Electronic Journal of Probability

Vol. 13 (2008), Paper no. 48, pages 1345–1361.

Journal URL
http://www.math.washington.edu/~ejpecp/

Random perturbations of stochastic processes with unbounded variable length memory*

PIERRE COLLET
Centre de Physique Théorique, CNRS UMR 7644,
Ecole Polytechnique, 91128 Palaiseau Cedex, France
collet@cpht.polytechnique.fr

ANTONIO GALVES
Instituto de Matemática e Estatística
Universidade de São Paulo,
BP 66281, 05315-970 São Paulo, Brasil
galves@ime.usp.br

FLORENCIA LEONARDI
Instituto de Matemática e Estatística
Universidade de São Paulo,
BP 66281, 05315-970 São Paulo, Brasil
florencia@usp.br

Abstract

We consider binary infinite order stochastic chains perturbed by a random noise. This means that at each time step, the value assumed by the chain can be randomly and independently flipped with a small fixed probability. We show that the transition probabilities of the perturbed chain are uniformly close to the corresponding transition probabilities of the original chain. As a consequence, in the case of stochastic chains with unbounded but otherwise finite variable length memory, we show that it is possible to recover the context tree of the original chain, using a suitable version of the algorithm Context, provided that the noise is small enough.

*This work is part of PRONEX/FAPESP's project *Stochastic behavior, critical phenomena and rhythmic pattern identification in natural languages* (grant number 03/09930-9), CNRS-FAPESP project *Probabilistic phonology of rhythm* and CNPq's projects *Stochastic modeling of speech* (grant number 475177/2004-5) and *Rhythmic patterns, prosodic domains and probabilistic modeling in Portuguese Corpora* (grant number 485999/2007-2). AG is partially supported by a CNPq fellowship (grant 308656/2005-9) and FL is supported by a FAPESP fellowship (grant 06/56980-0).

Key words: chains of infinite order, variable length Markov chains, chains with unbounded variable length memory, random perturbations, algorithm Context, context trees.

AMS 2000 Subject Classification: Primary 62M09, 60G99.

Submitted to EJP on July 23, 2007, final version accepted July 17, 2008.

1 Introduction

The original motivation of this paper is the following question. Is it possible to recover the context tree of a variable length Markov chain from a noisy sample of the chain. We recall that in a variable length Markov chain the conditional probability of the next symbol, given the past, depends on a variable portion of the past whose length depends on the past itself. This class of models were first introduced by Rissanen (1983) who called them *finite memory sources* or *tree machines*. They recently became popular in the statistics literature under the name of *variable length Markov chains* coined by Bühlmann and Wyner (1999).

The notion of variable memory model can be naturally extended to a non Markovian situation where the contexts are still finite, but their lengths are no longer bounded. We refer the reader to Galves and Löcherbach (2008) for a recent survey of the topic. This leads us to consider not only randomly perturbed unbounded variable length memory models, but more generally randomly perturbed infinite order stochastic chains.

We will consider binary chains of infinite order in which at each time step the value assumed by the chain can be randomly and independently flipped with a small fixed probability. Even if the original chain is Markovian, the perturbed chain is in general a chain of infinite order. (we refer the reader to Fernández et al. (2001) for a self contained introduction to chains of infinite order). We show that the transition probabilities of the perturbed chain are uniformly close to the corresponding transition probabilities of the original chain. More precisely, we prove that the difference between the conditional probabilities of the next symbol given a finite past of any fixed length is uniformly bounded above by the probability of flipping, multiplied by a fixed constant. This is the content of our first theorem.

Using this result we are able to solve our original problem of recovering the context tree of a chain with unbounded variable length from a noisy sample. To make this point clear, we recall the notion of *context*. In his original paper, Rissanen used the word context to designate the minimal suffix of the string of past symbols which is enough to define the probability of the next symbol. Rissanen also observed that the set of all contexts satisfies the suffix property, which means that no context is a proper suffix of another context. This property allows to represent the set of all contexts as the set of leaves of a rooted labeled tree, henceforth called the *context tree* of the chain. With this representation the process is described by the tree of all contexts and an associated family of probability measures over the set of symbols, indexed by the leaves of the tree. Given a context, its associated probability measure gives the probability of the next symbol for any past having this context as a suffix.

In his paper Rissanen not only introduced the class of variable memory models but he also introduced the algorithm *Context* to estimate both the context tree and the associated family of probability transition. The way the algorithm Context works can be summarized as follows. Given a sample produced by a chain with variable memory, we start with a maximal tree of candidate contexts for the sample. The branches of this first tree are then pruned until we obtain a minimal tree of contexts well adapted to the sample.

From Rissanen (1983) to Galves et al. (2008), passing by Ron et al. (1996) and Bühlmann and Wyner (1999), several variants of the algorithm Context have been presented in the literature. In all the variants the decision to prune a branch is taken by considering a *cost* function. A branch is pruned if the cost function assumes a value smaller than a given threshold. The estimated context tree is the smallest tree satisfying this condition. The estimated family of probability transitions is the one associated to the minimal tree of contexts.

The proof of the weak consistency of the algorithm Context when the tree of contexts is finite was done in Rissanen (1983). This result was extended in Bühlmann and Wyner (1999) where the weak consistency of the algorithm was proved in the finite case, but allowing the maximal length of the memory to grow with the size of the sample. In both papers the cost function was defined using the log likelihood ratio test to compare two candidate trees and the main ingredient of the consistency proofs was the chi-square approximation to the log likelihood ratio test for Markov chains of fixed order.

The unbounded case was considered by Ferrari and Wyner (2003), Duarte et al. (2006) and Csiszár and Talata (2006). The first two papers essentially extend to the unbounded case the original chi-square approach introduced by Rissanen. Instead of the chi-square, the last paper uses penalized likelihood algorithms, related to the Bayesian Information Criterion (BIC), to estimate the context tree. We refer the reader to Csiszár and Talata (2006) for a nice description of other approaches and results in this field, including the context tree maximizing algorithm by Willems et al. (1995).

In the present paper we use a variant of the algorithm Context introduced in Galves et al. (2008) for finite trees and extended to unbounded trees in Galves and Leonardi (2008). In this variant, the decision of pruning a branch is taken by considering the difference between the estimated conditional probabilities of the original branch and the pruned one, using a suitable threshold. Using exponential inequalities for the estimated transition probabilities associated to the candidate contexts, these papers not only show the consistency of this variant of the algorithm Context, but also provide an exponential upper bound for the rate of convergence.

This version of the algorithm Context does not distinguish transition probabilities which are closer than the threshold level used in the pruning decision. Our first theorem proves that the conditional probabilities of the original variable memory chain and of the perturbed one are uniformly close if the probability of random flipping is small enough. Hence it is natural to expect that with this version of the algorithm Context, one should be able to retrieve the original context tree out from the noisy sample. This is actually the case, as we prove in the second theorem.

The paper is organized as follows. In section 2 we give the definitions and state the main results. Section 3 and 4 are devoted to the proof of Theorem 1 and 2, respectively.

2 Definitions and results

Let A denote the binary alphabet $\{0,1\}$, with size $|A| = 2$. Given two integers $m \le n$ we denote by w_m^n the sequence $w_m, \ldots, w_n$ of symbols in A and A_m^n denotes the set of such sequences. The length of the sequence w_m^n is denoted by $\ell(w_m^n)$ and is defined by $\ell(w_m^n) = n - m + 1$. Any sequence w_m^n with $m > n$ represents the empty string. The same notation is extended to the case $m = -\infty$.

Given two sequences w and v, with $\ell(w) < \infty$, we will denote by vw the sequence of length $\ell(v) + \ell(w)$ obtained by concatenating the two strings. We say that the sequence s is a *suffix* of the sequence w if there exists a sequence u, with $\ell(u) \ge 1$, such that $w = us$. In this case we write $s \prec w$. When $s \prec w$ or $s = w$ we write $s \preceq w$. Given a finite sequence w we denote by $\mathrm{suf}(w)$ the largest suffix of w.

We consider a stationary ergodic stochastic process $(X_t)_{t\in\mathbb{Z}}$ over $A = \{0,1\}$. Given a sequence $w \in A_{-\infty}^{-1}$ and a symbol $a \in A$, we denote by

$$p(a|w) = \mathbb{P}(X_0 = a \mid X_{-1} = w_{-1}, X_{-2} = w_{-2}, \dots)$$

the regular version of the conditional probability of the process. Given a finite sequence $w \in A_{-j}^{-1}$ we denote by

$$p(w) = \mathbb{P}(X_{-j}^{-1} = w)$$

the stationary probability of the cylinder defined by the sequence w.

We assume the process (X_t) satisfies the following conditions

1. *Non-nullness*, that is
$$\alpha := \inf\{p(a|w) \colon a \in A, w \in A_{-\infty}^{-1}\} > 0\ ,$$

2. *Summable continuity rate*, that is
$$\beta := \sum_{k\in\mathbb{N}} \beta_k < +\infty,$$
where the sequence $\{\beta_k\}_{k\in\mathbb{N}}$ is defined by
$$\beta_k := \sup\left\{ \left|1 - \frac{p(a|w)}{p(a|v)}\right| \colon a \in A, v, w \in A_{-\infty}^{-1} \text{ with } w \stackrel{k}{=} v \right\}.$$
Here, $w \stackrel{k}{=} v$ means that there exists a sequence u with $\ell(u) = k$ such that $u \prec w$ and $u \prec v$. The sequence $\{\beta_k\}_{k\in\mathbb{N}}$ is called the *continuity rate*.

In this paper we are interested on the effect of a Bernoulli noise flipping, independent from the successive symbols of the process (X_t). Namely, let $(\xi_t)_{t\in\mathbb{Z}}$ be an i.i.d. sequence of random variables taking values in $\{0,1\}$, independent of (X_t), with

$$\mathbb{P}(\xi_t = 0) = 1 - \epsilon,$$

where ϵ is a fixed noise parameter in $(0,1)$. For a and b in $\{0,1\}$, we define

$$a \oplus b = a + b \pmod 2,$$

and $\bar{a} = 1 \oplus a$. We now define the stochastically perturbed chain $(Z_t)_{t\in\mathbb{Z}}$ by

$$Z_t = X_t \oplus \xi_t.$$

In the case $\epsilon = 1/2$, (Z_t) is an i.i.d. uniform Bernoulli. However, generically it is a process of infinite order.

In what follows we will use the shorthand notation $q(w_{-j}^{-1})$ to denote $\mathbb{P}(Z_{-j}^{-1} = w_{-j}^{-1})$. For any sequence $w = w_{-\infty}^{-1}$ denote by

$$q(a|w) = \mathbb{P}(Z_0 = a \mid Z_{-1} = w_{-1}, Z_{-2} = w_{-2}, \dots)$$

the transition probabilities corresponding to the process (Z_t). We can now state our first theorem.

Theorem 1. *For (X_t) and (Z_t) as above, for any $\epsilon \in (0,1)$ and for any $k \geq 0$,*

$$\sup\{|q(a|w) - p(a|w)|: a \in A, w \in A_{-k}^{-1}\} \;\leq\; \Big[1 + \frac{4\beta}{\min(\alpha\beta^*, 1)}\Big]\,\epsilon\,,$$

where $\beta^* = \prod_{k=0}^{+\infty}(1-\beta_k) \;< +\infty$.

Remark. Here and throughout the rest of the paper we accept conditional events defined by empty sequences, for example the ones appearing in Theorem 1 when $k = 0$. In these cases the convention is that these events are removed from the conditional expressions.

Definition 2.1. A sequence $w \in A_{-j}^{-1}$ is a *context* for the process (X_t) if it satisfies

1. For any semi-infinite sequence $x_{-\infty}^{-1}$ having w as a suffix

$$\mathbb{P}(X_0 = a \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) \,=\, p(a|w), \quad \text{for all } a \in A.$$

2. No suffix of w satisfies (1).

An *infinite context* is a semi-infinite sequence $w_{-\infty}^{-1}$ such that any of its suffixes w_{-j}^{-1}, $j = 1, 2, \ldots$ is a context.

Definition 2.1 implies that the set of all contexts (finite or infinite) can be represented as a rooted tree. This tree is called *the context tree* of the process (X_t) and will be denoted by $\mathscr{T}$. The non-nullness hypothesis implies that the context tree of the process (X_t) is complete, i.e., any sequence in $A_{-\infty}^{-1}$ belongs to $\mathscr{T}$ or has a suffix that belongs to $\mathscr{T}$. We say that the context tree $\mathscr{T}$ is *bounded* if it has a finite number of sequences. In the infinite case we say that $\mathscr{T}$ is *unbounded*. Examples of bounded and unbounded context trees related to renewal processes are presented in Csiszár and Talata (2006).

Given an integer K we will denote by $\mathscr{T}|_K$ the tree $\mathscr{T}$ *truncated* to level K, that is

$$\mathscr{T}|_K = \{w \in \mathscr{T} : \ell(w) \leq K\} \cup \{w : \ell(w) = K \text{ and } w \prec u, \text{ for some } u \in \mathscr{T}\}.$$

Our interest is to recover the truncated context tree of the process (X_t) from a sample of the noisy process (Z_t). We will assume $z_1, z_2 \ldots, z_n$ is a sample of the process (Z_t). For any finite string w with $\ell(w) \leq n$, we denote by $N_n(w)$ the number of occurrences of the string in the sample, that is

$$N_n(w) = \sum_{t=0}^{n-\ell(w)} \mathbf{1}\{z_{t+1}^{t+\ell(w)} = w\}.$$

For any element $a \in A$ and any finite sequence w, the empirical transition probability $\hat{q}_n(a|w)$ is defined by

$$\hat{q}_n(a|w) = \frac{N_n(wa) + 1}{N_n(w\cdot) + |A|}.$$

where

$$N_n(w\cdot) = \sum_{b \in A} N_n(wb)\,.$$

The variant of Rissanen's context tree estimator is defined as follows. First of all, let us define for any finite string w,

$$\Delta_n(w) = \max_{a\in A} |\hat{q}_n(a|w) - \hat{q}_n(a|\mathrm{suf}(w))|.$$

The $\Delta_n(w)$ operator computes a distance between the empirical transition probabilities associated to the sequence w and the one associated to the sequence suf(w).

Definition 2.2. Given $\delta > 0$ and $d < n$, the context tree estimator $\hat{\mathscr{T}}_n^{\delta,d}$ is the set containing all sequences $w \in A_{-d}^{-1}$ such that $\Delta_n(a\,\mathrm{suf}(w)) > \delta$ for some $a \in A$ and $\Delta_n(uw) \le \delta$ for all $u \in A_{-d}^{-\ell(w)}$.

In order to state our second theorem we need some definitions. Given an integer $k \ge 1$, define $\mathscr{C}_k = \{u \in \mathscr{T}|_k : p(a|u) \ne p(a|\mathrm{suf}(u))$ for some $a \in A\}$ and

$$D_k = \min_{u\in\mathscr{C}_k} \max_{a\in A} \{|p(a|u) - p(a|\mathrm{suf}(u))|\}.$$

From the definition we can see that $D_k > 0$ for all $k \ge 1$.

The second main result in this paper is the following.

Theorem 2. *Let K be an integer and let $z_1, z_2 \ldots, z_n$ be a sample of the perturbed process (Z_t). Then, there exist constants c, n_0 and an integer d depending on the process (X_t) such that for any $\epsilon \in (0, D_d/2c)$, any $\delta \in (c\epsilon, D_d - c\epsilon)$ and any $n \ge n_0$ we have*

$$\mathbb{P}(\hat{\mathscr{T}}_n^{\delta,d}|_K \ne \mathscr{T}|_K) \le c_1 \exp\big[-c_2(n-d)\big].$$

The constants are all explicit and are given by

1. $c = 2\big[1 + \frac{4\beta}{\min(\alpha\beta^*,1)}\big]$.
2. $d = \max_{u\notin\mathscr{T},\ell(u)<K} \min\{k\colon$ *there exists* $w \in \mathscr{C}_k$ *with* $w \succ u\}$.
3. $n_0 = \frac{6}{(\min(\delta, D_d-\delta)-c\epsilon)\alpha^d} + d$.
4. $c_1 = 12\, e^{\frac{1}{e}}\, 2^d$ *and* $c_2 = \frac{[\min(\delta,D_d-\delta)-c\epsilon-6/(n-d)\alpha^d]^2\alpha^{2d}}{256e(1+\frac{\beta}{\alpha})(d+1)}$.

As a consequence we obtain the following strong consistency result.

Corollary 3. *For any integer K and for almost all infinite sample $z_1, z_2 \ldots$ there exists a $\bar{n}$ such that, for any $n \ge \bar{n}$ we have*

$$\hat{\mathscr{T}}_n^{\delta,d}|_K = \mathscr{T}|_K,$$

where d and δ are chosen as in Theorem 2.

3 Proof of Theorem 1

We start by proving three preparatory lemmas.

Lemma 4. *For any $\epsilon \in (0,1)$, any $k > j \ge 0$, any $w_{-\infty}^0$ and any $a, b \in A$,*

$$\Big|\mathbb{P}\big(X_0 = w_0 \mid X_{-j}^{-1} = w_{-j}^{-1}, X_{-j-1} = a, Z_{-j-1} = b, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big) - p\big(w_0|w_{-\infty}^{-1}\big)\Big| \le \beta_j.$$

Proof. We observe that for $j \geq 0$ it follows from conditioning on the values of X_{-k}^{-j-2} and the independence of the flipping procedure that

$$\begin{aligned}\mathbb{P}\big(X_0 = & w_0 \,\big|\, X_{-j}^{-1} = w_{-j}^{-1}, X_{-j-1} = a,\, Z_{-j-1} = b,\, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big) \\ &= \frac{\sum_{u_{-k}^{-j-2}} p(u_{-k}^{-j-2} a\, w_{-j}^{-1} w_0)\mathbb{P}\big(Z_{-k}^{-j-1} = w_{-k}^{-j-2} b \mid X_{-k}^{-j-1} = u_{-k}^{-j-2} a\big)}{\sum_{u_{-k}^{-j-2}} p(u_{-k}^{-j-2} a\, w_{-j}^{-1})\mathbb{P}\big(Z_{-k}^{-j-1} = w_{-k}^{-j-2} b \mid X_{-k}^{-j-1} = u_{-k}^{-j-2} a\big)}.\end{aligned}$$

It is easy to see using conditioning on the infinite past that

$$\inf_{v_{-\infty}^{-j-1}} p(w_0 | v_{-\infty}^{-j-1} w_{-j}^{-1}) \;\leq\; p(w_0 | u_{-k}^{-j-2} a w_{-j}^{-1}) \;\leq\; \sup_{v_{-\infty}^{-j-1}} p(w_0 | v_{-\infty}^{-j-1} w_{-j}^{-1}). \tag{3.1}$$

Then, using continuity we have

$$p(w_0 | w_{-\infty}^{-1}) - \beta_j \;\leq\; p(w_0 | u_{-k}^{-j-2} a\, w_{-j}^{-1}) \;\leq\; p(w_0 | w_{-\infty}^{-1}) + \beta_j$$

and the assertion of the Lemma follows immediately. □

Lemma 5. *For any $\epsilon \in (0,1)$, any $k \geq 0$ and any w_{-k}^{0},*

$$q(w_0 | w_{-k}^{-1}) \;\geq\; \alpha$$

and

$$\mathbb{P}\big(X_0 = w_0 \,\big|\, Z_{-k}^{-1} = w_{-k}^{-1}\big) \;\geq\; \alpha.$$

Moreover, for any $0 \leq j \leq k$ we have

$$\mathbb{P}\big(X_{-j-1} = w_{-j-1} \mid X_{-j}^{-1} = w_{-j}^{-1},\, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big) \;\geq\; \alpha\beta^*.$$

Proof. We first observe that

$$q(w_0 | w_{-k}^{-1}) = (1-\epsilon)\mathbb{P}\big(X_0 = w_0 \,\big|\, Z_{-k}^{-1} = w_{-k}^{-1}\big) + \epsilon\, \mathbb{P}\big(X_0 = \bar{w}_0 \,\big|\, Z_{-k}^{-1} = w_{-k}^{-1}\big).$$

It is therefore enough to prove the second assertion. From conditioning on the value of X_{-l}^{-1}, the independence of the flipping procedure and the inequalities in (3.1) we have

$$\begin{aligned}&\mathbb{P}\big(X_0 = w_0 \,\big|\, Z_{-k}^{-1} = w_{-k}^{-1}\big) \;= \\ &\lim_{l\to\infty} \frac{(1-\epsilon)^k \sum_{u_{-l}^{-1}} p\big(w_0 | w_{-\infty}^{-l-1} u_{-l}^{-1}\big)\mathbb{P}\big(X_{-l}^{-1} = u_{-l}^{-1} \,\big|\, X_{-\infty}^{-l-1} = w_{-\infty}^{-l-1}\big)(\epsilon/(1-\epsilon))^{\sum_{j=-k}^{-1} u_j \oplus w_j}}{(1-\epsilon)^k \sum_{u_{-l}^{-1}} \mathbb{P}\big(X_{-l}^{-1} = u_{-l}^{-1} \,\big|\, X_{-\infty}^{-l-1} = w_{-\infty}^{-l-1}\big)(\epsilon/(1-\epsilon))^{\sum_{j=-k}^{-1} u_j \oplus w_j}}\end{aligned}$$

and for each l, the expression in the right hand side is lower bounded by α. Then, the same holds for the limit when $l \to \infty$. For the last assertion we first observe that

$$\begin{aligned}&\mathbb{P}\big(X_{-j-1} = w_{-j-1} \mid X_{-j}^{-1} = w_{-j}^{-1},\, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big) \\ &\frac{\sum_{x_{-k}^{-j-2}} \mathbb{P}\big(Z_{-k}^{-j-2} = w_{-k}^{-j-2} \mid X_{-k}^{-j-2} = x_{-k}^{-j-2}\big)\mathbb{P}\big(X_{-j-1}^{-1} = w_{-j-1}^{-1}, X_{-k}^{-j-2} = x_{-k}^{-j-2}\big)}{\sum_{x_{-k}^{-j-2}} \mathbb{P}\big(Z_{-k}^{-j-2} = w_{-k}^{-j-2} \mid X_{-k}^{-j-2} = x_{-k}^{-j-2}\big)\mathbb{P}\big(X_{-j}^{-1} = w_{-j}^{-1}, X_{-k}^{-j-2} = x_{-k}^{-j-2}\big)}.\end{aligned}$$

Moreover,

$$
\begin{aligned}
&\frac{\mathbb{P}\big(X_{-j-1}^{-1}=w_{-j-1}^{-1},X_{-k}^{-j-2}=x_{-k}^{-j-2}\big)}{\mathbb{P}\big(X_{-j}^{-1}=w_{-j}^{-1},X_{-k}^{-j-2}=x_{-k}^{-j-2}\big)}\\
&\quad= \frac{\prod_{l=1}^{j+1} p(w_{-l}|x_{-k}^{-j-2}w_{-j-1}^{-l-1})\prod_{l=j+2}^{k} p(x_{-l}|x_{-k}^{-l-1})}{\prod_{l=1}^{j}\mathbb{P}(X_{-l}=w_{-l}|X_{-j}^{-l-1}=w_{-j}^{-l-1},X_{-k}^{-j-2}=x_{-k}^{-j-2})\prod_{l=j+2}^{k} p(x_{-l}|x_{-k}^{-l-1})}\\
&\quad= p(w_{-j-1}|x_{-k}^{-j-2})\prod_{l=1}^{j}\frac{p(w_{-l}|x_{-k}^{-j-2}w_{-j-1}^{-l-1})}{\mathbb{P}\big(X_{-l}=w_{-l}|X_{-j}^{-l-1}=w_{-j}^{-l-1},X_{-k}^{-j-2}=x_{-k}^{-j-2}\big)}
\end{aligned}
$$

and using non-nullness and log-continuity this can be bounded below by

$$
\alpha\prod_{l=1}^{j}(1-\beta_{j-l}) \;\geq\; \alpha\beta^*.
$$

This finishes the proof of the Lemma. □

Lemma 6. *For any $\epsilon\in(0,1)$, any $k>j\geq 0$ and any w_{-k}^{0},*

$$
\mathbb{P}\big(X_{-j-1}=\bar{w}_{-j-1}\,\big|\,X_{-j}^{-1}=w_{-j}^{-1},\,Z_{-k}^{-j-1}=w_{-k}^{-j-1}\big)\;\leq\;\frac{\epsilon}{\alpha\beta^*}.
$$

Proof. We have

$$
\begin{aligned}
&\mathbb{P}\big(X_{-j-1}=\bar{w}_{-j-1}\mid X_{-j}^{-1}=w_{-j}^{-1},\,Z_{-k}^{-j-1}=w_{-k}^{-j-1}\big)\\
&\quad= \frac{\mathbb{P}\big(X_{-j-1}=\bar{w}_{-j-1},Z_{-j-1}=w_{-j-1}\mid X_{-j}^{-1}=w_{-j}^{-1},Z_{-k}^{-j-2}=w_{-k}^{-j-2}\big)}{\mathbb{P}\big(Z_{-j-1}=w_{-j-1}\mid X_{-j}^{-1}=w_{-j}^{-1},Z_{-k}^{-j-2}=w_{-k}^{-j-2}\big)}\\
&\quad= \frac{\epsilon\,\mathbb{P}\big(X_{-j-1}=\bar{w}_{-j-1}\mid X_{-j}^{-1}=w_{-j}^{-1},Z_{-k}^{-j-2}=w_{-k}^{-j-2}\big)}{\mathbb{P}\big(Z_{-j-1}=w_{-j-1}\mid X_{-j}^{-1}=w_{-j}^{-1},Z_{-k}^{-j-2}=w_{-k}^{-j-2}\big)}.
\end{aligned}
$$

It follows from Lemma 5 that

$$
\begin{aligned}
&\mathbb{P}\big(Z_{-j-1}=w_{-j-1}\mid X_{-j}^{-1}=w_{-j}^{-1},Z_{-k}^{-j-2}=w_{-k}^{-j-2}\big)\\
&\qquad= (1-\epsilon)\,\mathbb{P}\big(X_{-j-1}=w_{-j-1}\mid X_{-j}^{-1}=w_{-j}^{-1},Z_{-k}^{-j-2}=w_{-k}^{-j-2}\big)\\
&\qquad\quad+\epsilon\,\mathbb{P}\big(X_{-j-1}=\bar{w}_{-j-1}\mid X_{-j}^{-1}=w_{-j}^{-1},Z_{-k}^{-j-2}=w_{-k}^{-j-2}\big)\\
&\qquad\geq \alpha\beta^*.
\end{aligned}
$$

This concludes the proof of Lemma 6. □

Proof of Theorem 1. We first observe that for any $a\in A$ and any $w_{-k}^{-1}\in A_{-k}^{-1}$

$$
q(a|w_{-k}^{-1})=(1-\epsilon)\,\mathbb{P}\big(X_0=a\,\big|\,Z_{-k}^{-1}=w_{-k}^{-1}\big)+\epsilon\,\mathbb{P}\big(X_0=\bar{a}\,\big|\,Z_{-k}^{-1}=w_{-k}^{-1}\big).
$$

Therefore,

$$\left|q(a|w_{-k}^{-1}) - \mathbb{P}\big(X_0 = a \,\big|\, Z_{-k}^{-1} = w_{-k}^{-1}\big)\right| \leq \epsilon$$

and if $k = 0$ the Theorem is proved. We will now assume $k \geq 1$ and we write

$$\begin{aligned}\mathbb{P}\big(X_0 = a \,\big|\, Z_{-k}^{-1} = w_{-k}^{-1}\big) &- \mathbb{P}\big(X_0 = a \,\big|\, X_{-k}^{-1} = w_{-k}^{-1}\big)\\ &= \sum_{j=0}^{k-1}\Big[\mathbb{P}\big(X_0 = a \,\big|\, X_{-j}^{-1} = w_{-j}^{-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big)\\ &\qquad - \mathbb{P}\big(X_0 = a \,\big|\, X_{-j-1}^{-1} = w_{-j-1}^{-1}, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big)\Big].\end{aligned}$$

We will bound each term in the sum separately. We can write

$$\begin{aligned}&\mathbb{P}\big(X_0 = a \mid X_{-j}^{-1} = w_{-j}^{-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big) - \mathbb{P}\big(X_0 = a \mid X_{-j-1}^{-1} = w_{-j-1}^{-1}, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big)\\ &= \sum_{b\in\{0,1\}}\Big[\mathbb{P}\big(X_0 = a \mid X_{-j}^{-1} = w_{-j}^{-1}, X_{-j-1} = b, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big)\\ &\qquad - \mathbb{P}\big(X_0 = a \mid X_{-j-1}^{-1} = w_{-j-1}^{-1}, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big)\Big]\\ &\quad \times \mathbb{P}\big(X_{-j-1} = b \mid X_{-j}^{-1} = w_{-j}^{-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big).\end{aligned}$$

The above sum has two terms. For the term with $b = \bar{w}_{-j-1}$ we can use Lemma 4, Lemma 6 and the inequalities in (3.1) to obtain

$$\begin{aligned}&\Big|\mathbb{P}\big(X_0 = a \mid X_{-j}^{-1} = w_{-j}^{-1}, X_{-j-1} = \bar{w}_{-j-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big)\\ &\qquad - \mathbb{P}\big(X_0 = a \,\big|\, X_{-j-1}^{-1} = w_{-j-1}^{-1}, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big)\Big|\\ &\quad \times \mathbb{P}\big(X_{-j-1} = \bar{w}_{-j-1} \,\big|\, X_{-j}^{-1} = w_{-j}^{-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big) \leq \frac{2\beta_j}{\alpha\beta^*}\epsilon.\end{aligned}$$

For the other term with $b = w_{-j-1}$ we can write

$$\begin{aligned}&\Big|\mathbb{P}\big(X_0 = a \mid X_{-j}^{-1} = w_{-j}^{-1}, X_{-j-1} = w_{-j-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big)\\ &\qquad - \mathbb{P}\big(X_0 = a \,\big|\, X_{-j-1}^{-1} = w_{-j-1}^{-1}, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big)\Big|\\ &\quad \times \mathbb{P}\big(X_{-j-1} = w_{-j-1} \,\big|\, X_{-j}^{-1} = w_{-j}^{-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big)\\ &\leq \sum_{c\in\{0,1\}}\Big|\mathbb{P}\big(X_0 = a \,\big|\, X_{-j}^{-1} = w_{-j}^{-1}, X_{-j-1} = w_{-j-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big)\\ &\qquad - \mathbb{P}\big(X_0 = a \,\big|\, X_{-j-1}^{-1} = w_{-j-1}^{-1}, Z_{-j-1} = c, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big)\Big|\\ &\quad \times \mathbb{P}\big(Z_{-j-1} = c \,\big|\, X_{-j-1}^{-1} = w_{-j-1}^{-1}, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big)\\ &\quad \times \mathbb{P}\big(X_{-j-1} = w_{-j-1} \,\big|\, X_{-j}^{-1} = w_{-j}^{-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big).\end{aligned}$$

Note that the term with $c = w_{-j-1}$ vanishes. For $c = \bar{w}_{-j-1}$ we can use Lemma 4 and the inequalities in (3.1) to bound above the last sum with

$$\begin{aligned}&\Big|\mathbb{P}\big(X_0 = a \,\big|\, X_{-j}^{-1} = w_{-j}^{-1}, X_{-j-1} = w_{-j-1}, Z_{-k}^{-j-1} = w_{-k}^{-j-1}\big)\\ &\qquad - \mathbb{P}\big(X_0 = a \,\big|\, X_{-j-1}^{-1} = w_{-j-1}^{-1}, Z_{-j-1} = \bar{w}_{-j-1}, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big)\Big|\\ &\quad \times \mathbb{P}\big(Z_{-j-1} = \bar{w}_{-j-1} \,\big|\, X_{-j-1}^{-1} = w_{-j-1}^{-1}, Z_{-k}^{-j-2} = w_{-k}^{-j-2}\big) \leq 2\beta_j\,\epsilon.\end{aligned}$$

Putting all the above bounds together we get

$$\left|\mathbb{P}\big(Z_0 = w_0 \,\big|\, Z_{-k}^{-1} = w_{-k}^{-1}\big) - \mathbb{P}\big(X_0 = w_0 \,\big|\, X_{-k}^{-1} = w_{-k}^{-1}\big)\right| \le \epsilon + \frac{2\beta}{\alpha\beta^*}\epsilon + 2\beta\epsilon$$

and the Theorem follows. $\square$

4 Proof of Theorem 2

The proof relies on five lemmas. The first one is Lemma 3.4 from Galves and Leonardi (2008). For the convenience of the reader we recall this result.

Lemma 7 (Galves, Leonardi)**.** *Let (X_t) be a stationary stochastic process satisfying the non-nullness and the summable continuity rate hypotheses. Then, there exists a summable sequence $(\rho_l)_{l\in\mathbb{N}}$, satisfying*

$$\sum_{l\ge 1}\rho_l \le 1 + \frac{2\beta}{\alpha}$$

such that for any $i \ge 1$, any $k > i$, any $j \ge 1$ and any finite sequence w_1^j, the following inequality holds

$$\sup_{x_1^i \in A^i} |\mathbb{P}(X_k^{k+j-1} = w_1^j \mid X_1^i = x_1^i) - p(w_1^j)| \le \sum_{l=0}^{j-1}\rho_{k-i+l}\,.$$

The constants α and β appearing in the statement of the lemma were defined in Section 2. For a probabilistic interpretation of the sequence $(\rho_l)_{l\in\mathbb{N}}$ we refer the reader to Galves and Leonardi (2008).

The above lemma will be used in the proof of the following result involving the same quantities α, β and $(\rho_l)_{l\in\mathbb{N}}$.

Lemma 8. *There exists a summable sequence $(\rho_l)_{l\in\mathbb{N}}$, satisfying*

$$\sum_{l\in\mathbb{N}}\rho_l \le 2\big(1 + \frac{\beta}{\alpha}\big),$$

such that for any $i \ge 1$, any $k \ge i$, any $j \ge 1$ and any finite sequence w_1^j, the following inequality holds

$$\sup_{x_1^i,\, \theta_1^i \in A_1^i} |\mathbb{P}(Z_k^{k+j-1} = w_1^j \mid X_1^i = x_1^i,\, \xi_1^i = \theta_1^i) - q(w_1^j)| \le \sum_{l=0}^{j-1}\rho_{k-i+l}\,.$$

Proof. Observe that for any x_1^i, $\theta_1^i \in A_1^i$, by the independence of the flipping procedure we have

$$\begin{aligned}
&|\mathbb{P}(Z_k^{k+j-1} = w_1^j \mid X_1^i = x_1^i,\, \xi_1^i = \theta_1^i) - q(w_1^j)| = \\
&\quad |\sum_{x_k^{k+j-1}\in A_1^j} \mathbb{P}(X_k^{k+j-1} = x_k^{k+j-1}, Z_k^{k+j-1} = w_1^j \mid X_1^i = x_1^i, \xi_1^i = \theta_1^i) - q(w_1^j)| = \\
&\quad |\sum_{x_k^{k+j-1}\in A_1^j} \mathbb{P}(Z_k^{k+j-1} = w_1^j \mid X_k^{k+j-1} = x_k^{k+j-1})\mathbb{P}(X_k^{k+j-1} = x_k^{k+j-1} \mid X_1^i = x_1^i) - q(w_1^j)|.
\end{aligned}$$

The last term can be rewritten as

$$\Big|\sum_{x_k^{k+j-1}\in A_1^j} \mathbb{P}(Z_k^{k+j-1} = w_1^j \mid X_k^{k+j-1} = x_k^{k+j-1}) \\ \big[\mathbb{P}(X_k^{k+j-1} = x_k^{k+j-1} \mid X_1^i = x_1^i) - \mathbb{P}(X_k^{k+j-1} = x_k^{k+j-1})\big]\Big|.$$

Using lemma 7, this last expression is bounded above by

$$\sum_{x_k^{k+j-1}\in A_1^j} \mathbb{P}(\xi_k^{k+j-1} = w_1^j \oplus x_k^{k+j-1}) \sum_{l=0}^{j-1} \rho_{k-i+l} = \sum_{l=0}^{j-1} \rho_{k-i+l}.$$

This concludes the proof of Lemma 8. □

The proof of the next lemma uses Proposition 4 from Dedecker and Doukhan (2003). For the convenience of the reader, we recall this result.

Proposition 9 (Dedecker, Doukhan). *Let $(Y_t)_{t\in\mathbb{N}}$ be a sequence of centered and square integrable random variables and let $\mathcal{M}_i$ denote the σ-algebra generated by $Y_0, \dots, Y_i$. Define $S_n = Y_1 + \ldots + Y_n$ and*

$$b_{i,n} = \max_{i\le l\le n} \|Y_i \sum_{k=i}^{l} \mathbb{E}(Y_k \mid \mathcal{M}_i)\|_{p/2}.$$

Then, for any $p \ge 2$,

$$\|S_n\|_p \le \big(2p\sum_{i=1}^{n} b_{i,n}\big)^{1/2}.$$

Lemma 10. *For any finite sequence w and any $t > 0$ we have*

$$\mathbb{P}(\,|N_n(w) - (n-\ell(w)+1)q(w)| > t\,) \le e^{\frac{1}{e}} \exp\Big[\frac{-t^2}{4e[n-\ell(w)+1]\ell(w)(1+\frac{\beta}{\alpha})}\Big].$$

Moreover, for any $a \in A$ and any $n > \frac{|A|+1}{tq(w)} + \ell(w)$ we have

$$\mathbb{P}\big(|\hat{q}_n(a|w) - q(a|w)| > t\big) \le 3\, e^{\frac{1}{e}} \exp\Big[-(n-\ell(w))\frac{[t - \frac{3}{(n-\ell(w))q(w)}]^2 q(w)^2}{64e\ell(wa)[1+\frac{\beta}{\alpha}]}\Big].$$

Proof. Observe that for any finite sequence $w_1^j \in A_1^j$

$$N_n(w_1^j) = \sum_{t=0}^{n-j} \prod_{i=1}^{j} [\,\mathbf{1}_{\{X_{t+i}=w_i\}}\mathbf{1}_{\{\xi_{t+i}=0\}} + \mathbf{1}_{\{X_{t+i}=\bar{w}_i\}}\mathbf{1}_{\{\xi_{t+i}=1\}}\,].$$

Define the process $\{U_t : t \in \mathbb{Z}\}$ by

$$U_t = \prod_{i=1}^{j} [\,\mathbf{1}_{\{X_{t+i}=w_i\}}\mathbf{1}_{\{\xi_{t+i}=0\}} + \mathbf{1}_{\{X_{t+i}=\bar{w}_i\}}\mathbf{1}_{\{\xi_{t+i}=1\}}\,] - q(w_1^j)$$

and denote by $\mathscr{M}_i$ the σ-algebra generated by $U_0, \ldots, U_i$. Applying Proposition 9 we obtain for any $r \geq 2$

$$\|N_n(w_1^j) - (n-j+1)q(w_1^j)\|_r \leq \Big(2r \sum_{t=0}^{n-j} \max_{t \leq \ell \leq n-j} \|U_t \sum_{k=t}^{\ell} \mathbb{E}(U_k | \mathscr{M}_t)\|_{\frac{r}{2}} \Big)^{\frac{1}{2}}$$
$$\leq \Big(2r \sum_{t=0}^{n-j} \|U_t\|_{\frac{r}{2}} \sum_{k=t}^{n-j} \|\mathbb{E}(U_k | \mathscr{M}_t)\|_\infty \Big)^{\frac{1}{2}}.$$

Note that $\|U_t\|_{\frac{r}{2}} \leq 1$ for any $r \geq 2$. On the other hand we have

$$\sup_{\sigma_0^t \in A_0^t} |\mathbb{E}(U_k | U_0^t = \sigma_0^t)| = \sup_{x_1^{t+j}, \theta_1^{t+j} \in A_1^{t+j}} |\mathbb{E}(U_k | X_1^{t+j} = x_1^{t+j}, \xi_1^{t+j} = \theta_1^{t+j})|$$
$$= \sup_{x_1^{t+j}, \theta_1^{t+j} \in A_1^{t+j}} |\mathbb{P}(Z_{k+1}^{k+j} = w_1^j | X_1^{t+j} = x_1^{t+j}, \xi_1^{t+j} = \theta_1^{t+j}) - q(w_1^j)|.$$

Therefore, using Lemma 8 we obtain the bound

$$\|N_n(w) - (n-\ell(w)+1)q(w)\|_r \;\leq\; \big[4r\ell(w)(n-\ell(w)+1)(1+\frac{\beta}{\alpha})\big]^{\frac{1}{2}}.$$

Let $B = 4\ell(w)(n-\ell(w)+1)(1+\frac{\beta}{\alpha})$. Then, as in Dedecker and Prieur (2005) we obtain that, for any $t > 0$,

$$\mathbb{P}(|N_n(w) - (n-\ell(w)+1)q(w)| > t) \leq \min\Big(1, \frac{\mathbb{E}(|N_n(w) - (n-\ell(w)+1)q(w)|^r)}{t^r}\Big)$$
$$\leq \min\Big(1, \Big[\frac{rB}{t^2}\Big]^{\frac{r}{2}}\Big).$$

The function $r \to (cr)^{\frac{r}{2}}$ has a minimum at $r_0 = \frac{1}{ec}$. Then, comparing the value of this function with 1 and r_0 with 2 we can infer that

$$\min\Big(1, \Big[\frac{rB}{t^2}\Big]^{\frac{r}{2}}\Big) \;\leq\; \exp(-\frac{t^2}{eB} + e^{-1}).$$

We conclude that

$$\mathbb{P}(|N_n(w) - (n-\ell(w)+1)q(w)| > t) \leq e^{\frac{1}{e}} \exp\Big[\frac{-t^2}{4e[n-\ell(w)+1]\ell(w)(1+\frac{\beta}{\alpha})}\Big].$$

To prove the second assertion observe that

$$\Big| q(a|w) - \frac{(n-\ell(w))q(wa)+1}{(n-\ell(w))q(w)+|A|} \Big| \;\leq\; \frac{|A|+1}{(n-\ell(w))q(w)}.$$

Then, for all $n \geq (|A|+1)/tq(w) + \ell(w)$ we have that

$$\mathbb{P}\big(\big|\hat{q}_n(a|w) - q(a|w)\big| > t \big)$$
$$\leq \mathbb{P}\Big(\Big| \frac{N_n(wa)+1}{N_n(w\cdot)+|A|} - \frac{(n-\ell(w))q(wa)+1}{(n-\ell(w))q(w)+|A|} \Big| > t - \frac{|A|+1}{(n-\ell(w))q(w)} \Big)$$

Denote by $t' = t - (|A|+1)/(n-\ell(w))q(w)$. Then

$$\mathbb{P}\Big(\Big|\frac{N_n(wa)+1}{N_n(w\cdot)+|A|} - \frac{(n-\ell(w))q(wa)+1}{(n-\ell(w))q(w)+|A|}\Big| > t'\Big)$$
$$\leq \mathbb{P}\Big(\Big|N_n(wa)-(n-\ell(w))q(wa)\Big| > \frac{t'}{2}[(n-\ell(w))q(w)+|A|]\Big)$$
$$+\sum_{b\in A}\mathbb{P}\Big(\Big|N_n(wb)-(n-\ell(w))q(wb)\Big| > \frac{t'}{2|A|}[(n-\ell(w))q(w)+|A|]\Big).$$

Now, we can apply the bound in the first assertion of the Lemma to bound above the last sum by

$$3\,e^{\frac{1}{e}}\exp\Big[-(n-\ell(w))\,\frac{[t-\frac{3}{(n-\ell(w))q(w)}]^2q(w)^2}{64e\ell(wa)[1+\frac{\beta}{\alpha}]}\Big].$$

This concludes the proof of Lemma 10. $\square$

Lemma 11. *For any $\delta > 2(1+\frac{4\beta}{\min(\alpha\beta^*,1)})\epsilon$, any $w\in\mathscr{T}$, $uw\in\hat{\mathscr{T}}_n^{\delta,d}$ and*

$$n \;>\; \frac{6}{(\delta-2[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]\epsilon)\alpha^d}+d$$

we have that

$$\mathbb{P}(\Delta_n(uw)>\delta)\;\leq\;12\,e^{\frac{1}{e}}\exp\Big[-(n-d)\,\frac{[\frac{\delta}{2}-[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]\epsilon-\frac{3}{(n-d)\alpha^d}]^2\alpha^{2d}}{64e(1+\frac{\beta}{\alpha})(d+1)}\Big].$$

Proof. Recall that

$$\Delta_n(uw) = \max_{a\in A}|\hat{q}_n(a|uw)-\hat{q}_n(a|\mathrm{suf}(uw))|.$$

Note that the fact $w\in\mathscr{T}$ implies that for any finite sequence u and any symbol $a\in A$ we have $p(a|uw) = p(a|\mathrm{suf}(uw))$. Hence,

$$\begin{aligned}|\hat{q}_n(a|uw)-\hat{q}_n(a|\mathrm{suf}(uw))| \;\leq\; & |\hat{q}_n(a|uw)-q(a|uw)| \;+\; |q(a|uw)-p(a|uw)|\\ &+\;|q(a|\mathrm{suf}(uw))-p(a|\mathrm{suf}(uw))|\\ &+\;|\hat{q}_n(a|\mathrm{suf}(uw))-q(a|\mathrm{suf}(uw))|.\end{aligned}$$

Then, using Theorem 1 we have that

$$\begin{aligned}\mathbb{P}(\Delta_n(uw)>\delta)\;\leq\;\sum_{a\in A}\Big[&\mathbb{P}\big(|\hat{q}_n(a|uw)-q(a|uw)|>\frac{\delta}{2}-\epsilon\,[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]\big)\\ &+\mathbb{P}\big(|\hat{q}_n(a|\mathrm{suf}(uw))-q(a|\mathrm{suf}(uw))|>\frac{\delta}{2}-\epsilon\,[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]\big)\Big].\end{aligned}$$

Now, for

$$n \;>\; \frac{6}{(\delta-2[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]\epsilon)\alpha^d}+d$$

we can bound above the right hand side of the expression above using Lemma 10 by

$$12\,e^{\frac{1}{e}}\exp\Big[-(n-d)\,\frac{[\frac{\delta}{2}-[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]\epsilon-\frac{3}{(n-d)\alpha^d}]^2\alpha^{2d}}{64e(1+\frac{\beta}{\alpha})(d+1)}\Big].$$

□

Lemma 12. *There exists d such that for any* $\delta < D_d - 2(1+\frac{4\beta}{\min(\alpha\beta^*,1)})\epsilon$*, any* $w \in \hat{\mathscr{T}}_n^{\delta,d}$*,* $\ell(w) < K$*,* $w \notin \mathscr{T}$*, and any*

$$n > \frac{6}{[D_d-\delta-2(1+\frac{4\beta}{\min(\alpha\beta^*,1)})\epsilon]\alpha^d}+d$$

we have

$$\mathbb{P}(\bigcap_{uw\in\mathscr{T}|_d}\{\Delta_n(uw)\le\delta\}) \le 6\,e^{\frac{1}{e}}\exp\Big[-(n-d)\,\frac{[\frac{D_d-\delta}{2}-(1+\frac{4\beta}{\min(\alpha\beta*,1)})\epsilon-\frac{3}{(n-d)\alpha^d}]^2\alpha^{2d}}{64e(1+\frac{\beta}{\alpha})(d+1)}\Big].$$

Proof. Let

$$d = \max_{u\notin\mathscr{T},\,\ell(u)<K}\min\{k\colon \text{there exists } w\in\mathscr{C}_k \text{ with } w\succ u\}.$$

Then there exists $u\bar{w}\in\mathscr{T}|_d$ such that $p(a|u\bar{w})\neq p(a|\mathrm{suf}(u\bar{w}))$ for some $a\in A$. We have

$$\mathbb{P}(\bigcap_{uw\in\mathscr{T}|_d}\{\Delta_n(uw)\le\delta\}) \le \mathbb{P}(\Delta_n(u\bar{w})\le\delta).$$

Observe that for any $a\in A$,

$$\begin{aligned}|\hat{q}_n(a|\mathrm{suf}(u\bar{w}))-\hat{q}_n(a|u\bar{w})| \ge\; & |p(a|\mathrm{suf}(u\bar{w}))-p(a|u\bar{w})|-|\hat{q}_n(a|\mathrm{suf}(u\bar{w}))-q(a|\mathrm{suf}(u\bar{w}))|-\\ & |\hat{q}_n(a|u\bar{w})-q(a|u\bar{w})|-|q(a|\mathrm{suf}(u\bar{w}))-p(a|\mathrm{suf}(u\bar{w}))|-\\ & |q(a|u\bar{w})-p(a|u\bar{w})|.\end{aligned}$$

Hence, we have that for any $a\in A$

$$\Delta_n(u\bar{w}) \ge D_d-2\epsilon[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]-|\hat{q}_n(a|\mathrm{suf}(u\bar{w}))-q(a|\mathrm{suf}(u\bar{w}))|-|\hat{q}_n(a|u\bar{w})-q(a|u\bar{w})|.$$

Therefore,

$$\begin{aligned}\mathbb{P}(\Delta_n(u\bar{w})\le\delta) \le\; & \mathbb{P}(\bigcap_{a\in A}\{\,|\hat{q}_n(a|\mathrm{suf}(u\bar{w}))-q(a|\mathrm{suf}(u\bar{w}))|\ge\frac{D_d-2\epsilon[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]-\delta}{2}\}) \\ & +\mathbb{P}(\bigcap_{a\in A}\{\,|\hat{q}_n(a|u\bar{w})-q(a|u\bar{w})|\ge\frac{D_d-2\epsilon[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]-\delta}{2}\}).\end{aligned}$$

As $\delta < D_d-2\epsilon[1+\frac{4\beta}{\min(\alpha\beta^*,1)}]$ and

$$n > \frac{6}{[D_d-\delta-2(1+\frac{4\beta}{\min(\alpha\beta^*,1)})\epsilon]\alpha^d}+d$$

we can use Lemma 10 to bound above the right hand side of the last probability by

$$6\,e^{\frac{1}{e}}\exp\Big[-(n-d)\,\frac{[\frac{D_d-\delta}{2}-(1+\frac{4\beta}{\min(\alpha\beta*,1)})\epsilon-\frac{3}{(n-d)\alpha^d}]^2\alpha^{2d}}{64e(1+\frac{\beta}{\alpha})(d+1)}\Big].$$

This concludes the proof of Lemma 12 $\square$

Now we proceed with the proof of our main result.

Proof of Theorem 2. Define

$$O_{n,\delta}^{K,d}=\bigcup_{\substack{w\in\mathscr{T}\\ \ell(w)<K}}\ \bigcup_{uw\in\hat{\mathscr{T}}_n^{\delta,d}}\{\Delta_n(uw)>\delta\},$$

and

$$U_{n,\delta}^{K,d}=\bigcup_{\substack{w\in\hat{\mathscr{T}}_n^{\delta,d}\\ \ell(w)<K}}\ \bigcap_{uw\in\mathscr{T}|_d}\{\Delta_n(uw)\le\delta\}.$$

Then, if $d<n$ we have

$$\{\hat{\mathscr{T}}_n^{\delta,d}|_K\neq\mathscr{T}|_K\}\subseteq O_{n,\delta}^{K,d}\cup U_{n,\delta}^{K,d}.$$

Therefore,

$$\mathbb{P}(\hat{\mathscr{T}}_n^{\delta,d}|_K\neq\mathscr{T}|_K)\ \le\ \sum_{\substack{w\in\mathscr{T}\\ \ell(w)<K}}\ \sum_{uw\in\hat{\mathscr{T}}_n^{\delta,d}}\mathbb{P}(\Delta_n(uw)>\delta)+\sum_{\substack{w\in\hat{\mathscr{T}}_n^{\delta,d}\\ \ell(w)<K}}\mathbb{P}\Big(\bigcap_{uw\in\mathscr{T}|_d}\Delta_n(uw)\le\delta\Big).$$

Applying Lemma 11 and Lemma 12 we obtain, for

$$n\ >\ \frac{6}{[\min(\delta,D_d-\delta)-2(1+\frac{4\beta}{\min(\alpha\beta^*,1)})\epsilon]\alpha^d}+d,$$

the inequality

$$\mathbb{P}(\hat{\mathscr{T}}_n^{\delta,d}|_K\neq\mathscr{T}|_K)\ \le\ c_1\exp[-c_2(n-d)],$$

where $c_1=12\,e^{\frac{1}{e}}\,2^d$ and $c_2=\frac{[\min(\delta,D_d-\delta)-2(1+\frac{4\beta}{\min(\alpha\beta^*,1)})\epsilon-\frac{6}{(n-d)\alpha^d}]^2\alpha^{2d}}{256e(1+\frac{\beta}{\alpha})(d+1)}$. We conclude the proof of Theorem 2. $\square$

Proof of Corollary 3. It follows from Theorem 2, using the first Borel-Cantelli Lemma and the fact that the bounds for the error in the estimation of the truncated context tree are summable in n for appropriate choices of d and δ. $\square$

Acknowledgment

We thank two anonymous referees whose remarks and suggestions helped us to improve the presentation of this paper.

References

P. Bühlmann and A. J. Wyner. Variable length Markov chains. *Ann. Statist.*, 27:480–513, 1999. MR1714720

I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory*, 52(3):1007–1016, 2006. MR2238067

J. Dedecker and P. Doukhan. A new covariance inequality and applications. *Stochastic Process. Appl.*, 106(1):63–80, 2003. MR1983043

J. Dedecker and C. Prieur. New dependence coefficients. examples and applications to statistics. *Probab. Theory Relatated Fields*, 132:203–236, 2005. MR2199291

D. Duarte, A. Galves, and N.L. Garcia. Markov approximation and consistent estimation of unbounded probabilistic suffix trees. *Bull. Braz. Math. Soc.*, 37(4):581–592, 2006. MR2284889

R. Fernández, P.A. Ferrari, and A. Galves. Coupling, renewal and perfect simulation of chains of infinite order, 2001. URL http://www.ime.unicamp.br/∼ebp5. Notes for a minicourse at the Vth Brazilian School of Probability.

F. Ferrari and A. Wyner. Estimation of general stationary processes by variable length Markov chains. *Scand. J. Statist.*, 30(3):459–480, 2003. MR2002222

A. Galves and F.G. Leonardi. *Exponential inequalities for empirical unbounded context trees*, volume 60 of *Progress in Probability*, pages 257–270. Birkhauser, 2008.

A. Galves and E. Löcherbach. Stochastic chains with memory of variable length. *TICSP Series*, 38: 117–133, 2008.

A. Galves, V. Maume-Deschamps, and B. Schmitt. Exponential inequalities for VLMC empirical trees. *ESAIM Probab. Stat*, 12:43–45, 2008. MR2374639

J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 29(5):656–664, 1983. MR0730903

D. Ron, Y. Singer, and N. Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25(2-3):117–149, 1996.

F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Trans. Inform. Theory*, IT-44:653–664, 1995.