Markov approximation of chains of infinite order in the $\bar{d}$-metric

Sandro Gallo, Matthieu Lerasle, D.Y. Takahashi

MARKOV APPROXIMATION OF CHAINS OF INFINITE ORDER IN THE $\bar{d}$-METRIC

S. GALLO, M. LERASLE, AND D. Y. TAKAHASHI

ABSTRACT. We obtain explicit upper bounds for the $\bar{d}$-distance between a chain of infinite order and its canonical k-steps Markov approximation. Our proof is entirely constructive and involves a "coupling from the past" argument. The new method covers non-necessarily continuous probability kernels, and chains with null transition probabilities. These results imply in particular the Bernoulli property for these processes.

1. INTRODUCTION

Chains of infinite order are random processes specified by probability kernels (conditional probabilities), which may depend on the whole past. They constitute a wide class of flexible models that are very useful in different areas of applied probability and statistics, from bioinformatics (Bejerano & Yona, 2001; Busch *et al.*, 2009) to linguistics (Galves *et al.*, 2012). They are also models of considerable theoretical interest in ergodic theory (Coelho & Quas, 1998; Hulse, 1991; Quas, 1996; Walters, 2007) and in the general theory of stochastic process (Bramson & Kalikow, 1993; Comets *et al.*, 2002; Fernández & Maillard, 2005). A natural approach to study chains of infinite order is to approximate the original process by Markov chains of growing orders. In this article, we obtain new upper-bounds on the $\bar{d}$-distance between a chain and its canonical k-steps Markov approximation.

Introduced by Ornstein (1974) to study the isomorphism problem for Bernoulli shifts, the $\bar{d}$-metric is of fundamental importance in ergodic theory where chains of infinite order are also known as g-measures. The $\bar{d}$-distance between two processes can be informally described as the minimal proportion of sites we have to change in a *typical* realization of one process in order to obtain a *typical* realization of the other. Ornstein (1974) showed that the set of processes which are measure theoretic isomorphic to Bernoulli shifts, *i.e.*, have the *Bernoulli property*, is $\bar{d}$-closed. Ergodic Markov chains are examples of processes that are isomorphic to Bernoulli shifts. Therefore, if a process can be approximated arbitrary well under the $\bar{d}$-metric by a sequence of ergodic Markov chains, then this process has the Bernoulli property. In this article we prove the existence of Markov approximation schemes for classes of chains of infinite order with non-necessary continuous and with possibly null transition probabilities. Several of these processes were not considered before. For example, Coelho & Quas (1998), Fernández & Galves (2002), and Johansson *et al.* (2010) required the continuity of the probability kernels. Our results show that these new

2000 *Mathematics Subject Classification.* Primary 60G10; Secondary 60K99.

Key words and phrases. Chains of infinite order, coupling from the past algorithms, canonical Markov approximation, $\bar{d}$-distance.

SG was supported by FAPESP grant 2009/09809-1. ML was supported by FAPESP grant 2009/09494-0. DYT was supported by FAPESP grant 2008/08171-0 and Pew Latin American Fellowship. This work is part of USP project "Mathematics, computation, language and the brain", FAPESP project "NeuroMat" (grant 2011/51350-6), and CNPq project "Stochastic modeling of the brain activity" (grant 480108/2012-9). The three authors thank NUMEC-USP for hospitality during the elaboration of the paper.

examples are isomorphic to Bernoulli shifts and provide explicit upper bounds for the Markov approximation in several important cases, giving therefore information on *how good* these approximations are.

Additionally, the $\bar{d}$-distance is useful in statistics and information theory. Rissanen (1983) proposed to model data as realizations of stochastic processes, and proved that these data can be optimally compressed using the (unknown) probability kernel of the chain. The statistical problem is then to recover this probability kernel from the observation of typical data. Since the number of parameters to estimate is infinite, this task is impossible in general. A possible strategy to overcome this problem is the following. (1) Couple the original chain with a Markov approximation and (2) work with the approximating Markov chain. The $\bar{d}$-distance between the chain and its Markov approximation controls the error made in step (1). The idea is that, if this control is good enough, the good properties of the approximating Markov chain proved in step (2) can be used to study the original chain. For instance, Duarte *et al.* (2006) and Csiszár & Talata (2010) obtained consistency results for chains of infinite order from the consistency of BIC estimators for Markov chains proved in Csiszár & Talata (2006). This "two steps" procedure was also used in Collet *et al.* (2005) to obtain a bootstrap central limit theorem for chains of infinite order from the renewal property of the approximating Markov chains.

Our main results use coupling arguments. We first introduce a flexible class of *Coupling from the past* algorithms (CFTP algorithms, see Section 2.3). CFTP algorithms constitute an important class of perfect simulation algorithms popularized by Propp & Wilson (1996). Our main assumption on the chain is that the original chain of infinite order can be perfectly simulated *via* such CFTP algorithms. We state a technical result, Lemma 4.1, which provides an abstract upper bound for the $\bar{d}$-distance between the chain and its canonical Markov approximation. This bound is then made explicit under various extra assumptions on the process used in the study of the CFTP algorithms (Comets *et al.*, 2002; De Santis & Piccioni, 2012; Gallo, 2011; Gallo & Garcia, 2011).

To our knowledge, Fernández & Galves (2002) provide the best explicit bounds in the literature for the $\bar{d}$-distance between a chain of infinite order and its canonical Markov approximation, depending only on the continuity rate of the probability kernels. Their result applies to weakly non-null chains having summable continuity rates. Our method recovers the same bounds under weaker assumption, substituting weak non-nullness by a weaker assumption (see Theorem 4.1). Assuming weak non-nullness, we also obtain explicit upper bounds in some non-summable continuity regimes and other not even necessarily continuous, but satisfying certain types of localized continuity, as introduced in De Santis & Piccioni (2012), Gallo (2011) and Gallo & Garcia (2011). This is the content of Theorems 4.2 and 4.3 which provide, as far as we know, the first results for non-continuous chains. Our results should also be compared with the results in Johansson *et al.* (2010), where they prove the Bernoulli property for square summable continuity regime assuming strong non-nullness, although they don't provide an explicit upper bound for the approximations.

The article is organized as follows. In Section 2, we introduce the notation and basic definitions used all along the paper. In Section 3, we construct the coupling between the original chain and its canonical Markov approximation and we introduce the class of CFTP algorithms perfectly simulating the chains. Our main results are stated in Section 4. We postpone the proofs to Section 5. For convenience of the reader, we leave in Appendix

some extensions and technical results on the "house of cards" processes that are useful in our applications and are of independent interest.

2. NOTATION, DEFINITIONS AND BACKGROUND

2.1. **Notation.** We use the conventions that $\mathbb{N} = \{1, 2, \ldots\}$ and $\overline{\mathbb{N}} = \mathbb{N} \cup \{0, \infty\}$. Let A be the set $\{1, 2, \ldots, N\}$ for some $N \in \overline{\mathbb{N}}$. Given two integers $m \leq n$, let a_m^n be the string $a_m \ldots a_n$ of symbols in A. For any $m \leq n$, the length of the string a_m^n is denoted by $|a_m^n|$ and is equal to $n - m + 1$. Let $\emptyset$ denote the empty string, with length $|\emptyset| = 0$. For any $n \in \mathbb{Z}$, we will use the convention that $a_{n+1}^n = \emptyset$, and naturally $|a_{n+1}^n| = 0$. Given two strings v and v', we denote by vv' the string of length $|v| + |v'|$ obtained by concatenating the two strings. If $v' = \emptyset$, then $v\emptyset = \emptyset v = v$. The concatenation of strings is also extended to the case where $v = \ldots a_{-2}a_{-1}$ is a semi-infinite sequence of symbols. If $n \in \mathbb{N}$ and v is a finite string of symbols in A, $v^n = v \ldots v$ is the concatenation of n times the string v. In the case where $n = 0$, v^0 is the empty string $\emptyset$. Let

$$A^{-\mathbb{N}} = A^{\{\ldots,-2,-1\}} \quad \text{and} \quad A^\star = \bigcup_{j=0}^{+\infty} A^{\{-j,\ldots,-1\}},$$

be, respectively, the set of all infinite strings of past symbols and the set of all finite strings of past symbols. The case $j = 0$ corresponds to the empty string $\emptyset$. Finally, we denote by $\underline{a} = \ldots a_{-2}a_{-1}$ the elements of $A^{-\mathbb{N}}$.

2.2. **Kernels, chains and coupling.**

Definition 2.1. *A family of transition probabilities, or* kernel*, on an alphabet A is a function*

$$\begin{array}{rccc} P: & A \times A^{-\mathbb{N}} & \to & [0,1] \\ & (a, \underline{x}) & \mapsto & P(a|\underline{x}) \end{array}$$

such that

$$\sum_{a \in A} P(a|\underline{x}) = 1 \ , \quad \forall \underline{x} \in A^{-\mathbb{N}}.$$

P is called a Markov kernel if there exists k such that $P(a|\underline{x}) = P(a|\underline{y})$ when $x_{-k}^{-1} = y_{-k}^{-1}$. In the present article we are mostly interested in *non*-Markov kernels, in which $P(a|\underline{x})$ may depend on the whole past $\underline{x}$. When we consider a Markov kernel of order k, we will make explicit the dependence on the k past values and use the notation $P(\cdot|x_{-k}^{-1})$ to indicate $P(\cdot|\underline{x})$.

Definition 2.2. *A stationary stochastic process $\mathbf{X} = \{X_n\}_{n \in \mathbb{Z}}$ with distribution μ on $A^{\mathbb{Z}}$ is said to be* compatible *with a family of transition probabilities P if the later is a regular version of the conditional probabilities of the former, that is*

$$\mu(X_0 = a | X_{-\infty}^{-1} = \underline{x}) = P(a|\underline{x}) \tag{1}$$

for every $a \in A$ and μ-a.e. $\underline{x}$ in $A^{-\mathbb{N}}$.

If P is non-Markov, it may be hard to prove the existence of a stationary process $\mathbf{X}$ compatible with it. In order to solve this issue, we will assume the existence of *coupling from the past algorithms* for the chain (see Section 2.3). This "constructive argument" garantees the existence and uniqueness of the stationary process $\mathbf{X}$ compatible with P.

Definition 2.3 (Canonical k-steps Markov approximation)**.** *Assume that* $\mathbf{X}$ *is a stationary process with distribution* μ*. Let* $P_\mu^{[k]}$ *be the kernel defined by*

$$P_\mu^{[k]}(a|x_{-k}^{-1}) = \mu(X_0 = a|X_{-k}^{-1} = x_{-k}^{-1}).$$

The canonical k-steps Markov approximation *of* $\mathbf{X}$ *is the stationary* k*-step Markov chain* $\mathbf{X}^{[k]}$ *compatible with the kernel* $P_\mu^{[k]}(a|x_{-k}^{-1})$*.*

Since, in all cases considered in this article, μ is uniquely determined by P, we will omit in what follows the subscript μ in $P_\mu^{[k]}$, and it will be understood that $P^{[k]} = P_\mu^{[k]}$.

Let us recall that a coupling between two chains $\mathbf{X}$ and $\mathbf{Y}$ taking values in the same alphabet A is a stochastic process $\mathbf{Z} = \{Z_n\}_{n\in\mathbb{Z}} = \{(\bar{X}_n, \bar{Y}_n)\}_{n\in\mathbb{Z}}$ on $(A \times A)^{\mathbb{Z}}$ such that $\bar{\mathbf{X}}$ has the same distribution as $\mathbf{X}$ and $\bar{\mathbf{Y}}$ has the same distribution as $\mathbf{Y}$. For any pair of stationary processes $\mathbf{X}$ and $\mathbf{Y}$, let $\mathcal{C}(\mathbf{X}, \mathbf{Y})$ be the set of couplings between $\mathbf{X}$ and $\mathbf{Y}$.

Definition 2.4 ($\bar{d}$-distance)**.** *The* $\bar{d}$*-distance between two stationary processes* $\mathbf{X}$ *and* $\mathbf{Y}$ *is defined by*

$$\bar{d}(\mathbf{X}, \mathbf{Y}) = \inf_{(\bar{\mathbf{X}},\bar{\mathbf{Y}})\in\mathcal{C}(\mathbf{X},\mathbf{Y})} \mathbb{P}(\bar{X}_0 \neq \bar{Y}_0).$$

For the class of ergodic processes, this distance has another interpretation which is more intuitive: it is the minimal proportion of sites we have to change in a *typical* realization of $\mathbf{X}$ in order to obtain a *typical* realization of $\mathbf{Y}$. Formally,

$$\bar{d}(\mathbf{X}, \mathbf{Y}) = \inf_{(\bar{\mathbf{X}},\bar{\mathbf{Y}})\in\mathcal{C}(\mathbf{X},\mathbf{Y})} \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\bar{X}_i \neq \bar{Y}_i\}.$$

2.3. **Coupling from the past algorithm (CFTP).** Our CFTP algorithm constructs a sample of the stationary process compatible with a given kernel P, using a sequence $\mathbf{U} = \{U_n\}_{n\in\mathbb{Z}}$ of i.i.d. random variables uniformly distributed in $[0, 1[$. We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the probability space associated to $\mathbf{U}$. The CFTP is completely determined by its *update function* $F : A^{-\mathbb{N}} \times [0, 1[\to A$ which satisfies, for any $\underline{a} \in A^{-\mathbb{N}}$ and for any $a \in A$, $\mathbb{P}(F(\underline{a}, U_0) = a) = P(a|\underline{a})$. Using this function, we define the *set of coalescence times* Θ and the *reconstruction function* Φ associated to F. For any pair of integers m, n such that $-\infty < m \leq n < +\infty$, let $F_{\{m,n\}}(\underline{a}, U_m^n) \in A^{n-m+1}$ be the sample obtained by *applying recursively* F *on the fixed past* $\underline{a}$, i.e, let $F_{\{m,m\}}(\underline{a}, U_m) := F(\underline{a}, U_m)$ and

$$F_{\{m,n\}}(\underline{a}, U_m^n) := F_{\{m,n-1\}}(\underline{a}, U_m^{n-1})F(\underline{a}F_{\{m,n-1\}}(\underline{a}, U_m^{n-1}), U_n)\ .$$

Secondly, let $F_{[m,m]}(\underline{a}, U_m) := F(\underline{a}, U_m)$ and

$$F_{[m,n]}(\underline{a}, U_m^n) = F\left(\underline{a}F_{\{m,n-1\}}(\underline{a}, U_m^{n-1}), U_n\right)\ . \tag{2}$$

$F_{[m,n]}(\underline{a}, U_m^n)$ is the last symbol of the sample $F_{\{m,n\}}(\underline{a}, U_m^n)$. The set

$$\Theta[n] := \{j \leq n : F_{[j,n]}(\underline{a}, U_j^n) = F_{[j,n]}(\underline{b}, U_j^n) \text{ for all } \underline{a}, \underline{b} \in A^{-\mathbb{N}}\} \tag{3}$$

is called the *set of coalescence times* for the time index n. Finally, the reconstruction function of time n is defined by

$$[\Phi(\mathbf{U})]_n = F_{[\theta[n],n]}(\underline{a}, U_{\theta[n]}^n) \tag{4}$$

where $\theta[n]$ is any element of $\Theta[n]$. Given a kernel P, if $\Theta[0] \neq \emptyset$ a.s. and therefore $\Theta[n] \neq \emptyset$ a.s. for any $n \in \mathbb{Z}$, then $([\Phi(\mathbf{U})]_n)_{n\in\mathbb{Z}}$ is distributed according to the unique stationary measure compatible with P, see De Santis & Piccioni (2012).

3. CONSTRUCTION OF THE COUPLING

For any $\underline{a} \in A^{-\mathbb{N}}$, let $\mathcal{I}(\underline{a}) := \{I_k(a|a_{-k}^{-1})\}_{k\in\overline{\mathbb{N}},\, a\in A}$ be any partition of $[0,1[$ having the following properties:

(1) For any $k \in \overline{\mathbb{N}}$, the Lebesgue measure or length $|I_k(a|a_{-k}^{-1})|$ of $I_k(a|a_{-k}^{-1})$ only depends on a and a_{-k}^{-1},

(2) for any $\underline{a}$ and a
$$\sum_{k\in\overline{\mathbb{N}}} |I_k(a|a_{-k}^{-1})| = P(a|\underline{a}),$$

(3) the intervals are disposed as represented in the upper part of Figure 1.

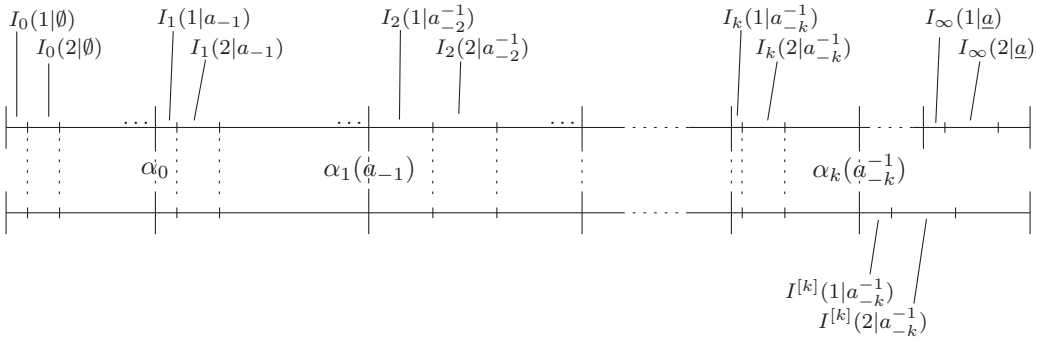

FIGURE 1. Illustration of a range partition related to some infinite past $\underline{a}$. The upper partition is the one used for the original kernel P, whereas the one below is used for the approximating kernel $P^{[k]}$.

Definition 3.1. *We call* range partitions *the partitions of* $[0,1[$ *satisfying (1), (2) and (3) for some kernel* P.

The following lemma is proved in Section 5.1.

Lemma 3.1. *A set of range partitions satisfies, for any* $\underline{a}$ *and* $a \in A$,
$$\sum_{i=0}^{k} |I_i(a|a_{-i}^{-1})| \le \inf_{\underline{z}} P(a|a_{-k}^{-1}\underline{z}), \quad \forall k \ge 0 \ .$$

Given a range partition $\mathcal{I}(\underline{a})$, the following F is an update function, due to property (2):
$$F(\underline{a}, U_0) := \sum_{a\in A} a.\mathbf{1}\left\{U_0 \in \bigcup_{k\in\overline{\mathbb{N}}} I_k(a|a_{-k}^{-1})\right\}. \tag{5}$$

This function F explains the name "range partition": for a given past $\underline{a}$, when the uniform r.v. U_0 belongs to $\bigcup_{a\in A}\bigcup_{i=0}^{k} I_i(a|a_{-i}^{-1})$, then F constructs a symbol looking at a range $\le k$ in the past.
Let $L : A^{-\mathbb{N}} \times [0,1[\to \{0,1,2,\ldots\}$ be the *range function* defined by
$$L(\underline{a}, u) := \sum_{k\in\overline{\mathbb{N}}} k.\mathbf{1}\{u \in \cup_{a\in A} I_k(a|a_{-k}^{-1})\} \ . \tag{6}$$

L associates to a past $\underline{a}$ and a real number $u \in [0,1[$ the length of the suffix of $\underline{a}$ that F needs in order to construct the next symbol when $U_0 = u$.

Using these functions, define, as in Section 2.3, the related coalescence sets $\Theta[i]$, $i \in \mathbb{Z}$, and the reconstruction function $\Phi(\mathbf{U})$, which is distributed according to the unique stationary distribution compatible with P whenever $\Theta[0]$ is a.s. non-empty.

Let us now define the functions $F^{[k]}$ and $L^{[k]}$ that we will use for the construction of $\mathbf{X}^{[k]}$. Observe that, on the one hand, by definition of the canonical k-steps Markov approximation we have for any $a \in A$ and $a_{-k}^{-1} \in A^k$

$$P^{[k]}(a|a_{-k}^{-1}) := \mu(X_0 = a|X_{-k}^{-1} = a_{-k}^{-1}) = \int_{A^{-\mathbb{N}}} P(a|a_{-k}^{-1}\underline{z})d\mu(\underline{z}|a_{-k}^{-1}) \geq \inf_{\underline{z}} P(a|a_{-k}^{-1}\underline{z}) \ .$$

On the other hand, by Lemma 3.1, $\inf_{\underline{z}} P(a|a_{-k}^{-1}\underline{z}) \geq \sum_{j=0}^{k} |I_k(a|a_{-k}^{-1})|$. Thus we can define, for any a_{-k}^{-1}, the set of intervals $\{I^{[k]}(a|a_{-k}^{-1})\}_{a\in A}$ having length $|I^{[k]}(a|a_{-k}^{-1})| = P^{[k]}(a|a_{-k}^{-1}) - \sum_{j=0}^{k} |I_k(a|a_{-k}^{-1})|$ and disposed as in Figure 1. The functions $F^{[k]}$ and $L^{[k]}$ are defined as follows

$$F^{[k]}(\underline{a}, U_0) := \sum_{a\in A} a\mathbf{1}\{U_0 \in \cup_{j=0}^{k} I_j(a|a_{-j}^{-1}) \cup I^{[k]}(a|a_{-k}^{-1})\} \tag{7}$$

and

$$L^{[k]}(\underline{a}, U_0) := \sum_{j=0}^{k} j.\mathbf{1}\{U_0 \in \cup_{a\in A} I_j(a|a_{-j}^{-1})\} + k.\mathbf{1}\{U_0 \in \cup_{a\in A} I^{[k]}(a|a_{-k}^{-1})\}. \tag{8}$$

Observe that, since $F^{[k]}$ and $L^{[k]}$ are derived from k-steps Markov kernels, these functions depend on $\underline{a}$ only through the k last value a_{-k}^{-1}. For this reason we will use the notation $F^{[k]}(a_{-k}^{-1}, U_0)$ to indicate $F^{[k]}(\underline{a}, U_0)$; the same for $L^{[k]}$.

Now, using these functions, define, as in Section 2.3, the related coalescence sets $\Theta^{[k]}[i]$, $i \in \mathbb{Z}$, and the reconstruction function $\Phi^{[k]}(\mathbf{U})$, which is distributed according to the unique stationary distribution compatible with $P^{[k]}$ whenever $\Theta^{[k]}[0]$ is a.s. non-empty.

Using the same sequence of uniforms $\mathbf{U}$ and assuming that $\Theta[0]$ and $\Theta^{[k]}[0]$ are a.s. non-empty, $(\Phi(\mathbf{U}), \Phi^{[k]}(\mathbf{U}))$ is a $(A \times A)$-valued chain with coordinates distributed as $\mathbf{X}$ and $\mathbf{X}^{[k]}$ respectively. It follows that $(\Phi(\mathbf{U}), \Phi^{[k]}(\mathbf{U}))$ is a coupling between both chains. Hence, we have constructed a CFTP algorithm for perfect simulation of the coupled chains.

4. STATEMENTS OF THE RESULTS

4.1. **A key lemma.** Let us first state a technical lemma that is central in the proof of our main results.

Lemma 4.1. *Assume that there exists a set of range partitions $\{\mathcal{I}(\underline{a})\}_{\underline{a}}$ such that the sets of coalescence times $\Theta[0] \cap \Theta^{[k]}[0]$ is $\mathbb{P}$-a.s. non-empty. Then, for any $\theta[0] \in \Theta[0]$,*

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq \mathbb{P}\left(\bigcup_{\underline{a}} \bigcup_{i=\theta[0]}^{0} \left\{L\left(\underline{a}F_{\{\theta[0],i-1\}}(\underline{a}, U_{\theta[0]}^{i-1}), U_i\right) > k\right\}\right). \tag{9}$$

where, for $i = \theta[0]$, the event reads $\{L\left(\underline{a}, U_{\theta[0]}\right) > k\}$.

Examples of range partitions satisfying the conditions of this lemma have already been built, for example in Comets *et al.* (2002), Gallo (2011), Gallo & Garcia (2011) and De Santis & Piccioni (2012). These works assume some regularity conditions on P and

some non-nullness hypothesis which are presented in Sections 4.2, 4.3, 4.4. In these sections, we obtain explicit upper bounds for (9) under the respective assumptions. Before that, let us give an interesting remark on Bernoullicity.

Observation 4.1 (A remark on Bernoullicity)**.** *In the conditions of each works cited above, we will exhibit $\theta[0] \in \Theta[0]$ which belongs to $\Theta^{[k]}[0]$ for any sufficiently large k's, and we will prove that*

$$\mathbb{P}\left(\bigcup_{\underline{a}}\bigcup_{i=\theta[0]}^{0}\left\{L\left(\underline{a}F_{\{\theta[0],i-1\}}(\underline{a},U_{\theta[0]}^{i-1}),U_i\right)>k\right\}\right)\overset{k\to\infty}{\longrightarrow}0. \tag{10}$$

It follows, by Lemma 4.1, that in this case

$$\lim_{k\to\infty}\bar{d}(\mathbf{X},\mathbf{X}^{[k]})=0.$$

We also have, for any sufficiently large k's, that $\mathbf{X}^{[k]}$ is an ergodic Markov chain since $\Theta^{[k]}[0]$ is non-empty. Now, by the $\bar{d}$-closure of the set of processes isomorphic to a Bernoulli shift (see for example Ornstein (1974)) and the fact that ergodic Markov processes have the Bernoulli property (Ornstein (1974), p.45), we conclude using the results of following sections that the processes considered in Comets et al. *(2002), Gallo (2011), Gallo & Garcia (2011) and De Santis & Piccioni (2012) have the Bernoulli property.*

4.2. **Kernels with summable continuity rate.**

Let us first define continuity.

Definition 4.1 (Continuity points and continuous kernels)**.** *For any $k\in\mathbb{N}$, a and a_{-k}^{-1}, let $\alpha_k(a|a_{-k}^{-1}):=\inf_{\underline{z}}P(a|a_{-k}^{-1}\underline{z})$ and $\alpha_0(a):=\inf_{\underline{z}}P(a|\underline{z})$. A past $\underline{a}$ is called a* continuity point *for P or P is said to be continuous in $\underline{a}$ if*

$$\alpha_k(a_{-k}^{-1}):=\sum_{a\in A}\alpha_k(a|a_{-k}^{-1})\overset{k\to+\infty}{\longrightarrow}1.$$

We say that P is continuous when

$$\alpha_k:=\inf_{a_{-k}^{-1}}\alpha_k(a_{-k}^{-1})\overset{k\to+\infty}{\longrightarrow}1.$$

We say that P has summable continuity rate when $\sum_{k\geq 0}(1-\alpha_k)<\infty$.

We also define weak non-nullness.

Definition 4.2. *We say that a kernel P is* weakly non-null *if $\alpha_0>0$, where $\alpha_0:=\sum_{a\in A}\alpha_0(a)$.*

De Santis & Piccioni (2012) have introduced a more general assumption that we call *very weak non-nullness*, see Definition 5.2. We postpone this definition to Section 5.3 in order to avoid technicality at this stage.

Theorem 4.1. *Assume that P has summable continuity rate and is very weakly non-null. Then, there exists a constant $C<+\infty$ such that, for any sufficiently large k,*

$$\bar{d}(\mathbf{X},\mathbf{X}^{[k]})\leq C(1-\alpha_k)\ .$$

Remark 4.1. *This upper bound is new since we do not assume weak non-nullness. Fernández & Galves (2002) showed, under weak non-nullness, that for any sufficiently large k, there exists a positive constant C such that*

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq C\beta_k$$

where

$$\beta_k := \sup\{|P(a|a_{-k}^{-1}\underline{x}) - P(a|a_{-k}^{-1}\underline{y})| \ : \ a \in A,\, a_{-k}^{-1} \in A^k,\, \underline{x}, \underline{y} \in A^{-\mathbb{N}}\}.$$

This later quantity is related to α_k through the inequalities $1-\alpha_k \geq \frac{|A|}{2}\beta_k$ and $1-\alpha_k \leq D\beta_k$ for some $D > 1$ and sufficiently larges k's. Moreover, $1 - \alpha_k = \beta_k$ for binary alphabets. Thus, Theorem 4.1 extends the bound in Fernández & Galves (2002).

4.3. **Using a prior knowledge of the histories that occur.** De Santis & Piccioni (2012) introduced the following assumptions on kernels. Define

$$\forall k \geq 1, \quad J_k(U_{-k}^{-1}) := \{\underline{x} \in A^{-\mathbb{N}} \text{ s.t. } \forall 1 \leq l \leq k,\ x_{-l} = a \quad \text{if } U_{-l} \in I(a|\emptyset) \text{ for some } a \in A\},$$

$$\forall k \geq 1, \quad A_k(U_{-k}^{-1}) := \inf\{\alpha_k(x_{-k}^{-1}) : \underline{x} \in J_k(U_{-k}^{-1})\} \ .$$

Finally, let

$$\ell(U_{-\infty}^{0}) := \inf\{j \geq 0 : U_0 < A_j(U_{-j}^{-1})\} \text{ where } A_0(U_0^1) := \alpha_0. \tag{11}$$

Theorem 4.2. *If $\mathbf{X}$ has a kernel that satisfies $\mathbb{E}\left(\prod_{k\geq 0} A_k(U_{-k}^{-1})^{-1}\right) < \infty$, then there exists a positive constant $C < +\infty$ such that*

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq C\mathbb{P}(\ell(U_{-\infty}^{0}) > k).$$

In order to illustrate the interest of this result, let us give two simple examples. Other examples can be found in De Santis & Piccioni (2012) and Gallo & Garcia (2011).

Summable continuity regime with weak non-nullness. Theorem 4.2 allows to recover the result of Theorem 4.1 in the weakly non-null case. To see this, it is enough to observe that, for any U_{-k}^{-1}, $A_k(U_{-k}^{-1}) \geq \alpha_k$ (see Definition 4.1 for α_k). It follows that $\prod_{k\geq 0}\alpha_k > 0$ (which is equivalent to $\sum_{k\geq 0}(1-\alpha_k) < +\infty$), implies that $\prod_{k\geq 0} A_k(U_{-k}^{-1})$ is bounded away from zero, hence, its inverse has finite expectation. Hence, Theorem 4.2 applies and gives

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq C\mathbb{P}(\ell(U_{-\infty}^{0}) > k) \leq C(1-\alpha_k) \ .$$

A simple discontinuous kernel on $A = \{1, 2\}$. Let $\epsilon \in (0, 1/2)$ and let $\{p_i\}_{i\geq 0}$ be any sequence such that, $\epsilon \leq p_i < 1 - \epsilon$ for any $i \geq 0$. Let $t(\underline{a}) := \inf\{i \geq 0 : a_{-i-1} = 2\}$ let $\bar{P}$ be the following kernel:

$$\forall \underline{a} \in \{1, 2\}^{-\mathbb{N}}, \ \ \bar{P}(2|\underline{a}) = p_{t(\underline{a})} \ . \tag{12}$$

The existence of a unique stationary process compatible with this kernel is proven in Gallo (2011) for instance. This chain is the renewal sequence, that is, a concatenation of blocks of the form $1 \ldots 12$ having random length with finite expectation. It is clearly weakly non-null, however, it is not necessarily continuous. In fact, a simple calculation shows that $\alpha_k = 1 - \sup_{l,m\geq k}|p_l - p_m|$, which needs not to go to 1. Nevertheless, if we assume furthermore that $\sup_{k\geq 0}\alpha_k > 1 - \alpha_0(2)$, we have $\mathbb{E}\left(\prod_{k\geq 0} A_k(U_{-k}^{-1})^{-1}\right) < \infty$. The proof of this fact was originally done in an unpublished preliminary version of the paper of De Santis & Piccioni (2012). We include it here for the sake of completeness. Define

$$N(U_{-\infty}^{-1}) := \inf\{n \geq 1 : U_{-n} \in I_0(2)\},$$

and observe that (1) $A_k(U_{-k}^{-1}) = 1$ for any $k \geq N(U_{-\infty}^{-1})$, and (2) $N(U_{-\infty}^{-1})$ has geometric distribution, with probability of success $\alpha_0(2)$. Now, for any $\delta \in (0, \sup_k \alpha_k - 1 + \alpha_0(2))$, we choose $n_0(\delta)$ such that $\alpha_{n_0} \geq 1 - \alpha_0(2) + \delta$, and compute

$$\left(\prod_{k\geq 0} A_k(U_{-k}^{-1})^{-1}\right) = \left(\prod_{k=0}^{N} A_k(U_{-k}^{-1})^{-1}\right) \leq (1-\alpha_0(2)+\delta)^{-N} \prod_{k=0}^{n_0} \alpha_k^{-1}.$$

The last equality follows from the fact that $A_k(U_{-k}^{-1}) \geq \alpha_k$ for any k and any sample U_{-k}^{-1}. Since the probability generating function of N has radius of convergence $\frac{\alpha_0(2)}{1-\alpha_0(2)}$ which is strictly smaller than $(1-\alpha_0(2)+\delta)^{-1}$, it follows that $\left(\prod_{k\geq 0} A_k(U_{-k}^{-1})^{-1}\right)$ is integrable.

We can now obtain an upper bound for $\mathbb{P}(\ell(U_{-\infty}^0) > k)$

$$\begin{aligned}\mathbb{P}(\ell(U_{-\infty}^0) > k) &= \mathbb{P}(\inf\{j \geq 0 : U_0 < A_j(U_{-j}^{-1})\} > k) \\ &\leq \mathbb{P}(U_0 \geq A_k(U_{-k}^{-1})) \leq \mathbb{P}(A_k(U_{-k}^{-1}) < 1) \leq \mathbb{P}(N(U_{-\infty}^{-1}) > k) \leq (1-\epsilon)^k.\end{aligned}$$

Observation 4.2. *The preceding theorems yield explicit upper bounds. However, they hold under restrictions we would like to surpass.*

First, in the continuous regime, we have assumed that $\sum_{k\geq 0}(1-\alpha_k) < +\infty$. *Nevertheless, CFTP are known to exist with the weaker assumption* $\sum_{k\geq 1}\prod_{i=0}^{k-1}\alpha_i = +\infty$, *and it is known that* $\bar{d}(\mathbf{X}, \mathbf{X}^{[k]})$ *goes to zero in this case. We will be interested in upper bounds for the rate of convergence to zero under these weak conditions.*

Second, in Theorem 4.2, the assumption $\mathbb{E}\left(\prod_{k\geq 0} A_k(U_{-k}^{-1})^{-1}\right) < \infty$ *is generally difficult to check: this is particularly clear for the example of* $\bar{P}$ *where, moreover, it requires the (unnecessary) extra-assumption* $\sup_{k\geq 0}\alpha_k > 1 - \alpha_0(2)$.

The next section will solve part of these objections.

4.4. **A simple upper bound under weak non-nullness.** Hereafter, we assume that P is weakly non-null. Let $\Theta'[0]$ be the following subset of $\Theta[0]$:

$$\Theta'[0] := \{i \leq 0 : \text{ for any } \underline{a},\ L(\underline{a}F_{\{i,j-1\}}(\underline{a}, U_i^{j-1}), U_j) \leq j - i\,,\ j = i, \ldots, 0\}. \tag{13}$$

We have the following theorem in which a priori nothing is assumed on the continuity.

Theorem 4.3. *Assume that* P *is weakly non-null and that we can construct a set of range partitions* $\{\mathcal{I}(\underline{a})\}_{\underline{a}}$ *for which* $\Theta'[0] \neq \emptyset$, $\mathbb{P}$*-a.s. Then, for any* $\theta[0] \in \Theta'[0]$

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq \mathbb{P}(\theta[0] < -k). \tag{14}$$

In order to illustrate this result, let us consider the examples of continuous kernels and of the kernel $\bar{P}$. Gallo & Garcia (2011) proposed a unified framework, including these examples and several other cases, which provides more examples of applications of this theorem. This is postponed to Appendix A in order to avoid technicality.

Application to the continuity regime. Let us first introduce the following range partition.

Definition 4.3. *Let $\{\mathcal{I}^{(1)}(\underline{a})\}_{\underline{a}}$ be the range partition such that, for any a and $\underline{a}$*

$$\forall k \geq 0, \quad \left|I_k^{(1)}(a|a_{-k}^{-1})\right| := \alpha_k(a|a_{-k}^{-1}) - \alpha_{k-1}(a|a_{-(k-1)}^{-1}) \ ,$$
$$\left|I_\infty^{(1)}(a|\underline{a})\right| := P(a|\underline{a}) - \lim_{k\to\infty} \alpha_k(a|a_{-k}^{-1}) \ ,$$

with the convention $\alpha_{-1}(a|\emptyset) = 0$.

Let $F^{(1)}$ and $L^{(1)}$ be the associated functions defined in (5) and (6). Let

$$\theta[0] := \max\{i \leq 0 : U_j \leq \alpha_{j-i}\,,\ j = i, \ldots, 0\}.$$

Observe that $L^{(1)}$ satisfies $L^{(1)}(\underline{a}, U_0) \leq k$ whenever $U_0 \leq \alpha_k$. Hence, $\theta[0]$ belongs to $\Theta'^{(1)}[0]$, the set defined by (13) using $F^{(1)}$ and $L^{(1)}$. Moreover, Comets *et al.* (2002) proved that, if $\sum_{k\geq 1}\prod_{i=0}^{k-1}\alpha_i = +\infty$ (that is, under weak non-nullnes but not necessarily summable continuity)

$$\mathbb{P}(\theta[0] < -k) \leq v_k := \sum_{j=1}^{k} \sum_{\substack{t_1,\ldots,t_j \geq 1 \\ t_1+\ldots+t_j = k}} \prod_{m=1}^{j}(1-\alpha_{t_m-1}) \prod_{l=0}^{t_m-2}\alpha_l \tag{15}$$

which goes to 0. This upper bound is not very satisfactory since it is difficult to handle in general. Nevertheless, Propositions B.1 and B.2, given in Appendix B, shed light on the behavior of this vanishing sequence. In particular, under the summable continuity assumption $\sum_{k\geq 0}(1-\alpha_k) < +\infty$, Proposition B.1 states that (15) essentially recovers the rates of Theorems 4.1 and 4.2. Also, if there exists a constant $r \in (0,1)$ and a summable sequence $(s_k)_{k\geq 1}$ such that, $\forall k \geq 1$, $1-\alpha_k = \frac{r}{k} + s_k$, then, from Proposition B.2, there exists a positive constant C such that

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq C\frac{(\log k)^{3+r}}{k^{2-(1+r)^2}}. \tag{16}$$

Application to the kernel $\bar{P}$. As a second direct application of Theorem 4.3, let us consider the kernel $\bar{P}$ defined in (12). Let $\{\mathcal{I}^{(2)}(\underline{a})\}_{\underline{a}}$ be the set of range partitions, such that $|I^{(2)}(2|\emptyset)| = \alpha_0(2)$, $|I^{(2)}(1|\emptyset)| = \alpha_0(1)$ and $I_k^{(2)}(a|a_{-k}^{-1}) = \emptyset$ for any $k \geq 1$ except for $k = t(\underline{a}) + 1$ for which $|I_k^{(2)}(1|a_{-k}^{-1})| = 1 - p_k - \alpha_0(1)$ and $|I_k^{(2)}(2|a_{-k}^{-1})| = p_k - \alpha_0(2)$. It satisfies

$$L^{(2)}(\underline{a}, U_0) = (t(\underline{a}) + 1)\mathbf{1}\{U_0 > \alpha_0\}.$$

Hence, $\theta[0] := \max\{i \leq 0 : U_i \in I(2)\}$ belongs to $\Theta'[0]$ ($\Theta'[0]$ is defined by (13) with the functions $F^{(2)}$ and $L^{(2)}$ obtained from the set of range partitions $\{\mathcal{I}^{(2)}(\underline{a})\}_{\underline{a}}$). Therefore,

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq \mathbb{P}(\theta[0] < -k) \leq (1-\epsilon)^k \tag{17}$$

independently of the value $\sup_{k\geq 0}\alpha_k$. For this simple example, Theorem 4.3 is then less restrictive than Theorem 4.2.

5. PROOFS OF THE RESULTS

5.1. **Proof of Lemma 3.1.** Assume that for some $k \geq 0$ we have

$$\sum_{i=0}^{k} |I_i(a|a_{-i}^{-1})| > \inf_{\underline{z}} P(a|a_{-k}^{-1}\underline{z}) \ .$$

Then, consider a past $\underline{z}^\star$ such that $|I(a)| + \sum_{i=1}^k |I_i(a|a_{-i}^{-1})| > P(a|a_{-k}^{-1}\underline{z}^\star)$. As, for all $l \geq k+1$, $|I_l(a|a_{-k}^{-1}z_{-l}^{-1})| \geq 0$, we have

$$\sum_{i=0}^{k} |I_i(a|a_{-i}^{-1})| + \sum_{l\geq k+1} |I_l(a|a_{-k}^{-1}z_{-l}^{-1})| > P(a|a_{-k}^{-1}\underline{z}^\star) \ .$$

This is a contradiction with the second properties of the partition. This concludes the proof. □

5.2. **Proof of Lemma 4.1.** We assume that $\Theta[0] \cap \Theta^{[k]}[0]$ is $\mathbb{P}$-a.s. non-empty, and we therefore have a coupling $(\Phi(\mathbf{U}), \Phi^{[k]}(\mathbf{U}))$ of both chains. By definitions of $F^{[k]}$ and $L^{[k]}$, we observe that

$$L(\underline{b}a_{-k}^{-1}, U_0) \leq k \Rightarrow \text{ for any } \underline{b} \text{ we have } \begin{cases} L(\underline{b}a_{-k}^{-1}, U_0) = L^{[k]}(a_{-k}^{-1}, U_0) \text{ and,} \\ F(\underline{b}a_{-k}^{-1}, U_0) = F^{[k]}(a_{-k}^{-1}, U_0). \end{cases} \tag{18}$$

Assume that, $\forall \underline{a} \in A^{-\mathbb{N}}$ and $\forall i = \theta[0], \ldots, 0$, $L(\underline{a}F_{\{\theta[0],i-1\}}(\underline{a}, U_{\theta[0]}^{i-1}), U_i) \leq k$. Then, using recursively (18), $F_{\{\theta[0],0\}}(\underline{a}, U_{\theta[0]}^0) = F_{\{\theta[0],0\}}^{[k]}(a_{-k}^{-1}, U_{\theta[0]}^0)$. In particular, $\theta[0] \in \Theta^{[k]}[0]$ and $[\Phi(\mathbf{U})]_0 = [\Phi^{[k]}(\mathbf{U})]_0$. Therefore,

$$\begin{aligned} \bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) &\leq \mathbb{P}([\Phi(\mathbf{U})]_0 \neq [\Phi^{[k]}(\mathbf{U})]_0) \\ &\leq \mathbb{P}\left(\bigcup_{\underline{a}} \bigcup_{i=\theta[0]}^{0} \left\{ L\left(\underline{a}F_{\{\theta[0],i-1\}}(\underline{a}, U_{\theta[0]}^{i-1}), U_i\right) > k \right\} \right). \end{aligned}$$

5.3. **Proof of Theorem 4.1.** This section is divided in three parts. First, as mentioned before the statement of the theorem, we define *very weak non-nullness*. Then, we prove some technical lemmas allowing to apply Lemma 4.1. Finally, we prove the theorem.

5.3.1. *Definition of* very weak non-nullness. Consider the set of range partitions $\{\mathcal{I}^{(1)}(\underline{a})\}_{\underline{a}}$ of Definition 4.3. As observed by De Santis & Piccioni (2012), in the continuous case, since $\{\alpha_k\}_{k\geq 0}$ increases monotonically to 1, there exists $k \geq 0$ such that $\alpha_k > 0$. Let $k^\star$ be the smallest of these integers and let $F^\star$ be the following update function

$$F^\star(a_{-k^\star}^{-1}, U_0) := F^{(1)}(\underline{b}, U_0\alpha_{k^\star}) \qquad \forall \underline{b} \text{ s.t. } a_{-k^\star}^{-1} = b_{-k^\star}^{-1} \ .$$

$F^\star$ is well defined, since $U_0\alpha_{k^\star} \leq \alpha_{k^\star}$, hence $L^{(1)}(\underline{b}, U_0\alpha_{k^\star}) \leq k^\star$. In the case where $k^\star = 0$, $F^\star$ is simply defined as

$$F^\star(\emptyset, U_0) := F^{(1)}(\underline{b}, U_0\alpha_0) = \sum_{a\in A} \mathbf{1}\{U_0\alpha_0 \in I_0(a|\emptyset)\} \qquad \forall \underline{b} \ .$$

Definition 5.1 (Coalescence set)**.** *For $m \geq k^\star + 1$, let E_m, the* coalescence set *(different from the set of coalescence times), be defined as the set of all $u_{-m+1}^0 \in A^m$ such that*

$$F_{\{-k^\star+1,0\}}^\star\left(a_{-k^\star}^{-1} F_{\{-m+1,-k^\star\}}^\star\left(a_{-k^\star}^{-1}, \frac{u_{-m+1}^{-k^\star}}{\alpha_{k^\star}}\right), \frac{u_{-k^\star+1}^0}{\alpha_{k^\star}}\right) \text{ does not depend on } a_{-k^\star}^{-1} \ .$$

When $k^\star = 0$ and $m = 1$, we have $E_1 := \cup_{a\in A} I_0(a|\emptyset)$.

Definition 5.2. *We say that P is very weakly non-null if*

$$\exists m \geq k^\star + 1 \qquad s.t. \qquad \mathbb{P}(U_{-m+1}^0 \in E_m) > 0 \ . \tag{19}$$

Weak non-nullness corresponds to $\mathbb{P}(U_0 \in E_1) > 0$, hence, it implies very weak non-nullness.

5.3.2. *Technical lemmas.* Let $\Theta^{(1)}[0]$ be the set of coalescence times defined by (3) for the function $F^{(1)}$. In a first part of the proof, we define a random time $\theta[0]$ (see (20)) and we show that it belongs to $\Theta^{(1)}[0]$ and that it has finite expectation whenever $\sum_{k\geq k^\star}(1-\alpha_k) < +\infty$. This random variable is defined in the proof of Theorem 2 in De Santis & Piccioni (2012).

Recall that, by construction of the range partition $\{\mathcal{I}^{(1)}(\underline{a})\}_{\underline{a}}$, for any $\underline{a}$, $L(\underline{a}, U_i) = k$ whenever $\alpha_{k-1} \leq U_i < \alpha_k$. This means that the sequence of ranges forms a sequence $\{L_i\}_{i\in\mathbb{Z}} := \{L(\underline{a}, U_i)\}_{i\in\mathbb{Z}}$ of i.i.d. $(\mathbb{N}\cup\{0\})$-valued r.v.'s. We now introduce two sequences of random times in the past, which are represented on Figure 2, in the particular case where $k^\star = 2$. Let

$$W_1 := \sup\{m \leq 0 : U_j < \alpha_{j-m+k^\star}\ ,\ j = m, \ldots, 0\},$$

and for any $i \geq 1$

$$Y_i := \inf\{m < W_i : U_n < \alpha_{k^\star}\ ,\ n = m+1, \ldots, W_i\}$$

and

$$W_{i+1} := \sup\{m \leq Y_i : U_j < \alpha_{j-m+k^\star}\ ,\ j = m, \ldots, Y_i\}.$$

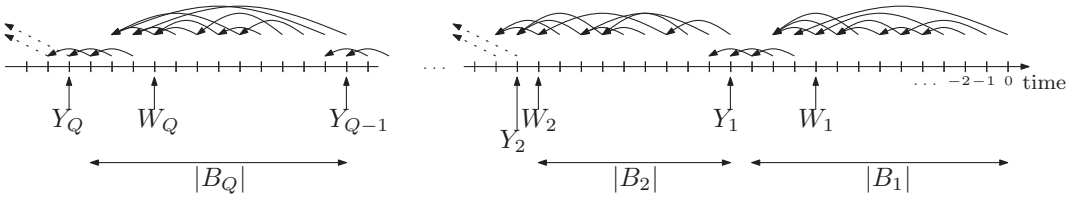

FIGURE 2. We consider a realization of $L^0_{-\infty}$ in the particular case $k^\star = 2$, that is, the arrows, which represent the length function at each time index, have length larger or equal to 2.

Consider now the random variable

$$Q := \inf\{i \geq 1 : (U_{Y_i+1}, \ldots, U_{W_i-1}) \in E_{W_i-Y_i-1}\}$$

(see Definition 5.1 for E_m) and put

$$\theta[0] := Y_Q. \tag{20}$$

Lemma 5.1. $\theta[0] \in \Theta^{(1)}[0]$.

Proof. If $\theta[0] = -k$, then there exists some $l(= -W_Q + 1) \leq k$ such that $U_{-i} \leq \alpha_{k^\star}$, $i = l, \ldots, k$, and moreover, $U^{-l}_{-k} \in E_{k-l+1}$, that is

$$F^\star_{\{-l-k^\star+1,-l\}}\left(a^{-1}_{-k^\star}F^\star_{\{-k,-l-k^\star\}}(a^{-1}_{-k^\star}, \frac{1}{\alpha_{k^\star}}U^{-l-k^\star}_{-k}), \frac{1}{\alpha_{k^\star}}U^{-l}_{-l-k^\star+1}\right) \quad \text{is independent of } a^{-1}_{-k^\star}.$$

Since $U_{-i} \leq \alpha_{k^\star}$, $i = l, \ldots, k$, it follows that

$$F_{\{-l-k^\star+1,-l\}}\left(\underline{b}F_{\{-k,-l-k^\star\}}(\underline{b}, U^{-l-k^\star}_{-k}), U^{-l}_{-l-k^\star+1}\right) \quad \text{is independent of } \underline{b}.$$

By definition of the random times W_i, all the symbols in times $\{W_Q, \ldots, 0\}$ can then be built using those in times $\{W_Q - k^\star, \ldots, W_Q - 1\}$ since none of the arrows from time

W_Q until 0 go further time $W_Q - k^\star$, see the Figure 2. Therefore, the construction of the symbol at times 0 does not depend on the symbols before $\theta[0]$, i.e $\theta[0] \in \Theta^{(1)}[0]$. □

Lemma 5.2. $\mathbb{E}|W_1| < +\infty$ *whenever* $\sum_{k\geq k^\star}(1-\alpha_k) < +\infty$.

Proof. Letting $\bar{\alpha}_{l-k^\star} = \alpha_l$ for any $l \geq k^\star$, we have

$$W_1 = \sup\{m \leq 0 : U_j < \bar{\alpha}_{j-m}\,,\ j = m, \dots, 0\}.$$

Thus W_1 is defined exactly as $\tau[0]$ of display (4.2) in Comets *et al.* (2002), substituting their a_k's by our $\bar{\alpha}_k$'s. They proved (see display (4.6) and item (ii) Proposition 5.1 therein) that $\mathbb{E}|\tau[0]| < +\infty$ whenever $\sum_{k\geq 0}(1-\bar{\alpha}_k) < +\infty$. It follows that $\mathbb{E}|W_1| < +\infty$ whenever $\sum_{k\geq k^\star}(1-\alpha_k) < +\infty$. □

Lemma 5.3. $\mathbb{E}|\theta[0]| < +\infty$ *whenever* $\sum_{k\geq k^\star}(1-\alpha_k) < +\infty$.

Proof. As observed in De Santis & Piccioni (2012), $\{W_i - Y_i - 1\}_{i\geq 1}$ is a sequence of i.i.d. geometric r.v.'s with success probability $1-\alpha_{k^\star}$ and $\{Y_i - W_{i+1}\}_{i\geq 0}$ (with $Y_0 := 0$) is a sequence of i.i.d. r.v.'s distributed as $-W_1$, conditional to be non-zero. Moreover, Lemma 5.2 states that $\mathbb{E}|W_1| < +\infty$. It follows that $\{B_i\}_{i\geq 1} := \{Y_{i-1} - Y_i - 1\}_{i\geq 1}$ is a sequence of i.i.d. $\mathbb{N}$-valued r.v.'s with finite expectation. Thus $\sum_{k=1}^n B_i - n\mathbb{E}B_1$ forms a martingale with respect to the filtration $\mathcal{F}(B_1, \dots, B_i\ :\ i \geq 1)$ and we have by the optional sampling theorem

$$\mathbb{E}|\theta[0]| := \mathbb{E}|Y_Q| = \mathbb{E}\left(\sum_{i=1}^{Q} B_i\right) = \mathbb{E}Q.\mathbb{E}B_1 < +\infty.$$

□

We finally need the following lemma.

Lemma 5.4. *For any* $k \geq k^\star$, $\theta[0] \in \Theta^{(1),[k]}[0]$.

Proof. For any $k \geq k^\star$, $F^{(1),[k]}$ and $L^{(1),[k]}$ satisfy (18). This implies that, in the interval $\{Y_Q, \dots, W_Q - 1\}$, coalescence occurs as well for $F^{(1),[k]}$, i.e. $U_{-\theta[0]}^{W_Q-1} \in E^{[k]}_{\theta[0]-W_Q}$. Both constructed chains are equals until the first time $F^{(1)}$ uses a range larger than k. But at this moment, due to the definition of the W_i's, we have already perfectly simulated at least k symbols of both chains, and therefore, we can continue constructing until time 0 because the ranges of $F^{(1),[k]}$ are smaller of equal to k. It follows that Y_Q is a coalescence time for $F^{(1),[k]}$, and therefore, $\theta[0] \in \Theta^{(1),[k]}[0]$ for any $k \geq k^\star$. □

5.3.3. *Proof of Theorem 4.1.* By definition of $F^{(1)}$ and $L^{(1)}$, we have for any sufficiently large k,

$$\bigcup_{\underline{a}} \bigcup_{i=\theta[0]}^{0} \left\{ L^{(1)}\left(\underline{a} F^{(1)}_{\{\theta[0],i-1\}}(\underline{a}, U_{\theta[0]}^{i-1}), U_i\right) > k \right\} \subset \bigcup_{i=\theta[0]}^{0} \{U_i > \alpha_k\}.$$

By Lemmas 5.1 and 5.4, Lemma 4.1 applies and gives, for sufficiently large k's

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq \mathbb{P}\left(\bigcup_{i=\theta[0]}^{0} \{U_i > \alpha_k\}\right) = \mathbb{P}\left(\sum_{i=0}^{|\theta[0]|} \mathbf{1}\{U_i > \alpha_k\} \geq 1\right) \leq \mathbb{E}\left(\sum_{i=0}^{|\theta[0]|} \mathbf{1}\{U_i > \alpha_k\}\right)$$

where we used the Markov inequality for the last inequality. Using the fact that $\theta[0]$ is a stopping time in the past for the sequence $U_i, i \leq 0$, and that it has finite expectation by Lemma 5.3, we can apply the Wald's equality to obtain

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq \mathbb{E}|\theta[0]|.\mathbb{E}\mathbf{1}\{U_i > \alpha_k\} = \mathbb{E}|\theta[0]|.(1-\alpha_k).$$

5.4. **Proof of Theorem 4.2.** We divide this proof into two parts. First, we prove technical lemmas allowing to use Lemma 4.1. Then, we prove the theorem.

5.4.1. *Technical lemmas.* Using the quantity $\ell(U_{-\infty}^0)$ defined by (11), we define

$$\theta[0] := \sup\{j \leq 0 : \ell(U_{-\infty}^i) \leq i - j\ ,\ i = j, \dots, 0\}. \tag{21}$$

Lemma 5.5. *$\theta[0]$, defined by (21), belongs to $\Theta^{(1)}[0] \cap \Theta^{(1),[k]}[0]$ for any $k \geq 0$ and*

$$\mathbb{P}\left(\bigcup_{\underline{a}} \bigcup_{i=\theta[0]}^{0} \left\{L^{(1)}\left(\underline{a}F^{(1)}_{\{\theta[0],i-1\}}(\underline{a}, U_{\theta[0]}^{i-1}), U_i\right) > k\right\}\right) \leq \mathbb{P}\left(\bigcup_{i=\theta[0]}^{0} \left\{\ell\left(U_{-\infty}^i\right) > k\right\}\right). \tag{22}$$

Proof. For any U_{-k}^{-1}, the way the sets of strings $\{\underline{z}F^{(1)}_{\{-k,-1\}}(\underline{z}, U_{-k}^{-1})\}_{\underline{z}}$ and $J_k(U_{-k}^{-1})$ are defined ensure that the former is included in the later. It follows that, for any U_{-k}^{-1},

$$A_k(U_{-k}^{-1}) := \inf_{x_{-k}^{-1} : \underline{x} \in J_k(U_{-k}^{-1})} \sum_{a \in A} \inf_{\underline{z}} P(a|x_{-k}^{-1}\underline{z}) \leq \sum_{a \in A} \inf_{\underline{z}} P(a|F^{(1)}_{\{-k,-1\}}(\underline{z}, U_{-k}^{-1})\underline{z}).$$

As the inequality $A_0 \leq \sum_{a \in A} \inf_{\underline{z}} P(a|\underline{z})$ is also true, we deduce that, for any $k \geq 0$, any $\underline{z} \in A^{-\mathbb{N}}$, and any $U_{-\infty}^0$

$$\ell(U_{-\infty}^0) \leq k \Rightarrow L^{(1)}\left(\underline{z}F^{(1)}_{\{-k,-1\}}(\underline{z}, U_{-k}^{-1}), U_0\right) \leq k\ .$$

By recurrence, this means that, for all $\theta[0] \leq i \leq 0$, $F^{(1)}_{\{\theta[0],i\}}(\underline{z}, U_{\theta[0]}^i)$ does not depend on $\underline{z}$. Hence, $\theta[0]$ is also a coalescence time for the update function $F^{(1)}$, that is $\theta[0] \in \Theta^{(1)}[0]$. Observe that we have proved, more specifically, that $\theta[0] \in \Theta'^{(1)}[0]$, where $\Theta'^{(1)}[0] \subset \Theta^{(1)}[0]$ is defined by 13 using $F^{(1)}$ and $L^{(1)}$. By Lemma 5.6 below, this implies that $\theta[0] \in \Theta^{(1),[k]}[0]$ for any $k \geq 0$ as well.

We now prove the second statement of the lemma. If there exist $i \geq k$, U_{-i}^{-1} and $\underline{z}$ such that $L^{(1)}\left(\underline{z}F^{(1)}_{\{-i,-1\}}(\underline{z}, U_{-i}^{-1}), U_0\right) > k$, then there exists some past $\underline{a}$ (take $\underline{a} = \underline{z}F^{(1)}_{\{-i,-k-1\}}(\underline{z}, U_{-i}^{-k-1})$ for instance) such that $L^{(1)}\left(\underline{a}F^{(1)}_{\{-k,-1\}}(\underline{a}, U_{-k}^{-1}), U_0\right) > k$.

We now have the following sequence of inclusions

$$\begin{aligned}
&\bigcup_{\underline{a}} \bigcup_{i=\theta[0]}^{0} \left\{L^{(1)}\left(\underline{a}F^{(1)}_{\{\theta[0],i-1\}}(\underline{a}, U_{\theta[0]}^{i-1}), U_i\right) > k\right\} \\
&\qquad = \bigcup_{\underline{a}} \bigcup_{i=\theta[0]}^{0} \left\{L^{(1)}\left(\underline{a}F^{(1)}_{\{\theta[0],i-1\}}(\underline{a}, U_{\theta[0]}^{i-1}), U_i\right) > k\right\} \cap \{\theta[0] \leq i-k-1\} \\
&\qquad \subset \bigcup_{\underline{a}} \bigcup_{i=\theta[0]}^{0} \left\{L^{(1)}\left(\underline{a}F^{(1)}_{\{i-k,i-1\}}(\underline{a}, U_{i-k}^{i-1}), U_i\right) > k\right\} \cap \{\theta[0] \leq i-k-1\} \\
&\qquad \subset \bigcup_{i=\theta[0]}^{0} \left\{\ell\left(U_{-\infty}^i\right) > k\right\}.
\end{aligned}$$

This concludes the proof of the lemma. □

Recall the definition (13) of $\Theta'[0]$ for generic range partitions of a weakly non-null kernel P. We will need the following lemma.

Lemma 5.6. *For any $k \geq 0$, $\Theta'[0] \subset \Theta^{[k]}[0]$.*

Proof. Let $\theta[0] \in \Theta'[0]$. For any fixed $k \geq 0$, we separate two cases.

(1) If $\theta[0] \geq -k$, then, by the definition of $\Theta'[0]$, the ranges used by F from $\theta[0]$ to 0 are all smaller than or equals to k, and therefore using (18), we have that the length used by $F^{[k]}$ in the same interval of indexes are the same and the constructed symbols are the same as well. Thus $\theta[0] \in \Theta^{[k]}[0]$.
(2) If $\theta[0] < -k$, then, by the definition of $\Theta'[0]$, we can apply the same method as in the preceding case, and obtain that $\theta[0]$ is a coalescence time for $F^{[k]}$ for the time indexes from $\theta[0]$ up to $\theta[0] + k$. But $\theta[0]$ is also a coalescence time for the time indexes from $\theta[0] + k + 1$ up to 0, since the ranges used by $F^{[k]}$ are always smaller than or equal to k. Thus, in this case also, $\theta[0] \in \Theta^{[k]}[0]$.

□

5.4.2. *Proof of Theorem 4.2.* In the conditions of this theorem, by Theorem 1 in De Santis & Piccioni (2012), $\theta[0]$ is $\mathbb{P}$-a.s. finite. Moreover, by Lemma 5.5, $\theta[0] \in \Theta^{(1)}[0] \cap \Theta^{(1),[k]}[0]$ for any $k \geq 0$. Thus we can apply Lemma 4.1, and obtain, using Lemma 5.5

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq \mathbb{P}\left(\bigcup_{i=\theta[0]}^{0} \left\{ \ell\left(U_{-\infty}^{i}\right) > k \right\} \right)$$

and moreover

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=\theta[0]}^{0} \left\{ \ell\left(U_{-\infty}^{i}\right) > k \right\} \right) &= \mathbb{P}\left(\sum_{i=\theta[0]}^{0} \mathbf{1}\left\{ \ell\left(U_{-\infty}^{i}\right) > k \right\} \geq 1 \right) \\ &\leq \mathbb{E}\left(\sum_{i=\theta[0]}^{0} \mathbf{1}\left\{ \ell\left(U_{-\infty}^{i}\right) > k \right\} \right) \ . \end{aligned}$$

Consider the σ-algebra $\mathcal{F}_k$ generated by U_{-k}^{0}, $k \geq 0$. Then, $\ell(U_{-\infty}^{0})$ is a stopping time with respect to $\mathcal{F}_k$ and, by definition, so is $\theta[0]$. Moreover, $\ell(U_{-\infty}^{i})$ is independent of U_{i+1}^{0} by independence of the U_j's. Finally, by stationarity, $\ell\left(U_{-\infty}^{i}\right) \stackrel{\mathcal{D}}{=} \ell\left(U_{-\infty}^{0}\right)$, hence $\mathbb{E}\left(\mathbf{1}\left\{\ell\left(U_{-\infty}^{i}\right) > k\right\}\right) = \mathbb{E}\left(\mathbf{1}\left\{\ell\left(U_{-\infty}^{0}\right) > k\right\}\right)$, for any $i \in \mathbb{Z}$. By Theorem 1 in De Santis & Piccioni (2012), $\theta[0]$ has finite expectation, hence we can use Wald equality to obtain

$$\bar{d}(\mathbf{X}, \mathbf{X}^{[k]}) \leq \mathbb{E}|\theta[0]|\mathbb{P}(\ell\left(U_{-\infty}^{0}\right) > k) \ . \tag{23}$$

This concludes the proof of Theorem 4.2.

5.5. **Proof of Theorem 4.3.** Recall the definition of the set $\Theta'[0]$ given by (13). If $\theta[0] \in \Theta'[0]$ and $\theta[0] \geq -k$, then we know that the range $L\left(\underline{a}F_{\{\theta[0],i-1\}}(\underline{a}, U_{\theta[0]}^{i-1}), U_i\right) \leq k$ for any $i = \theta[0], \ldots, 0$ and any $\underline{a}$, therefore

$$\bigcup_{\underline{a}} \bigcup_{i=\theta[0]}^{0} \left\{ L\left(\underline{a}F_{\{\theta[0],i-1\}}(\underline{a}, U_{\theta[0]}^{i-1}), U_i\right) > k \right\} \subset \{\theta[0] < -k\}.$$

By Lemma 5.6, any $\theta[0] \in \Theta'[0]$ also belongs to $\Theta^{[k]}[0]$ for any $k \geq 0$. We can thus apply Lemma 4.1 and conclude the proof of the theorem.

References

Bejerano, G. & Yona, G. (2001). Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics* **17**(1), 23–43.

Bramson, M. & Kalikow, S. (1993). Nonuniqueness in g-functions. *Israel J. Math.* **84**(1-2), 153–160.

Bressaud, X., Fernández, R. & Galves, A. (1999). Decay of correlations for non-Hölderian dynamics. A coupling approach. *Electron. J. Probab.* **4**, no. 3, 19 pp. (electronic).

Busch, J. R., Ferrari, P. A., Flesia, A. G., Fraiman, R., Grynberg, S. P. & Leonardi, F. (2009). Testing statistical hypothesis on random trees and applications to the protein classification problem. *Annals of applied statistics* **3**(2).

Coelho, Z. & Quas, A. (1998). Criteria for $\bar{d}$-continuity. *Transactions of the American Mathematical Society* **350**, 3257–3268.

Collet, P., Duarte, D. & Galves, A. (2005). Bootstrap central limit theorem for chains of infinite order via Markov approximations. *Markov Process. Related Fields* **11**(3), 443–464.

Comets, F., Fernández, R. & Ferrari, P. A. (2002). Processes with long memory: regenerative construction and perfect simulation. *Ann. Appl. Probab.* **12**(3), 921–943. URL `http://dx.doi.org/10.1214/aoap/1031863175`.

Csiszár, I. & Talata, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory* **52**(3), 1007–1016.

Csiszár, I. & Talata, Z. (2010). On rate of convergence of statistical estimation of stationary ergodic processes. *IEEE Trans. Inform. Theory* **56**(8), 3637–3641.

De Santis, E. & Piccioni, M. (2012). Backward coalescence times for perfect simulation of chains with infinite memory. *J. Appl. Probab.* **49**(2), 319–337.

Duarte, D., Galves, A. & Garcia, N. L. (2006). Markov approximation and consistent estimation of unbounded probabilistic suffix trees. *Bull. Braz. Math. Soc. (N.S.)* **37**(4), 581–592.

Fernández, R. & Galves, A. (2002). Markov approximations of chains of infinite order. *Bull. Braz. Math. Soc. (N.S.)* **33**(3), 295–306. Fifth Brazilian School in Probability (Ubatuba, 2001).

Fernández, R. & Maillard, G. (2005). Chains with complete connections: general theory, uniqueness, loss of memory and mixing properties. *J. Stat. Phys.* **118**(3-4), 555–588. URL `http://dx.doi.org/10.1007/s10955-004-8821-5`.

Gallo, S. (2011). Chains with unbounded variable length memory: perfect simulation and a visible regeneration scheme. *Adv. in Appl. Probab.* **43**(3), 735–759. URL `http://dx.doi.org/10.1239/aap/1316792668`.

Gallo, S. & Garcia, N. L. (2011). General context-tree-based approch to perfect simulation for chains of infinite order. *Submitted, arXiv: 1103.2058v3* .

Galves, A., Galves, C., García, J. E., Garcia, N. L. & Leonardi, F. (2012). Context tree selection and linguistic rhythm retrieval from written texts. *Ann. Appl. Stat.* **6**(1), 186–209. URL `http://dx.doi.org/10.1214/11-AOAS511`.

Hulse, P. (1991). Uniqueness and ergodic properties of attractive g-measures. *Ergodic Theory Dynam. Systems* **11**(1), 65–77. URL `http://dx.doi.org/10.1017/S0143385700006015`.

Johansson, A., Öberg, A. & Pollicott, M. (2010). Unique bernoulli g-measures. *arXiv:1004.0650v1* , 1–18.

KALLENBERG, O. (2002). *Foundations of modern probability.* Probability and its Applications (New York). New York: Springer-Verlag, second ed.

ORNSTEIN, D. S. (1974). *Ergodic theory, randomness, and dynamical systems.* New Haven, Conn.: Yale University Press. James K. Whittemore Lectures in Mathematics given at Yale University, Yale Mathematical Monographs, No. 5.

PROPP, J. G. & WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. In: *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, vol. 9.

QUAS, A. (1996). Non-ergodicity for c^1 expanding maps and g-measures. *Ergodic Theory and Dynamical Systems* **16**, 531–544.

RISSANEN, J. (1983). A universal data compression system. *IEEE Trans. Inform. Theory* **29**(5), 656–664.

WALTERS, P. (2007). A natural space of functions for the ruelle operator theorem. *Ergodic Theory and Dynamical Systems* **27**, 1323–1348.

APPENDIX A. LOCAL CONTINUITY WITH RESPECT TO THE PAST $\underline{1}$

In this section, we assume that $A = \{1, 2\}$, and that P has only one discontinuity point, the point $\underline{1} = \ldots 111$. We refer the interested reader to Gallo & Garcia (2011) for examples with discontinuities in more complicated set of pasts. To begin, we need the following definition.

Definition A.1 (Local continuity with respect to the past $\underline{1}$)**.** *We say that a kernel P on $\{1, 2\}$ is* locally continuous with respect to the past $\underline{1}$ *if*

$$\forall i \geq 0, \qquad \inf_{a_{-k}^{-1}} \sum_{a \in A} \inf_{\underline{z}} P(a|1^i 2 a_{-k}^{-1} \underline{z})$$

converges to 1 *as k diverges. We distinguish two particular situations of interest.*

- *We say that P is* strongly locally continuous with respect to $\underline{1}$ *if there exists an integer function $\ell : (\mathbb{N} \cup \{0\}) \to (\mathbb{N} \cup \{0\})$ such that*

$$\forall i \geq 0, \qquad \inf_{a_{-k}^{-1}} \sum_{a \in A} \inf_{\underline{z}} P(a|1^i 2 a_{-k}^{-1} \underline{z}) = 1 \tag{24}$$

  for any $k \geq \ell(i)$, and
- *we say that P is* uniformly locally continuous with respect to $\underline{1}$ *if*

$$\alpha_k^{\underline{1}} := \inf_{i \geq 0} \inf_{a_{-k}^{-1}} \sum_{a \in A} \inf_{\underline{z}} P(a|1^i 2 a_{-k}^{-1} \underline{z}) \tag{25}$$

  converges to 1 *as k diverges.*

Strongly locally continuous kernels are known as *probabilistic context trees*, a model that have been introduced by Rissanen (1983) as a *universal data compression model.* It was first consider, from the "CFTP point of view", by Gallo (2011). The kernel $\bar{P}$ is a simple example which is strongly and uniformly locally continuous with respect to $\underline{1}$.

Assumption 1: P is strongly locally continuous with respect to $\underline{1}$.

Assumption 2: P is uniformly locally continuous with respect to $\underline{1}$.

Notation A.1. *Let us introduce the following notation.*

- *Stationary processes compatible with kernels satisfying Assumptions i=1 and 2 are denoted* $\mathbf{X}^{(i)}$*, and the corresponding canonical* k*-steps Markov approximations are denoted* $\mathbf{X}^{(i),[k]}$*.*
- *We use the notations* $r_0^{(i)} := \alpha_0$ *for i=1 and 2, and for* $k \geq 1$*,*
$$r_k^{(1)} := r_{k-1}^{(1)} \vee (1 - (1-\alpha(2))^{\ell^{-1}(k)})$$
$$r_k^{(2)} := r_{k-1}^{(2)} \vee (1 - (1-\alpha_{\frac{1}{k}})/\alpha(2))$$
where ℓ *and* $\alpha_{\frac{1}{k}}$ *are the parameters of the kernels under assumptions 1 and 2 respectively.*
- *For i=1 and 2*
$$v_k^{(i)} := \sum_{j=1}^{k} \sum_{\substack{t_1,\dots,t_j \geq 1 \\ t_1+\ldots+t_j=k}} \prod_{m=1}^{j} (1 - r_{t_m-1}^{(i)}) \prod_{l=0}^{t_m-2} r_l^{(i)} \tag{26}$$
where $\prod_{l=0}^{-1} := 1$*.*
- *And finally, for any* $k \geq 0$*, let*
$$u_k := \lfloor k\alpha(2)/2 \rfloor \mathbb{P}\left(\left| \sum_{j=0}^{\lfloor k\alpha(2)/2 \rfloor} \xi_j - \frac{\lfloor k\alpha(2)/2 \rfloor}{\alpha(2)} \right| > k/2 \right) \tag{27}$$
It is well-known that this sequence goes exponentially fast to 0 *(see Kallenberg (2002) for instance). An explicit upper bound is obtained in Appendix C.*

Corollary A.1. *Under the weak non-nullness assumption, we have for i=1 and 2 that, if* $\sum_{k\geq 1} \prod_{i=0}^{k-1} r_k^{(i)} = \infty$*,*
$$\bar{d}(\mathbf{X}^{(i)}, \mathbf{X}^{(i),[k]}) \leq u_k + v^{(i)}_{\lfloor k\alpha(2)/2 \rfloor} \to 0. \tag{28}$$

The quantity defined on display (26) is related to the house of card process presented in Section B (see equation (30)). We provide in Propositions B.1 and B.2 explicit upper-bounds on the term (26) that can be plugged in (28). The term (27) is studied in Corollary C.1. It follows in particular from these propositions that, whenever $r_k^{(i)}$ is not exponentially decreasing, the leading term in (28) is $v_k^{(i)}$, and therefore, we obtain for some constant $C > 1$ and any sufficiently large k
$$\bar{d}(\mathbf{X}^{(i)}, \mathbf{X}^{(i),[k]}) \leq C v^{(i)}_{\lfloor k\alpha(2)/2 \rfloor}.$$
For instance, Proposition B.2, states that, if $1 - r_k^{(i)} = \frac{r}{k} + s_k$, $k \geq 1$ with $r \in (0,1)$ and $\{s_k\}_{k\geq 1}$ is any summable sequence, we obtain for some constant $C > 1$
$$\bar{d}(\mathbf{X}^{(i)}, \mathbf{X}^{(i),[k]}) \leq C \frac{(\log k)^{3+r}}{k^{2-(1+r)^2}}. \tag{29}$$

Proof. Under Assumptions 1 and 2 with weak non-nullness, Gallo & Garcia (2011) constructed a set of range partitions generating a set of coalescence times $\Theta'[0]$ which is a.s. non-empty. This is what is stated in Corollaries 5.1 and 5.2 (and the discussions following them) for respectively Assumption 1 and 2. They defined a random time $\Lambda^{(i)}[0]$ (see

display (26) therein) which belongs to $\Theta'[0]$, as stated by Lemma 6.2 therein. They also prove that $\mathbb{P}(\Lambda^{(i)}[0] < -k)$ is upper bounded by $u_k^{(i)} + v_{\lfloor k\alpha(2)/2 \rfloor}^{(i)}$. This is the content of display (29) therein, making the necessary changes from their notation to our.

By Theorem 4.3, these upper bounds are therefore upper bounds for $\bar{d}(\mathbf{X}, \mathbf{X}^{[k]})$.

□

Appendix B. Some results on the House of Cards Markov chain

Fix a non-decreasing sequence $\{r_k\}_{k\geq 0}$ of $[0,1]$-valued real numbers converging to 1. The house of Cards Markov chain $\mathbf{H} = \{H_n\}_{n\geq 0}$ related to this sequence is the $(\mathbb{N}\cup\{0\})$-valued Markov chain starting from state 0 and having transition matrix $Q = \{Q(i,j)\}_{i\geq 0,\, j\geq 0}$ where $Q(i,j) := r_i \mathbf{1}\{j = i+1\} + (1-r_i)\mathbf{1}\{j=0\}$. Let us denote $v_k := \Pr(H_k = 0)$, the probability that the house of cards is at state 0 at time k. We want to obtain explicit rates of convergence to 0 of this sequence when $\mathbf{H}$ is not positive recurrent. These results will be used in the next section in order to obtain explicit upper bounds for $\bar{d}(\mathbf{X}, \mathbf{X}^{[k]})$ under several types of assumptions. Decomposing the event $\{H_k = 0\}$ into the possible come back of the process $\{H_\ell\}_{\ell=0,\ldots,k}$ to 0 yields, for any $n \geq 1$

$$v_k := \sum_{j=1}^{k} \sum_{\substack{t_1,\ldots,t_j \geq 1 \\ t_1+\ldots+t_j = k}} \prod_{m=1}^{j} (1 - r_{t_m - 1}) \prod_{l=0}^{t_m - 2} r_l, \tag{30}$$

where $\prod_{l=0}^{-1} := 1$. Although explicit, this bound cannot be used directly and has to be simplified. As a first insight, we borrow the following Proposition of Bressaud *et al.* (1999).

Proposition B.1. *(i) v_k goes to zero as k diverges if $\sum_{m\geq 1} \prod_{l=0}^{m-1} r_l = +\infty$,*
(ii) v_k is summable in k if $1 - r_k$ is summable in k,
(iii) v_k behaves as $O(1-r_k)$ if $1-r_k$ is summable in k and $\sup_j \limsup_{k\to+\infty} (\frac{1-r_j}{1-t_{kj}}) \leq 1$
(iv) v_k goes to zero exponentially fast if $1 - r_k$ decreases exponentially.

As observe in Bressaud *et al.* (1999), the conditions of item (iii) are satisfied if, for example, $1 - r_k^{(i)} \sim (\log k)^\eta k^{-\zeta}$ for some $\zeta > 1$, and for any η. However, this is one of the only cases in which this proposition yields explicit rates. In the present paper, we will prove the following proposition.

Proposition B.2. *We have the following explicit upper bounds.*

(i) A non summable case: if $1 - r_k = \frac{r}{k} + s_k$, $k \geq 1$ where $r \in (0,1)$ and $\{s_n\}_{n\geq 1}$ is a summable sequence, there exists a constant $C > 0$ such that

$$v_k \leq C \frac{(\ln k)^{3+r}}{(k)^{2-(1+r)^2}}.$$

(ii) Generic summable case: if $t_\infty := \prod_{k\geq 0} r_k > 0$, then

$$v_k \leq \inf_{K=1,\ldots,k} \left\{ K^2 (1 - r_{k/K}) + (1 - t_\infty)^K \right\}.$$

(iii) Exponential case: if $1 - r_k \leq C_r r^k$, $k \geq 1$, for some $r \in (0,1)$ and a constant $C_r \in (0, \log \frac{1}{r})$ then

$$v_k \leq \frac{1}{C_r} (e^{C_r} r)^k.$$

B.1. **Proof of Proposition B.2.** Before we come into the proofs of each item of this proposition, let us collect some simple remarks on the House of Cards Markov chain.

Let $\{T_k\}_{k\geq 0}$ be a sequence of the stopping times defined as $T_0 := 0$ and, recursively, for any $k \geq 1$, $T_k := \inf\{\, l \geq T_{k-1}+1 \text{ s.t. } H_l = 0\,\}$. The Markov property ensures that the random variables $I_k := T_{k+1} - T_k$ are i.i.d., valued in $\mathbb{N}$ and it is easy to check that

$$\forall k \geq 1, \qquad t_k := \Pr(I_1 = k) = (1 - r_{k-1}) \prod_{i=0}^{k-2} r_i \ ,$$

where $\prod_{l=0}^{-1} := 1$. We have, for any $n \geq 0$,

$$\Pr(H_n = 0) = \Pr(\exists k \geq 0, \text{ s.t. } T_k = n) = \sum_{k=0}^{\infty} \Pr(T_k = n) \ .$$

We write $T_k = \sum_{l=0}^{k-1} I_l$. As all the $I_l \geq 1$, we have $\Pr(T_k = n) = 0$ for all $k > n$. Therefore, for all $K \in [1, n]$,

$$\Pr(H_n = 0) = \sum_{k=0}^{n} \Pr(T_k = n) = \sum_{k=0}^{K} \Pr(T_k = n) + \sum_{k=K+1}^{n} \Pr(T_k = n) \ . \tag{31}$$

Fact B.1. *Let $K \in [1, n]$, we have $\Pr(\forall l \in [1, K],\ I_l \leq n) = (1 - \nu_{n+1})^K$. In particular, if $K \in [1, n]$, then*

$$\Pr\left(\exists j \in [K, n], \text{ s.t. } \sum_{l=0}^{j} I_l = n\right) \leq \Pr(\forall l \in [1, K], I_l \leq n)$$
$$= (1 - \nu_{n+1})^K \ .$$

In order to control $\sum_{k=0}^{K} \Pr(T_k = n) = \Pr(\exists k = 0, \ldots, K,\ T_k = n)$, we can simply remark that, if there exists $k \in 1, \ldots K$ such that $\sum_{i=1}^{k} I_l = n$, there exists necessarily $i \in [1, K]$ and $r \in [1, \ldots, K]$ such that $I_i = n/r$. This implies that

$$\Pr(\exists k = 0, \ldots, K,\ T_k = n) \leq \Pr\left(\exists i \in [1, K],\ \exists r \in [1, \ldots, K], \text{ s.t. } I_i = \frac{n}{r}\right)$$
$$\leq \sum_{i=1}^{K} \sum_{j=1}^{K} \Pr\left(I_1 = \frac{n}{r}\right) \leq K^2 t_{n/K} \ .$$

We obtained the following result.

Fact B.2. *Let $K \in [1, n]$, we have*

$$\sum_{k=0}^{K} \Pr(T_k = n) \leq K^2 t_{n/K} \ .$$

Restricting our attention to the summable case (that is, when $\sum_{k\geq 0}(1 - r_k) < +\infty$), the following fact is fundamental. Its proof is immediate.

Fact B.3. *If $\sum_{n\geq 0}(1 - r_n) < \infty$, then $t_\infty := \Pr(I_1 = \infty) = \prod_{i=0}^{\infty} r_i > 0$, in particular, $\nu_n := \Pr(I_1 \geq n) \geq t_\infty > 0$. Moreover, for all $n \geq 0$, $t_\infty(1 - r_n) \leq t_n \leq (1 - r_n)$*

Using Facts B.1, B.2 and B.3, we are ready to prove items (i) and (ii) of Proposition B.2.

Proof of Item (i) of Proposition B.2. As far as we know, all the results on the house of card process hold in the summable case. When $\sum_{k\geq 0}(1-r_k)=\infty$, it is only known that $\sum_{n\geq 0}\Pr(H_n=0)=\infty$. It is interesting to notice that we can still obtain some rate of convergence for $\Pr(H_n=0)$ from our elementary facts, at least in the following example. Let us assume that there exists $r<1$ and a summable sequence s_n such that, for all $n\geq 1$, $1-r_n=\frac{r}{n}+s_n$. In this case, we have $\sum_{n\geq 0}(1-r_n)=\infty$, therefore $t_\infty=0$. Nevertheless,

$$\prod_{i=0}^{n} r_i \leq \prod_{i=1}^{n} e^{-(1-r_i)} = e^{-r\ln n+O(1)} \leq Cn^{-r} \ .$$

Therefore $t_n\leq Cn^{-(1+r)}$. Moreover, using the inequality $(1-u)\geq e^{-u-u^2}$, valid for all $u<1/8$, we see that $t_n\geq cn^{-(1+r)}$. Therefore, $\nu_n=\sum_{k\geq n}t_k\geq cn^{-r}$. It follows from Fact B.1 that, for large K and n,

$$\sum_{k=K+1}^{n}\Pr(T_k=n)\leq(1-\nu_{n+1})^K\leq e^{-cKn^{-r}} \ .$$

Using Fact B.1, we also have

$$\sum_{k=0}^{K}\Pr(T_k=n)\leq CK^2t_{n/K}\leq CK^{3+r}n^{-(1+r)} \ . \tag{32}$$

We deduce then from (31) that, for all $K\in[0,n]$,

$$\Pr(H_n=0)\leq C\left(\frac{K^{3+r}}{n^{1+r}}+e^{-cKn^{-r}}\right) \ .$$

For $K=2n^r\ln n$, we obtain

$$\Pr(H_n=0)\leq C\frac{(\ln n)^{3+r}}{n^{1-2r-r^2}}=C\frac{(\ln n)^{3+r}}{n^{2-(1+r)^2}} \ .$$

If $0<r<1$, we have $2-(1+r)^2>0$. This bound may not be optimal, but it is interesting to see that we still can obtain rates of convergence from our basic remarks even in this pathological example. □

Proof of Item (ii) of Proposition B.2. We deduce from Facts B.1 and B.3 that, in the summable case

$$\sum_{k=K+1}^{n}\Pr(T_k=n)\leq(1-t_\infty)^K \ .$$

Therefore, from Facts B.2 and B.3,

$$\Pr(H_n=0)\leq\inf_{K=1,\dots,n}\left\{K^2(1-r_{n/K})+(1-t_\infty)^K\right\} \ . \tag{33}$$

□

Proof of Item (iii) of Proposition B.2. In this section, we assume that, for all k, $1-r_k\leq C_rr^k$, for some $r\in(0,1)$ and a constant $C_r>0$. In that case, for all k, we have, by

independence,

$$\begin{aligned}\Pr\left(\sum_{l=1}^{k} I_l = n\right) &= \sum_{i_1+\ldots+i_k=n} \Pr\left(\bigcap_{l=1}^{k} I_l = i_l\right) \\ &= \sum_{i_1+\ldots+i_k=n} \prod_{l=1}^{k} \Pr\left(I_l = i_l\right) \\ &\leq \sum_{i_1+\ldots+i_k=n} C_r^k r^{i_1+\ldots+i_k} = C_r^k r^n \sum_{i_1+\ldots+i_k=n} 1 \ .\end{aligned}$$

Let us evaluate the numbers $p_{k,n} = \sum_{i_1+\ldots+i_k=n} 1$. We have $p_{1,n} = 1$ and

$$\begin{aligned}p_{k,n} &= \sum_{l=1}^{n-k+1} \sum_{i_k=l} \sum_{i_1+\ldots+i_{k-1}=n-l} 1 = \sum_{l=1}^{n-k+1} p_{1,l} p_{k-1,n-l} \\ &= \sum_{l=1}^{n-k+1} p_{k-1,n-l} \ .\end{aligned}$$

Let us then assume that, for some k, we have, for all $n \geq k-1$, $p_{k-1,n} \leq n^{k-2}/(k-2)!$. Notice that this is the case for $k = 2$, then, for all $n \geq k$,

$$p_{k,n} \leq \sum_{l=1}^{n-k+1} \frac{(n-l)^{k-2}}{(k-2)!} = \sum_{l=k-1}^{n-1} \frac{l^{k-2}}{(k-2)!} \leq \int_{k-1}^{n} \frac{x^{(k-2)}}{(k-2)!} \leq \frac{n^{k-1}}{(k-1)!} \ .$$

We deduce that

$$\sum_{k=1}^{n} C_r^k \sum_{i_1+\ldots+i_k=n} 1 \leq \frac{1}{C_r} \sum_{k=1}^{n} \frac{(C_r n)^{k-1}}{(k-1)!} \leq \frac{e^{C_r n}}{C_r} \ .$$

Therefore,

$$\Pr\left(H_n = 0\right) = \sum_{k=1}^{n} \Pr\left(T_k = n\right) \leq \frac{1}{C_r} (e^{C_r} r)^n \ .$$

Hence, when $C_r < \ln(1/r)$, $e^{C_r} r < 1$ and $\Pr\left(H_n = 0\right)$ decreases exponentially fast. □

Appendix C. Concentration of geometric random variables

Let $\xi, \xi_{1:n}$ be i.i.d. geometric random variables with parameter α, i.e., $\forall k \geq 1, \mathbb{P}\left(\xi = k\right) = (1-\alpha)^{k-1}\alpha$. We obtain in this section the following upper bounds.

Proposition C.1. *let* $C_{1,\alpha} = \frac{1-\alpha}{\alpha} + 4\left(\frac{1-\alpha}{\alpha}\right)^2$, $C_{2,\alpha} = \ln\left(\frac{2-\alpha}{2(1-\alpha)} \wedge 2\right)$. *Then,* $\forall x > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{\alpha} > x\right) \leq e^{-n\left(\frac{x^2}{2C_{1,\alpha}} \wedge \frac{C_{2,\alpha}}{2} x\right)} \ . \tag{34}$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{\alpha} < -x\right) \leq e^{-n\left(\frac{x^2}{2C_{1,\alpha}} \wedge \frac{x}{2}\right)} \ .$$

As a corollary of this result, we obtain the following bound when $n = \lfloor k\alpha/2 \rfloor$ and $x = 1/\alpha$.

Corollary C.1. *Let $k \in \mathbb{N}$, $\alpha \in (0,1)$, $n = \lfloor k\alpha/2 \rfloor$, $x = k/(2n) \geq 1/\alpha$, $\xi_{1:n}$ be i.i.d. random variables with parameters α, and*

$$u_k := n\mathbb{P}\left(\left| \sum_{j=1}^{n} \xi_j - \frac{n}{\alpha} \right| > nx \right) \ .$$

Then, we have, for $C_{3,\alpha} := \frac{\alpha}{4(1-\alpha)(4-3\alpha)} \wedge \frac{1}{4} \ln\left(\frac{2-\alpha}{2(1-\alpha)} \wedge 2 \right)$, for all $\epsilon > 0$ and all $k > k(\epsilon)$,

$$u_k \leq \alpha e^{-k(C_{3,\alpha}-\epsilon)} \ .$$

C.1. **Chernov's bound.** Let $Y, Y_{1:n}$ be i.i.d. random variables such that $\forall a < \lambda < b$, $\mathbb{E}\left(e^{\lambda Y}\right) < \infty$, then,

$$\forall x > 0, \qquad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} Y_i > x \right) \leq \inf_{na<\lambda<nb} e^{-\lambda x} \left(\mathbb{E}\left(e^{\frac{\lambda}{n} Y} \right) \right)^n \ . \tag{35}$$

Proof. We have, by independence of the Y_i and Markov's inequality, for all $na < \lambda < nb$,

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} Y_i > x \right) = \mathbb{P}\left(e^{\frac{\lambda}{n} \sum_{i=1}^{n} Y_i} > e^{\lambda x} \right) \leq e^{-\lambda x} \mathbb{E}\left(e^{\frac{\lambda}{n} \sum_{i=1}^{n} Y_i} \right) = e^{-\lambda x} \left(\mathbb{E}\left(e^{\frac{\lambda}{n} Y} \right) \right)^n \ .$$

□

C.2. **Exponential moments of geometric random variables.** Let ξ be a geometric random variable with parameter α, then

$$\begin{aligned} \forall \lambda < -\ln(1-\alpha), \qquad & \mathbb{E}\left(e^{\lambda \xi} \right) \leq \frac{\alpha e^{\lambda}}{1-(1-\alpha)e^{\lambda}} \ , \\ \forall \lambda > \ln(1-\alpha), \qquad & \mathbb{E}\left(e^{\lambda(-\xi)} \right) \leq \frac{\alpha e^{-\lambda}}{1-(1-\alpha)e^{-\lambda}} \ . \end{aligned} \tag{36}$$

Proof. By definition, we have, $\forall \lambda < -\ln(1-\alpha)$,

$$\mathbb{E}\left(e^{\lambda \xi} \right) = \sum_{k \geq 1} e^{\lambda k} (1-\alpha)^{k-1} \alpha = \alpha e^{\lambda} \sum_{k \geq 0} \left((1-\alpha) e^{\lambda} \right)^k = \frac{\alpha e^{\lambda}}{1-(1-\alpha)e^{\lambda}} \ .$$

Moreover, for all $\lambda > \ln(1-p)$,

$$\mathbb{E}\left(e^{-\lambda \xi} \right) = \alpha e^{-\lambda} \sum_{k \geq 0} \left((1-\alpha) e^{-\lambda} \right)^k = \frac{\alpha e^{-\lambda}}{1-(1-\alpha)e^{-\lambda}} \ .$$

C.3. **Proof of the deviation bounds.** Plugging (36) in (35), we obtain, for all $\lambda < -n\ln(1-\alpha)$,

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \xi_i > \frac{1}{\alpha} + x \right) &\leq e^{-\lambda\left(\frac{1}{\alpha}+x\right)} \left(\frac{\alpha e^{\lambda/n}}{1-(1-\alpha)e^{\lambda/n}} \right)^n \\ &= \alpha^n e^{-\lambda\left(\frac{1}{\alpha}+x-1\right)} e^{-n\ln\left(1-(1-\alpha)e^{\lambda/n}\right)} \end{aligned}$$

Choosing $\lambda = n\epsilon$ for $\epsilon \leq \ln\Big(\frac{2-\alpha}{2(1-\alpha)} \wedge 2\Big)$, using the inequalities $e^\epsilon \leq 1+\epsilon+\epsilon^2$ for all $\epsilon \leq \ln 2$ and $-\ln(1-u) \leq 1+u+u^2$ when $u \leq 1/2$, this last bound is equal to

$$\Big(\alpha e^{-\epsilon\left(\frac{1}{\alpha}+x-1\right)} e^{-\ln\left(1-(1-\alpha)e^\epsilon\right)}\Big)^n$$

$$\leq \left(\alpha e^{-\epsilon\left(\frac{1}{\alpha}+x-1\right)} e^{-\ln(\alpha)-\ln\left(1-\frac{(1-\alpha)}{\alpha}(e^\epsilon-1)\right)}\right)^n \leq \left(e^{-\epsilon\left(\frac{1}{\alpha}+x-1\right)} e^{\frac{(1-\alpha)}{\alpha}(e^\epsilon-1)+\left(\frac{(1-\alpha)}{\alpha}(e^\epsilon-1)\right)^2}\right)^n$$

$$\leq e^{-n\epsilon\left(x-\epsilon\left(\frac{1-\alpha}{\alpha}+4\left(\frac{1-\alpha}{\alpha}\right)^2\right)\right)} \ .$$

Let $C_\alpha = \frac{1-\alpha}{\alpha} + 4\left(\frac{1-\alpha}{\alpha}\right)^2$, choosing $\epsilon \leq x/(2C_\alpha)$, we have $x - \epsilon C_\alpha \geq x/2$, hence, choosing $\epsilon = \frac{x}{2C_\alpha} \wedge \ln\Big(\frac{2-\alpha}{2(1-\alpha)} \wedge 2\Big)$, we conclude the proof. Plugging (36) in (35), we obtain, for all $\lambda > n\ln(1-\alpha)$,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \xi_i < \frac{1}{\alpha} - x\right) &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n(-\xi_i) > -\frac{1}{\alpha} + x\right) \\ &\leq e^{-\lambda\left(-\frac{1}{\alpha}+x\right)}\left(\frac{\alpha e^{-\lambda/n}}{1-(1-\alpha)e^{-\lambda/n}}\right)^n \\ &= \alpha^n e^{-\lambda\left(-\frac{1}{\alpha}+x+1\right)} e^{-n\ln\left(1-(1-\alpha)e^{-\lambda/n}\right)}\end{aligned}$$

Choosing $\lambda = n\epsilon$, with $\epsilon \leq 1$, this last bound is equal to

$$\Big(\alpha e^{-\epsilon\left(-\frac{1}{\alpha}+x+1\right)} e^{-\ln\left(1-(1-\alpha)e^{-\epsilon}\right)}\Big)^n$$

$$= \left(\alpha e^{-\epsilon\left(-\frac{1}{\alpha}+x+1\right)} e^{-\ln(\alpha)-\ln\left(1-(e^{-\epsilon}-1)\frac{1-\alpha}{\alpha}\right)}\right)^n \leq \left(e^{-\epsilon\left(-\frac{1}{\alpha}+x+1\right)} e^{(e^{-\epsilon}-1)\frac{1-\alpha}{\alpha}+\left((e^{-\epsilon}-1)\frac{1-\alpha}{\alpha}\right)^2}\right)^n$$

$$\leq \left(e^{-\epsilon\left(-\frac{1}{\alpha}+x+1\right)} e^{(-\epsilon+\epsilon^2)\frac{1-\alpha}{\alpha}+\left((-\epsilon+\epsilon^2)\frac{1-\alpha}{\alpha}\right)^2}\right)^n \leq e^{-n\left[\epsilon x-\epsilon^2\left(\frac{1-\alpha}{\alpha}+4\left(\frac{1-\alpha}{\alpha}\right)^2\right)\right]} \ .$$

Let $C_\alpha = \frac{1-\alpha}{\alpha} + 4\left(\frac{1-\alpha}{\alpha}\right)^2$, choosing $\epsilon \leq x/(2C_\alpha)$, we have $x - \epsilon C_\alpha \geq x/2$, hence, choosing $\epsilon = \frac{x}{2C_\alpha} \wedge 1$, we conclude the proof. □

Departamento de métodos estatísticos, Instituto de Matématica, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, 21945-970, Rio de Janeiro, Brasil

E-mail address: sandro@im.ufrj.br

Laboratoire J.A.Dieudonné UMR CNRS 6621, Université de Nice Sophia-Antipolis, Parc Valrose, 06108 Nice Cedex 2, France

E-mail address: mlerasle@unice.fr

Institute of Neuroscience and Psychology Department, Princeton University, Green Hall, NJ, 08540

E-mail address: takahashiyd@gmail.com