

Large Sample Properties of Higher Order Markov Models

Tuhin Majumder

Department of Statistical Sciences, Wake Forest University

and

Donald E.K. Martin

Department of Statistics, North Carolina State University

and

Soumendra N. Lahiri

Department of Mathematics and Statistics, Washington University in St. Louis

Abstract

We study large-sample properties of higher-order Markov chains on a finite alphabet Σ when the order m_n is allowed to grow with the sequence length n . By embedding the process into a first-order chain on Σ^{m_n} and exploiting return-time decompositions, we establish a central limit theorem for additive functionals $\sum_t g_n(Y_t^{(n)})$ under natural ergodicity and sparsity conditions. The normalization involves the stationary return time to a suitably chosen state and accommodates triangular arrays with $m_n \rightarrow \infty$ and $m_n/n \rightarrow 0$. We further illustrate the assumptions in a binary variable length Markov chain (VLMC), deriving explicit lower bounds on stationary masses that yield a concrete growth regime (e.g., $m_n \log m_n/n \rightarrow 0$) ensuring the CLT. These results provide asymptotic foundations for inference in sparse/partitioned higher-order models; including VLMCs and sparse Markov models (SMMs) where the effective dimensionality grows with the sample size.

1 Introduction

Large-sample properties of first-order Markov chains are well understood and have wide applications. However, in many applications, the present state depends on the previous m states. Let $\Phi = \{X_t : t = 0, 1, 2, \dots\}$ be a sequence of random variables taking values in a finite state space Σ of size $|\Sigma| = d$. For $t \geq m$, if $P(X_t = x_t \mid X_{t-1} = x_{t-1}, \dots, X_0 = x_0) = P(X_t = x_t \mid X_{t-1} = x_{t-1}, \dots, X_{t-m} = x_{t-m})$, then Φ is called a Markov chain of order m . Studying a Markov chain of order m is not substantially different from the first-order case, as we can transform it into a first-order Markov chain by taking consecutive m -tuples as states. Denote this new chain by $\Phi' = \{\mathbf{Y}_t : t = 0, 1, 2, \dots\}$, where $\mathbf{Y}_t = (X_{t+m-1}, \dots, X_t)$. The new state space is $\Sigma' = \Sigma^m$, i.e., the set of all possible m -tuples from Σ . The corresponding transition matrix P' has dimension $d^m \times d^m$. In each row of P' , at most d entries can be non-zero, since $P(\mathbf{Y}_{t+1} = (j_1, \dots, j_m) \mid \mathbf{Y}_t = (i_1, \dots, i_m)) > 0$ only if $i_k = j_{k+1}$ for $k = 1, \dots, m-1$. In that case, $P(\mathbf{Y}_{t+1} = (j_1, \dots, j_m) \mid \mathbf{Y}_t = (i_1, \dots, i_m)) = P(X_{t+m} = j_1 \mid X_{t+m-1} = i_1, \dots, X_t = i_m)$. Hence, the total number of entries to be estimated is $d^m(d-1)$, which grows exponentially with increasing m .

To reduce this model complexity, several dimension-reduction techniques have been proposed for higher-order Markov chains. For instance, [Raftery \[1985\]](#) introduced a linear mixture model of the form

$$P(X_t = j_0 \mid X_{t-1} = j_1, \dots, X_{t-m} = j_m) = \sum_{i=1}^m \lambda_i q_{j_i j_0},$$

where $\sum_{i=1}^m \lambda_i = 1$ and $Q = ((q_{ij}))$ is a known transition matrix on Σ . The weights $(\lambda_1, \dots, \lambda_m)$ are estimated via maximum likelihood. Another widely used strategy is the Variable Length Markov Chain (VLMC) model, originally proposed by [Rissanen \[1983\]](#), where the context length varies depending on the observed past. [Bühlmann and Wyner \[1999\]](#) and [Bühlmann \[2000\]](#) developed model selection methods and studied the asymptotic behavior of VLMCs. In these models, a context tree determines the relevant past suffixes

for predicting future states, allowing for substantial reduction in the number of parameters when long dependencies are sparse. Recent work on Bayesian Context Trees (BCT) by [Kontoyiannis et al. \[2020\]](#) and [Papageorgiou and Kontoyiannis \[2022\]](#) has provided posterior inference and predictive distributions for discrete-time sequence models.

A more general approach to sparse modeling of higher-order Markov data involves grouping the m -tuples in Σ^m into partitions such that tuples within the same group share identical transition probabilities. These models are referred to as sparse Markov models (SMMs), first proposed by [Garcia et al. \[2011\]](#) and later extended in a Bayesian framework by [Jääskinen et al. \[2014\]](#). [Xiong et al. \[2016\]](#) introduced a recursive partitioning algorithm, while [Bennett et al. \[2023\]](#) used Gibbs sampling for posterior estimation. More recently, [Majumder et al. \[2026\]](#) developed a convex clustering approach for learning the underlying partition structure in the SMM set-up.

As long as the state space Σ is finite, the central limit theorems (CLTs) and other large-sample results for SMMs mirror those for first-order Markov chains, provided certain ergodicity conditions hold. For example, [Bühlmann and Wyner \[1999\]](#) derived large-sample properties of VLMC estimators, including bootstrap-based CLTs. However, the scenario becomes more intriguing when the Markov chain order grows with the sequence length, i.e., when the order is a function of n , denoted by m_n . Proving CLTs in this setup is considerably more challenging. In this paper, we give a central limit theorem for this case.

The rest of the chapter is organized as follows. Section (2) presents our main theoretical result. In Section (3), we illustrate the assumptions and conclusions for the special case of a binary VLMC. Section (4) concludes the chapter with final remarks.

2 Central Limit Theorem

Consider a triangular array of variable-order Markov chains as follows. The n^{th} chain is presented as $\Phi_n = \{X_0^{(n)}, \dots, X_n^{(n)}\}$, where $X_j^{(n)} \in \Sigma$. Recall that Σ is a finite state space

with $|\Sigma| = d$. However, the order of the Markov chain Φ_n (denoted by m_n) varies with n . One can easily represent this Markov chain as a first-order Markov chain in the following manner. Let $\mathbf{Y}_t^{(n)} = (X_{t+m_n-1}^{(n)}, \dots, X_t^{(n)})$, $t = 0, 1, \dots, n - m_n + 1$. Then $\Phi'_n = \{\mathbf{Y}_t^{(n)}\}$ is a first-order Markov chain with state space $S_n = \Sigma^{m_n}$. We work with the Markov chain Φ'_n in the sequel.

Our goal is to establish a CLT-type result for a sequence of real-valued functions over Σ in this triangular-array set up. Before going into the main result, we define some useful notation. Assume that the chain Φ'_n is aperiodic and irreducible for each n . Let π_n be its corresponding stationary probability vector of length d^{m_n} . For a state $\alpha \in S_n$, define $\sigma_\alpha^{(n)}(0) \equiv \inf\{j \geq 0 : \mathbf{Y}_j^{(n)} = \alpha\}$. Thus $\sigma_\alpha^{(n)}(0)$ is the first hitting time of state α for the chain Φ'_n . In a similar fashion, successively define the k^{th} hitting time of α as follows:

$$\sigma_\alpha^{(n)}(k) = \inf\{j > \sigma_\alpha^{(n)}(k-1) : \mathbf{Y}_j^{(n)} = \alpha\}.$$

Also define

$$\ell_{n,\alpha} \equiv \max\{k : \sigma_\alpha^{(n)}(k) \leq n - m_n + 1\} = \sum_{j=0}^{n-m_n+1} \mathcal{I}(\mathbf{Y}_j^{(n)} = \alpha) - 1. \quad (1)$$

Then $\ell_{n,\alpha}$ is the number of returns of Φ'_n to state α after the first hitting time.

If the order of the Markov chains in each array is fixed, i.e $m_n = m$ for some $m \in \mathbb{N}$, then any state $\alpha \in \Sigma^m$ is recurrent, so that $\ell_{n,\alpha} \rightarrow \infty$ as $n \rightarrow \infty$. However, in our set-up, $m_n \rightarrow \infty$ as $n \rightarrow \infty$. Note that for $\alpha \in \Sigma^{m_n}$, $E_{\pi_n}(\ell_{n,\alpha}) = (n - m_n + 2)\pi_n(\alpha) - 1$, where $E_{\pi_n}[\cdot]$ is the expectation under the stationarity assumption. If there exists a sequence of states $\{\alpha_n : \alpha_n \in \Sigma^{m_n}\}$ such that $n\pi_n(\alpha_n) \rightarrow \infty$ as $n \rightarrow \infty$, then $E_{\pi_n}(\ell_{n,\alpha_n}) \rightarrow \infty$, i.e. the expected number of returns to α_n diverges.

Next, we present our main result, which demonstrates the large sample properties of a triangular array of sparse Markov chains.

Theorem 2.1. *Consider the following triangular array of higher order Markov chains over*

finite alphabet Σ of cardinality d . The n^{th} array $\Phi_n = \{X_0^{(n)}, \dots, X_n^{(n)}\}$ has order m_n . Define $\mathbf{Y}_t^{(n)} \equiv (X_{t+m_n-1}^{(n)}, \dots, X_t^{(n)})$, $t = 0, 1, \dots, n - m_n + 1$. Denote the stationary probability vector of the n^{th} chain by π_n . Suppose the following assumptions hold.

- (i) For each $n \in \mathbb{N}$, $\Phi'_n = \{\mathbf{Y}_t^{(n)}\}$ is an aperiodic and irreducible Markov chain over Σ^{m_n} .
- (ii) $m_n \rightarrow \infty$ as $n \rightarrow \infty$ with $\frac{m_n}{n} \rightarrow 0$.
- (iii) There exists a sequence of states $\{\alpha_n : \alpha_n \in \Sigma^{m_n}\}$ such that $n\pi_n(\alpha_n) \rightarrow \infty$ as $n \rightarrow \infty$.
- (iv) $\text{Var}_{\pi_n}(l_{n,\alpha_n}/n\pi_n(\alpha_n)) \rightarrow 0$ as $n \rightarrow \infty$.

Define $E_{\alpha_n}[\cdot]$ as the expectation under the assumption that the initial state of the Markov chain is α_n . Consider a sequence of functions $g_n : \Sigma^{m_n} \rightarrow \mathbb{R}$, such that

- (i) For each n , $E_{\alpha_n} \left[\sum_{j=1}^{\tau_{\alpha_n}} \bar{g}_n(\mathbf{Y}_j^{(n)}) \right]^2 < \infty$;
- (ii) $\sup_n \left\{ \frac{E_{\alpha_n} \left[\sum_{j=1}^{\tau_{\alpha_n}} |\bar{g}_n(\mathbf{Y}_j^{(n)})| \right]^2}{E_{\alpha_n} \left[\sum_{j=1}^{\tau_{\alpha_n}} \bar{g}_n(\mathbf{Y}_j^{(n)}) \right]^2} \right\} < \infty$;

where $\bar{g}_n(x) = g_n(x) - E_{\pi_n}(g_n)$.

Under the above conditions,

$$\frac{1}{\sqrt{n}} \sum_{j=0}^{n-m_n+1} \frac{\bar{g}_n(\mathbf{Y}_j^{(n)})}{\sqrt{\frac{E_{\alpha_n}(\tau_{\alpha_n})}{E_{\alpha_n} \left[\sum_{j=1}^{\tau_{\alpha_n}} \bar{g}_n(\mathbf{Y}_j^{(n)}) \right]^2}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Proof: From now on, we will denote ℓ_{n,α_n} as ℓ_n for simplicity. First observe that $E_{\pi_n}(\ell_n) = (n - m_n + 2)\pi_n(\alpha_n) - 1$, i.e. $E_{\pi_n} \left(\frac{\ell_n}{n\pi_n(\alpha_n)} \right) \rightarrow 1$ as $n \rightarrow \infty$. Also, by assumption (iv), $\text{Var}_{\pi_n} \left(\frac{\ell_n}{n\pi_n(\alpha_n)} \right) \rightarrow 0$. Hence $\frac{\ell_n}{n\pi_n(\alpha_n)} \xrightarrow{p} 1$ as $n \rightarrow \infty$, eventually leading to $\ell_n \xrightarrow{p} \infty$.

For any function $f : \Sigma^{m_n} \rightarrow \mathbb{R}$, define

$$s_j^{(n)}(f) = \sum_{t=\sigma_{\alpha_n}(j)+1}^{\sigma_{\alpha_n}(j+1)} f(\mathbf{Y}_t^{(n)}), \quad j = 0, 1, \dots, \ell_n - 1.$$

By the strong Markov property, $s_j^{(n)}(g_n)$ are iid random variables with mean $E_{\alpha_n} \left[\sum_{j=1}^{\tau_{\alpha_n}} g_n(\mathbf{Y}_j^{(n)}) \right] = E_{\alpha_n} \left[s_0^{(n)}(g_n) \right]$ and variance $Var_{\alpha_n} \left[s_0^{(n)}(g_n) \right] = E_{\alpha_n} \left[(s_0^{(n)}(g_n))^2 \right] - E_{\alpha_n}^2 \left[s_0^{(n)}(g_n) \right]$. Define $\bar{g}_n(x) = g_n(x) - E_{\pi_n}(g_n(X))$, where $X \sim \pi_n$. For simplicity, let's define $E_{\pi_n}(g_n(X)) = E_{\pi_n}[g_n]$. Note that $E_{\pi_n}[g_n] = \pi_n(\alpha_n) E_{\alpha_n} \left[\sum_{j=1}^{\tau_{\alpha_n}} g_n(\mathbf{Y}_j^{(n)}) \right]$. Hence

$$\begin{aligned} E_{\alpha_n} \left[s_j^{(n)}(\bar{g}_n) \right] &= E_{\alpha_n} \left[\sum_{j=1}^{\tau_{\alpha_n}} (g_n(\mathbf{Y}_j^{(n)}) - E_{\pi_n}(g_n)) \right] \\ &= E_{\alpha_n} \left[\sum_{j=1}^{\tau_{\alpha_n}} g_n(\mathbf{Y}_j^{(n)}) - \tau_{\alpha_n} E_{\pi_n}(g_n) \right] \\ &= E_{\alpha_n} \left[\sum_{j=1}^{\tau_{\alpha_n}} g_n(\mathbf{Y}_j^{(n)}) \right] - E_{\alpha_n} \left[\tau_{\alpha_n} \right] E_{\pi_n}(g_n) \\ &= 0, \quad \text{since } \pi_n(\alpha_n) = 1/E_{\alpha_n} \left[\tau_{\alpha_n} \right]. \end{aligned}$$

The normalized and centralized random variables are defined as follows:

$$s_j'^{(n)}(\bar{g}_n) = \frac{s_j^{(n)}(\bar{g}_n)}{\sqrt{Var_{\alpha_n} \left[s_0^{(n)}(\bar{g}_n) \right]}} = \frac{s_j^{(n)}(\bar{g}_n)}{\sqrt{E_{\alpha_n} \left[s_0^{(n)}(\bar{g}_n) \right]^2}}.$$

Thus, $s_j'^{(n)}(\bar{g}_n), j = 0, 1, \dots, \ell_n - 1$ are i.i.d random variables with mean 0 and variance 1. Denote $v_n^2 = E_{\alpha_n} \left[s_0^{(n)}(\bar{g}_n) \right]^2$. Fix $0 < \epsilon < 1$. Define $\underline{n} = [(1 - \epsilon)(n - m_n + 2)\pi_n(\alpha_n)]$, $\bar{n} = [(1 + \epsilon)(n - m_n + 2)\pi_n(\alpha_n)]$, and $n^* = [(n - m_n + 2)\pi_n(\alpha_n)]$, where $[\cdot]$ is the greatest integer function.

Note that

$$\begin{aligned} \sum_{j=0}^{n-m_n+1} \bar{g}_n(\mathbf{Y}_j^{(n)}) &= \sum_{t=0}^{\sigma_{\alpha_n}(0)} \bar{g}_n(\mathbf{Y}_t^{(n)}) + \sum_{j=0}^{\ell_n-1} s_j^{(n)}(\bar{g}_n) + \sum_{t=\sigma_{\alpha_n}(\ell_n)+1}^{n-m_n+1} \bar{g}_n(\mathbf{Y}_t^{(n)}) \\ \implies \left| \sum_{j=0}^{n-m_n+1} \bar{g}_n(\mathbf{Y}_j^{(n)}) - \sum_{j=0}^{\ell_n-1} s_j^{(n)}(\bar{g}_n) \right| &\leq \sum_{t=0}^{\sigma_{\alpha_n}(0)} \left| \bar{g}_n(\mathbf{Y}_t^{(n)}) \right| + s_{\ell_n}^{(n)}(|\bar{g}_n|) \\ \implies \left| \frac{1}{\sqrt{n^*v_n}} \sum_{j=0}^{n-m_n+1} \bar{g}_n(\mathbf{Y}_j^{(n)}) - \frac{1}{\sqrt{n^*}} \sum_{j=0}^{\ell_n-1} s_j'^{(n)}(\bar{g}_n) \right| &\leq \frac{1}{\sqrt{n^*v_n}} \sum_{t=0}^{\sigma_{\alpha_n}(0)} \left| \bar{g}_n(\mathbf{Y}_t^{(n)}) \right| + \frac{s_{\ell_n}^{(n)}(|\bar{g}_n|)}{\sqrt{n^*v_n}}. \end{aligned}$$

Define

$$Z_{n,j} = \frac{s_j^{(n)}(|\bar{g}_n|)}{v_n}.$$

By assumption (ii), there exists a constant $C < \infty$ such that

$$\sup_n E_{\alpha_n} [Z_{n,0}^2] = \sup_n \frac{E_{\alpha_n} \left[\left(s_0^{(n)}(|\bar{g}_n|) \right)^2 \right]}{v_n^2} \leq C.$$

Recall that $\ell_n/n^* \xrightarrow{p} 1$. Fix $\eta > 0$ and $0 < \delta < 1$. Then

$$\begin{aligned} P \left(\frac{Z_{n,\ell_n}}{\sqrt{n^*}} > \eta \right) &\leq P \left(\left| \frac{\ell_n}{n^*} - 1 \right| > \delta \right) \\ &\quad + P \left(\max_{(1-\delta)n^* \leq j \leq (1+\delta)n^*} Z_{n,j} > \eta\sqrt{n^*} \right). \end{aligned}$$

By the union bound and Markov's inequality,

$$\begin{aligned} P \left(\max_{(1-\delta)n^* \leq j \leq (1+\delta)n^*} Z_{n,j} > \eta\sqrt{n^*} \right) &\leq \sum_{j=\lfloor (1-\delta)n^* \rfloor}^{\lceil (1+\delta)n^* \rceil} P \left(Z_{n,j} > \eta\sqrt{n^*} \right) \\ &\leq \sum_{j=\lfloor (1-\delta)n^* \rfloor}^{\lceil (1+\delta)n^* \rceil} \frac{E_{\alpha_n} [Z_{n,j}^2]}{\eta^2 n^*} \\ &\leq \frac{(2\delta n^* + 2)C}{\eta^2 n^*}. \end{aligned}$$

Consequently,

$$\limsup_{n \rightarrow \infty} P \left(\frac{Z_{n,\ell_n}}{\sqrt{n^*}} > \eta \right) \leq \frac{2C\delta}{\eta^2}.$$

Since $\delta > 0$ is arbitrary, letting δ arbitrarily small

$$\frac{Z_{n,\ell_n}}{\sqrt{n^*}} \xrightarrow{p} 0.$$

Therefore,

$$\frac{s_{\ell_n}^{(n)}(|\bar{g}_n|)}{\sqrt{n^*}v_n} \xrightarrow{p} 0.$$

By similar logic,

$$\frac{1}{\sqrt{n^*v_n}} \sum_{t=0}^{\sigma_{\alpha_n}(0)} \left| \bar{g}_n(\mathbf{Y}_t^{(n)}) \right| \xrightarrow{p} 0.$$

Combining these results, we obtain

$$\left| \frac{1}{\sqrt{n^*v_n}} \sum_{j=0}^{n-m_n+1} \bar{g}_n(\mathbf{Y}_j^{(n)}) - \frac{1}{\sqrt{n^*}} \sum_{j=0}^{\ell_n-1} s'_j(\bar{g}_n) \right| \xrightarrow{p} 0. \quad (2)$$

Since $\frac{\ell_n}{n\pi_n(\alpha_n)} \xrightarrow{p} 1$ as $n \rightarrow \infty$, $\exists \quad n_0 \in \mathbb{N}$ s.t. for $n \geq n_0$

$$P(\underline{n} \leq \ell_n - 1 \leq \bar{n}) \geq 1 - \epsilon.$$

Hence, by Kolmogorov's inequality, for any arbitrary $\beta > 0$,

$$\begin{aligned} P\left(\left| \frac{1}{\sqrt{n^*}} \sum_{j=0}^{\ell_n-1} s'_j(\bar{g}_n) - \frac{1}{\sqrt{n^*}} \sum_{j=0}^{n^*} s'_j(\bar{g}_n) \right| > \beta\right) &\leq \epsilon + P\left(\left| \max_{\underline{n} \leq l \leq n^*} \sum_{j=l}^{n^*} s'_j(\bar{g}_n) \right| > \beta\sqrt{n^*}\right) \\ &\quad + P\left(\left| \max_{n^* \leq l \leq \bar{n}} \sum_{j=n^*}^l s'_j(\bar{g}_n) \right| > \beta\sqrt{n^*}\right) \\ &\leq \epsilon + \frac{2\epsilon\pi_n(\alpha_n)(n - m_n + 2)E_{\alpha_n}\left[s_0^{(n)}(\bar{g}_n)\right]^2}{\beta^2 n^*} \\ &\leq \epsilon + \frac{4\epsilon}{\beta^2}. \end{aligned} \quad (3)$$

From equation (3), we can conclude that

$$\left| \frac{1}{\sqrt{n^*}} \sum_{j=0}^{\ell_n-1} s'_j(\bar{g}_n) - \frac{1}{\sqrt{n^*}} \sum_{j=0}^{n^*} s'_j(\bar{g}_n) \right| \xrightarrow{p} 0. \quad (4)$$

Combining this with equation (2), we have

$$\frac{1}{\sqrt{n^*v_n}} \sum_{j=0}^{n-m_n+1} \bar{g}_n(\mathbf{Y}_j^{(n)}) - \frac{1}{\sqrt{n^*}} \sum_{j=0}^{n^*} s'_j(\bar{g}_n) \xrightarrow{p} 0. \quad (5)$$

Using the CLT for row-wise i.i.d random variables with mean 0 and variance 1, we obtain

$$\frac{1}{\sqrt{n^*}} \sum_{j=0}^{n^*} s'_j(\bar{g}_n) \xrightarrow{d} \mathcal{N}(0, 1). \quad (6)$$

Hence, from equation (5), we conclude

$$\begin{aligned} & \frac{1}{\sqrt{n^*} v_n} \sum_{j=0}^{n-m_n+1} \bar{g}_n(\mathbf{Y}_j^{(n)}) \xrightarrow{d} \mathcal{N}(0, 1) \\ \Rightarrow & \frac{1}{\sqrt{n \pi_n(\alpha_n)} v_n} \sum_{j=0}^{n-m_n+1} \bar{g}_n(\mathbf{Y}_j^{(n)}) \xrightarrow{d} \mathcal{N}(0, 1) \\ \Rightarrow & \frac{1}{\sqrt{n}} \sum_{j=0}^{n-m_n+1} \frac{\bar{g}_n(\mathbf{Y}_j^{(n)})}{\sqrt{\frac{E_{\alpha_n}(\tau_{\alpha_n})}{E_{\alpha_n}[s_0^{(n)}(\bar{g}_n)]^2}}} \xrightarrow{d} \mathcal{N}(0, 1). \end{aligned}$$

3 Example

In the previous section, we discussed requirements for a large sample CLT. One of the conditions to satisfy the CLT is $n\pi_n(\alpha_n) \rightarrow \infty$ as $n \rightarrow \infty$. It is easy to see that there is always a state $\alpha_n \in \Sigma^{m_n}$ with $\pi_n(\alpha_n) \geq 1/d^{m_n}$, hence assuming $n/d^{m_n} \rightarrow \infty$ will suffice. However, this bound is the most obvious one. If we can make the bound tighter, we can hope for $m_n \rightarrow \infty$ at a faster rate. One such example is given below.

Example 3.1. Suppose $\Sigma = \{0, 1\}$. Consider a VLMC of length n and order m with the contexts $\{0, 10, 110, \dots, 1^{m-1}0, 1^m\}$. Hence there are $m+1$ leaves in the context tree. Suppose that for a fixed m , $P(0|1^j0) = 1 - p_j$ and $P(1|1^j0) = p_j$, $j = 0, 1, \dots, m-1$; and $P(0|1^m) = 1 - p_m$. In this set-up, we find that

$$\frac{1}{\pi_n(1^{m-1}0)} = \prod_{j=0}^{m-2} \frac{1}{p_j} + \frac{p_{m-1} + 1 - p_m}{1 - p_m} + \sum_{j=1}^{m-2} \prod_{k=j}^{m-2} \frac{1}{p_j}.$$

Proof: We use recursive relations for stationary probabilities to determine $\pi_n(1^{m-1}0)$.

For any history $(i_1, \dots, i_m) \in \{0, 1\}^m$, denote the index by $w(i_1, \dots, i_m) = 1 + \sum_{j=0}^{m-1} 2^j i_{m-j}$. Then the function w will provide a unique index, ranging from 1 to 2^m . For simplicity, we write $\pi_n(j)$ as the stationary probability of the m -tuple $(i_1, \dots, i_m) \in \{0, 1\}^m$ having $w(i_1, \dots, i_m) = j$. Using the definition of stationary probability, we find that

$$\sum_{j=0}^{2^{m-k}-1} \pi_n(2^k j + 2^{k-1}) = \sum_{j=0}^{2^{m-k}-1} \pi_n(2^k j + 2^k - 1); \quad k = 2, 3, \dots, m. \quad (7)$$

Also define the following quantities and obtain the relations

$$A_{k-1,1} := \sum_{j=0}^{2^{m-k}-1} \pi_n(2^k j + 2^{k-1} - 1) = c_{k-1} \sum_{j=0}^{2^{m-k}-1} \pi_n(2^k j + 2^{k-1}) =: c_{k-1} A_{k-1,2} \quad (8)$$

and $c_{k-1} = (1 - p_{k-2})/p_{k-2}$ for $k = 2, 3, \dots, m$ and $c_m = (1 - p_{m-1})/p_m$. Define $A_{m,1} = \pi_n(2^m - 1)$, $A_{m,2} = \pi_n(2^m)$. Now, observe that

$$\begin{aligned} A_{m-2,1} &= c_{m-2} A_{m-2,2} = c_{m-2} (A_{m-1,1} + A_{m,1}) \\ &= c_{m-2} (1 + c_{m-1}) A_{m,1}; \\ A_{m-3,1} &= c_{m-3} (1 + c_{m-2}) (1 + c_{m-1}) A_{m,1}; \\ &\dots \\ A_{1,1} &= c_1 (1 + c_2) \dots (1 + c_{m-1}) A_{m,1} \end{aligned}$$

and

$$\begin{aligned} A_{1,2} &= (1 + c_2) \dots (1 + c_{m-1}) A_{m,1} \\ &= c_{m-2} (1 + c_{m-1}) A_{m,1}; \\ A_{2,2} &= (1 + c_3) \dots (1 + c_{m-1}) A_{m,1}; \\ &\dots \\ A_{m-2,2} &= (1 + c_{m-1}) A_{m,1}; \\ A_{m-1,2} &= A_{m,1}. \end{aligned}$$

Combining these relations, we have

$$A_{m,1} = \pi_n(2^m - 1) = \frac{1}{\prod_{i=1}^{m-1} a_i + a_m + \sum_{i=2}^{m-1} a_i a_{i+1} \dots a_{m-1}},$$

where $a_i = 1 + c_i$. Replacing the values of c_i , we find that

$$\frac{1}{\pi_n(1^{m-1}0)} = \prod_{j=0}^{m-2} \frac{1}{p_j} + \frac{p_{m-1} + 1 - p_m}{1 - p_m} + \sum_{j=1}^{m-2} \prod_{k=j}^{m-2} \frac{1}{p_j} := q(\mathbf{p}, m).$$

If the transition probabilities satisfy $q(\mathbf{p}, m_n) = o(n)$, we can claim $n\pi_n(1^{m-1}0) \rightarrow \infty$.

Corollary 3.1.1. *Suppose $p_j = (j+1)/(j+2)$. Then $n\pi_n(1^{m_n-1}0) \rightarrow \infty$ if $m_n \log(m_n)/n \rightarrow 0$.*

Proof: $q(p, m_n) \leq 2m_n + m_n \sum_{j=1}^{m_n-2} \frac{1}{j+1} = \mathcal{O}(m_n \log(m_n))$. Hence, the required condition simplifies to $m_n \log(m_n)/n \rightarrow 0$ as $n \rightarrow \infty$.

4 Conclusion

The theoretical results given guarantee central limit theorems for certain functionals of a higher order Markov chain, whose order changes with the length of the chain. We also simplify the conditions of the theorem (2.1) for a particular type of binary variable length Markov chain and determined how the order m_n varies with n . Similar reductions may be possible for Sparse Markov Models (SMMs) by exploiting their underlying partition structure. Investigating such conditions for SMMs and other sparse higher-order Markov models is an interesting direction for future work.

References

Iris Bennett, Donald EK Martin, and Soumendra Nath Lahiri. Fitting sparse Markov models through a collapsed gibbs sampler. *Computational Statistics*, 38(4):1977–1994, 2023.

- Peter Bühlmann. Model selection for variable length Markov chains and tuning the context algorithm. *Annals of the Institute of Statistical Mathematics*, 52(2):287–315, 2000.
- Peter Bühlmann and Abraham J Wyner. Variable length Markov chains. *The Annals of Statistics*, 27(2):480–513, 1999.
- Jesús E Garcia, Verónica A González-López, Rua Sergio Buarque de Holanda, and Cidade Universitária-Barao Geraldo. Minimal Markov models. In *Fourth Workshop on Information Theoretic Methods in Science and Engineering*, page 25, 2011.
- Väinö Jääskinen, Jie Xiong, Jukka Corander, and Timo Koski. Sparse Markov chains for sequence data. *Scandinavian Journal of Statistics*, 41(3):639–655, 2014.
- Ioannis Kontoyiannis, Lambros Mertzanis, Athina Panotopoulou, Ioannis Papageorgiou, and Maria Skoularidou. Bayesian context trees: Modelling and exact inference for discrete time series. *arXiv preprint arXiv:2007.14900*, 2020.
- Tuhin Majumder, Soumendra N Lahiri, and Donald EK Martin. Fitting sparse Markov models to categorical time series using convex clustering. *Journal of Computational and Graphical Statistics*, pages 1–13, 2026.
- Ioannis Papageorgiou and Ioannis Kontoyiannis. Posterior representations for bayesian context trees: Sampling, estimation and convergence. *arXiv preprint arXiv:2202.02239*, 2022.
- Adrian E Raftery. A model for high-order Markov chains. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(3):528–539, 1985.
- Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of statistics*, 11(2):416–431, 1983.
- Jie Xiong, Väinö Jääskinen, Jukka Corander, et al. Recursive learning for sparse Markov models. *Bayesian analysis*, 11(1):247–263, 2016.