{h}(x,y):\ (x,y)\in E^2\}$).

Proposition 10 (Asymptotic bias) *Suppose that assumptions* $\mathbf{A_0}$ *with* $q=4+\delta$ *for some* $\delta>0$, $\mathbf{A_1}$ *with* $k=2$, $\mathbf{A_4}$ *with* $p=2$, $\mathbf{A_6}$ *and* $\mathbf{A_7}$ *are fulfilled. Then, as* $n\to\infty$, *we have:*

$$\mathbb{E}_\nu\left[U_n(h)\right]=\mu(h)+2\cdot\frac{\Delta+\phi_\nu-2\beta/\alpha+\gamma}{n}+O(n^{-3/2}),$$

where $\Delta=\mathbb{E}_A\left[\sum_{1\leq k<j\leq\tau_A}h(X_k,X_j)\right]/\alpha$, $\beta=\mathbb{E}_A\left[\tau_A\sum_{i=1}^{\tau_A}h_0(X_i)\right]$, $\phi_\nu=\mathbb{E}_\nu[\sum_{i=1}^{\tau_A}h_0(X_i)]$ *and* $\gamma=\mathbb{E}_A[\sum_{i=1}^{\tau_A}(\tau_A-j)h_0(X_i)]/\alpha$, *with* $h_0(x)=\int_{y\in E}\{h(x,y)-\mu(h)\}\mu(dx)$ *for all* $x\in E$.

Remark 11 (On asymptotic bias components) *As may be shown by a careful examination of Lemma 19's proof, the component* Δ *of the first order term corresponds to the contribution to the bias of the block-diagonal sum* $\sum_{j=1}^{l_n-1}\sum_{1+\tau_A(j)\leq k<l\leq\tau_A(j+1)}h(X_k,X_l)$, ϕ_ν *to that of the sum* $\sum_{j=1}^{l_n-1}\omega_h(\mathcal{B}_0,\mathcal{B}_j)$ *which involves the first (nonregenerative) data segment,* $-2\beta/\alpha$ *to the one of* $\sum_{1\leq j<k\leq l_n-1}\omega_h(\mathcal{B}_j,\mathcal{B}_k)$, *while the component* γ *is induced by the quantity* $\sum_{j=1}^{l_n-1}\sum_{i=1+\tau_A(l_n)}^{n}\omega_h(\mathcal{B}_j,X_i)$ *involving the last (nonregenerative) data block.*

The next result provides a Berry-Esseen bound for the U-statistic (6), generalizing the result obtained in [13] for sample mean statistics.

Theorem 12 (A Berry-Esseen bound) *Under assumptions* $\mathbf{A_0}$ *with* $q=2$, $\mathbf{A_1}$ *with* $k=2$, $\mathbf{A_2}$ *with* $l=2$, $\mathbf{A_3}$ *with* $m=3$ *and* $\mathbf{A_5}$, *there exists a constant* $K<\infty$ *such that, for all* $n\geq 1$*:*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}_\nu\left\{\sqrt{n}\sigma(h)^{-1}\left(U_n(h)-\mu(h)\right)\leq x\right\}-\Phi(x)\right|\leq Kn^{-1/2}. \tag{9}$$

The proof of the theorem above combines the Stein's approach and the partitioning technique mentioned in §2.3 to a Berry-Esseen bound for Markovian sample mean statistics (see Theorem 17). Incidentally, it should be noticed that the latter result improves upon that of [13], insofar as an estimate of the constant K involved in the upper bound obtained by the proof technique we used can be exhibited (refer to Theorem 12's proof in the Appendix).

It is not the purpose of this paper to extend all probabilistic results proved for i.i.d. U-statistics to the markovian setup, but rather to illustrate, through the statement of a few important theorems, how the regenerative method combined with the standard projection technique enables a precise study of the asymptotic behavior of markovian U-statistics. The same approach can be used for establishing a functional version of the CLT for instance, as well as probability/moment bounds, following in the footsteps of [11], even though these are not asymptotic results. However, we point out that proving an Edgeworth expansion up to $O(n^{-1})$ for $U_n(h)$, as in the case of sample mean statistics (see [31, 6]), is not as straightforward. Even if one tries to reproduce the argument in [6], consisting in partitioning the underlying probability space according to every possible realization of the regeneration time sequence between 0 and n, the problem boils down to control, as $m \to \infty$, the asymptotic behavior of the distribution

$$\mathbb{P}\left\{ \sum_{1 \leq i \neq j \leq m} \omega_h(\mathcal{B}_i, \mathcal{B}_j)/\sigma^2_{U,\,m} \leq y,\ \sum_{j=1}^{m} l(\mathcal{B}_j) = k \right\},$$

where $l(\mathcal{B}_j) = \tau_A(j+1) - \tau_A(j)$ denotes the length of the block $\mathcal{B}_j$, $j \geq 1$. Precisely, a local Edgeworth expansion of this probability distribution is required (analogous to the one obtained by [24] in the case of the sample mean). The major barrier lies in the simultaneous presence of the lattice component and the degenerate part of the U-statistics. To our knowledge, no result of this type has been established in the literature and we leave this question for further research.

4 Studentization and bootstrap confidence intervals

We now turn to the problem of constructing confidence intervals based on markovian U-statistics or their regenerative versions.

4.1 Normalization of markovian U-statistics

Here we show how one may benefit from the underlying regenerative structure of the data for computing a proper standardization of markovian U-statistics. This is of crucial importance, insofar as it permits the construction of asymptotic (gaussian) confidence intervals.

4.1.1 Regenerative case

Let $L \geq 1$. Consider the empirical counterpart of the conditional expectation based on all the first L regenerative data blocks, except $\mathcal{B}_j$, $j \in \{1, \ldots, L\}$:

$$\hat{h}_{1,-j}(b) = \frac{1}{L-1} \sum_{k=1,\ k \neq j}^{L} \omega_h(b, \mathcal{B}_k) - \frac{2}{L(L-1)} \sum_{1 \leq k < l \leq L} \omega_h(\mathcal{B}_k, \mathcal{B}_l), \tag{10}$$

as well as the *Jacknife estimator* of the asymptotic variance $s^2(h)$, see [17]:

$$\hat{s}_L^2(h) = \frac{1}{L}\sum_{k=1}^{L} \hat{h}_{1,-k}^2(\mathcal{B}_k). \tag{11}$$

We may now introduce the norming constant related to the U-statistic $U_n(h)$:

$$\hat{\sigma}_n^2(h) = 4\,(l_n/n)^3\,\hat{s}_{l_n-1}^2(h). \tag{12}$$

The next results shows that, equipped with this simple normalization, it is possible to construct asymptotically pivotal quantities in order to produce limiting confidence intervals for the parameter $\mu(h)$.

Proposition 13 (STUDENTIZATION) *Suppose that assumptions* $\mathbf{A_0} - \mathbf{A_2}$ *with* $q = k = l = 2$ *and* $\mathbf{A_5}$ *are fulfilled. Then, the next asymptotic results hold.*

(i) The statistic $\hat{\sigma}_n^2(h)$ *is a strongly consistent estimator of* $\sigma^2(h)$*:*

$$\hat{\sigma}_n^2(h) \to \sigma^2(h) \ \ \mathbb{P}_\nu\text{-almost-surely, as } n \to \infty.$$

(ii) In addition, when recentered and renormalized by $\sqrt{n/\hat{\sigma}_n^2(h)}$*, the statistic* $U_n(h)$ *is asymptotically normal:*

$$\frac{\sqrt{n}}{\hat{\sigma}_n(h)}(U_n(h) - \mu(h)) \Rightarrow \mathcal{N}(0,1) \ \text{ in } \mathbb{P}_\nu\text{-distribution as } n \to \infty.$$

4.1.2 General case - the "plug-in" approach

As pointed out in subsection 2.2, a positive Harris chain X possesses no regenerative set in general. Even though it can be viewed as a marginal of a regenerative Nummelin extension $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ built from parameters (s, Φ, m) of a minorization condition (3) fulfilled by the original chain, it should be noticed that the split chain is a theoretical construction and the Y_n's cannot be observed in practice. However, it has been suggested in [7] to extend regeneration-based inference techniques the following way: generate first a sequence $(\widehat{Y}_1, \ \ldots, \ \widehat{Y}_n)$ from the supposedly known parameters (s, Φ) in a way that $((X_1, Y_1), \ \ldots, \ (X_n, Y_n))$ and $((X_1, \widehat{Y}_1), \ \ldots, \ (X_n, \widehat{Y}_n))$ have close distributions in the Mallows sense and then apply adequate statistical procedures to the data blocks thus defined $\widehat{\mathcal{B}}_1, \ \ldots, \ \widehat{\mathcal{B}}_{\widehat{N}_n}$ (corresponding to the successive times when $\widehat{Y}$ visits the state 1), as if they were really regenerative. Here we briefly recall the basic principle underlying this approach.

For simplicity, we assume that condition (3) is fulfilled with $m = 1$ (see §2.2). Observe first that, conditioned upon $X^{(n)} = (X_1, \ldots, X_n)$, the random variables $Y_1, \ \ldots, \ Y_n$ are

mutually independent and, for all $i \in \{1, \ldots, n\}$, Y_i is drawn from a Bernoulli distribution with parameter given by:

$$\delta(X_i, X_{i+1}) = s(X_i) \times \frac{d\Phi}{d\Pi(X_i, .)}(X_{i+1}), \tag{13}$$

where $d\Phi/d\Pi(x, .)$ denotes (a version of) the density of $\Phi(dy)$ with respect to $\Pi(x, dy)$. Suppose that a transition kernel $\widehat{\Pi}(x, y)$ providing an accurate estimation of $\Pi(x, dy)$ and such that $\forall x \in E$, $\widehat{\Pi}(x, y) \geq \delta\phi(y)$, is available. Given $X^{(n)}$, the construction of the $\widehat{Y}_i$'s boils down to drawing mutually independent Bernoulli random variables, the parameter of $\widehat{Y}_i$'s conditional distribution being obtained by replacing the unknown quantity $d\Phi/d\Pi(X_i, .)(X_{i+1})$ by its empirical counterpart $d\Phi/d\widehat{\Pi}(X_i, .)(X_{i+1})$ in (13). A more detailed description of this plug-in approximation is available in [10] together with a discussion of numerical issues regarding its practical implementation. In particular, special attention is paid to the problem of selecting the parameters (s, Φ) in a data-driven fashion. As shown in [33] (see Chapter 5 therein), one may always choose $s(x)$ of the form $s(x) = \delta \cdot \mathbb{I}_{\{x \in S\}}$ where $\delta \in]0, 1[$, S is a measurable set such that $\mu(S) > 0$ in which Φ's support is included. Such a subset S is generally called a *small set* for the chain X. Considering λ a measure of reference on $(E, \mathcal{E})$ dominating the collection of probability distributions $\{\Pi(x, dy),\ x \in E\}$ and, consequently, the minorizing probability $\Phi(dy)$, one may then re-write the Bernoulli parameters as $\delta(x, y) = \delta\mathbb{I}_{\{x \in S\}} \times d\phi/d\pi(x, y)$, $(x, y) \in S^2$, where $\pi(x, .) = d\Pi(x, .)/d\lambda$ and $\phi = d\Phi/d\lambda$.

The accuracy of the resulting approximation, measured in terms of Mallows distance between the random vectors $(Y_1, \ldots, Y_n)$ and $(\widehat{Y}_1, \ldots, \widehat{Y}_n)$, mainly depends on the quality of the transition density $\widehat{\pi}(x, y) = d\widehat{\Pi}(x, .)/d\lambda(y)$ as an estimate of $\pi(x, y)$ over S^2. A sharp bound, based on a coupling argument, is given in Theorem 3.1 of [7] under the following assumptions.

$\mathbf{A_8}$. The Mean Squared Error of $\widehat{\pi}$ is of order α_n when error is measured by the sup-norm over S^2:

$$\mathbb{E}_\nu \left[\sup_{(x,y) \in S^2} |\widehat{\pi}(x, y) - \pi(x, y)|^2 \right] = O(\alpha_n),$$

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

$\mathbf{A_9}$. The parameters S and Φ of condition (3) are chosen so that: $\inf_{x \in S} \phi(x) > 0$.

$\mathbf{A_{10}}$. We have $\sup_{(x,y) \in S^2} \pi(x, y) < \infty$ and $\sup_{n \in \mathbb{N}} \sup_{(x,y) \in S^2} \widehat{\pi}_n(x, y) < \infty$ $\mathbb{P}_\nu$-a.s. .

In addition, the following hypotheses guarantee that block-length moment assumptions hold for the split chain, see Chapter 11 in [33] (notice that the latter do not depend on the chosen small set S). Consider the hitting time to the set S: $\tau_S = \inf\{n \geq 1,\ X_n \in S\}$.

$\mathbf{A_{11}}$. Let $q \geq 1$, we have: $\sup_{x_0 \in S} \mathbb{E}_{x_0} \left[\tau_S^q\right] < \infty$.

$\mathbf{A_{12}}$. Let $l \geq 1$, we have: $\mathbb{E}_\nu\left[\tau_S^l\right] < \infty$.

In the general case, the plug-in approach to $U_n(h)$'s standardization thus consists in computing the quantities (10), (11) and (12) by using the pseudo-blocks $\widehat{\mathcal{B}}_1,\ \ldots,\ \widehat{\mathcal{B}}_{\widehat{N}_n}$ instead of (X, Y)'s regeneration blocks. We denote by $\tilde{\sigma}_n^2(h)$ the resulting estimate of the asymptotic variance. The next theorem reveals that the plug-in approximation step does not spoil the studentization of the statistic.

Proposition 14 (STUDENTIZATION (BIS)) *Suppose that hypotheses* $\mathbf{A_5}$ *and* $\mathbf{A_7} - \mathbf{A_{12}}$, *with* $q = 4$ *and* $l = 2$, *are fulfilled. Assume also that the kernel* h *is bounded, i.e.* $||h||_\infty = \sup_{(x,y)\in E^2} |h(x, y)| < \infty$. *Then, we have, as* $n \to \infty$,

$$\frac{\sqrt{n}}{\tilde{\sigma}_n(h)}\left(U_n(h) - \mu(h)\right) \Rightarrow \mathcal{N}(0, 1) \text{ in } \mathbb{P}_\nu\text{-distribution.}$$

Remark 15 (ON THE BOUNDEDNESS ASSUMPTION) *From a careful examination of Proposition 14's argument, one may shows that the convergence stated above extends to a more general framework, including cases where* $h(x, y)$ *is not bounded. Assumptions of the form* $\sup_{x_0\in S} \mathbb{E}[(\sum_{i=1}^{\tau_S} \sum_{j=1+\tau_S}^{\tau_S(2)} |h(X_i, X_j))|)^k] < \infty$ *with* $k \geq 1$ *suitably chosen and* $\tau_S(2) = \inf\{n > \tau_S,\ X_n \in S\}$ *would then be required, as well as much technicality (in order to extend the coupling results involved in the proof mainly, see also Theorem 3.2's proof in [7]). For brevity's sake, here we restrict ourselves to the bounded case and leave extensions to the reader.*

4.2 Regenerative block-bootstrap for markovian U-statistics

This subsection is devoted to extend the (approximate) regenerative block-bootstrap methodology, *(A)RBB* in abbreviated form, originally proposed in [7] for bootstrapping standard markovian sample means $n^{-1}\sum_{i\leq n} f(X_i)$, to markovian U-statistics. We start off with describing the resampling algorithm in this context and then establish the asymptotic validity of the bootstrap distribution estimate thus produced.

4.2.1 The (A)RBB algorithm

Suppose that a sample path $X^{(n)} = (X_1, \ldots, X_n)$ drawn from a general Harris chain X is observed, from which a random number $N_n = l_n - 1$ of regeneration blocks, or pseudo-regeneration blocks using the plug-in method described in §4.1.2, are formed: $\mathcal{B}_1,\ \ldots,\ \mathcal{B}_{N_n}$. Consider then a U-statistic $U_n(h) = U_n(X_1,\ \ldots,\ X_n)$ with kernel $h(x, y)$ and standardization $\widehat{\sigma}_n(h) = \sigma_n(\mathcal{B}_1,\ \ldots,\ \mathcal{B}_{N_n})$ built as described in section 4, estimating the parameter $\mu(h) \in \mathbb{R}$. The (A)RBB algorithm below produces an estimate of the sampling distribution of $\widehat{\sigma}_n(h)^{-1}\{U_n(h) - \mu(h)\}$. It is performed in three steps, as follows.

(Approximate) Regenerative Block-Bootstrap

1. Generate sequentially bootstrap data blocks $\mathcal{B}_1^*, ..., \mathcal{B}_k^*$ by drawing with replacement from the initial blocks $\mathcal{B}_1, ..., \mathcal{B}_{N_n}$ until the length $l^*(k)$ of the bootstrap data series $(\mathcal{B}_1^*, \ \ldots, \ \mathcal{B}_k^*)$ is larger than n. Let $N_n^* = \inf\{k \geqslant 1 \mid l^*(k) > n\} - 1$.

2. From the bootstrap data blocks generated at step 1, build a trajectory of length $n^* = l^*(N_n^*)$ by binding the blocks together

$$X^{*(n)} = (X_1^*, ..., X_{n^*}^*),$$

and compute the (A)RBB version of the U-statistic

$$U_n^*(h) = \frac{2}{n^*(n^*-1)} \sum_{1 \leq i < j \leq N_n^*} h(X_i^*, \ X_j^*),$$

and of its standardization as well: $\sigma_n^{*2}(h) = 4(N_n^*/n^*)^3 s_{N_n^*}^{*2}(h)$ with

$$\begin{aligned} s_{N_n^*}^{*2}(h) &= \frac{1}{N_n^*} \sum_{k=1}^{N_n^*} h_{1,-k}^{*2}(\mathcal{B}_k^*), \\ h_{1,-j}^*(b) &= \frac{1}{N_n^* - 1} \sum_{k=1, \ k \neq j}^{N_n^*} \omega_h(b, \mathcal{B}_k^*) - \frac{2}{N_n^*(N_n^*-1)} \sum_{1 \leq k < l \leq N_n^*} \omega_h(\mathcal{B}_k^*, \mathcal{B}_l^*), \end{aligned}$$

for all $b \in \mathbb{T}$.

3. Eventually, the (A)RBB estimate of the root $\mathbb{P}\{(U_n(h) - \mu(h))/\sigma_n \leq x\}$ is the distribution given by

$$H_{(A)RBB}(x) = \mathbb{P}^* \left\{ \sigma_n^{*-1}(U_n^*(h) - U_n(h)) \leq x \right\},$$

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

Of course, the bootstrap distribution estimate is in practice approximated by a Monte-Carlo scheme, by iterating the steps of the algorithm above.

4.3 Asymptotic validity of the (A)RBB for markovian U-statistics

The next theorem reveals that the (A)RBB is asymptotically correct under the hypotheses previously listed, paving the way for non Gaussian confidence interval construction in a valid asymptotic framework. It straightforwardly derives from the results stated in §3.3.

Theorem 16 ((A)RBB ASYMPTOTIC VALIDITY) *We have the following convergences in distribution.*

(i) (REGENERATIVE CASE) *Under the same assumptions as Theorem 12, the RBB distribution is asymptotically valid: as* $n \to \infty$,

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}^* \left\{ \sqrt{n^*} \frac{U_n^*(h) - U_n(h)}{\widehat{\sigma}_n(h)} \leq x \right\} - \mathbb{P}_\nu \left\{ \sqrt{n} \frac{U_n(h) - \mu(h)}{\sigma(h)} \leq x \right\} \right| = O_{\mathbb{P}_\nu}(n^{-1/2}),$$

(ii) (PSEUDO-REGENERATIVE CASE) *If, in addition, assumptions* $\mathbf{A_8} - \mathbf{A_{10}}$ *are fulfilled, we also have: as* $n \to \infty$,

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}^* \left\{ \sqrt{n^*} \frac{U_n^*(h) - U_n(h)}{\widehat{\sigma}_n^*(h)} \leq x \right\} - \mathbb{P}_\nu \left\{ \sqrt{n} \frac{U_n - \mu(h)}{\widehat{\sigma}_n(h)} \leq x \right\} \right| = O_{\mathbb{P}_\nu}(n^{-1/2}).$$

Remark 17 (HIGHER-ORDER ACCURACY) *We point out that, in contrast to the sample mean case (see [7]), we have not been able to prove the second order accuracy of the (A)RBB procedure, even up to* $o_{\mathbb{P}_\nu}(n^{-1/2})$ *only, when applied to a Markovian* U*-statistic. The argument would indeed involve an Edgeworth expansion for the distribution of such a statistic, a result that cannot be asserted unless major advances in the asymptotic analysis of i.i.d. sequences of* 1*-lattice random vectors have been made, refer to the final discussion of §3.3.*

5 An illustrative simulation result

In this section, we now illustrate the inference principles described and studied above through a simple numerical example: sampling data are drawn from a regenerative chain, such as the ones encountered in Operations Research for modeling queuing or storage systems, see [2]. Precisely, we analyze a simulated sequence $X_1, \ldots, X_n$ of waiting times in a $GI/G/1$ queuing system, in the absence of prior knowledge on the underlying model except the regenerative Markovian structure. Such a model may be classically viewed as a random walk on the half line, since one may write

$$X_{n+1} = (X_n + S_n - \Delta T_{n+1})_+,$$

where $x_+ = \max(x,\ 0)$ denotes the positive part of any real number $x \in \mathbb{R}$, $(\Delta T_n)_{n \geq 1}$ and $(S_n)_{n \geq 1}$ the sequences of interarrival and service times respectively, assumed i.i.d. and independent from each other. Suppose in addition that the mean interarrival time $\mathbb{E}[\Delta T_n] = 1/\lambda$ and the mean service time $\mathbb{E}[S_n] = 1/\theta$ are both finite and that the *load condition* "$\lambda/\theta < 1$" is fulfilled. The discrete-time process X is then a positive recurrent regenerative Markov chain with the "empty file" $A = \{0\}$ as a Harris recurrent atom, see §14.4.1 in [33]. In the case where interarrival and service times are both exponentially distributed, X is classically geometrically ergodic and has a limiting distribution μ with

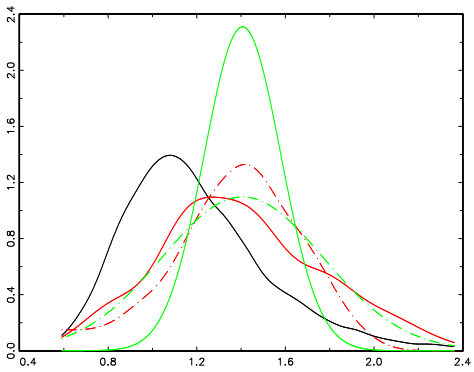

Figure 1: Comparison of bootstrap, gaussian and true distribution estimates of the distribution of σ_μ^2's U-estimator for the M/M/1 model: Monte-Carlo in black, percentile RBB in solid red, studentized RBB in dashed red, Gaussian with standard deviation σ_n in solid green, Gaussian with standard deviation σ_{BOOT} in dashed green.

exponential tails, refer to §16.1.3 in [33] (in particular, the moment assumption A_0 is thus fulfilled for any $q \geq 1$).

We considered the problem of computing confidence intervals for the variance $\sigma_\mu^2 = var_\mu(X_0)$ of the waiting time distribution in steady-state, from a sample path of length $n = 5000$ simulated from such a $M/M/1$ model, with parameters $\lambda = 0.2$ and $\theta = 0.8$. Based on the simulated trajectory, we computed the U-estimate of σ_μ^2 and the related normalizing constant, yielding $U_n = 1.407$ and $\widehat{\sigma}_n = 5.462$. We also generated $B = 199$ RBB replications, leading to bootstrap versions U_n^{*b}, with $b \in \{1, \ldots, B\}$, of the U-statistic, as well as the corresponding normalization constants $\widehat{\sigma}_n^{*b}$. A "bootstrap norming constant"

$$\widehat{\sigma}_{BOOT}^2 = \frac{1}{B} \sum_{b=1}^{B} \left(U_n^{*b} - \overline{U_n^*} \right)^2 ,$$

with $\overline{U_n^*} = \frac{1}{B} \sum_{b=1}^{B} U_n^{*b}$, has also been computed. The studentized bootstrap replicates are given by: $\forall b \in \{1, \ldots, B\}$,

$$t_n^{*b} = \frac{U_n^{*b} - U_n}{\widehat{\sigma}_n^{*b} / \sqrt{n^*}}$$

The four distribution estimates are displayed in Fig. 1, together with a Monte-Carlo estimate based on $M = 500$ replications of the simulation scheme. On this exemple, the RBB estimate (percentile version) clearly provides the best approximation, reflecting the high skewness of the target distribution. For instance, the corresponding two-sided confidence interval with level 95% leads to a coverage probability of 91.2%.

Appendix - Technical proofs

Proof of Proposition 4

The proof is based on the next result, a slight adaptation of Proposition 8 in [19].

Lemma 18 (Rosenthal inequality) *Let $p \geq 2$ and $f : E \to \mathbb{R}$ be a measurable function such that the expectations $\mathbb{E}_A[(\sum_{i=1}^{\tau_A} |f|(X_i))^p]$ and $\mathbb{E}_\nu[(\sum_{i=1}^{\tau_A} f(X_i))^p]$ are both finite. Rather than replacing f by $f - \mu(f)$, suppose that $\mu(f) = 0$. Then, there exists a constant $C_p < \infty$ such that: $\forall n \geq 1$,*

$$\frac{n^p}{3^{p-1}}\mathbb{E}_\nu\left[|S_n(f)|^p\right] \leq \left(\frac{p}{p-1}\right)^p C_p \left\{ n\mathbb{E}_A\left[\left|\sum_{i=1}^{\tau_A} f(X_i)\right|^p\right] + \left(n\mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A} f(X_i)\right)^2\right]\right)^{p/2} \right\}$$
$$\mathbb{E}_\nu\left[\left|\sum_{i=1}^{\tau_A} f(X_i)\right|^p\right] + \mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A} |f|(X_i)\right)^p\right].$$

One may take $C_p = 2^p$ or $7.35p/(1 \vee \log p)$ when $p > 2$ and $C_2 = 1/2$.

Proof. Based on the decompostion (4), we have

$$\frac{n^p}{3^{p-1}} \cdot \mathbb{E}_\nu\left[|S_n(f)|^p\right] \leq \mathbb{E}_\nu\left[\left|\sum_{i=1}^{\tau_A} f(X_i)\right|^p\right] + \mathbb{E}\left[\max_{1\leq l\leq n}\left|\sum_{j=1}^{l} f(\mathcal{B}_j)\right|^p\right] + \mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A} |f|(X_i)\right)^p\right]. \tag{14}$$

By L_p-Doob's inequality applied to the submartingale $\{|\sum_{j\leq l} f(\mathcal{B}_j)|\}_{l\geq 1}$ combined with Rosenthal inequality (see Theorem 2.9 in [35] for instance), we obtain that the middle term on the left hand side is bounded by $(p/(p-1))^p C_p\{n\mathbb{E}[|f(\mathcal{B}_1)|^{2p}] + (n\mathbb{E}[(f(\mathcal{B}_1))^2])^{p/2}\}$, while the first and last terms are finite by assumption. ■

For notational simplicity, we abusively denote the (possibly empty) non regenerative blocks of observations by $\mathcal{B}_0 = (X_1, \ldots, X_{\tau_A})$ and $\mathcal{B}_{l_n} = (X_{1+\tau_A(l_n)}, \ldots, X_n)$. Observe first that the event $\{l_n \leq 2\}$ occurs with a probability of order $O(n^{-2})$. Indeed, as it is included in $\{\tau_A > n/3\} \cup \{\tau_A(2) - \tau_A > n/3\} \cup \{\tau_A(3) - \tau_A(2) > n/3\}$, it immediately follows from the union bound combined with the moment assumptions $\mathbf{A_1} - \mathbf{A_2}$ that $\mathbb{P}_\nu(l_n \leq 2) = O(n^{-2})$. On the complementary event $\{l_n > 2\}$, one may then write

$$W_n(h) = (I) + (II),$$

where

$$(I) = \frac{2}{n(n-1)} \left\{ \sum_{k=0}^{l_n} \omega_{\tilde{h}}(\mathcal{B}_0, \mathcal{B}_k) + \sum_{k=1}^{l_n} \omega_{\tilde{h}}(\mathcal{B}_{l_n}, \mathcal{B}_k) \right\},$$
$$(II) = \frac{2}{n(n-1)} \left\{ \sum_{k=1}^{l_n - 1} \omega_{\tilde{h}}(\mathcal{B}_k, \mathcal{B}_k) - \sum_{i=1}^{n} \tilde{h}(X_i, X_i) \right\},$$

Therefore, we have $l_n/n \to \alpha^{-1}$ $\mathbb{P}_\nu$-a.s. and it follows from the SLLN for sums of i.i.d. random variables that $n^{-1} \sum_{k=1}^{l_n - 1} \omega_{\tilde{h}}(\mathcal{B}_k, \mathcal{B}_k) \to (\mathbb{E}[\omega_h(\mathcal{B}_1, \mathcal{B}_1)] - \mu(h)\mathbb{E}_A[\tau_A^2])/\alpha$ as $n \to \infty$ with probability one under $\mathbb{P}_\nu$. Observe in addition that $\mathbb{E}_A[\sum_{i=1}^{\tau_A} |h(X_i, X_i)|] \leq \mathbb{E}[\omega_{|h|}(\mathcal{B}_1, \mathcal{B}_1)] < \infty$, we may thus apply the SLLN for positive Harris chains (see Theorem 17.1.7 in [33]) and obtain that $n^{-1} \sum_{i=1}^{n} \tilde{h}(X_i, X_i) \to \int_{x \in E} h(x, x)\mu(dx) - \mu(h)$ $\mathbb{P}_\nu$-almost surely as $n \to \infty$. Hence, the term (II) almost-surely goes to zero under $\mathbb{P}_\nu$. Additionally, under the stipulated moment assumptions, we also have:

$$n^{-2}\mathbb{E}_\nu \left[\left(\sum_{k=1}^{l_n} \omega_{\tilde{h}}(\mathcal{B}_0, \mathcal{B}_k) \right)^2 \right] \leq \mathbb{E}_\nu \left[\left(\omega_{|\tilde{h}|}(\mathcal{B}_0, \mathcal{B}_1) \right)^2 \right],$$
$$\leq 2 \left\{ \mathbb{E}_\nu \left[\left(\omega_{|h|}(\mathcal{B}_0, \mathcal{B}_1) \right)^2 \right] + \mu(h)^2 \mathbb{E}_\nu[\tau_A]\mathbb{E}_A[\tau_A] \right\} < \infty.$$

Similarly, we have

$$n^{-2}\mathbb{E}_\nu \left[\left(\sum_{k=1}^{l_n} \omega_{\tilde{h}}(\mathcal{B}_k, \mathcal{B}_{l_n}) \right)^2 \right] \leq 2 \left\{ \mathbb{E} \left[\left(\omega_{|h|}(\mathcal{B}_1, \mathcal{B}_2) \right)^2 \right] + \mu(h)^2 (\mathbb{E}_A[\tau_A])^2 \right\} < \infty,$$

By a straightforward Borel-Cantelli argument, we obtain that the term (I) almost-surely converges to zero as $n \to \infty$ under $\mathbb{P}_\nu$.

The L_2-bound then follows from (the argument of) Lemma 18 applied to $f(x) = h(x, x)$.

Proof of Theorem 5

On the event $\{l_n \geq 2\}$, which occurs with probability $1 - O(n^{-2})$, we may write

$$U_n(\tilde{h}) = W_n(h) + \frac{(l_n - 1)(l_n - 2)}{n(n-1)} R_{l_n - 1}.$$

As $L \to \infty$, by virtue of the SLLN for U-statistics, we have $R_L(h) \to 0$ $\mathbb{P}_\nu$-almost surely, see [39] for instance. The desired result then follows from Proposition 4 combined with the fact that $\mathbb{P}_\nu\{\lim_{n \to \infty} l_n/n \to \alpha^{-1}\} = 1$.

Proof of Theorem 6

Observe that one may decompose $U_n(\tilde{h})$ as follows

$$W_n(h) + \frac{(l_n-1)(l_n-2)}{n(n-1)}\left\{2S_{l_n-1}(h) + D_{l_n-1}(h)\right\}.$$

Following in the footsteps of the argument of Theorem 17.2.2 in [33], we have, under $\mathbb{P}_\nu$, $(l_n/\sqrt{n})S_{l_n-1} \Rightarrow \mathcal{N}(0, s_h^2/\alpha)$ as $n \to \infty$, provided that $s_h^2 > 0$ (*cf* assumption $\mathbf{A_5}$). By virtue of assertion (ii) in Proposition 4, we have $\sqrt{n}W_n(h) \to 0$ in $\mathbb{P}_\nu$-probability as $n \to \infty$. We shall now prove that $D_{l_n-1} = o_{\mathbb{P}_\nu}(1/\sqrt{n})$ as $n \to \infty$. Let $\epsilon > 0$ and $\kappa \in]0, 1/2[$, write

$$\begin{aligned}\mathbb{P}_\nu\left\{\sqrt{n}|D_{l_n-1}| \geq \epsilon\right\} &\leq \sum_{l:\,|l-n|<n^{\kappa+1/2}} \mathbb{P}_\nu\left\{\sqrt{n}|D_l(h)| \geq \epsilon\right\} + \mathbb{P}_\nu\left\{|l_n - n/\alpha| \geq n^{\kappa+1/2}\right\} \\ &\leq \sum_{l:\,|l-n|<n^{\kappa+1/2}} \frac{n\mathbb{E}[D_l^2]}{\epsilon^2} + C\cdot n^{-2\kappa}\end{aligned}$$

by applying Lemma 18 with $f(x) = \mathbb{I}_{\{x\in A\}}$. Therefore, for all $l \geq 2$,

$$\mathbb{E}[D_l^2] = \frac{2}{l(l-1)}\mathbb{E}[(\omega_{\tilde{h}}(\mathcal{B}_1, \mathcal{B}_2))^2],$$

which, when combined to the previous bound, yields

$$\mathbb{P}_\nu\left\{\sqrt{n}|D_{l_n-1}| \geq \epsilon\right\} \leq 2n^{\kappa+1/2}\mathbb{E}[(\omega_{\tilde{h}}(\mathcal{B}_1, \mathcal{B}_2))^2]/((n - n^{\kappa+1/2})\epsilon^2) + C\cdot n^{-2\kappa}.$$

Since the bound on the right hand side goes to zero when $n \to \infty$, we obtain that $\sqrt{n}D_{l_n-1} \to 0$ in $\mathbb{P}_\nu$-probability. Eventually, one concludes the proof by observing that $l_n/n \to \alpha$ $\mathbb{P}_\nu$-almost-surely and applying Slutsky's lemma.

Proof of Theorem 9

Since $l_n \sim n/\alpha$ $\mathbb{P}_\nu$-almost-surely as $n \to \infty$, it follows from the LIL for non-degenerate U-statistics that

$$\limsup_{n\to\infty} \frac{\sqrt{n}R_{l_n-1}(h)}{\sqrt{8\alpha s_h^2 \log\log n}} = +1$$

with probability one under $\mathbb{P}_\nu$. Therefore, Proposition 4 combined with the fact that $\sigma_h^2 = 4s_h^2/\alpha^3$ entails that

$$\limsup_{n\to\infty} \frac{\sqrt{n}R_{l_n-1}(h)}{\sqrt{8\alpha s_h^2 \log\log n}} = \limsup_{n\to\infty} \frac{\sqrt{n}U_n(h)}{\sqrt{2\sigma_h^2 \log\log n}} \ \ \mathbb{P}_\nu\text{-almost surely},$$

which proves the desired result.

Proof of Proposition 10

For clarity's sake, we recall the following result, established at length in [30].

Lemma 19 *(*Malinovskii, 1985*) Let $f : E \to \mathbb{R}$ be a measurable function such that $\mu(|f|) < \infty$. Suppose in addition that $\mu(f) = 0$, the expectations $\mathbb{E}_A[(\sum_{i=1}^{\tau_A} f(X_i))^4]$, $\mathbb{E}_\nu[(\sum_{i=1}^{\tau_A} f(X_i))^2]$ and $\mathbb{E}_A[\tau_A^4]$ are finite and the Cramer condition*

$$\limsup_{t\to\infty} \mathbb{E}_A\left[\left|\exp\left(it\sum_{i=1}^{\tau_A} f(X_i)\right)\right|\right] < 1$$

is fulfilled. Then, there exists a constant C independent from the initial distribution ν such that: $\forall n \in \mathbb{N}^*$,

$$\mathbb{E}_\nu\left[\sum_{i=1}^{n} f(X_i)\right] = \phi_\nu(f) - \beta(f)/\alpha + \gamma(f) + C\cdot n^{-1/2},$$

where $\phi_\nu(f) = \mathbb{E}_\nu[\sum_{i=1}^{\tau_A} f(X_i)]$, $\gamma(f) = \mathbb{E}_A[\sum_{i=1}^{\tau_A}(\tau_A - i)f(X_i)]/\alpha$, $\beta(f) = \mathbb{E}_A[\tau_A \sum_{i=1}^{\tau_A} f(X_i)]$.

It should be noticed that, in the subsequent calculations, the constant C is not necessarily the same at each appearance. Rather than replacing h by $h - \mu(h)$, assume that $\mu(h) = 0$. One may write

$$\mathbb{E}_\nu\left[U_n(h)\right] = \frac{2}{n(n-1)}\cdot\mathbb{E}_\nu\left[\sum_{i=1}^{n-1}\mathbb{E}_{X_i}\left[\sum_{j=i+1}^{n} h(X_i, X_j)\right]\right]$$

We set, for all $x \in E$, $\int_{y\in E} h(x,y)\mu(dy) = h_1(x)/\alpha$ and let $\bar{h}(x,y) = h(x,y) - h_0(x)$. Taking $f(y) = \bar{h}(x,y)$, by virtue of Lemma 19 we obtain that: $\forall x \in E$, $\forall k \geq 1$,

$$\mathbb{E}_x\left[\sum_{j=1}^{k}\bar{h}(x, X_j)\right] = \mathbb{E}_x\left[\sum_{j=1}^{k} h(x, X_j)\right] - kh_0(x) = H(x) + C\cdot k^{-1/2},$$

where

$$H(x) = \mathbb{E}_x\left[\sum_{j=1}^{\tau_A}\bar{h}(x, X_j)\right] - \alpha^{-1}\mathbb{E}_A\left[\tau_A\sum_{j=1}^{\tau_A}\bar{h}(x, X_j)\right] + \alpha^{-1}\mathbb{E}_A\left[\sum_{j=1}^{\tau_A}(\tau_A - j)\bar{h}(x, X_j)\right].$$

Applying again Lemma 19 (to $f(x) = H(x) - \mu(H)$ this time), this yields: $\forall n \geq 1$,

$$\mathbb{E}_\nu[\sum_{1\leq i<j\leq n} h(X_i, X_j)] = \mathbb{E}_\nu\left[\sum_{i=1}^{n-1}(n-i)h_0(X_i)\right] + n\mu(H) + \phi_\nu(H) + \gamma(H) - \beta(H)/\alpha + C\cdot n^{1/2}, \tag{15}$$

where, by virtue of Kac's lemma and using the fact that $\mu(h_0) = 0$, we have:

$$\begin{aligned}
\mu(H) &= \frac{1}{\alpha}\mathbb{E}_A\left[\sum_{1\le k<j\le\tau_A} h(X_k, X_j)\right] - \frac{1}{\alpha}\mathbb{E}_A\left[\sum_{j=1}^{\tau_A}(\tau_A - j)\, h_0(X_j)\right] \\
&- \frac{1}{\alpha}\mathbb{E}_A\left[\tau_A\sum_{j=1}^{\tau_A} h_0(X_j)\right] + \frac{1}{\alpha}\mathbb{E}_A\left[\sum_{j=1}^{\tau_A}(\tau_A - j)h_0(X_j)\right] \\
&= \frac{1}{\alpha}\mathbb{E}_A\left[\sum_{1\le k<j\le\tau_A} h(X_k, X_j)\right] - \frac{1}{\alpha}\mathbb{E}_A\left[\tau_A\sum_{j=1}^{\tau_A} h_0(X_j)\right].
\end{aligned}$$

By applying Lemma 19 once again with $f(x) = h_0(x)$, we obtain that

$$\mathbb{E}_\nu\left[\sum_{i=1}^{n-1} h_0(X_i)\right] = \phi_\nu(h_0) - \beta(h_0)/\alpha + \gamma(h_0) + C\cdot n^{-1/2}. \tag{16}$$

Now, in order to prove that $\sum_{n\ge1}\mathbb{E}_\nu[nh_0(X_n)]$ is summable, notice that the condition

$$\mathbb{E}_A\left[\sum_{n=1}^{\tau_A} r(n)f(X_n)\right] < \infty$$

is fulfilled, when taking $f(x) = 1 + |h_0(x)|$ and $r(n) = n^3$. Indeed, by Hölder inequality we have:

$$\mathbb{E}_A\left[\sum_{i=1}^{\tau_A} r(i)f(X_i)\right] \le \mathbb{E}_A\left[\tau_A^4\right] + \left(\mathbb{E}\left[\tau_A^4\right]\right)^{3/4}\cdot\left(\mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}|h_0|(X_i)\right)^4\right]\right)^{1/4}.$$

The upper bound being finite under our assumptions, it follows from Theorem 2.1 in [44] that: as $n\to\infty$,

$$|\mathbb{E}_\nu\left[h_0(X_n)\right]| = |\mathbb{E}_\nu\left[h_0(X_n)\right] - \mu(h_0)| = o(r(n)).$$

Hence, we have $\sum_{i=1}^{\infty}|\mathbb{E}_\nu[ih_0(X_i)]| < \infty$. Eventually, combined with Eq. (15), (16) and (16), this permits us to conclude the proof.

Proof of Theorem 12

With no restriction, we may assume that $\mu(h) = 0$ for notational simplicity. We first recall the following result, which provides a very convenient way of establishing a Berry-Esseen bound for functionals of i.i.d. random variables through linearization. Refer to Lemma 1.3 in [40] for a detailed proof based on Stein's method.

Lemma 20 (Stein's lemma) *Let L_n be a random variable such that $\mathbb{E}\left[L_n^2\right] = 1$ and such that there exists some constant $C < \infty$ (possibly depending on L_n's distribution), the following Berry-Esseen type bound holds: $\forall n \geq 1$,*

$$\sup_{x\in\mathbb{R}} |\mathbb{P}\left\{L_n \leq x\right\} - \Phi(x)| \leq \frac{C}{\sqrt{n}}.$$

Then, for any sequence of random variables $\{\Delta_n\}_{n\geq 1}$, we have: $\forall n \geq 1$,

$$\sup_{x\in\mathbb{R}} |\mathbb{P}\{L_n + \Delta_n \leq x\} - \Phi(x)| \leq \frac{C}{\sqrt{n}} + 8\left(\mathbb{E}\left[\Delta_n^2\right]\right)^{1/2}.$$

We first state the following Berry-Esseen theorem, improving upon the result stated in [13] for sample mean statistics in the markovian setup (see Theorem 1 therein), insofar as the constant involved in the bound can be (over-) estimated, as may be shown by examining carefully its proof, which relies on Lemma 20 too and is postponed to the next subsection.

Theorem 21 (A Berry-Esseen bound for the sample mean statistic) *Suppose that the expectations $\mathbb{E}_\nu\left[\tau_A\right]$, $\mathbb{E}_A\left[\tau_A^3\right]$, $\mathbb{E}_\nu\left[\sum_{i=1}^{\tau_A} |f(X_i)|\right]$ and $\mathbb{E}_A\left[\sum_{i=1}^{\tau_A} |f(X_i)|^3\right]$ are all finite. Then, there exists a constant $K < \infty$ such that for all $n \geq 1$:*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}_\nu \left\{ \frac{1}{\sqrt{\gamma_f^2 n}} \sum_{i=1}^{n} \{f(X_i) - \mu(f)\} \leq x \right\} - \Phi(x) \right| \leq K n^{-1/2}, \tag{17}$$

where $\gamma_f^2 = \alpha^{-1}\mathbb{E}_A\left[(\sum_{i=1}^{\tau_A}\{f(X_i) - \mu(f)\})^2\right]$ is assumed to be strictly positive.

Consider now the following decomposition

$$\frac{\sqrt{n}}{\sigma(h)} U_n(h) = L_n + \Delta_n,$$

where

$$L_n = \frac{2}{\alpha} \frac{n^{-1/2}}{\sigma(h)} \sum_{i=1}^{n} h_1(X_i)$$

and $\sigma(h) n^{-1/2} \Delta_n = I + II + III + IV$ with

$$\begin{aligned}
I &= W_n(h), \\
II &= \frac{2}{n(n-1)} \sum_{1\leq k<l\leq l_n-1} h_2(\mathcal{B}_k, \mathcal{B}_l), \\
III &= 2\left(\frac{l_n - 2}{n-1} - \frac{1}{\alpha}\right) \times \frac{1}{n} \sum_{i=1+\tau_A(1)}^{\tau_A(l_n)} h_1(X_i), \\
IV &= -\frac{2}{\alpha}\frac{1}{n}\left\{ \sum_{i=1}^{\tau_A} h_1(X_i) + \sum_{i=1+\tau_A(l_n)}^{n} h_1(X_i) \right\}.
\end{aligned}$$

The second order moments of the r.v.'s I and IV are bounded by C/n for some properly chosen constant $C < \infty$, refer to Proposition 4's argument. Additionally, by examining Lemma 18's proof, one may establish that those of the two factors involved in the term III are bounded by C'/n for some constant $C' < \infty$ and, consequently, that of the r.v. III as well, by a straightforward Cauchy-Schwarz argument. Observe also that the second order moment of the variable II is bounded by $\mathbb{E}[\max_{1\leq l\leq n}(D_l(h))^2]$. Now, writing the (degenerate) U-statistic $D_l(h)$ as a martingale (see §5.1.5 in [39] for instance), L_2 Doob's inequality permits to show that the latter quantity is of order $O(n^{-2})$. Finally, we obtained that $\mathbb{E}[\Delta_n^2] \leq C''/n$ for some constant $C'' < \infty$.

The desired Berry-Esseen bound then follows from this variance control combined with Lemma 20 and Theorem 21.

Proof of Theorem 21

Repeating the argument of Theorem 12, combined with the partitioning technique mentioned in §2.3, one may obtain an explicit constant in the markovian version of the Berry-Esseen inequality for sample mean statistics, in contrast to the result established in [13] (see also [6]).

For notational simplicity, we suppose $\mu(f) = 0$. We write

$$S_n(f) = \frac{1}{n}\sum_{i=1}^{\tau_A} f(X_i) + \frac{1}{n}\sum_{j=1}^{l_n-1} f(\mathcal{B}_j) + \frac{1}{n}\sum_{i=1+\tau_A(l_n)+1}^{n} f(X_i).$$

Since $l_n \sim n/\alpha$ as $n \to \infty$ with $\mathbb{P}_\nu$-probability one, consider the r.v.

$$Z_n = n^{-1/2}\sum_{j=1}^{\lfloor n/\alpha\rfloor} f(\mathcal{B}_j)/\sigma(f)$$

and write $\sqrt{n}S_n(f)/\sigma(f) = Z_n + \Delta_n$. Since Z_n is a recentered and standardized sum of i.i.d. random variables with finite moment of order 3, the usual Berry-Esseen theorem applies and yields: $\forall n \geq 1$,

$$\sup_{x\in\mathbb{R}} |\mathbb{P}\{Z_n \leq x\} - \Phi(x)| \leq \gamma_B \cdot \frac{\kappa_3}{\sqrt{n}}. \tag{18}$$

The constant $\gamma_B = 0.7056$ has been recently obtained in [41]. Now, observe that

$$\frac{\sigma(f)}{n^{1/2}}\Delta_n = \frac{1}{n}\sum_{i=1}^{\tau_A} f(X_i) + \frac{1}{n}\sum_{\lfloor n/\alpha\rfloor-1}^{l_n-1} f(\mathcal{B}_j) + \frac{1}{n}\sum_{i=\tau_A(l_n-1)+1}^{n} f(X_i)$$

with the convention that $\sum_{\lfloor n/\alpha\rfloor-1}^{l_n-1} f(\mathcal{B}_j)$ is equal to $\sum_{j=\lfloor n/\alpha\rfloor}^{l_n-1} f(\mathcal{B}_j)$ if $l_n > \lfloor n/\alpha\rfloor$, to 0 if $l_n = \lfloor n/\alpha\rfloor$ and to $\sum_{j=l_n}^{\lfloor n/\alpha\rfloor-1} f(\mathcal{B}_j)$ if $l_n < \lfloor n/\alpha\rfloor$, when $l_n \geq 2$. We have

$$\frac{\sigma^2(f)}{3}\mathbb{E}\left[\Delta_n^2\right] \leq \frac{1}{n}\mathbb{E}_\nu\left[(f(\mathcal{B}_0))^2\right] + \mathbb{E}_\nu\left[\left(n^{-1/2}\sum_{\lfloor n/\alpha\rfloor-1}^{l_n-1} f(\mathcal{B}_j)\right)^2\right] + \frac{1}{n}\mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}|f|(X_i)\right)^2\right] \quad (19)$$

Thus, it essentially remains to control the term

$$\mathbb{E}_\nu\left[\left(\frac{1}{\sigma(f)n^{1/2}}\sum_{\lfloor n/\alpha\rfloor-1}^{l_n-1} f(\mathcal{B}_j)\right)^2\right] = I + II, \quad (20)$$

where

$$\begin{aligned} I &= \mathbb{E}_\nu\left[\left(\frac{1}{\sigma(f)n^{1/2}}\sum_{j=\lfloor n/\alpha\rfloor}^{l_n-1} f(\mathcal{B}_j)\cdot\mathbb{I}_{\{l_n>\lfloor n/\alpha\rfloor\}}\right)^2\right] \\ II &= \mathbb{E}_\nu\left[\left(\frac{1}{\sigma(f)n^{1/2}}\sum_{j=l_n}^{\lfloor n/\alpha\rfloor-1} f(\mathcal{B}_j)\cdot\mathbb{I}_{\{l_n<\lfloor n/\alpha\rfloor\}}\right)^2\right] \end{aligned}$$

The second term on the right hand side of Eq. (20) can be bounded as follows

$$II \leq \mathbb{E}_\nu\left[\left|\frac{l_n-\lfloor n/\alpha\rfloor}{n}\right|\right]\mathbb{E}_\nu\left[\left(|l_n-\lfloor n/\alpha\rfloor|^{-1/2}\sum_{\lfloor n/\alpha\rfloor-1}^{l_n-1}\frac{f(\mathcal{B}_j)}{\sigma(f)}\cdot\mathbb{I}_{\{l_n<\lfloor n/\alpha\rfloor\}}\right)^2\right]. \quad (21)$$

We control the first factor in the upper bound stated above using the continuous inclusion $L_2 \hookrightarrow L_1$ and Lemma 18 applied to $f(x) = \mathbb{I}_{\{x\in A\}} - \mu(A)$, as follows

$$\begin{aligned} \mathbb{E}_\nu\left[\left(\frac{l_n-\lfloor n/\alpha\rfloor}{n}\right)^2\right] &\leq 2\left\{\mathbb{E}_\nu\left[n^{-2}(l_n-n/\alpha)^2\right]+n^{-2}\right\} \\ &\leq \frac{24}{n}\mathbb{E}_A\left[\left(1-\frac{\tau_A}{\alpha}\right)^2\right] + \frac{6}{n^2}\mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}\left|\mathbb{I}_{\{X_i\in A\}}-1/\alpha\right|\right)^2\right] \\ &+ 2n^{-2}. \end{aligned} \quad (22)$$

The second factor involved in the upper bound (21) can be estimated using the partitioning technique used in [30] (see also [6]). This consists in viewing the event $\{l_n < \lfloor n/\alpha\rfloor\}$

as the union of the events $\mathcal{E}_{l,r,s} = \{\tau_A = r,\ \sum_{j=1}^{l-1} l(\mathcal{B}_j) = n-r-s,\ l(\mathcal{B}_l) > s\}$ with $l < \lfloor n/\alpha \rfloor$ and $1 \leq r,\ s \leq n$ (recall that we set $l(\mathcal{B}_j) = \tau_A(j+1) - \tau_A(j)$ for $j \geq 1$). Using the Markov property, this yields

$$\mathbb{E}_\nu \left[\left(|l_n - \lfloor n/\alpha \rfloor|^{-1/2} \sum_{\lfloor n/\alpha \rfloor - 1}^{l_n - 1} f(\mathcal{B}_j)/\sigma(f) \cdot \mathbb{I}_{\{l_n < \lfloor \frac{n}{\alpha} \rfloor\}} \right)^2 \right] = \sum_{l=1}^{\lfloor n/\alpha \rfloor - 1} \sum_{r=1}^{n} \sum_{s=1}^{n} a_{l,r,s}$$

with, for $1 \leq l < \lfloor n/\alpha \rfloor$ and $1 \leq r,\ s \leq n$,

$$\begin{aligned}
a_{l,r,s} &= \mathbb{P}_\nu\{\tau_A = r\}\mathbb{P}_A\{\tau_A > s\} \\
&\times \int x^2 \mathbb{P}_A \left\{ |l - \lfloor n/\alpha \rfloor|^{-1/2} \sum_{j=l}^{\lfloor n/\alpha \rfloor - 1} f(\mathcal{B}_j)/\sigma(f) \in dx, \sum_{j=1}^{l-1} l(\mathcal{B}_j) = n-r-s \right\} \\
&= \mathbb{P}_\nu\{\tau_A = r\}\mathbb{P}_A\{\tau_A > s\} \\
&\times \int x^2 \mathbb{P}_A \left\{ |l - \lfloor n/\alpha \rfloor|^{-1/2} \sum_{j=l}^{\lfloor n/\alpha \rfloor - 1} f(\mathcal{B}_j)/\sigma(f) \in dx \right\} \mathbb{P}_A \left\{ \sum_{j=1}^{l-1} l(\mathcal{B}_j) = n-r-s \right\} \\
&\leq \mathbb{P}_\nu\{\tau_A = r\}\mathbb{P}_A\{\tau_A > s\}\mathbb{P}_A \left\{ \sum_{j=1}^{l-1} l(\mathcal{B}_j) = n-r-s \right\}
\end{aligned}$$

using the fact that the random variable $|l - [n/\alpha]|^{-1/2} \sum_{j=l}^{\lfloor n/\alpha \rfloor - 1} f(B_j)/\sigma(f)$ is independent from $\sum_{i=1}^{l-1} l(B_j)$ and has variance 1. We thus obtained that the second factor involved in the upper bound (21) is smaller than 1. Combined with (22), this yields a control of the term II.

Turning now to the term I in (20), we use Cauchy-Schwarz's inequality again to get:

$$I \leq \mathbb{E}_\nu \left[\left| \frac{n - \tau_A(\lfloor n/\alpha \rfloor)}{n} \right| \right] \mathbb{E}_\nu \left[\left(\sum_{\lfloor n/\alpha \rfloor - 1}^{l_n - 1} \frac{f(\mathcal{B}_j)}{\sigma(f)|n - \tau_A(\lfloor n/\alpha \rfloor)|^{1/2}} \cdot \mathbb{I}_{\{l_n \geq \lfloor n/\alpha \rfloor\}} \right)^2 \right]. \quad (23)$$

For the second factor, by conditioning upon $\tau_A(\lfloor n/\alpha \rfloor)$ and using next the Markov prop-

erty, one gets:

$$\begin{aligned}
\mathbb{E}_\nu\left[\left(\sum_{\lfloor n/\alpha\rfloor-1}^{l_n-1}\frac{f(\mathcal{B}_j)}{\sigma(f)|l_n-\lfloor n/\alpha\rfloor|^{1/2}}\cdot\mathbb{I}_{\{l_n\geq\lfloor n/\alpha\rfloor\}}\right)^2\right] &= \sum_{k=0}^{n-\lfloor n/\alpha\rfloor}\mathbb{E}_A\left[\left(\sum_{j=1}^{l_k-1}\frac{f(\mathcal{B}_j)}{k^{1/2}\sigma(f)}\right)^2\right]\\
&\times\ \mathbb{P}_\nu\left\{\tau_A(\lfloor n/\alpha\rfloor)=n-k\right\}\\
&\leq \sum_{k=0}^{n-\lfloor n/\alpha\rfloor}\mathbb{E}_A\left[\max_{1\leq l\leq k}\left(\sum_{j=1}^{l}\frac{f(\mathcal{B}_j)}{k^{1/2}\sigma(f)}\right)^2\right]\\
&\times\ \mathbb{P}_\nu\left\{\tau_A(\lfloor n/\alpha\rfloor)=n-k\right\}\\
&\leq 2\sum_{k=0}^{n-\lfloor n/\alpha\rfloor}\mathbb{P}_\nu\left\{\tau_A(\lfloor n/\alpha\rfloor)=n-k\right\}\leq 2,
\end{aligned}$$

using L_2-Doob's inequality.

For showing that the first factor is smaller than $C/\sqrt{n}$ for some constant $C<\infty$, it suffices to observe that $\tau_A(\lfloor n/\alpha\rfloor)=\tau_A+\sum_{j=1}^{\lfloor n/\alpha\rfloor-1}\{\tau_A(j+1)-\tau_A(j)\}$ can be written as a sum of independent r.v.'s and use the continuous inclusion $L_2\hookrightarrow L_1$.

Now, noticing that

$$n^{1/2}(\mu_n(f)-\mu(f))=W_n+\Delta_n,$$

apply lemma 18 with the bounds (18), (19) to get the desired result.

Proof of Proposition 13

It suffice to observe that the preliminary results stated in [17] imply that, as $L\to\infty$, $\hat{s}_L^2(h)\to s^2(h)=\mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}h_1(X_i)\right)^2\right]$ with probability one. Assertion (i) thus follows from the fact that l_n/n almost surely converges to $1/\alpha$ as $n\to\infty$. The second assertion immediately results from the first one combined with Theorem 6 and Slutsky's lemma.

Proof of Proposition 14

The proof immediately results from the CLT stated in Theorem 6 applied to the split chain combined with the following result.

Proposition 22 *Suppose that assumptions of Proposition 14 are fulfilled. We then have: as $n\to\infty$,*

$$\widetilde{\sigma}_n^2(h)\to\sigma^2(h)\ \textit{in}\ \mathbb{P}_\nu\textit{-probability.}$$

Proof. For simplicity, we assume that $\mu(h)=0$. Let $\widehat{l}_n=\widehat{N}_n+1$ be the number of times $\widehat{Y}$ visits the state 1 between time 1 and n. We also introduce versions of the quantities

(10) and (11) based on the pseudo-regenerative blocks: $\forall b \in \mathbb{T}$, $\forall j \in \{1, \ldots, \widehat{l}_n - 1\}$,

$$\tilde{h}_{1,-j}(b) = \frac{1}{\widehat{l}_n - 2} \sum_{k=1,\, k\neq j}^{\widehat{l}_n - 2} \omega_h(b, \widehat{B}_k) - \frac{2}{(\widehat{l}_n - 1)(\widehat{l}_n - 2)} \sum_{1 \leq k < l \leq \widehat{l}_n - 1} \omega_h(\widehat{B}_k, \widehat{B}_l),$$

and

$$\tilde{s}^2_{\widehat{l}_n - 1}(h) = \frac{1}{\widehat{l}_n - 1} \sum_{k=1}^{\widehat{l}_n - 1} \tilde{h}_{1,-k}(\widehat{B}_k).$$

The proof re-uses the coupling-based argument of Theorems 3.1 and 3.2 in [7]. Recall first that, by virtue of Lemma 4.3 in [7], we have $\widehat{l}_n / n = 1/\alpha + o_{\mathbb{P}_\nu}(\alpha_n^{1/2})$ as $n \to \infty$, where α is the mean length of the split chain cycles $\mathcal{B}_j$, $j \geq 1$. We next establish the following result.

Lemma 23 *Under the assumptions of Proposition 14, we have, for all $j \leq \widehat{l}_n - 1$,*

$$\frac{1}{\widehat{l}_n - 1} \sum_{j=1}^{\widehat{l}_n - 1} \left\{ \widetilde{h}^2_{1,-j}(\widehat{\mathcal{B}}_j) - h_1^2(\widehat{\mathcal{B}}_j) \right\} \to 0 \text{ in } \mathbb{P}_\nu\text{-probability as } n \to \infty.$$

Proof. Recall that $\omega_h(b_1, b_2) = h_1(b_1) + h_1(b_2) + h_2(b_1, b_2)$ for all $(b_1, b_2) \in \mathbb{T}^2$, since we assumed $\mu(h) = 0$ here. Hence, we have: $\forall j \leq \widehat{l}_n - 1$,

$$\widetilde{h}_{1,-j}(\widehat{\mathcal{B}}_j) = h_1(\widehat{\mathcal{B}}_j) + \frac{1}{\widehat{l}_n - 2} \sum_{k \neq j} h_1(\widehat{\mathcal{B}}_k) + \frac{1}{\widehat{l}_n - 2} \sum_{k \neq j} h_2(\widehat{\mathcal{B}}_j, \widehat{\mathcal{B}}_k) - \frac{2}{(\widehat{l}_n - 1)(\widehat{l}_n - 2)} \sum_{k<l} \omega_h(\widehat{\mathcal{B}}_l, \widehat{\mathcal{B}}_k).$$

Lemma 4.2 in [7] entails that, as $n \to \infty$, the sum of the second and third terms on the right hand side of the equation above is equal to

$$\frac{1}{l_n - 2} \sum_{k \neq j} h_1(\mathcal{B}_k) + \frac{1}{l_n - 2} \sum_{k \neq j} h_2(\widehat{\mathcal{B}}_j, \mathcal{B}_k) + o_{\mathbb{P}_\nu}(n^{-1} \alpha_n^{1/2}),$$

and thus tends to zero in $\mathbb{P}_\nu$-probability as $n \to \infty$, by virtue of the LLN and using the fact that $\mathbb{E}_A[h_1(\mathcal{B}_1)] = \mathbb{E}[h_2(b, \mathcal{B}_1)] = 0$ for any $b \in \mathbb{T}$. In addition, notice that

$$\frac{1}{n^2} \sum_{k<l} \omega_h(\widehat{\mathcal{B}}_l, \widehat{\mathcal{B}}_k) = \frac{1}{n^2} \sum_{\widehat{\tau}_1 < i < m \leq \widehat{\tau}_2} h(X_i, X_m) - \frac{1}{n^2} \sum_{k=1}^{\widehat{l}_n - 1} \omega_h(\widehat{\mathcal{B}}_k, \widehat{\mathcal{B}}_k),$$

where $\widehat{\tau}(1) = \inf\{k \in \{1, \ldots, n\} : \widehat{Y}_n = +1\}$ and $\widehat{\tau}(\widehat{l}_n) = \max\{k \in \{1, \ldots, n\} : \widehat{Y}_n = +1\}$. Using Lemma 4.1 combined with the boundedness assumption for h, one gets that the

first term on the right hand side of the equation above is equal to $U_n(h) + O_{\mathbb{P}_\nu}(\alpha_n^{1/2}/n)$, and thus converges to $\mu(h) = 0$ in $\mathbb{P}_\nu$-probability. The second term is bounded in absolute value by $||h||_\infty n^{-2} \sum_{k=1}^{\widehat{l}_n - 1} \mathcal{L}^2(\widehat{\mathcal{B}}_j)$, where $\mathcal{L}(b)$ denotes the length of any block $b \in \mathbb{T}$. The second assertion of Lemma 4.2 in [7] applied to $g(x) \equiv 1$ implies that this bound is equal to $||h||_\infty n^{-2} \sum_{k=1}^{l_n - 1} \mathcal{L}^2(\mathcal{B}_j) + O_{\mathbb{P}_\nu}(\alpha_n/n)$ as $n \to \infty$ and thus tends to zero, since $n^{-1} \sum_{k=1}^{l_n - 1} \mathcal{L}^2(\mathcal{B}_j) \to \mathbb{E}[\mathcal{L}^2(\mathcal{B}_1)]$ $\mathbb{P}_\nu$-almost-surely. The result stated in the lemma then immediately follows. ■

Using again Lemma 4.2 in [7], we obtain that

$$\frac{1}{\widehat{l}_n - 1} \sum_{j=1}^{\widehat{l}_n - 1} h_1^2(\widehat{\mathcal{B}}_j) = \frac{1}{l_n - 1} \sum_{j=1}^{l_n - 1} h_1^2(\mathcal{B}_j) + O_{\mathbb{P}_\nu}(\alpha_n).$$

The desired convergence then results from Lemma 23 combined with the fact that, as $n \to \infty$, $(l_n - 1)^{-1} \sum_{j=1}^{l_n - 1} h_1^2(\mathcal{B}_j) \to \mathbb{E}[h_1^2(\mathcal{B}_1)] = \alpha^3 \sigma^2(h)/4$ $\mathbb{P}_\nu$-almost-surely. ■

Proof of Theorem 16

This is a consequence of Theorem 12. We follow the standard argument used in the i.i.d. case when a Berry-Esseen bound is available under adequate moment assumptions, see [29] for instance. A Berry-Esseen bound naturally holds for the bootstrap distribution too. As the constant involved in this bound only depends on moments taken with respect to the empirical distribution of the (pseudo-) blocks, which converge to the empirical counterparts at the rate $O(n^{-1/2})$, the rate of convergence to the asymptotic distribution is of the same order and the theorem follows.

References

[1] J. Aaronson, R. Burton, H.G. Dehling, D. Gilat, T. Hill, and B. Weiss. Strong laws for L- and U-statistics. *Trans. Amer. Math. Soc.*, 348:2845–2866, 1996.

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

[3] K.B. Athreya and G.S. Atuncar. Kernel estimation for real-valued Markov chains. *Sankhya*, 60(1):1–17, 1998.

[4] N.G. Becker and S. Utev. Threshold results for U-statistics of dependent binary variables. *J. Theoret. Probab.*, 14(1):97–114, 2001.

[5] W. Bednorz and K. Latuszynski and R. Latala. A Regeneration Proof of the Central Limit Theorem for Uniformly Ergodic Markov Chains *Elect. Comm. Probab.*, 13:85-98, 2008.

[6] P. Bertail and S. Clémençon. Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains. *Prob. Th. Rel. Fields*, 130(3):388–414, 2004.

[7] P. Bertail and S. Clémençon. Regenerative-block bootstrap for Markov chains. *Bernoulli*, 12(4), 2005.

[8] P. Bertail and S. Clémençon. Regeneration-based statistics for Harris recurrent Markov chains. In P. Doukhan P. Bertail and P. Soulier, editors, *Probability and Statistics for dependent data*, number 187 in Lecture notes in Statistics, pages 1–54. Springer, 2006.

[9] P. Bertail, S. Clémençon, and J. Tressou. Extreme value statistics for Markov chains via the (pseudo-)regenerative method. *Extremes*, 2009. In Press. Preprint available at `http://hal.archives-ouvertes.fr/hal-00165652`.

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

[11] P. Bertail and S. Clémençon. Sharp bounds for the tails of functionals of Markov chains. *Th. Prob. Appl.*, 53(3), 2009. In press.

[12] P. Bertail, S. Clémençon, and J. Tressou. Regenerative Block-Bootstrap Confidence Intervals for the Extremal Index of Markov Chains. In *Proceedings of the International Workshop in Applied Probability*, 2008. Available at `http://hal.archives-ouvertes.fr/hal-00214306/fr/`.

[13] E. Bolthausen. The Berry-Esseen theorem for functionals of discrete Markov chains. *Z. Wahrsch. Verw. Geb.*, 54(1):59–73, 1980.

[14] S. Borovkova, R. Burton, and H.G. Dehling. Consistency of the Takens estimator for the correlation dimension. *Ann. Appl. Probab.*, 9(2):376–390, 1999.

[15] S. Borovkova, R. Burton, and H.G. Dehling. From dimension estimation to asymptotics of dependent U-statistics. In I. Berkes, E. Csaki, and M. Csorgo, editors, *Limit Theorems in Probability and Statistics I*, pages 201–234, Budapest, 1999.

[16] S. Borovkova, R. Burton, and H.G. Dehling. Limit theorems for functionals of mixing processes with applications to U -statistics and dimension estimation. *Trans. Amer. Math. Soc.*, 353:4261–4318, 2001.

[17] H. Callaert and N. Veraverbeke. The order of the normal approximation for a studentized U-statistic. *Ann. Statist.*, 9(1):194–200, 1981.

[18] S. Clémençon. Adaptive estimation of the transition density of a regular Markov chain by wavelet methods. *Math. Meth. Statist.*, 9(4):323–357, 2000.

[19] S. Clémençon. Moment and probability inequalities for sums of bounded additive functionals of a regular Markov chains via the Nummelin splitting technique. *Statistics and Probability Letters*, 55:227–238, 2001.

[20] H.G. Dehling. *Limit Theorems for Dependent U-statistics*, volume 187 of *Lecture Notes in Statistics*, pages 65–86. 2006.

[21] M. Denker and G. Keller. On U-statistics and von Mises statistics for weakly dependent processes. *Z. Wahrsch. Verw. Gebiete*, 64:505–522, 1983.

[22] M. Denker and G. Keller. Rigorous statistical procedures for data from dynamical systems. *J. Stat. Phys.*, 44:67–93, 1986.

[23] R. Douc, A. Guillin, and E. Moulines. Bounds on regeneration times and limit theorems for subgeometric markov chains. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 44(2):239–257, 2008. 0246-0203.

[24] Dubinskaite. Limit theorems in $\mathbb{R}^k$. *Lithuanian Math. J.*, 24:256–265, 1984.

[25] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

[26] J. Hajek. Asymptotically most powerful rank tests. *Ann. Math. Statist.*, 33:1124–1147, 1962.

[27] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Statist.*, 19:293–325, 1948.

[28] V.V. Kalashnikov. *Topics on regenerative processes*. CRC Press, 1994.

[29] S.N. Lahiri. *Resampling Methods for Dependent Data*. Springer-Verlag, 2003.

[30] V.K. Malinovskiĭ. On some asymptotic relations and identities for Harris recurrent Markov chains. In *Statistics and Control of Stochastic Processes*, pages 317–336, 1985.

[31] V.K. Malinovskiĭ. Limit theorems for Harris Markov chains I. *Theory Prob. Appl.*, 31:269–285, 1987.

[32] V.K. Malinovskiĭ. Limit theorems for Harris Markov chains II. *Theory Prob. Appl.*, 34:252–265, 1989.

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

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

[35] V.V. Petrov *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford Science Publications, 1995.

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

[37] E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*. Mathématiques et Applications 31. Springer. 2000.

[38] C.P. Robert and R. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 1999.

[39] R.J. Serfling. *Approximation theorems of mathematical statistics*. Wiley, New York, 1980.

[40] G. Shorack. *Probability for Statisticians*. Springer, 2000.

[41] I.G. Shevtsova Sharpening of the Upper Bound of the Absolute Constant in the BerryEsseen Inequality. *Theory Probab. Appl.*, 51(3): 549-553, 2007.

[42] W. L. Smith. Regenerative stochastic processes. *Proc. Royal Stat. Soc.*, 232:6–31, 1955.

[43] H. Thorisson. *Coupling, Stationarity, and Regeneration*. Springer, New York, 2000.

[44] P. Tuominen and R. L. Tweedie. Subgeometric rates of convergence of f-ergodic Markov chains. *Adv. Appl. Probab.*, 26:775–798, 1994.

[45] K. Yoshihara. Limiting behavior of U-statistics for stationary, absolurely regular processes. *Z. Wahrsch. Verw. Gebiete*, 35:237–252, 1976.