Sequences conveying information: Unbounded Variable Length Markov Chains

Denise Duarte, *Universidade Federal de Goiás*

Antonio Galves, *Universidade de São Paulo*

Nancy L. Garcia, *Universidade Estadual de Campinas*

05/12/2005

Abstract

Key words: AMS Classification: Primary: 60D05, Secondary: 60G55

1 Unbounded Variable Length Markov Chains

The term Variable Length Markov Chain (VLMC) was introduced by Bühlmann and Wyner (1999) and Bühlmann (1999) to refer to finite order Markov chains where the memory of the chain is allowed to be a function of the past values. However, there is no need to restrict oneself to finite memory chains. The definitions make perfect sense for chains with finite but unbounded memory.

More precisely, consider $(X_t)_{t\in\mathbb{Z}}$ a stationary process with values on a finite alphabet $\mathcal{A}$ such that $|\mathcal{A}| < \infty$ with transition probabilities given by a function $P : \mathcal{A} \times \mathcal{A}^\infty \to [0, 1]$ such that

$$\mathbb{P}(X_0 = x_0 \mid X_{-\infty}^{-1} = x_{-\infty}^{-1}) = p(x_0 \mid x_{-\infty}^{-1}) \tag{1.1}$$

for all $x \in \mathcal{A}^\infty$. We call such process a *chain of infinite order* with transition probabilities $p(\cdot \mid \cdot)$. These processes have been introduced by Onicescu and Mihoc (1935) who called them chain with complete connections. Doeblin and Fortet (1937) studied speed of convergence towards the invariant measure. The name chains of infinite order was coined by Harris (1955).

Definition 1.2 *Let $(X_t)_{t\in\mathbb{Z}}$ be a chain of infinite order with transition probability $p(\cdot \mid \cdot)$. A function $c : \mathcal{A}^\infty \to \cup_{m=0}^{\infty}\mathcal{A}^m \cup \mathcal{A}^\infty$ is called the context function of the process if*

$$c : \mathbf{x} \mapsto x_{-\ell(\mathbf{x})+1}^{-1} \tag{1.3}$$

for $\mathbf{x} = x_{\infty}^{-1}$ where $\ell : \mathcal{A}^\infty \to \mathbb{N} \cup \{0\}$ is the length of the context function and it is given by

$$\ell(\mathbf{x}) = \min\{k; p(a \mid \mathbf{x}) = p(a \mid x_{-k+1}^{-1}),\ \textit{for all } a \in \mathcal{A}\}. \tag{1.4}$$

There are 3 types of chains of infinite order depending upon what is required from the transition probabilities in terms of (i) continuity with respect to histories, and (ii) strict positivity. See Fernández and Galves (2002) for details. In this paper we will work with just the so-called **Type A** chains.

Remark.: This type of chains was already considered by Doeblin and Fortet (1937). They also considered a more restrictive class of chains called **Type B**. [We will present the definitions below]

Definition 1.5 *A system of transition probabilities is* **continuous** *if the functions $p(a \mid \cdot)$ are continuous for each $a \in \mathcal{A}$ or, equivalently if*

$$\beta_s \quad := \quad \sup_{x,y} |p(x \mid x_{\infty}^{-1}) - p(x \mid x_{-s}^{-1}y_{-\infty}^{-s-1})| \quad \to \quad 0 \tag{1.6}$$

as $s \to \infty$. The sequence $(\beta_s)_s \in \mathbb{N}$ is called the **continuity rate**.

Definition 1.7 *A system of transition probabilities is* **log-continuous** *if*

$$\gamma_s \quad := \quad \sup_{x,y} \left| \frac{p(x \mid x_{\infty}^{-1})}{p(x \mid x_{-s}^{-1}y_{-\infty}^{-s-1})} - 1 \right| \quad \to \quad 0 \tag{1.8}$$

as $s \to \infty$. The sequence $(\gamma_s)_s \in \mathbb{N}$ is called the **log-continuity rate**.

Definition 1.9 *A system of transition probabilities is* **weakly non-null** *if*

$$\sum_{a\in\mathcal{A}} \inf_x p(a \mid \mathbf{x}) > 0. \tag{1.10}$$

Definition 1.11 *A system of transition probabilities is* **strongly non-null** *if*

$$\inf_{a\in\mathcal{A}} \inf_x p(a \mid \mathbf{x}) > 0. \tag{1.12}$$

Based on these definitions we can define the two types of chains:

Definition 1.13 *A stochastic process is a* **chain of infinite order**
(i) **of Type A** *if its system of transition probabilities is continous and weakly non-null.*
(ii) **of Type B** *if its system of transition probabilities is log-continous and strongly non-null.*

Notice that Type B chains are a special case of Type A chains.

When we allow the momory of the chain to vary according to the past, one major challenge is to find a "good" estimator for the context function. Rissanen (1983) introduced a version of the so-called "context algorithm" which is described in Algorithm 1.24 in the framework of data compression. Later Bühlmann and Wyner (1999) proved that such algorithm is consistent in the case that the context function has a maximum length, that is, there exists a constant $k \in \mathbb{N}$ such that

$$|c(\mathbf{x})| = \ell(\mathbf{x}) \leq k, \quad \text{for all } \mathbf{x} \in \mathcal{A}^{\infty}. \tag{1.14}$$

Ferrari and Wyner (2003) use a sieve estimator which yields a version of the context algorithm that is asymptotically consistent for strongly log-continuous chains (geometrically α-mixing stationary processes) with unlimited memory. Although they claim an approximation through finite memory chains, in fact their approach is to compute directly the probabilities of underestimating and overestimating the model and shows that these probabilities converge to zero as the sample size increases.

In this work, we weaken the mixing requirements by using the maximal coupling between chains of Type A with finite order Markov chains introduced by Fernández and Galves (2002). In this case, with a further assumption on nun-nullness we prove that the context algorithm yields a consistent version of the context function. With some abuse on the nomenclature, we will call the stationary processes with limited but variable and unbounded memory as Unbounded Variable Length Markov Chains (UVLMC). By requiring stationarity, a UVLMC is completely specified by its transition probabilities

$$p_c(a \mid c(\mathbf{x})) = \mathbb{P}_{p_c}\left[X_0 = a \mid c(X_{-\infty}^{-1}) = c(\mathbf{x})\right], \quad \text{for all } \mathbf{x} \in \mathcal{A}^{\infty}. \tag{1.15}$$

Example 1.16 *Sparce binary trees.*

Let $\mathcal{A} = \{0, 1\}$ and a set of transtion probabilities is defined by

$$p(1 \mid \mathbf{x}) = q_{\ell(\mathbf{x})} \tag{1.17}$$

where

$$\ell(x_{-\infty}^{-1}) = \min\{k \leq -1; x_{-k} + 1 = 1\}. \tag{1.18}$$

In this case, the context function is given by

$$c(x_{-\infty}^{-1}) = \begin{cases} 1 & \text{if } x_{-1} = 1 \\ 0,1 & \text{if } x_{-1} = 0, x_{-2} = 1 \\ 0,0,1 & \text{if } x_{-1} = 0, x_{-2} = 0, x_{-3} = 1 \\ \dots & \end{cases} \tag{1.19}$$

Assume that

$$1 > q_n \searrow q_\infty > 0. \tag{1.20}$$

Given a context function $c(\cdot) : \mathcal{A}^\infty \to \cup_{m=0}^{\infty} \mathcal{A}^m$ each value of this function can be represented as a branch of a tree constructed as:

- Root on the top;
- Branches grow downwards;
- Every internal node has at most $|\mathcal{A}|$ offsprings.;
- The context $w = c(x_{-\infty}^{-1})$ is represented by a branch, whose sub-branch on top is determined by x_{-1}, the next subbranch is determined by x_{-2} and so on;
- If $\ell(x_{-\infty}^{-1}) < \infty$, the terminal sub-branch is determined by $x_{-\ell(x_{-\infty}^{-1})+1}$.

Definition 1.21 *Let $c(\cdot)$ be a context function of a stationary UVLCM. The corresponding $|\mathcal{A}|$-ary context tree τ and the terminal node context tree τ^t are defined as*

$$\tau = \tau_c = \{w; w = c(\mathbf{x}), \mathbf{x} \in \mathcal{A}^\infty\} \tag{1.22}$$

$$\tau^t = \tau_c^t = \{w; w \in \tau_c, and wu \notin \tau_c \text{ for all } u \in \cup_{m=1}^{\infty} \mathcal{A}^m\} \tag{1.23}$$

where τ^t are defined only when possible.

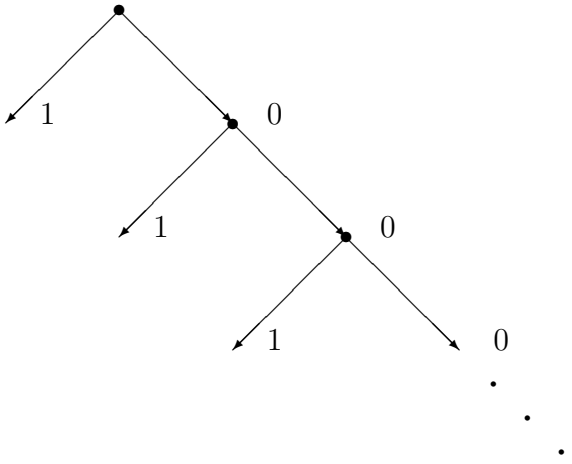

Figure 1.20: Context tree for the sparse binary tree

Algorithm 1.24 (Context algorithm)

Given a data set $X_1, X_2, \ldots, X_n$ from a UVLMC p_c, the aim is to find the underlying context function $c(\cdot)$ and an estimate of p_c.

Let

$$N(w) = \sum_{t=1}^{n} \mathbf{1}\left(X_t^{t+|w|-1}\right) \tag{1.25}$$

be the number of occurrences of string w in sequence X_1^n. Define

$$\hat{p}(w) = \frac{N(w)}{n}, \text{ for } v \in \cup_{m=1}^{\infty} \mathcal{A}^m \tag{1.26}$$

and

$$\hat{p}(v|w) = \frac{N(vw)}{N(w)}, \text{ for } v, w \in \cup_{m=1}^{\infty} \mathcal{A}^m. \tag{1.27}$$

Construct the estimate context tree $\hat{\tau}$ to be the biggest context tree such that

$$\Delta_{wu} = \sum_{x \in \mathcal{A}} \hat{p}(x \mid wu) \log \frac{\hat{p}(x \mid wu)}{\hat{p}(x \mid w)} N(wu) \geq K, \quad \text{for all } w, u \in \hat{\tau}^t \tag{1.28}$$

where $K = K_n \to \infty$ as $n \to \infty$ a cut-off chosen by the user.

Step 1 Given the data fit a maximal $|\mathcal{A}|$-ary context tree, that is, search for the context function $c_{\max}(\cdot)$ with terminal node context tree representation $\tau^t_{\max}$, where $\tau^t_{\max}$ is the biggest tree such that every element (terminal node) in $\tau^t_{\max}$ has been observed at least twice in the data.

- $w \in \tau^t_{\max} \Rightarrow N(w) \geq 2$;
- every τ^t such that $w \in \tau^t \Rightarrow N(w) \geq 2$, we have $\tau^t \leq \tau^t_{\max}$;
- $\tau^t_{(0)} = \tau^t_{\max}$.

Step 2 Examine every element (terminal node) of $\tau^t_{(0)}$.

Let c be the corresponding context function to $\tau^t_{(0)}$ and let

$$wu = x^{-1}_{-\ell(x^{-1}_{-\infty})+1} = c(x^{-1}_{-\infty}), \quad u = x_{-\ell(x^{-1}_{-\infty})+1} \text{ and } w = x_{-\ell(x^{-1}_{-\infty})+2}. \tag{1.29}$$

Replace the context wu by w if

$$\Delta_{wu} = \sum_{x \in \mathcal{A}} \hat{p}(x \mid wu) \log \frac{\hat{p}(x \mid wu)}{\hat{p}(x \mid w)} N(wu) < K. \tag{1.30}$$

Prune every terminal node in $\tau^t_{(0)}$ to obtain a smaller tree $\tau_{(1)}$.

$$\tau^t_{(1)} = \{w; w \in \tau_{(1)} and wu \notin \tau_{(1)} \text{ for all } u \in \cup_{m=1}^{\infty} \mathcal{A}^m\}. \tag{1.31}$$

Step 3 Repeat until possible to get $\hat{\tau}$ and $\hat{\tau}^t$ and the corresponding $\hat{c}_n$.

Step 4 Estimate the transition probabilities $p_c(x \mid c(\mathbf{x}))$ by $\hat{p}_c(x \mid c(\mathbf{x}))$ given by (1.27).

2 Markov approximations

The main tool for the proofs of this work is a Markovian approximation of the chains Type A chains given by Fernández and Galves (2002). Furthermore, they provide bounds to the distance between the infinite order chain and the Markovian approximation.

Definition 2.1 *The* **canonical Markov approximation of order** k *of a process* $(X_t)_{t \in \mathbb{Z}}$ *is the Markov chain of order* k, $X^{[k]} = (X^{[k]}_t)_{t \in \mathbb{Z}}$ *having as transition probabilities,*

$$p^{[k]}(a \mid x^{-1}_{-k}) := \mathbb{P}(X_0 = a \mid X^{-1}_{-k} = x^{-1}_{-k}) \tag{2.2}$$

for all $k \geq 1$ *and all* $a \in \mathcal{A}$ *and* $x^{-1}_{-k} \in \mathcal{A}^k$.

Definition 2.3 *The distance $\bar{d}$ between two processes $X = (X_t)$ and $Y = (Y_t)$ is defined as*

$$\bar{d}(X,Y) \;=\; \inf\left\{\sup_{t\in\mathbb{N}} \mathbb{P}(\tilde{\mathrm{X}}_{\mathrm{t}} \neq \tilde{\mathrm{Y}}_{\mathrm{t}} : (\tilde{\mathrm{X}},\tilde{\mathrm{Y}}) \textit{ is a coupling of } (\mathrm{X},\mathrm{Y})\right\}. \tag{2.4}$$

Their main result to be used in our work is

Theorem 2.5 *Let $X = (X_t)_{t\in\mathbb{N}}$ be a chain of infinite order of Type A with summable continuity rates $(\beta(s))_{s\geq 1}$. Then, there is a constant $C > 0$ such that, for all $k \geq 1$,*

$$\bar{d}(X, X^{[k]}) \leq C\beta(k), \tag{2.6}$$

where $X^{[k]} = (X_t^{[k]})_{t\in\mathbb{N}}$ is the canonical approximation of order k of the process X.

Given $(X_t)_{t\in\mathbb{Z}}$ a UVLCM of type A, let $(X_t^{[k(n)]})_{t\in\mathbb{Z}}$ be a sequence of canonical approximations for corresponding VLMC with context tree τ_n and context function c_n with maximal length $k(n) = \log n^{\alpha}$. In fact, the canonical coupling $(X_t, X_t^{[k]})_{t\in\mathbb{Z}}$ is such that $\mathbb{P}(X_t = X_t^{[k]})$ is as big as possible. Define, for $r \geq 1$ and $s \geq 1$

$$D_{r,s}^{(n)} \;:=\; \{X_t^{[k(n)]} = X_t,\, t = r\ldots, r+s\}. \tag{2.7}$$

Lemma 2.8

$$\mathbb{P}\left(D_{1,k(n)}^{(n)}\right) \to 1 \tag{2.9}$$

as $n \to \infty$.

Proof. According to Fernández and Galves (2002)

$$\mathbb{P}\left((D_{1,k(n)}^{(n)}\right) \geq 1 - \frac{1-(1-\beta^*(k))^k}{\prod_{p=0}^{+\infty}(1-\beta^*(p))} \to 1 \tag{2.10}$$

as $n \to \infty$, where $(\beta^*(p))_{p\in\mathbf{N}}$ is defined by,

$$\begin{cases} \beta^*(0) &=& 1 - \inf_{a\in A, \underline{u}\in\underline{A}} \mathbb{P}(a|\underline{u}), \\ \beta^*(p) &=& \min\left(\beta_0^*, \frac{\beta(p)}{2}\right). \end{cases}$$

3 Consistency

Theorem 3.1 *Consider the data $X_1, \ldots, X_n$ to be a finite realization of a UVLCM with context function c and a context tree τ (possibly infinite). Assume further that the system of transition probabilities p_c is continuous and weakly non-null (Type A chains) with summable $\beta(s)$ and that $(\min_{a\in\mathcal{A}} \min_{w\in\tau_n} p(a|w))^{-1} = O(n)$. Then*

1. *$\hat{c}_n(\cdot) \to c(\cdot)$ a.s. and equivalently*

2. *$\hat{\tau}_n \to \tau$ a.s.*

where $\hat{c}_n$ and $\hat{\tau}_n$ are, respectively, the estimated context function and context tree given by the context algorithm based on the sample $X_1, \ldots, X_n$.

Let $k = k(n) = \log n^\alpha$ and let $\hat{c}_n^*$ and $\hat{\tau}_n^*$ be the MLE given by the context algorithm based on a sample $X_1^{[k(n)]}, \ldots, X_n^{[k(n)]}$ of a finite Markov chain with c_n as a context function and τ_n as its context tree . In order to use the consistency result of Bühlmann and Wyner (1999) we have to verify conditions (A1)–(A3) for the VLMC $\{X^{[k(n)]}\}$.

Lemma 3.2 *Let $p^{[k(n)]}(\cdot,\cdot)$ be the transition probabilities of the canonical approximation $\{X^{[k(n)]}\}$. If $\{X\}$ is a chain of Type A, then $p^{[k(n)]}(\cdot,\cdot)$ satisfies: (A1) For some $r \in \mathbb{N}$, $p^{[k(n)]}$ satisfies*

$$\sup_{n\in\mathbb{N}} \sup_{A\subset\mathcal{A}^{k(n)};w,w'\in\mathcal{A}^{k(n)}} |p^{k(n),r}(A,w) - p^{k(n),r}(A,w')| < 1 - 2\kappa \tag{3.3}$$

for some $\kappa > 0$, where $p^{k(n),r}(A,w) = \mathbb{P}[(X_{r-k(n)+1}^{[k(n)]}, \ldots, X_r^{[k(n)]}) \in A \mid (X_{-k(n)+1}^{[k(n)]}, \ldots, X_0^{[k(n)]}) = w]$ denotes the r-step transition kernel of $p^{[k(n)]}$.
(A2) Let $b_n = \min_{w\in\tau_{k(n)}^T} p^{[k(n)]}(w)$ and $\epsilon_n = \min_{w,u\in\tau_{k(n)}^T, u\in\mathcal{A}} \|p^{k(n)}(\cdot \mid wu) - p^{k(n)}(\cdot \mid wu)\|_1$. Then

$$b_n^{-1} = O(\log(n)^{-(1/2+)} n^{1/2}) \quad \text{for some } 0 < b < \infty \tag{3.4}$$

$$\epsilon_n^{-1} = O(\log(n)^{-(1+\epsilon)} n b_n) \quad \text{for some } 0 < \epsilon < \infty \tag{3.5}$$

as $n \to \infty$.

Note: Since for Type A chains we assume weakly non-nullness, in order to use the consistence theorem we have to assume further that the minimal transition probabilities satisfy

$$\left[\min_{x\in\mathcal{A},w\in\tau_{k(n)}} p^{[k(n)]}(x \mid w)\right]^{-1} = O(n), \tag{3.6}$$

as $n \to \infty$.

Proof.

(A1)

$$|P^{k(n),r}(A,w) - P^{k(n),r}(A,w')| \leq \mathbb{P}(D^c_{r,r+k}|(w,w'))$$

where $\mathbb{P}$ is the coupling probability of two chains with different pasts and the conditioning event (w,w') is the past of the two chains between times -k and -1. But,

$$\begin{aligned}
\mathbb{P}(D^c_{r,r+k}|(w,w')) &= \mathbb{P}(D^c_{r,r+k}, D_{0,r}|(w,w')) + \mathbb{P}(D^c_{r,r+k}, D^c_{0,r}|(w,w')) \\
&\leq \mathbb{P}(D_{0,r}|(w,w'))\mathbb{P}(D^c_{r,r+k}|D^c_{r-k,r})\mathbb{P}(D^c_{0,r}|(w,w')) \\
&\leq \mathbb{P}(D_{0,r}|(w,w'))[1-\prod_{p=0}^{k}(1-\beta^*_{k\wedge(r+p)})] \\
&\quad +[1-(1-\beta^*_k)^k]\mathbb{P}(D^c_{0,r}|(w,w')) \\
&= 1-\mathbb{P}(D_{0,r}|(w,w'))\prod_{p=0}^{k}(1-\beta^*_{k\wedge(r+p)})] - \mathbb{P}(D^c_{0,r}|(w,w'))(1-\beta^*_k)^k \\
&< 1-\epsilon
\end{aligned} \tag{3.7}$$

for some $r \in \mathbb{N}$.

(A2) Let

$$b_n = \min_{w\in\tau^T_{k(n)}} P^{[k(n)]}(w). \tag{3.8}$$

Shannon-Macmillan-Breiman theorem says that

$$-\lim_{n\to\infty}\frac{1}{n}\log p(x_1^n) = h \text{ a.s.} \tag{3.9}$$

and

$$-\lim_{n\to\infty}\frac{1}{n}\log p^{(k_m)}(x_1^n) = h_{k(m)} \text{ a.s.} \tag{3.10}$$

for fixed m, where h and $h_{k(m)}$ are the entropy of the UVLMC $(X_n)_{n\in\mathbb{Z}}$ and $(X^{(k)}_n)_{n\in\mathbb{Z}}$ respectively. Moreover, $h_{k(n)} \nearrow h$ and $n \to \infty$.

Furthermore, it is possible to prove that for any $\epsilon > 0$

$$p^{[k(n)]}(x_1^n) \geq 2^{-n(h+\epsilon)} \text{ eventually, a.s..} \tag{3.11}$$

Therefore, for any $w \in \tau_{k(n)}^T$ we have

$$p^{[k(n)]}(w) \geq 2^{-k(n)(h_{k(n)}+\epsilon)} \text{ eventually, a.s.} \tag{3.12}$$

and condition (A2) is valid for any $\beta > 1/2$ and $\alpha h < 1/2$.

On the other hand,

$$\epsilon_n = \min_{wu \in \tau_{k(n)}^T, u \in \mathcal{A}} \sum_{a \in \mathcal{A}} |p^{[k(n)]}(a \mid wu) - p^{[k(n)]}(a \mid w)| \leq \beta_1^* |\mathcal{A}| \tag{3.13}$$

since

$$\inf_{x \overset{1}{=} y} |p^{[k(n)]}(a \mid x) - p^{[k(n)]}(a \mid y)| \geq (1 - \beta_1^*).$$

Therefore, by Bühlmann and Wyner (1999) we have

1. $\lim_{n\to\infty} \mathbb{P}\left[\hat{c}_n^*(\cdot) = c_n(\cdot)\right] = 1$;
2. $\lim_{n\to\infty} \mathbb{P}[\hat{\tau}_n^* = \tau_n] = 1$;
3. $\sup_{\mathbf{x}\in\mathcal{A}^\infty} |\hat{p}_n(x_1 \mid \hat{c}_n^*(\mathbf{x})) - p_n^*(x_1 \mid c_n(\mathbf{x}))| = o_P(1)$.

Theorem 3.1 is an immediate consequence of Lemmas 2.8 and 3.2.