

APPROXIMATE REGENERATIVE BLOCK-BOOTSTRAP FOR MARKOV CHAINS.

Patrice Bertail and Stephan Clémençon

Key words: Bootstrap, Markov Chains, Edgeworth expansions, Nummelin splitting, algorithms. *COMPSTAT 2004 section*: Resampling method.

1 Introduction

Prolongating ideas introduced in Datta & McCormick (1993), Bertail & Clémençon (2003a,b) proposed a general resampling method, namely the *Regenerative Block Bootstrap* (RBB in abbreviated form), for bootstrapping statistics based on data $X_1, ..., X_n$ drawn from (eventually nonstationary) regenerative Markov chains. When the chain (positive Harris recurrent) possesses a known atom, they proved that this resampling method is second order correct up to $O_P(n^{-1})$ in the case of the studentized sample mean statistic under specific Cramer and "block moment" conditions (less restrictive than the exponential strong mixing rate condition generally assumed when the matter is to deal with dependent data). This is the optimal rate that may be attained by the naive Bootstrap method in the i.i.d. case (see Hall (1992)). These results should be put in contrast with the usual rates that may be attained by the *Moving Block Bootstrap* (MBB), which are at best $O_P(n^{-3/4})$ (see Götze & Künsch (1996)). We emphasize that the RBB straightforwardly applies to numerous specific regenerative models, widely used in the modeling of queuing and storage systems, and to all countable Markov chains. Resting on the theoretical construction introduced by Nummelin (1978), namely the *Nummelin splitting technique*, which is based on the crucial notion of *small set* (cf Meyn & Tweedie (1996)), any general Harris Markov chain could be considered as regenerative in the sense of the existence of a regenerative extension. Bertail & Clémençon (2003b) proposed a resampling procedure, the *Approximate Regenerative Block Bootstrap* (ARBB), that generalizes the RBB method and applies to all Harris Markov chains. The method is based on the prior knowledge of a small set for the chain and a practical approximation of the *Nummelin splitting extension*. It thus consists in using an empirical method to build approximatively a realization drawn from a regenerative extension of the chain and in applying the RBB methodology to the latter.

In this paper we propose a modification of the original ARBB algorithm based on the "2-split" method considered by Schick (2001). We also show how the asymptotic results obtained for the RBB in the regenerative case may be extended to this modified ARBB procedure at the cost of some small loss in the Edgeworth expansions, which is closely linked to the uniform rate for estimating the transition kernel of the chain over a well chosen small set. The outline is as follows. In section 2 the principles of the ARBB are

briefly recalled and a modification of the original method using a variant of the "2-split" trick is presented. In section 3 an asymptotic result claiming the second order asymptotic validity of this ARBB method for studentized sample mean statistics is stated. Finally, in section 4, practical selection rules for the tuning parameters of the algorithm are proposed and some simulation results are presented.

2 Nummelin splitting approximation and ARBB

2.1 Notation and basic notions

Here and throughout we shall use the same notations as in section 2 of Bertail & Clémençon (2003b). Consider $X = (X_n)_{n\in\mathbb{N}}$ a positive recurrent Markov chain on a countably generated state space $(E, \mathcal{E})$ with transition probability $\Pi(.,.)$, stationary probability measure μ and initial distribution ν. We denote by P_ν (respectively P_x for x in E, resp. P_A for $A \in \mathcal{E}$) the probability measure on the underlying space such that $X_0 \sim \nu$ (resp. conditionally to $X_0 = x$, resp. conditionally to $X_0 \in A$), by $E_\nu(.)$ the P_ν-expectation (resp. by $E_x(.)$ the P_x-expectation, resp. by $E_A(.)$ the P_A-expectation) and by $I\{\mathcal{A}\}$ the indicator function of the event $\mathcal{A}$.

We recall that a set $S \in \mathcal{E}$ is said to be *small* (see Meyn & Tweedie (1996)) if there exist $k \in \mathbb{N}$, a probability measure Φ supported by S, and $\delta > 0$ such that $\forall x \in S, \forall A \in \mathcal{E}, \quad \Pi^k(x, A) \geq \delta\Phi(A)$, denoting by Π^k the k-th iterate of Π (recall that small sets always exist for irreducible chains). When this holds, we shall say that X satisfies the minorization condition $\mathcal{M}(k, S, \delta, \Phi)$. Even if it entails to replace the chain $(X_n)_{n\in\mathbb{N}}$ by $\big((X_{nk}, ..., X_{n(k+1)-1})\big)_{n\in\mathbb{N}}$, we suppose $k = 1$ in what follows. We 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 $x \in E$. In this case, the condition $\mathcal{M}(k, S, \delta, \Phi)$ entails that Φ is also absolutely continuous with respect to λ and $p(x, y) \geq \delta\phi(y)$, $\lambda(dy)$ a.s, for any $x \in S$, with $\Phi(dy) = \phi(y)d\lambda(y)$. We assume

$\mathcal{H}_0$: The chain X satisfies condition $\mathcal{M}(1, S, \delta, \Phi)$ for some known parameters $S \in \mathcal{E}$ such that $\mu(S) > 0$, $\delta > 0$ and probability $\Phi(dy) = \phi(y)d\lambda(y)$ supported by S such that $\inf_{y\in S} \phi(y) > 0$.

The Nummelin splitting technique consists in constructing a bivariate Markov chain $X^{\mathcal{M}} = ((X_n, Y_n))_{n\in\mathbb{N}}$, called the *split chain*, taking its values in the state space $E \times \{0, 1\}$. This construction entails that, conditionally to $X^{(n+1)} = (X_1, ..., X_{n+1})$, the Y_i's, $1 \leqslant i \leqslant n$, are independent Bernoulli r.v.'s. The Bernoulli parameter is δ, unless X has hit the small set S at time i. And in the case when $X_i \in S$, Y_i is drawn from the Bernoulli distribution with parameter $\delta\phi(X_{i+1})/p(X_i, X_{i+1})$. We denote by $\mathcal{L}^{(n)}(p, S, \delta, \phi, X^{(n+1)})$ the probability distribution of $Y^{(n)} = (Y_1, ..., Y_n)$ conditionally to $X^{(n+1)}$, which is simply the tensor product of these Bernoulli distributions. The whole point of the construction consists in the fact that $A_{\mathcal{M}} = S \times \{1\}$ is an atom for the

split chain $X^{\mathcal{M}}$, which inherits all the communication and stochastic stability properties from X. In particular, the sample path of $X^{\mathcal{M}}$ can be classically divided into *regeneration blocks* corresponding to the blocks of observations between successive visits of the split chain to $A_{\mathcal{M}}$, which are i.i.d. r.v.'s valued in the torus $T = \cup_{n=1}^{\infty} E^n$, by virtue of the strong Markov property. For a given time $m^* \in \mathbb{N}$ that will be fixed later, we shall here consider the regeneration times (*i.e.* the times i at which $X_i \in S$ and $Y_i = 1$) posterior to m^*, which are denoted by $\tau_{m^*} \doteq \tau_{m^*}(1) = \inf\{k \geqslant m^* + 1 / \ X_k \in S, Y_k = 1\}$, $\tau_{m^*}(j) = \inf\{k > \tau_{m^*}(j-1) / \ X_k \in S, Y_k = 1\}$ for $j \geqslant 2$. We denote by $l_{m^*,n} = \sum_{i=m^*+1}^{n} I\{X_i \in S, Y_i = 1)$ the number of visits to the set $A_{\mathcal{M}} = S \times \{1\}$ between time $m^* + 1$ and time n. The corresponding regeneration blocks are denoted by $\mathcal{B}_{0,m^*} = (X_{m+1}, ..., X_{\tau_{m^*}(1)})$, $\mathcal{B}_{1,m*} = (X_{\tau_{m^*}(1)+1}, ..., X_{\tau_{m^*}(2)}), ..., \mathcal{B}_{l_{m^*,n}-1,m^*} = (X_{\tau_{m^*}(l_{m^*,n}-1)+1}, ..., X_{\tau_{m^*}(l_{m^*,n})})$, $\mathcal{B}^{(n)}_{l_{m^*,n},m^*} = (X_{\tau_{m^*}(l_{m^*,n})+1}, ..., X_n)$.

2.2 Approximate Nummelin splitting construction

Of course these blocks are practically unknown since their construction explicitly depends on the unknown transition density $p(x, y)$ (see § 2.1). The proposal of Bertail & Clémençon (2003b) for approximating this construction consists in using an estimate $p_n(x, y)$ of the transition density computed from data $X_1, ..., X_n$ to generate a random vector $(\widehat{Y}_1, ..., \widehat{Y}_n)$, conditionally to $X^{(n+1)}$, drawn from the distribution $\mathcal{L}^{(n)}(p_n, S, \delta, \phi, X^{(n+1)})$. However this estimation step induces strong dependency problems that make the second order properties of the ARBB procedure very difficult to study, when applied to the data $(X_1, \widehat{Y}_1), ..., (X_n, \widehat{Y}_n)$. Here we propose a modification of the method based on the well known semiparametric "splitting trick".

Given the data $X^{(n+1)}$, keep the first m observations $X^{(m)} = (X_1, ..., X_m)$ only to compute an estimate $p_m(x, y)$ of $p(x,\ y)$ such that $p_m(x, y) \geq \delta\phi(y)$, $\lambda(dy)$ a.s. and $p_m(X_i, X_{i+1}) > 0$, $1 \leqslant i \leqslant n$. To ensure that the observations $X^{(m^*,n+1)} = (X_{m^*+1}, ..., X_{n+1})$, which shall be used for forming the pseudo-regeneration blocks to resample, are independent from the first m observations (*i.e.* that a regeneration, or equivalently a visit of $X^{\mathcal{M}}$ to $A_{\mathcal{M}}$, occurs between time $m+1$ and time m^*) with overwhelming probability, we separate them by a small gap of length p. We will typically choose $m^* = m + p$ with m, p and m^* depending on n such that $p = O(m)$ as $n \to \infty$. This procedure is very similar to the *2-split method* proposed in Schick (2001), except that the user is here free to pick the exact number p of deleted observations, within the limits of the previous asymptotic constraint. In the following, we take $m \to \infty$ as $n \to \infty$, so as to get a consistent estimator $p_m(x, y)$, at a rate sufficiently slow (typically such that $\frac{m}{n} \to 0$ as $n \to \infty$) to ensure that the number of pseudo-blocks to resample also tends to infinity as $n \to \infty$.

Conditionally to $X^{(n+1)}$, draw then a vector $(\widehat{Y}_{m^*+1}, ..., \widehat{Y}_n)$ from the dis-

tribution estimate $\mathcal{L}^{(n-m^*)}(p_m, S, \delta, \phi, X^{(m^*,n+1)})$. From a practical viewpoint, it actually suffices to draw the $\widehat{Y}_i$'s at times i when the chain visits the set S (*i.e.* when $X_i \in S$), which are the only time points at which the split chain may regenerate: at such a time i, draw $\widehat{Y}_i$ according to the Bernoulli law with parameter $\delta\phi(X_{i+1})/p_m(X_i, X_{i+1})$). Count then the number of visits $\widehat{l}_{m^*,n} = \sum_{i=m^*+1}^{n} I\{X_i \in S, \widehat{Y}_i = 1)$ to $A_{\mathcal{M}} = S \times \{1\}$ between time $m^* + 1$ and time n and divide the truncated sample path $X^{(m^*,n)}$ into $\widehat{l}_n + 1$ blocks, corresponding to the pieces of the data segment between consecutive visits to $A_{\mathcal{M}}$, $\widehat{\mathcal{B}}_{0,m^*} = (X_{m^*+1}, ..., X_{\widehat{\tau}_{m^*}(1)})$, $\widehat{\mathcal{B}}_{1,m^*} = (X_{\widehat{\tau}_{m^*}(1)+1}, ..., X_{\widehat{\tau}_{m^*}(2)}), ...,$ $\widehat{\mathcal{B}}^{(n)}_{\widehat{l}_{m^*,n},m^*} = (X_{\widehat{\tau}_{m^*}(\widehat{l}_{m^*,n})+1}, ..., X_n)$ with $\widehat{\tau}_{m^*}(0) = m^*$ and for any $j \geqslant 1$, $\widehat{\tau}_{m^*}(j) = \inf\left\{k > \widehat{\tau}_{m^*}(j-1),\ X_k \in S, \widehat{Y}_k = 1\right\}$. For convenience, denote by $l(\widehat{\mathcal{B}}_{j,m^*}) = \widehat{\tau}_{m^*}(j+1) - \widehat{\tau}_{m^*}(j)$ the length of the block $\widehat{\mathcal{B}}_{j,m^*}$, $j \geqslant 1$.

2.3 Approximate Regenerative Block Bootstrap

Let $T_{n+1} = T_{n+1}(X^{(n+1)})$ be a statistic of interest and $S_{n+1} = S_{n+1}(X^{(n+1)})$ be an adequate standardization of the latter. The modified ARBB algorithm (which we still call ARBB algorithm for the sake of the simplicity) consists then in applying the RBB procedure in the following manner.

1. Draw sequentially bootstrap data blocks $\mathcal{B}^*_1, ..., \mathcal{B}^*_k$ independently from the empirical distribution $F_{m^*,n} = (\widehat{l}_{\,m^*,n} - 1)^{-1} \sum_{j=1}^{\widehat{l}_{m^*,n}-1} \delta_{\widehat{\mathcal{B}}_{j,m^*}}$ of the blocks $\widehat{\mathcal{B}}_{1,m^*}, ..., \widehat{\mathcal{B}}_{\widehat{l}_{m^*,n}-1,m^*}$ conditioned on $X^{(n+1)}$, 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 these bootstrap data blocks, reconstruct a pseudo-trajectory by binding the blocks together, getting the reconstructed *ARBB sample path* $X^{*(n)} = (\mathcal{B}^*_1, ..., \mathcal{B}^*_{l^*_n - 1})$. Then compute the *ARBB statistic* $T^*_n = T_{L^*(l^*_n)}(X^{*(n)})$ and the *ARBB standardization* $S^*_n = S_{L^*(l^*_n)}(X^{*(n)})$.

3. The *ARBB distribution* is then given by $H_{ARBB}(x) = P^*(S^{*-1}_n(T^*_n - T_{n+1}) \leqslant x \mid X^{(n+1)})$, which may be approximated by a classical Monte-Carlo resampling scheme.

As shown in Bertail & Clémençon (2003b), the sequential resampling in step 1 allows to approximatively mimic the renewal property of the split chain and to efficiently reproduce the second order structure.

3 Second order properties for linear functionals

3.1 Basic estimators

Let $f : E \to \Re$ be a μ-integrable function. Our parameter of interest is now the unknown mean $\mu(f) = E_\mu(f(X_1))$. Although the sample mean $\mu_{n+1}(f) = (n+1)^{-1} \sum_{i=1}^{n+1} f(X_i)$ is an asymptotically normal estimator of $\mu(f)$ un-

der simple moment conditions, we shall consider the truncated sample mean based on the data segment $(X_{\widehat{\tau}_{m^*}(1)+1}, ..., X_{\widehat{\tau}_{m^*}(1_{m^*,n})})$ only (or equivalently on the blocks $\widehat{\mathcal{B}}_{1,m^*}, ..., \widehat{\mathcal{B}}_{\widehat{l}_{m^*,n}-1,m^*}$), since the matter is here to deal with estimators of which the distribution may be accurately approximated (refer to the discussions in Bertail & Clémençon (2003a, b)). Denote by $\widehat{n} = \widehat{\tau}_{m^*}(\widehat{l}_{m^*,n}) - \widehat{\tau}_{m^*}(1) = \sum_{j=1}^{\widehat{l}_{m^*,n}-1} l(\widehat{\mathcal{B}}_{j,m^*})$ the length of this segment. Set $f(\widehat{\mathcal{B}}_{j,m^*}) = \sum_{i=1+\widehat{\tau}_{m^*}(j)}^{\widehat{\tau}_{m^*}(j+1)} f(X_i)$, $j \geqslant 1$, $\widehat{\mu}_{m^*,n}(f) = \widehat{n}^{-1} \sum_{j=1}^{\widehat{l}_{m^*,n}-1} f(\widehat{\mathcal{B}}_{j,m^*})$, $\widehat{\sigma}^2_{m^*,n}(f) = \widehat{n}^{-1} \sum_{j=1}^{\widehat{l}_{m^*,n}-1} \{f(\widehat{\mathcal{B}}_{j,m^*}) - \widehat{\mu}_{m^*,n}(f) l(\widehat{\mathcal{B}}_{j,m^*})\}^2$. It can easily be shown by using the argument of Theorems 17.2.2 and 17.3.6 in Meyn & Tweedie (1996) that, under suitable block moment conditions, $\widehat{\mu}_{m^*,n}(f)$ is asymptotically normal and $\widehat{\sigma}^2_{m^*,n}(f)$ is a consistent estimator of the asymptotic variance of $\widehat{\mu}_{m^*,n}(f)$ (resp., of $\mu_{n+1}(f)$), namely $\sigma^2(f) = E_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-1} E_{A_{\mathcal{M}}}((\sum_{i=1}^{\tau_{A_{\mathcal{M}}}} \{f(X_i) - \mu(f)\})^2)$, where $\tau_{A_{\mathcal{M}}} = \inf\{k \geqslant 1 / \ X_k \in S, \ Y_k = 1\}$ and $E_{A_{\mathcal{M}}}(.)$ denotes the conditional expectation given $(X_0, Y_0) \in S \times \{1\}$. We then define the *unstudentized mean* $\widehat{\varsigma}_n = \widehat{n}^{1/2} \frac{\widehat{\mu}_{m^*,n}(f) - \mu(f)}{\sigma(f)}$ *and the studentized mean* $\widehat{t}_n = \widehat{n}^{1/2} \frac{\widehat{\mu}_n(f) - \mu(f)}{\widehat{\sigma}_{m^*,n}(f)}$. Bertail & Clémençon (2003a) have shown how to obtain Edgeworth expansions up to $O(n^{-1})$ for such quantities using the same technique as in Bolthausen (1982) and in Malinovskii (1987).

3.2 Asymptotic validity of the ARBB

Let $P^*(.)$ denote the conditional probability under the resampling scheme described in step 1 (see § 2.3) for given $X^{(n+1)}$. Consider now the ARBB counterparts of the statistics introduced above: $\mu^*_n(f) = n^{*-1} \sum_{j=1}^{l^*_n - 1} f(\mathcal{B}^*_j)$ and $\sigma^{*2}_n(f) = n^{*-1} \sum_{j=1}^{l^*_n - 1} \{f(\mathcal{B}^*_j) - \mu^*_n(f) l(\mathcal{B}^*_j)\}^2$ with $n^* = \sum_{j=1}^{l^*_n - 1} l(\mathcal{B}^*_j)$. Define also the ARBB version of the pseudo-regenerative unstudentized sample mean by $\widehat{\varsigma}^*_n = n^{*1/2} \sigma^*_n(f)^{-1} (\mu^*_n(f) - \widehat{\mu}_n(f))$ and the one of the pseudo-regenerative studentized mean by $\widehat{t}^*_n = n^{*1/2} \sigma^*_n(f)^{-1} (\mu^*_n(f) - \widehat{\mu}_n(f))$. We shall use the following assumptions. Let $k \geqslant 2$ and set $\tau_S = \inf\{i \geqslant 1 / \ X_i \in S\}$.

$\mathcal{H}_1(f, k)$: The small set S is such that $\sup_{x \in S} E_x((\sum_{i=1}^{\tau_S} |f(X_i)|)^k) < \infty$.

$\mathcal{H}_2(k)$: The small set S is such that $\sup_{x \in S} E_x(\tau_S^k) < \infty$.

These conditions may be classically replaced by some Liapounov's drift conditions (see Meyn & Tweedie (1996)). For a sequence of nonnegative real numbers $\alpha = (\alpha_n)_{n \in \mathbf{N}}$ converging to 0 as $n \to \infty$, consider

$\mathcal{H}_3 : p(x, y)$ is uniformly estimated by $p_m(x, y)$ based on $X^{(m)}$ at the rate α_m at least for the MSE when error is measured by the L^∞ loss over $S \times S$:

$$\lim_{m \to \infty} \alpha_m^{-1} (E(\sup_{(x,y) \in S \times S} |p_m(x, y) - p(x, y)|^2))^{1/2} = 0.$$

$\mathcal{H}_4(k)$: The sequences $m = m(n)$ and $p = p(n)$ are chosen such that $n^{1/k} \leq p \leq m$ and $m/n \to 0$ as $n \to \infty$.

$\mathcal{H}_5$: $\overline{lim}_{t\to\infty}$ $\sup_{x\in S}$ $|E_x(\exp(it\sum_{i=1}^{\tau_S}\{f(X_i)-\mu(f)\}))| < 1$ (Cramer type condition).

$\mathcal{H}_6$: There exists $N>0$ such that the N-fold convolution of the density of $(\sum_{i=1}^{\tau_S}\{f(X_i)-\mu(f)\})^2$ is uniformly bounded over any starting value $X_0 = x$ in S.

We then have the following results :

Theorem 3.1 *Under assumptions $\mathcal{H}_0$, $\mathcal{H}_1(f,k)$, $\mathcal{H}_2(k)$ $\mathcal{H}_3$, $\mathcal{H}_4(k)$ and $\mathcal{H}_5(k)$ with $k>6$, we have the second order validity of the ARBB distribution both in the standardized and unstandardized case:*

$$\sup_{x\in\mathbb{R}}|P^*(\widehat{\varsigma}_n^* \le x) - P_\nu(\widehat{\varsigma}_n \le x)| = O_{P_\nu}(n^{-1/2}\alpha_m \vee n^{-1/2}n^{-1}m\}) \ ,$$

as $n\to\infty$. And if these conditions holds for some $k>8$ and $\mathcal{H}_6$ hold, we have as $n\to\infty$:

$$\sup_{x\in\mathbb{R}}|P^*(\widehat{t}_n^* \le x) - P_\nu(\widehat{t}_n \le x)| = O_{P_\nu}(n^{-1/2}\alpha_m \vee n^{-1/2}n^{-1}m).$$

In particular if $\alpha_m = m^{-1/2}\log(m)$, by choosing $m = n^{2/3}$, the ARBB is second order correct up to $O(n^{-5/6}\log(n))$.

Proof: The proof is based on the same technical ideas as in Bertail & Clémençon (2003a, b) (refer to these papers for further details). It relies on establishing the closeness between the conditional distribution of the blocks $\mathcal{B}_{1,m^*},...,\mathcal{B}_{l_{m^*,n},m^*}$ dividing the segment $X^{(m^*,n)} = (X_{m^*+1},...,X_{n+1})$ according to the $l_{m^*,n}$ visits of $(X_i,Y_i)_{m^*<i\leqslant n}$ to the atom $A_{\mathcal{M}}$ between time m^*+1 and time n and the conditional distribution of the blocks $\widehat{\mathcal{B}}_{1,m^*},...,\widehat{\mathcal{B}}_{\widehat{l}_{m^*,n},m^*}$ dividing $X^{(m^*,n)}$ according to the $\widehat{l}_{m^*,n}$ successive visits of $(X_i,\widehat{Y}_i)_{m^*<i\leqslant n}$ to $A_{\mathcal{M}}$, for given $X^{(n+1)}$. By coupling arguments one may show that, under $\mathcal{H}_2(2\gamma)$, $\gamma\geqslant 2$ and $\mathcal{H}_3$, there exists a constant C such that for $i\in\{1,2\}$,

$$E_\nu(|\widehat{\tau}_i - \tau_i|^\gamma) \leqslant C\alpha_m, \tag{1}$$

with the further notations $\tau_1 = \tau_{m^*}(1)$, $\widehat{\tau}_1 = \widehat{\tau}_{m^*}(1)$, $\tau_2 = \tau_{m^*}(l_{m^*,n})$ and $\widehat{\tau}_2 = \widehat{\tau}_{m^*}(\widehat{l}_{m^*,n})$. Now set $T_n^{(k)}(f) = n^{-1}\sum_{j=1}^{l_{m^*,n}-1} f(\mathcal{B}_{j,m^*})^k$ and $\widetilde{T}_n^{(k)}(f) = n^{-1}\sum_{j=1}^{\widehat{l}_{m^*,n}-1} f(\widehat{\mathcal{B}}_{j,m^*})^k$ for $1\leqslant k\leqslant 3$, with by convention $T_n^{(k)}(f)=0$ (respectively, $\widetilde{T}_n^{(k)}(f)=0$) when $l_{m^*,n}\leqslant 1$ (resp., when $\widehat{l}_{m^*,n}\leqslant 1$) and set

$$D_n^{(k)}(f) = E_\nu\left|T_n^{(k)}(f) - \widetilde{T}_n^{(k)}(f)\right|.$$

Then, following line by line the argument in Bertail & Clémençon (2003b), we have as $n\to\infty$

$$D_n^{(1)}(f) = O((n-m-p)^{-1}\alpha_m), \tag{2}$$

$$D_n^{(k)}(f) = O(\alpha_m), \text{ for } k=2,\ 3. \tag{3}$$

Observing that, conditioned on $X^{(n+1)}$, the reconstructed ARBB sample path does not keep the markovian structure but still forms a regenerative sequence, the results in Malinovskii (1987) (resp. in Bertail & Clémençon (2003a) allow to derive an explicit Edgeworth expansion (E.E.) up to the second order for the unstudentized ARBB version (resp., for the studentized ARBB version). Given (2) and (3) it is straightforward to check that the conditions of validity of these E.E. hold and that the empirical moments appearing in the empirical E.E. of the ARBB distribution converges to their theoretical counterparts at the rate α_m at least. Moreover the bias induced by the first and last pseudo-regeneration blocks does not perturb the E.E. up to $O_P(n^{-1}\alpha_m)$. The main difficulty actually consists in establishing an E.E. for the original statistic. In the unstudentized case, since the functional is then linear, it simply amounts to control the error induced by a split at the "wrong place" for the first (resp. the last) block (*i.e.* the distance between τ_i and $\widehat{\tau}_i$, $i = 1, 2$): this is typically of the same order as the deviation (2). The unstandardized mean thus admits an E.E. on powers of $(n-m-p)^{-1/2}$, which in turn coincides with the E.E. of the empirical mean up to $O(n^{-1/2}(m/n))$. In the studentized case one must first check that the variance estimate computed from the pseudo-blocks is close to the variance estimate based on the regeneration blocks up to $O_P(n^{-1}\alpha_m)$, conditionally to the first m observations. Combining $\mathcal{H}_4(k)$ with $\mathcal{H}_2(k)$, for $k > 4$, it is straightforward that the probability that the split chain does not visit the regeneration set $S \times \{1\}$ between m and $m + p$ is typically of order $O(n^{-1})$. Subsequently to a regeneration occurring between $m + 1$ and $m + p$, the remaining observations may be then decomposed into true regeneration blocks (independent from the first m observations) using the same partitioning arguments as in Malinovskii (1987) or Bertail & Clémençon (2003). This yields the validity of the E.E. on powers of $(n - m - p)^{-1/2} = n^{-1/2} + O(n^{-1/2}(m/n))$. A straightforward optimization argument leads to the last statement. ■

4 Tuning parameters and simulation results

The main tuning parameter relies in the choice of the small set. If the transition density $p(x, y)$ is continuous on some neighborhood $V_{x_0}(\varepsilon)^2 = [x_0 - \varepsilon, x_0 + \varepsilon]^2$ of some fixed point $(x_0,\ x_0)$ such that $p(x_0, x_0) > 0$, then there exists $\delta = \delta(\varepsilon, p) \in]0,\ 1[$ such that $\inf_{(x,y)\in V_{x_0}^2} p(x, y) \geqslant \delta(2\varepsilon)^{-1}$. Such a compact interval $V_{x_0}(\varepsilon)$ is thus a small set for X. It satisfies condition $\mathcal{M}(1, V_{x_0}(\varepsilon), \delta, \mathcal{U}_{V_{x_0}(\varepsilon)})$, where $\mathcal{U}_{V_{x_0}(\varepsilon)}$ denotes the uniform distribution on $V_{x_0}(\varepsilon)$. Hence, in the case when one knows x_0, ε and δ such that (2) holds (this simply amounts to know a uniform lower bound estimate for the probability of returning to $V_{x_0}(\varepsilon)$ in one step), one may effectively apply the ARBB methodology to X. A possible selection rule for ε relies on fixing x_0 and searching for $\varepsilon > 0$ so as to maximize the expected number of regeneration-blocks conditionally to the observed trajectory $X^{(n+1)}$, that is

$$N_n(\varepsilon,p) = E(\sum_{i=m^*+1}^{n} I\{X_i \in V_{x_0}(\varepsilon),\, Y_i = 1\} \mid X^{(n+1)})$$
$$= \frac{\delta(\varepsilon,p)}{2\varepsilon} \sum_{i=m^*+1}^{n} I\{(X_i, X_{i+1}) \in V_{x_0}(\varepsilon)^2\} \frac{1}{p(X_i, X_{i+1})}.$$

Since the transition density p and its minimum over $V_{x_0}(\varepsilon)^2$ are unknown, a practical criterion $\widehat{N}_n(\varepsilon)$ to optimize is obtained by replacing p by p_m and $\delta(\varepsilon,p)/2\varepsilon$ by a sharp lower bound $\widehat{\delta}_m(\varepsilon,p_m)/2\varepsilon$ for p_m over $V_{x_0}(\varepsilon)^2$. The final procedure may be then implemented in 4 steps as follows. Let x_0 be fixed.

1. Compute an estimator $\widehat{p}_m$ of the transition density, for instance of Nadaraya-Watson's type, with $m = Cn^{2/3}, C > 0$.

2. Select the small set $V_{x_0}(\varepsilon)$ by maximizing the empirical criterion $\widehat{N}_n(\varepsilon)$ described above over $\varepsilon > 0$. This yields $\widehat{\varepsilon}_{m,opt}$ and a corresponding minimum value $\widehat{\delta}_{m,opt}$.

3. At each time $i > m^*$ when $(X_i, X_{i+1}) \in [-\widehat{\varepsilon}_{m,opt}, \widehat{\varepsilon}_{m,opt}]^2$, draw independent Bernoulli r.v.'s $\widehat{Y}_i$ with parameter $1-\widehat{\delta}_{m,opt}(2\varepsilon_{m,opt})^{-1}/\widehat{p}_m(X_i, X_{i+1})$. At each time i such that $\widehat{Y}_i = 1$, divide the trajectory, getting data blocks of random size.

4. Apply the ARBB procedure to the sample mean as previously described.

Because the tuning parameters p_m, $\widehat{\varepsilon}_{m,opt}$, $\widehat{\delta}_{m,opt}$ explicitly depends on the first m observations only, the "2-split" technique ensures that the ARBB resampling will not be asymptotically perturbed by the latter.

In the following tables, we compare the quantile of order γ of the true distribution (TD) of the mean respectively. We take $X_0 = 0$, ε_i *i.i.d.* $\sim N(0,1)$ and consider

-an AR(1) model : $X_i = \rho X_{i-1} + \varepsilon_i$, with $\rho = 0.95$ and $n = 200$, $m = 68 = [2 * n^{2/3})]$.

-an AR model with a ARCH(1) structure $X_i = \rho X_{i-1} + (1 + \alpha X_{i-1}^2)^{1/2}\varepsilon_i$, $\rho = 0.6$, $\alpha = 0.1$. See Bertail and Clémençon (2003b) for comparison with the ARBB without the double splitting trick. The performance are quite similar and suggest that the ARBB without the splitting trick enjoy the same second order properties.

	AR		AR-ARCH		
γ	TD	ARBB	TD	ARBB	ASY
1	-3.63	-3.72	-2.53	-2.65	-2.32
2.5	-2.77	-2.81	-2.02	-2.09	-1.96
5	-2.34	-2.36	-1.79	-1.84	-1.65
10	-1.74	-1.73	-1.42	-1.44	-1.28

	AR		AR-ARCH		
γ	TD	RBB	TD	ARBB	ASY
90	1.68	1.61	1.36	1.41	1.28
95	2.16	1.99	1.73	1.82	1.65
97.5	2.73	2.46	2.00	2.14	1.96
99	3.62	3.60	2.53	2.69	2.32

Table 1: Comparison of the tails of the true (TD), modified ARBB and gaussian (ASY) distributions for the two models.

References

[1] Bertail, P., Clémençon, S. (2003a). Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains. *Revised for Probability Theory and Related Fields.*

[2] Bertail, P., Clémençon, S. (2003b). Regenerative block bootstrap for Markov chains. *Submitted to Ann. Statist.*
[3] Bolthausen, E. (1982). The Berry-Esseen Theorem for strongly mixing Harris recurrent Markov Chains. *Z. Wahrsch. Verw. Gebiete*, **60**, 283-289.
[4] Datta, S., McCormick W.P. (1993). Regeneration-based bootstrap for Markov chains. *Canadian J. Statist.,* **21**, No.2, 181-193.
[5] Götze, F., Künsch, H.R. (1996). Second order correctness of the blockwise bootstrap for stationary observations. *Ann. Statist.*, **24**, 1914-1933.
[6] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer.
[7] Malinovskii, V. K. (1987). Limit theorems for Harris Markov chains I. *Theory Prob. Appl.*, **31**, 269-285.
[8] Meyn, S.P., Tweedie, R.L., (1996). *Markov chains and stochastic stability.* Springer.
[9] Nummelin, E. (1978). A splitting technique for Harris recurrent chains. *Z. Wahrsch. Verw. Gebiete,* **43**, 309-318.
[10] Schick, A (2001). Sample splitting with Markov Chains. *Bernoulli*, **7**, No. 1, 33-61.

Address: Patrice Bertail, CREST, Laboratoire de Statistique, Timbre J340, 1, Bd A. Pinard, 75675, Paris, email: Patrice.Bertail@ensae.fr
Stéphan Clémençon, MODAL'X, Université Paris X, email: sclemenc@u-paris10.fr