Journal of Machine Learning Research 24 (2023) 1-54 Submitted 3/22; Published 8/23

Sparse Markov Models for High-dimensional Inference

Guilherme Ost GUILHERMEOST@IM.UFRJ.BR
Institute of Mathematics
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil

Daniel Y. Takahashi TAKAHASHIYD@GMAIL.COM
Brain Institute
Federal University of Rio Grande do Norte
Natal, RN, Brazil

Editor: Aurélien Garivier

Abstract

Finite-order Markov models are well-studied models for dependent finite alphabet data. Despite their generality, application in empirical work is rare when the order d is large relative to the sample size n (*e.g.*, $d = \mathcal{O}(n)$). Practitioners rarely use higher-order Markov models because (1) the number of parameters grows exponentially with the order, (2) the sample size n required to estimate each parameter grows exponentially with the order, and (3) the interpretation is often difficult. Here, we consider a subclass of Markov models called Mixture of Transition Distribution (MTD) models, proving that when the set of relevant lags is sparse (*i.e.*, $\mathcal{O}(\log(n))$), we can consistently and efficiently recover the lags and estimate the transition probabilities of high-dimensional ($d = \mathcal{O}(n)$) MTD models. Moreover, the estimated model allows straightforward interpretation. The key innovation is a recursive procedure for a priori selection of the relevant lags of the model. We prove a new structural result for the MTD and an improved martingale concentration inequality to prove our results. Using simulations, we show that our method performs well compared to other relevant methods. We also illustrate the usefulness of our method on weather data where the proposed method correctly recovers the long-range dependence.

Keywords: Markov Chains, High-dimensional inference, Mixture Transition Distribution

1. Introduction

From the daily number of COVID-19 cases to the activity of neurons in the brain, discrete time series are ubiquitous in our life. A natural way to model these time series is by describing how the present events depend on the past events, *i.e.*, characterizing the transition probabilities. Therefore, finite-order Markov chains - models specified by transition probabilities that depend only on a limited portion of the past - are an obvious choice to model time series with discrete values. The length of the portion of the relevant past defines the order of the Markov chain. At first glance, estimating the transition probabilities of a Markov chain from the data is straightforward. Given a sample $X_{1:n} := (X_1, X_2, \ldots, X_n)$ of a stationary d-th order Markov chain on a discrete alphabet A, the empirical transition probabilities are computed, for all past $x_{-d:-1} := (x_{-d}, \ldots, x_{-1}) \in A^{\{-d,\ldots,-1\}}$ and symbol

$a \in A$, as

$$\hat{p}_n(a|x_{-d:-1}) := \frac{N_n(x_{-d:-1}, a)}{\sum_{b \in A} N_n(x_{-d:-1}, b)},$$

where $N_n(x_{-d:-1}, a)$ denotes the number of occurrences of the past $x_{-d:-1}$ followed by the symbol a in the sample $X_{1:n}$.

Nevertheless, some difficulties become apparent. First, for a Markov chain of order d, we have to estimate $|A|^d(|A|-1)$ transition probabilities (*parameters*), making the uniform control of estimation errors much harder when the order d increases. One solution to avoid the exponential increase in the number of parameters is to consider more parsimonious classes of models. One such popular class of models is the *variable length Markov chains* (VLMC), in which

$$\mathbb{P}(X_t = a|X_{t-d:t-1} = x_{-d:-1}) = \mathbb{P}(X_t = a|X_{t-\ell:t-1} = x_{-\ell:-1}),$$

where ℓ is a function of the past $x_{-d:-1}$ (Rissanen, 1983; Bühlmann and Wyner, 1999; Galves et al., 2012). The relevant portion $x_{-\ell:-1}$ of the past $x_{-d:-1}$ is called a *context*. The key feature of VLMC is that all transition probabilities with the same context have the same values. Therefore, denoting τ as the set of all contexts, the number of transition probabilities that need to be estimated reduces to $|\tau|(|A|-1)$. Another class of models that is even more parsimonious is the *Minimal Markov Models* - also known as *Sparse Markov Chains* (SMC) (Garcıa et al., 2011; Jääskinen et al., 2014). In SMC, we say that the pasts $x_{-d:-1}$ and $y_{-d:-1}$ are related if for all symbols $a \in A$,

$$\mathbb{P}(X_t = a|X_{t-d:t-1} = x_{-d:-1}) = \mathbb{P}(X_t = a|X_{t-d:t-1} = y_{-d:-1}).$$

This relation generates the equivalent classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$ that partition $A^{\{-d,\ldots,-1\}}$. Now, the number of transition probabilities that need to be estimated is $K(|A|-1)$. Both VLMC and SMC have the advantage of better balancing the bias and variance tradeoff. Nevertheless, in both models we still need to estimate the transition probability using $\hat{p}_n(a|x_{-d:-1})$, either because we need to estimate the largest context (for VLMC) or because we need first to calculate the transition probabilities to establish the partitions (for SMC). This creates a second difficulty. For the estimator $\hat{p}_n(a|x_{-d:-1})$ to have any meaning, we have to observe the sequence $x_{-d:-1}$ in the sample $X_{1:n}$ at least once. By ergodicity, the number of times that we will observe the sequence $x_{-d:-1}$ is roughly $n\mathbb{P}(X_{1:d} = x_{-d:-1})$. It is straightforward to show that, if the transition probabilities are bounded below from zero, there exists a constant $c > 0$ such that $\mathbb{P}(X_{1:d} = x_{-d:-1}) < e^{-cd}$. Therefore, in general, it is hopeless to have a reasonable estimator $\hat{p}_n(a|x_{-d:-1})$ if $d > (1+\varepsilon)\log n/c$, for some positive value ε. This imposes a fundamental limit to the size of the past that can be included in the description of the time series.

Moreover, Markov chains with small orders are not consistent with the known workings of several natural phenomena where the transition probabilities might depend on remote pasts. For example, in predicting whether today will be a warm or cold day, we might need to know remote past events like the corresponding weather approximately a year ago (Király et al., 2006; Yuan et al., 2013). Physiological phenomena in humans with cycles of different lengths might result from dependence on events at vastly different temporal scales (Gilden et al., 1995; Chen et al., 1997; Buzsaki and Draguhn, 2004). Importantly, not all portions of the past are necessarily relevant. These observations motivate us to explore sparser representations of the dependence on past events. The *mixture of transition distribution* model (MTD) is a subclass of finite order Markov chains that can be used to obtain such sparse representation. Like VLMC and SMC, MTD was initially introduced to overcome the problem of exponential increase in the number of transition probabilities for Markov chains (Raftery, 1985; Berchtold and Raftery, 2002). MTD represents higher-order Markov models as a convex mixture of single-step Markov chains, where each single-step Markov chain depends on a single time point in the past. If an MTD model is a mixture of only a few single step Markov chains, we naturally obtain a class of sparse Markov chains that depends only on a small portion of the past events. Nevertheless, available methods to consistently estimate the transition probabilities of MTD still need to consider all the past events up to the MTD order (Berchtold and Raftery, 2002), which might include irrelevant portions of the past. This fact still restricts the MTD order to $d = \mathcal{O}(\log n)$.

In this work, we introduce a simple method that consistently recovers the relevant part of the past even when the order d of the MTD model is proportional to the sample size n (*i.e.*, $d = \mathcal{O}(n)$) if the size of the relevant past is $\mathcal{O}(\log n)$. Consequently, we prove that we can consistently estimate the transition probabilities for high dimensional MTD under sparsity constraint. Our estimator is computationally efficient, consisting of a forward stepwise procedure that finds the candidates for relevant past portions and a cutting procedure to eliminate the irrelevant portions. The theoretical guarantees of our estimator are based on a novel structural result for MTD and an improved martingale concentration inequality. Both results might have an interest on their own. Moreover, we show that the estimator can be further improved when the alphabet is binary. We also prove that in several cases, our estimator is minimax rate optimal.

Finally, using simulated data, we show that our method's performance is, in general superior to a best subset selection method, where the lags with k largest weights are selected after estimating the model with a classical MTD estimation method (Berchtold, 2001), and similar to the performance of Conditional Tensor Factorization (CTF) based on higher-order Markov chain estimation when the order is moderate (Sarkar and Dunson, 2016). We also applied our method to weather data to model a binary sequence indicating days with and without rain. Our method successfully captures long-range dependencies (*e.g.* annual cycle) that were not detected either by VLMC algorithm with BIC order selection (Csiszár and Talata, 2006) or by the CTF based higher order Markov chain estimation. New Bayesian approaches for higher order VLMC and MTD selection were introduced in (Kontoyiannis et al., 2020; Heiner and Kottas, 2021), where a posteriori most likely model estimation is considered. These works provide interesting alternative approaches for modeling higher-order Markovian dependence in a Bayesian setting.

We organized the paper as follows. In Section 2 we introduce the main notations, definitions, and assumptions that we will use throughout the paper. In Section 3 we introduce the algorithms to select the relevant part of the past. In Section 3.4 we provide an estimate of the estimator's convergence rate for the transition probabilities. In Section 3.5, we show that our estimator achieves the optimal minimax rate. In Section 4, we illustrate the performance of the proposed estimators through simulations and an application on weather data.

2. Notation, Model Definition and Preliminary Remarks

2.1 General notation

We denote $\mathbb{Z} = \{\ldots - 1, 0, 1, \ldots\}$ and $\mathbb{Z}_+ = \{1, 2 \ldots\}$ the set of integers and positive integers respectively. For $s, t \in \mathbb{Z}$ with $s \leq t$, we write $[\![s, t]\!]$ to denote the discrete interval $\mathbb{Z} \cap [s, \ldots, t]$. Throughout the article A denotes a finite subset of $\mathbb{R}$, called *alphabet.* The elements of A will be denoted by the first letters of the alphabet a, b and c. Hereafter, we denote $\|A\|_\infty = \max_{a \in A} |a|$ and $Diam(A) = \max_{a,b \in A} |a - b|$. For each $S \subset \mathbb{Z}$, the set A^S denotes the set of all A-valued strings $x_S = (x_j)_{j \in S}$ indexed by the set S. To alleviate the notation, if $S = [\![s, t]\!]$ for $s, t \in \mathbb{Z}$ with $s \leq t$, we write $x_{s:t}$ instead of $x_{[\![s,t]\!]}$. For any non-empty subsets $U \subset S \subseteq \mathbb{Z}$ and any string $x_S \in A^S$, we denote $x_{(S \setminus U)} \in A^{(S \setminus U)}$ the string obtained from x_S by removing the string $x_U \in A^U$. For all $t \in \mathbb{Z}$ and $S \subset \mathbb{Z}$, we will write in some cases $t + S$ do denote the set $\{t + s : s \in S\}$.

The set of all finite A-valued strings is denoted by

$$\mathcal{A} = \bigcup_{S \subset \mathbb{Z} : S \text{ finite}} A^S.$$

For all $x \in \mathcal{A}$, we denote $S_x \subset \mathbb{Z}$ the set indexing the string x, i.e., such that $x \in A^{S_x}$.

Given two probability measures μ and ν on A, we denote $d_{TV}(\mu, \nu)$ the total variation distance between μ and ν, defined as

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{a \in A} |\mu(a) - \nu(a)|.$$

For $q \in \mathbb{Z}_+$, the $\| \cdot \|_q$-norm of vector $v \in \mathbb{R}^L$ is defined as

$$\|v\|_q = \left(\sum_{\ell=1}^{L} |v_\ell|^q \right)^{1/q}.$$

The dimension $L \in \mathbb{Z}_+$ will be implicit in most cases.

For two probability distributions P and Q on $A^{[\![1,k]\!]}$ where P is absolutely continuous with respect to Q, we denote $KL(P||Q)$ the *Kullback-Leibler* divergence between P and Q, given by

$$KL(P||Q) = \sum_{x_{1:k} \in A^{[\![1,k]\!]}} P(x_{1:k}) \log \left(\frac{P(x_{1:k})}{Q(x_{1:k})} \right).$$

2.2 Markov models

Let $\boldsymbol{X} = (X_t)_{t \in \mathbb{Z}}$ be a discrete time stochastic chain, defined in a suitable probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking values in an alphabet A. For a $d \in \mathbb{Z}_+$, we say that $\mathbf{X}$ is a *Markov chain of order* d if for all $k \in \mathbb{Z}_+$ with $k > d$, $t \in \mathbb{Z}$ and $x_{t-k:t} \in A^{[\![t-k,t]\!]}$ with $\mathbb{P}(X_{t-k:t-1} = x_{t-k:t-1}) > 0$, we have

$$\mathbb{P}\left(X_t = x_t | X_{t-k:t-1} = x_{t-k:t-1}\right) = \mathbb{P}\left(X_t = x_t | X_{t-d:t-1} = x_{t-d:t-1}\right). \tag{1}$$

We say that a Markov chain is *stationary* if $X_{s:t}$ and $X_{s+h:t+h}$ have the same distribution for all $t, s, h \in \mathbb{Z}$. Throughout the article, the distribution of a stationary Markov chain will be denoted by $\mathbf{P}$. For a finite $S \subset \mathbb{Z}$ and $x_S \in A^S$, we write $\mathbf{P}(x_S)$ to denote $\mathbb{P}(X_S = x_S)$. The *support* of a stationary Markov model is the set $\operatorname{supp}(\mathbf{P}) = \{x \in \mathcal{A} : \mathbf{P}(x_{S_x}) > 0\}$.

For stationary Markov chains, the conditional probabilities in (1) do not depend on the time index t. Therefore, for a stationary Markov chain of order d, for any $a \in A$, $x_S \in \operatorname{supp}(\mathbf{P})$ with $S \subseteq [\![-d, -1]\!]$ and $t \in \mathbb{Z}$, we denote

$$p(a|x_S) = \mathbb{P}\left(X_t = a | X_{t+S} = x_S\right).$$

Notice that $p(\cdot|x_S)$ is a probability measure on A, for each fixed past $x_S \in \operatorname{supp}(\mathbf{P})$. The set $\{p(\cdot|x_{-d:-1}) : x_{-d:-1} \in \operatorname{supp}(\mathbf{P})\}$ is called the family of *transition probabilities* of the chain. In this article, we consider only stationary Markov chains.

For a Markov chain of order d, the *oscillation* δ_j for $j \in [\![-d, -1]\!]$ is defined as

$$\delta_j = \max\{d_{TV}\left(p(\cdot|x_{-d:-1}), p(\cdot|y_{-d:-1})\right) : (x_{-d:-1}, y_{-d:-1}) \in A^{[\![-d,-1]\!]}, x_{-k} = y_{-k},\ \forall\ k \neq j\}.$$

The oscillation is useful to measure the influence of a j-th past value in the values of the transition probabilities.

2.3 Mixture transition distribution (MTD) models

A MTD model of order $d \in \mathbb{Z}_+$ is a Markov chain of order d for which the associated family of transition probabilities $\{p(\cdot|x_{-d:-1}) : x_{-d:-1} \in \operatorname{supp}(\mathbf{P})\}$ admits the following representation:

$$p(a|x_{-d:-1}) = \lambda_0 p_0(a) + \sum_{j=-d}^{-1} \lambda_j p_j(a|x_j),\ a \in A, \tag{2}$$

with $\lambda_0, \lambda_{-1}, \ldots, \lambda_{-d} \in [0, 1]$ satisfying $\sum_{j=-d}^{0} \lambda_j = 1$ and $p_0(\cdot)$ and $p_j(\cdot|b)$,$j \in [\![-d, -1]\!]$ and $b \in A$, being probability measures on A.

Following (Berchtold and Raftery, 2002), we call the index $j \in [\![-d, 0]\!]$ of the weight λ_j in (2) the *j-th lag* of the model. The representation in (2) has the following probabilistic interpretation. To sample a symbol from $p(\cdot|x_{-d:-1})$, we first choose a lag in $[\![-d, 0]\!]$ randomly, being λ_j the probability of choosing the lag j. Once the lag has been chosen, say lag j, we then sample a symbol from the probability measure $p_j(\cdot|x_j)$ which depends on the past $x_{-d:-1}$ only through the symbol x_j. Notice that a symbol is sampled independently from the past $x_{-d:-1}$, whenever the lag 0 is chosen.

For later use, let us define the conditional average at lag j as

$$m_j(b) = \sum_{a \in A} a p_j(a|b), \tag{3}$$

for each $j \in [\![-d, -1]\!]$ and $b \in A$.

For a MTD model of order d, we have that the oscillation δ_j of the lag $j \in [\![-d, -1]\!]$ can be written as,

$$\delta_j = \lambda_j \max_{b,c \in A} d_{TV}(p_j(\cdot|b), p_j(\cdot|c)). \tag{4}$$

Notice that in this case $\delta_j = 0$ if and only if either $\lambda_j = 0$ or $d_{TV}(p_j(\cdot|b), p_j(\cdot|c)) = 0$ for all $b, c \in A$.

In the sequel, we say that the lag j is *relevant* if $\delta_j > 0$, and *irrelevant* otherwise. We will denote Λ the set of all relevant lags, i.e.,

$$\Lambda = \{j \in [\![-d, -1]\!] : \delta_j > 0\}. \tag{5}$$

The set Λ captures the dependence structure of the MTD model. The size $|\Lambda|$ of the set Λ represents the degree of sparsity of the MTD model. The smaller the value of $|\Lambda|$, the sparser the MTD model.

The following quantities will appear in many of our results:

$$\delta_{min} = \min_{j\in\Lambda} \delta_j \text{ and } \tilde{\delta}_{min} = \min_{j\in\Lambda} \lambda_j \|m_j\|_{Lip}, \tag{6}$$

where $\|m_j\|_{Lip} = \max\{|m_j(b) - m_j(c)|/|b-c| : b, c \in A, b \neq c\}$ denotes the Lipschitz norm of the function m_j defined in (3). One can check easily that these quantities coincide when the alphabet A is *binary* (i.e. $A = \{0, 1\}$). For general alphabets, the following inequality holds:

$$\delta_{min} \geq \|A\|_\infty^{-1} \tilde{\delta}_{min} \min_{b,c\in A: b\neq c} |b - c|.$$

2.4 Statistical lag selection

Suppose that we are given a sample $X_{1:n}$ of a MTD model of known order $d < n$ and whose set of relevant lags Λ is unknown. The goal of *statistical lag selection* is to estimate the set Λ from the sample $X_{1:n}$. Our particular interest is in the high-dimensional setting in which the parameters $d = d_n$ and $|\Lambda| = |\Lambda_n|$ scale as a function of the sample size n. Let us write $\hat{\Lambda}_n$ to indicate an estimator of the set of relevant lags Λ computed from the sample $X_{1:n}$. We say that the estimator $\hat{\Lambda}_n$ is *consistent* if

$$\mathbb{P}(\hat{\Lambda}_n \neq \Lambda) \to 0 \text{ as } n \to \infty.$$

With respect to statistical lag selection, our goal is to exhibit sufficient conditions for each proposed estimator guaranteeing its consistency.

2.5 Empirical transition probabilities

Let n, m and d be positive integers such that $n - m > d$. We denote for each $a \in A$ and $x_S \in \mathcal{A}$ with $S \subseteq [\![-d, -1]\!]$ non-empty,

$$N_{m,n}(x_S, a) = \sum_{t=m+d+1}^{n} 1\{X_{t+j} = x_j, j \in S, X_t = a\}.$$

The random variable $N_{m,n}(x_S, a)$ indicates the number of occurrences of the string x_S "followed" by the symbol a, in the last $n - m$ symbols $X_{m+1:n}$ of the sample $X_{1:n}$. We also define $\bar{N}_{m,n}(x_S) = \sum_{a\in A} N_{m,n}(x_S, a)$. With this notation, the empirical transition probabilities computed from the last $n-m$ symbols $X_{m+1:n}$ of the sample $X_{1:n}$ are defined as,

$$\hat{p}_{m,n}(a|x_S) = \begin{cases} \frac{N_{m,n}(x_S,a)}{\bar{N}_{m,n}(x_S)}, \text{ if } \bar{N}_{m,n}(x_S) > 0 \\ \frac{1}{|A|}, \text{ otherwise} \end{cases} . \tag{7}$$

When the countings are made over the whole sample $X_{1:n}$, we denote $N_n(x_S,a)$ and $\bar{N}_n(x_S)$ the corresponding counting random variables, and $\hat{p}_n(a|x_S)$ the corresponding empirical transition probabilities.

In the next sections, the estimators for the set of relevant lags we propose in this paper rely on these empirical transition probabilities. If $\hat{\Lambda}_m$ denotes an estimator for the set of relevant lags Λ computed from $X_{1:m}$, we expect that under some assumptions (guaranteeing in particular the consistency of $\hat{\Lambda}_m$) the empirical transition probability $\hat{p}_{m,n}(a|x_{\hat{\Lambda}_m})$ converges (in probability) to $p(a|x_\Lambda)$ as $\min\{n,m\} \to \infty$, for any $x_{-d:-1} \in \text{supp}(\mathbf{P})$. To understand the convergence for the transition probabilities of high order Markov chains is crucial in our analysis.

2.6 Assumptions

We collect here the main assumptions used in the article.

Assumption 1 *The MTD model has full support, that is, supp*$(\mathbf{P}) = \mathcal{A}$*.*

In other words, Assumption 1 means that $\mathbb{P}(X_S = x_S) = \mathbf{P}(x_S) > 0$ for any string $x_S \in A^S$ with $S \subset \mathbb{Z}$ finite. This means that the marginal distributions of the distribution generating the data are strictly positive. Such a condition is usually assumed in the problem of estimating the graph structure underlying graphical models (see for instance Chapter 11 of (Wainwright, 2019)). Notice that this assumption implies, in particular, that

$$p_{min} = \min\{p(a|x_\Lambda) : a \in A,\ x_\Lambda \in A^\Lambda\} > 0, \tag{8}$$

where $p(\cdot|x_\Lambda)$ are the transition probabilities of MTD generating the data.

Assumption 2 *The quantity* $\Delta := 1 - \sum_{j\in\Lambda} \delta_j > 0$*, where* δ_j *is given by* (4)*.*

We have that $\lambda_0 > 0$ is a sufficient condition to Assumption 2 to hold. To check this, notice that

$$\sum_{j\in\Lambda} \delta_j = \sum_{j\in\Lambda} \lambda_j \max_{b,c\in A} d_{TV}(p_j(\cdot|b), p_j(\cdot|c)) \leq \sum_{j\in\Lambda} \lambda_j = 1 - \lambda_0,$$

where we have used that $d_{TV}(p_j(\cdot|b), p_j(\cdot|c)) \leq 1$ for all $b,c \in A$ and $j \in \Lambda$. Hence, it follows that $\Delta > 0$ whenever $\lambda_0 > 0$.

Assumptions 1 and 2 are used to obtain concentration inequalities for the counting random variables $N_{m,n}(x_S,a)$ and $\bar{N}_{m,n}(x_S)$ appearing in the definition of the empirical transition probabilities $\hat{p}_{m,n}(a|x)$.

The next assumption is as follows.

Assumption 3 *For each* $j \in \Lambda$*, there exists* $b^\star, c^\star \in A$ *such that* $m_j(b^\star) \neq m_j(c^\star)$*, where* $m_j(\cdot)$ *is defined in* (3)*.*

Notice that if $A = \{0,1\}$, then $m_j(1) - m_j(0) = p_j(1|1) - p_j(1|0)$, so that Assumption 3 holds whenever $d_{TV}(p_j(\cdot|1), p_j(\cdot|0)) = |p_j(1|1) - p_j(1|0)| > 0$ for each $j \in \Lambda$. In this case this is always true by the definition of the set Λ. As we will see in Section 3, the condition is crucial to prove a structural result about MTD models, presented in Proposition 6.

In what follows, $\mathbb{P}_{x_S}(X_j \in \cdot | X_k = b)$ denotes the conditional distribution of X_j given $X_S = x_S$ and $X_k = b$. We use the convention that, for $S = \emptyset$, these conditional probabilities correspond to the unconditional ones. Moreover, for any function $f : A \to \mathbb{R}$, we write $\mathbb{E}_{x_S}(f(X_j)|X_k = b)$ to denote the expectation of $f(X_j)$ with respect to $\mathbb{P}_{x_S}(X_j \in \cdot | X_k = b)$.

The next two assumptions are the following.

Assumption 4 (Inward weak dependence condition) *There exists $\Gamma_1 \in (0,1]$ such that the following condition holds: for all $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \not\subseteq S$, $k \in \Lambda \setminus S$ and $b, c \in A$ with $b \neq c$ satisfying $|m_k(b) - m_k(c)| > 0$,*

$$\max_{x_S \in A^S} \sum_{j \in \Lambda \setminus S \cup \{k\}} \frac{\lambda_j \left|\mathbb{E}_{x_S}(m_j(X_j)|X_k = b) - \mathbb{E}_{x_S}(m_j(X_j)|X_k = c)\right|}{\lambda_k |m_k(b) - m_k(c)|} \leq (1 - \Gamma_1). \tag{9}$$

Assumption 5 (Outward weak dependence condition) *The alphabet is binary, i.e. $A = \{0,1\}$. Moreover, there exists $\Gamma_2 \in (0,1]$ such that the following condition holds: for all $S \subseteq [\![-d,-1]\!]$ such that $S \subset \Lambda$ and $k \notin \Lambda$,*

$$\sum_{j \in \Lambda \setminus S} \max_{x_S \in \{0,1\}^S} |\mathbb{P}_{x_S}(X_k = 1|X_j = 1) - \mathbb{P}_{x_S}(X_k = 1|X_j = 0)| \leq \Gamma_2. \tag{10}$$

Both Assumptions 4 and 5 are conditions of weak dependence. In words, Assumption 4 says that no relevant lag j can be completely determined by any subset S containing only relevant lags or any other relevant lag k when combined with some irrelevant lags. Similarly, Assumption 5 says that irrelevant lags cannot be completely determined by some subset of relevant lags. These two assumptions will be only necessary to obtain a computationally very efficient algorithm.

3. Statistical Lag Selection

In this section, we address the problem of statistical lag selection for the MTD models. We will first introduce a statistical procedure called PCP estimator that is general and works well if there is a known small set S such that $\Lambda \subseteq S$. When such set S is not available, we will have to consider an alternative procedure called FSC estimator, which will be introduced later.

3.1 Estimator based on pairwise comparisons

Throughout this section we suppose that there is a known set $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \subseteq S$. Note that this is always satisfied in the worse case scenario in which the set S is the whole set $[\![-d,-1]\!]$. In some cases, however, we may have a prior knowledge on the set Λ and we can use this information to restrict our analysis to the lags in a known set S of size (possibly much) smaller than d.

The estimator discussed in this section is based on pairwise comparisons of empirical transition probabilities corresponding to compatible pasts. For this reason, we call it `PCP` estimator. The estimator is based on the following observation. For any $j \in S$, we say that the pasts $x_S, y_S \in A^S$ are $(S \setminus \{j\})$-*compatible*, if $y_{S\setminus\{j\}} = x_{S\setminus\{j\}}$. We have that if $j \in \Lambda$, then there exist a pair of $(S \setminus \{j\})$-compatible pasts $x_S, y_S \in A^S$ such that total variation distance between $p(\cdot|x_S)$ and $p(\cdot|y_S)$ is strictly positive. On the other hand, if $j \in S \setminus \Lambda$, then the total variation distance between $p(\cdot|x_S)$ and $p(\cdot|y_S)$ is 0 for all $(S\setminus\{j\})$-compatible pasts $x_S, y_S \in A^S$.

These remarks suggests to estimate Λ by the subset of all lags $j \in S$ for which the total variation distance between $\hat{p}_n(\cdot|x_S)$ and $\hat{p}_n(\cdot|y_S)$ is larger than a suitable positive threshold, for some pair of $(S\setminus\{j\})$-compatible pasts x_S and y_S. An uniform threshold over all possible realizations usually gives suboptimal results by either underestimating or overestimating for some configurations. The threshold we use here is adapted to each realization of the MTD, relying on improved martingale concentration inequalities that are of independent interest (see Appendix B).

Fix $\varepsilon > 0$, $\alpha > 0$ and $\mu \in (0, 3)$ such that $\mu > \psi(\mu) := e^{\mu} - \mu - 1$. For each $x_S, y_S \in A^S$, consider the random threshold $t_n(x_S, y_S)$ defined as,

$$t_n(x_S, y_S) = s_n(x_S) + s_n(y_S), \tag{11}$$

where $s_n(x_S)$ is given by

$$s_n(x_S) = \sqrt{\frac{\alpha(1+\varepsilon)}{2\bar{N}_n(x_S)}} \sum_{a\in A} \sqrt{\frac{\mu}{\mu - \psi(\mu)} \left(\hat{p}_n(a|x_S) + \frac{\alpha}{\bar{N}_n(x_S)}\right)} + \frac{\alpha|A|}{6\bar{N}_n(x_S)}. \tag{12}$$

With this notation, the `PCP` estimator $\hat{\Lambda}_{1,n}$ is defined as follows. A lag $j \in S$ belongs to $\hat{\Lambda}_{1,n}$ if and only if there exists a $(S \setminus \{j\})$-compatible pair of pasts $x_S, y_S \in A^S$ such that

$$d_{TV}(\hat{p}_n(\cdot|x_S), \hat{p}_n(\cdot|y_S)) \geq t_n(x_S, y_S). \tag{13}$$

In the sequel, the set $S \subseteq [\![-d, -1]\!]$ such that $\Lambda \subseteq S$ and the constants $\varepsilon > 0$, $\alpha > 0$ and $\mu \in (0, 3)$ such that $\mu > \psi(\mu)$ are called parameters of the `PCP` estimator $\hat{\Lambda}_{1,n}$.

Hereafter, for each $j \in S$ and any $b, c \in A$, let

$$\mathcal{C}_j(b,c) = \left\{(x,y) \in A^S \times A^S : x_{S\setminus\{j\}} = y_{S\setminus\{j\}},\ x_j = b \text{ and } y_j = c\right\},$$

and define

$$t_{n,j}(b,c) = \min_{(x_S,y_S)\in\mathcal{C}_j(b,c)} t_n\left(x_S, y_S\right),\ t_{n,j} = \max_{b,c\in A: b\neq c} t_{n,j}(b,c) \text{ and } \gamma_{n,j} = 2t_{n,j}. \tag{14}$$

Finally, consider the following quantity

$$\mathbf{P}_S = \min_{j\in\Lambda} \min_{b,c\in A: b\neq c} \max_{(x_S,y_S)\in\mathcal{C}_j(b,c)} \left(\mathbf{P}(x_S) \wedge \mathbf{P}(y_S)\right). \tag{15}$$

With these definitions, we have the following result.

Theorem 1 *Let $X_{1:n}$ be a sample of MTD model with set of relevant lags Λ, where $n > d$. If $\hat{\Lambda}_{1,n}$ is the* `PCP` *estimator defined in* (13) *with parameters $\mu \in (0,3)$ such that $\mu > \psi(\mu)$, $\alpha > 0$, $\varepsilon > 0$ and $\Lambda \subseteq S \subseteq [\![-d,-1]\!]$, we have that*

1. *For each $j \in S \setminus \Lambda$, we have that*

$$\mathbb{P}\left(j \in \hat{\Lambda}_{1,n}\right) \leq 8|A|(n-d)\left\lceil \frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}.$$

2. *For each $j \in \Lambda$, we have that*

$$\mathbb{P}\left(j \notin \hat{\Lambda}_{1,n}, \gamma_{n,j} \leq \delta_j\right) \leq 8|A|\left\lceil \frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha},$$

where $\gamma_{n,j}$ and and δ_j are defined in (14) *and* (4) *respectively.*

3. *Furthermore, if assumptions 1 and 2 hold, then there exits a constant $C = C(\varepsilon, \mu) > 0$ such that for n satisfying*

$$n \geq d + \frac{C|A|\alpha}{\delta_{min}^2 P_S}, \tag{16}$$

where δ_{min} and $\mathbf{P}_S$ are defined in respectively (6) *and* (15)*, we have that*

$$\mathbb{P}\left(\hat{\Lambda}_{1,n} \neq \Lambda\right) \leq 8|A|\left((|S|-|\Lambda|)(n-d)+|\Lambda|\right)\left\lceil \frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha} + 6|A|(|A|-1)|\Lambda| \exp\left\{-\frac{\Delta^2(n-d)^2\mathbf{P}_S^2}{8n(|S|+1)^2}\right\}. \tag{17}$$

The proof of Theorem 1 is given in Appendix A.1.1.

Remark 2 *(a) The sum over $j \in S \setminus \Lambda$ of the upper bound provided by Item 1 of Theorem 1 controls the probability that the* `PCP` *estimator $\hat{\Lambda}_{1,n}$ overestimates the set of relevant lags Λ. The sum over $j \in \Lambda$ of the upper bound given in Item 2 of Theorem 1 is as an upper bound for the probability that the* `PCP` *estimator underestimates the subset of relevant lags $j \in \Lambda$ whose oscillation δ_j is larger or equal than the "noise level" $\gamma_{n,j}$. Note that the sum of these upper bounds corresponds to the first term appearing on the right hand side of* (17).

(b) The second term on the right hand side of (17) *is an upper bound for the probability that there exists some relevant lag $j \in \Lambda$ whose oscillation δ_j is strictly smaller than the "noise level" $\gamma_{n,j}$.*

(c) (Computation of PCP estimator) As we show in Appendix (A.6)*, the PCP estimator can be implemented with at most $O(|A|^2|S|(n-d))$ computations.*

Remark 3 *By Assumption 1, we have that $\mathbf{P}_S \geq p_{min}/|A|^{|S|-1}$, where p_{min} is defined in* (8)*. As a consequence, it follows from* (16) *that if the sample size n is such that*

$$n \geq d + \frac{C|A|^{|S|}\alpha}{p_{min}\delta_{min}^2}, \tag{18}$$

then inequality (17) *holds with the second exponential term replaced by*

$$\exp\left\{-\frac{\Delta^2 p_{min}^2 (n-d)^2}{8n(|S|+1)^2|A|^{2(|S|-1)}}\right\}. \tag{19}$$

Combining Theorem 1 and Remark 3, one can deduce the following result.

Corollary 4 *For each n, consider a MTD model with set of relevant lags Λ_n and transition probabilities $p_n(a|x_{\Lambda_n})$ such that $p_{min,n} \geq p^\star_{min}$ and $\Delta_n \geq \Delta^\star_{min}$ for some positive constants $p^\star_{min}$ and $\Delta^\star_{min}$. Let $d_n = \beta n$ for some $\beta \in (0,1)$ and suppose that $\Lambda_n \subseteq S_n \subseteq [\![-d_n, -1]\!]$ with $|S_n| \leq ((1-\gamma)/2)\log_{|A|}(n)$ for some $\gamma \in (0,1)$. Let $X_{1:n}$ be a sample from the MTD specified by Λ_n and $p_n(a|x_{\Lambda_n})$, and denote $\hat{\Lambda}_{1,n}$ the* PCP *estimator defined in* (13) *computed from this sample with parameters $\mu_n = \mu \in (0,3)$ such that $\mu > \psi(\mu)$, $\varepsilon_n = \varepsilon > 0$, $\alpha_n = (1+\eta)\log(n)$ with $\eta > 0$ and S_n. Under these assumptions there exists a constant $C = C(\mu, \varepsilon, \beta, p^\star_{min}, \Delta^\star_{min}, \gamma, \eta) > 0$ such that if*

$$\delta^2_{min,n} \geq \frac{C\log(n)}{n^{(1+\gamma)/2}}, \tag{20}$$

then $\mathbb{P}(\hat{\Lambda}_{1,n} \neq \Lambda_n) \to 0$ as $n \to \infty$.

The proof of Corollary is given in Appendix A.1.2.

Remark 5 *(a) Under the assumptions of Corollary 4, if additionally we have $|\Lambda_n| \leq L$ for all values of n for some positive integer L, then one can choose a suitable sequence $\gamma_n \to 1$ as $n \to \infty$ to obtain that $\mathbb{P}(\hat{\Lambda}_{1,n} \neq \Lambda_n)$ vanishes as $n \to \infty$, as long as*

$$\delta^2_{min,n} \geq \frac{C\log(n)}{n}, \tag{21}$$

where the constant C here is larger than the one given in (20).

(b) Observe that in Corollary 4, the set of relevant lags can be either finite or grow very slowly with respect to the sample size n. On the other hand, no assumption on the orders d_n of the underlying sequence of MTD models is made. In particular, we could consider MTD models with very large orders, for example $d_n = \beta n$ with $\beta \in (0,1)$.

As Corollary 4 indicates, in the setting $d_n = \beta n$, the major drawback of the PCP estimator $\hat{\Lambda}_{1,n}$ is that it requires a prior knowledge of Λ_n in the form of a set S_n growing slowly enough and such that $\Lambda_n \subseteq S_n$. The main goal of the next two sections is to propose alternative estimators of Λ_n to deal with this issue.

3.2 Forward Stepwise and Cut estimator

In this section we introduce a second estimator of the set of relevant lags Λ, called *Forward Stepwise and Cut* (FSC) estimator. This estimator is based on a structural result about MTD models presented in Proposition 6 below. Before presenting this structural result, we need to introduce some notation.

In what follows, for each lag $k \in [\![-d,-1]\!]$, subset $S \subseteq [\![-d,-1]\!] \setminus \{k\}$, configuration $x_S \in A^S$ and symbols $b, c \in A$, let us denote

$$d_{k,S}(b,c,x_S) = d_{TV}\left(\mathbb{P}_{x_S}(X_0 \in \cdot | X_k = b), \mathbb{P}_{x_S}(X_0 \in \cdot | X_k = c)\right), \tag{22}$$

and

$$w_{k,S}(b,c,x_S) = \mathbb{P}_{x_S}(X_k = b)\mathbb{P}_{x_S}(X_k = c). \tag{23}$$

Recall that $\mathbb{P}_{x_S}(X_0 \in \cdot | X_k = b)$ and $\mathbb{P}_{x_S}(X_k = b)$ denote, respectively, the conditional distribution of X_0 given $X_S = x_S$ and $X_k = b$ and the conditional probability of $X_k = b$ given $X_S = x_S$, with the convention that these conditional probabilities for $S = \emptyset$ correspond to the unconditional ones.

Let us also denote for each lag $k \in [\![-d,-1]\!]$ and subset $S \subseteq [\![-d,-1]\!] \setminus \{k\}$,

$$\nu_{k,S}(x_S) = \sum_{b \in A} \sum_{c \in A} w_{k,S}(b,c,x_S) d_{k,S}(b,c,x_S), \tag{24}$$

and

$$\bar{\nu}_{k,S} = \mathbb{E}\left(\nu_{k,S}(X_S)\right). \tag{25}$$

The quantity $\nu_{k,S}(x_S)$ measures the influence of X_k on X_0, conditionally on the variables $X_S = x_S$. The average conditional influence of X_k on X_0 is measured through the quantity $\bar{\nu}_{k,S}$.

In the sequel, we write $\mathrm{Cov}_{x_S}(X_0, X_k)$ to denote the conditional covariance between the random variables X_0 and X_k given that $X_S = x_S$. Here, we also use the convention that the conditional covariance for $S = \emptyset$ corresponds to the unconditional one. With this notation, we can prove the following structural result about MTD models.

Proposition 6 *For any lag $k \in [\![-d,-1]\!]$ and subset $S \subseteq [\![-d,-1]\!] \setminus \{k\}$,*

$$Diam(A)\|A\|_\infty \bar{\nu}_{k,S} \geq \mathbb{E}\left(|Cov_{X_S}(X_0, X_k)|\right). \tag{26}$$

Moreover, if Assumptions 1 and 3 hold, then there exists a constant $\kappa > 0$ such that the following property holds: for any $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \nsubseteq S$ there exists $k \in \Lambda \setminus S$ satisfying

$$\mathbb{E}\left(|Cov_{X_S}(X_0, X_k)|\right) \geq \kappa. \tag{27}$$

Furthermore, if Assumption 3 is replaced by Assumption 4, then the constant κ satisfies

$$\kappa \geq \frac{p_{min}^2 \Gamma_1 \min\{|b-c|^2 : b \neq c\} \tilde{\delta}_{min}}{2\sqrt{|\Lambda|}}, \tag{28}$$

where $\tilde{\delta}_{min}$, p_{min} and Γ_1 are defined respectively in (6), (8) *and* (9).

The proof of Proposition 6 is given in A.2.

Remark 7 *Denote $f(S) = \max_{k \in S^c} \bar{\nu}_{k,S}$, for each $S \subseteq [\![-d,-1]\!]$. On one hand, we have that $f(S) = 0$ for any $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \subseteq S$. This follows immediately from the definition of Λ. On the other hand, Proposition 6 assures that $f(S) \geq \kappa / Diam(A)\|A\|_\infty > 0$ for any $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \nsubseteq S$. Putting together these facts, we deduce that the set of relevant lags can be written as $\Lambda = \arg\min\{|S| : f(S) = 0\}$. This observation motivates the **FSC** estimator defined below.*

In what follows, we split the data $X_{1:n}$ into two pieces. The first part is composed of the first m symbols $X_{1:m}$ where $1 \leq m < n$, whereas the second part is composed of the $n-m$ last symbols $X_{m+1:n}$. In the sequel, we write $\hat{\nu}_{m,k,S}$ to denote the empirical estimate of $\bar{\nu}_{k,S}$ computed from $X_{1:m}$. The formal definition of $\hat{\nu}_{m,k,S}$ involves extra notation and is postponed to Appendix A.

The FSC estimator is built in two steps. The first step is called Forward Stepwise (FS) and the second one is called CUT. In the FS step, we start with $S = \emptyset$ and add iteratively to the set S a lag $j \in \arg\max_{k \in S^c} \hat{\nu}_{m,k,S}$, as long as $|S| < \ell$, where $0 \leq \ell \leq d$ is a parameter of the estimator. We denote $\hat{S}_m$ the set obtained at the end of FS step, with the convention that $\hat{S}_m = \emptyset$ if the parameter $\ell = 0$. As we will see, if ℓ is properly chosen the candidate set $\hat{S}_m$ will contain the set of relevant lags Λ with high probability. It may, of course, include irrelevant lags j (those with $\delta_j = 0$). In the CUT step, for each $j \in \hat{S}_m$, we remove j from $\hat{S}_m$ unless $d_{TV}(\hat{p}_{m,n}(\cdot|x_{\hat{S}_m}), \hat{p}_{m,n}(\cdot|y_{\hat{S}_m})) \geq t_{m,n}(x_{\hat{S}_m}, y_{\hat{S}_m}) := s_{m,n}(x_{\hat{S}_m}) + s_{m,n}(y_{\hat{S}_m})$ for some $(\hat{S}_m \setminus \{j\})$-compatible pasts $x_{\hat{S}_m}, y_{\hat{S}_m} \in A^{\hat{S}_m}$, where $s_{m,n}(x_{\hat{S}_m})$ is given by (12) replacing $\bar{N}_n(\cdot)$ and $\hat{p}_n(\cdot|\cdot)$ by $\bar{N}_{m,n}(\cdot)$ and $\hat{p}_{m,n}(\cdot|\cdot)$ respectively. The FSC estimator is defined as the set $\hat{\Lambda}_{2,n}$ of all lags not removed in the CUT step. The pseudo-code of the algorithm to compute the FSC estimator is given in Algorithm 1.

Algorithm 1: FSC$(X_1, \ldots, X_n)$

```
FS Step;
1. $\hat{S}_m = \emptyset$;
2. While $|\hat{S}_m| < \ell$;
3. Compute $j_* = \arg\max_{j \in \hat{S}_m^c} \hat{\nu}_{m,j,\hat{S}_m}$ and include $j_*$ in $\hat{S}_m$;
CUT step;
6. For each $j \in \hat{S}_m$, remove $j$ from $\hat{S}_m$ unless

       $d_{TV}(\hat{p}_{m,n}(\cdot|x_{\hat{S}_m}), \hat{p}_{m,n}(\cdot|y_{\hat{S}_m})) \geq t_{m,n}(x_{\hat{S}_m}, y_{\hat{S}_m})$,

   for some $(\hat{S}_m \setminus \{j\})$-compatible pasts $x_{\hat{S}_m}, y_{\hat{S}_m} \in A^{\hat{S}_m}$ ;
7. Output $\hat{S}_m$;
```

Remark 8 *(a) It is worth mentioning the following alternative algorithm (henceforth called Algorithm 2) to estimate the set of relevant lags Λ. As Algorithm 1, Algorithm 2 has two steps as well. In the first step, we start with $S = \emptyset$ and add iteratively a lag $j \in \arg\max_{k \in S^c} \hat{\nu}_{n,k,S}$ as long as $\hat{\nu}_{n,j,S} > \tau$, where τ is a parameter of the algorithm and $\hat{\nu}_{n,j,S}$ is the empirical estimate of $\bar{\nu}_{j,S}$ computed from the entire data $X_{1:n}$. Let $\hat{S}_n$ denote the set obtained at the end of this step. Next, in the second step, for each $j \in \hat{S}_n$, we remove j from $\hat{S}_n$ unless $\hat{\nu}_{n,j,\hat{S}_n \setminus \{j\}} \geq \tau$. The output of Algorithm 2 is the set of all lags in $\hat{S}_n$ which were not removed in the second step. Algorithm 2 can be seen as a version adapted for our setting of the LearnNbhd algorithm, proposed in Bresler (2015), to estimate the interaction graph underlying an Ising model from i.i.d samples of the model.*

(b) *As opposed to Algorithm 2, notice that the data $X_{1:n}$ is split into two parts in Algorithm 1. The first m symbols $X_{1:m}$ of the sample are used in the* `FS` *step, whereas the last $n-m$ symbols $X_{m+1:n}$ are only used in the* `CUT` *step. Despite requiring to split the data into two parts, one nice feature of Algorithm 1 is that even if a large ℓ is chosen the* `CUT` *step would remove the non-relevant lags, whereas in Algorithm 2, we have to calibrate τ carefully to recover the relevant lags.*

(c) *(Computation of FSC estimator) As we show in the Appendix A.6, we need to perform at most $O(|A|^3\ell(m-d)(d-(\ell-1)/2)+|A|^2(n-m-d)\ell)$ computations to determine the FSC estimator. The first summand in the sum corresponds to the algorithmic complexity of the FS step, whereas that the second summand can be interpreted as the algorithmic complexity of the PCP estimator computed from a sample of size $n-m$ and whose set S has ℓ elements (recall item (c) of Remark 2).*

In what follows, for any $\xi > 0$ and $0 \le \ell \le d$, let us define the following event,

$$G_m(\ell,\xi) = \bigcap_{S\in\mathcal{S}_{d,\ell}} \left\{ \max_{j\in S^c} |\bar{\nu}_{j,S} - \hat{\nu}_{m,j,S}| \le \xi \right\}, \tag{29}$$

where $\mathcal{S}_{d,\ell} = \{S \subseteq [\![-d,-1]\!] : |S| \le \ell\}$. In the next result we show that whenever the event $G_m(\ell,\xi)$ holds with properly chosen parameters ξ and ℓ, the candidate set $\hat{S}_m$ constructed in the FS step with parameter ℓ contains Λ.

Theorem 9 *Suppose Assumptions 1 and 3 hold and let κ be the lower bound provided by Proposition 6. Let*

$$\xi_* = \frac{\kappa}{4\|A\|_\infty Diam(A)} \text{ and } \ell_* = \left\lfloor \frac{\log_2(|A|)}{8\xi_*^2} \right\rfloor = \left\lfloor \frac{2(Diam(A)\|A\|_\infty)^2\log_2(|A|)}{\kappa^2} \right\rfloor. \tag{30}$$

Let $\hat{S}_m$ denote the candidate set constructed in the FS step of Algorithm 1 with parameter ℓ_*. On the event $G_m(\ell_*,\xi_*)$, we have that $\Lambda \subseteq \hat{S}_m$.

The proof of Theorem 9 is given in Appendix A.2.2. Theorem 9 ensures that the candidate set $\hat{S}_m$ contains the set of relevant lags Λ whenever the event $G_m(\ell_*,\xi_*)$ holds. In this case, we can think of the `CUT` step as the `PCP` estimator discussed in the previous section applied to the $n-m$ last observations $X_{m+1:n}$ of the data, where $S = \hat{S}_m$. The main difference is that $\hat{S}_m$ is a random set, depending on the first m observations $X_{1:m}$ of the data.

In the sequel, let us denote

$$\mathbf{P}_{\mathcal{S}_{d,\ell}} = \min_{S\in\mathcal{S}_{d,\ell}} \mathbf{P}_S, \tag{31}$$

where $\mathbf{P}_S$ is defined in (15).

In the next result we estimate the error probability of the `FSC` estimator.

Theorem 10 *Suppose Assumptions 1, 2, and 3 hold. Let $\Delta > 0$ be the quantity defined in Assumption 2. Denote $\hat{\Lambda}_{2,n}$ the* `FSC` *estimator constructed by Algorithm 1 with parameter*

ℓ_*, as defined in (30). *Suppose also that $m > d \geq 2\ell_*$. Then there exits a constant $C = C(\varepsilon, \mu) > 0$ such that if*

$$n \geq m + d + \frac{C|A|\alpha}{\delta_{min}^2 \mathbf{P}_{S_{d,\ell_*}}}, \tag{32}$$

where δ_{min} and $\mathbf{P}_{S_{d,\ell_*}}$ are defined in (6) *and* (31)*, then we have that,*

$$\begin{aligned}\mathbb{P}(\hat{\Lambda}_{2,n} \neq \Lambda) \leq 2|A|(\ell_* + 1)\binom{d}{\ell_*}\Bigg[d|A|^{\ell_*+1}\exp\left\{-\frac{(\xi_*\Delta)^2(m-d)^2}{18m|A|^{2(\ell_*+2)}(\ell_*+2)^2}\right\} \\ +3(|A|-1)|\Lambda|\exp\left\{-\frac{\Delta^2(n-m-d)^2\mathbf{P}_{S_{d,\ell_*}}^2}{8(n-m)(\ell_*+1)^2}\right\}\Bigg] \\ +\, 8|A|\left((\ell_* - |\Lambda|)(n-m-d) + |\Lambda|\right)\left\lceil\frac{\log(\mu(n-m-d)/\alpha+2)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha},\end{aligned} \tag{33}$$

where ξ_* is defined in (30).

The proof of Theorem 10 is given in Appendix A.2.3.

Remark 11 *(a) Let us give some intuition about the three terms appearing on the right-hand side of* (33)*. The first one is an upper bound for $\mathbb{P}(G_m^c(\ell_*, \xi_*))$. The other two are related to the terms appearing in* (17)*. Indeed, by recalling that $|\hat{S}_m| = \ell_*$, one immediately sees that the third terms of* (33) *corresponds to the first term of* (17) *with $\hat{S}_m$ and $n-m$ in the place of S and n respectively. Besides, the second term of* (33) *is similar (modulo a factor which depends on d and ℓ_*) to the second term of* (17)*. This extra factor reflects the fact that we do not know a priori a set S containing the set of relevant lags Λ.*

(b) Similar to Remark 3, one can also show that $\mathbf{P}_{S_{d,\ell}} \geq p_{min}/|A|^{\ell-1}$, where p_{min} is defined in (8)*. Hence, we can deduce from* (32) *that when the sample size n satisfies*

$$n \geq d + \frac{C|A|^{\ell_*}\alpha}{p_{min}\delta_{min}^2}, \tag{34}$$

then inequality (33) *holds with the second exponential term replaced by*

$$\exp\left\{-\frac{(\Delta p_{min})^2(n-d)^2}{8n(\ell_*+1)^2|A|^{2(\ell_*-1)}}\right\}. \tag{35}$$

The next result is a corollary of Theorem 10.

Corollary 12 *For each n, consider a MTD model with set of relevant lags Λ_n and transition probabilities $p_n(a|x_{\Lambda_n})$ satisfying $p_{min,n} \geq p_{min}^\star$ $\Delta_n \geq \Delta_{min}^\star$ for some positive constants $p_{min}^\star$ and $\Delta_{min}^\star$, and such that Assumption 3 holds. Let $m_n = n/2$ and $d_n = m_n\beta$ with $\beta \in (0,1)$. Let $X_{1:n}$ be a sample from the MTD specified by Λ_n and $p_n(a|x_{\Lambda_n})$, and denote $\hat{\Lambda}_{2,n}$ the* ***FSC*** *estimator constructed by Algorithm 1 with parameters m_n, $\mu_n = \mu \in (0,3)$ such that $\mu > \psi(\mu)$, $\varepsilon_n = \varepsilon > 0$, $\alpha_n = (1+\eta)\log(n)$ with $\eta > 0$ and $\ell_{*,n}$ as defined in* (30).

Assume that $\ell_{*,n} \leq ((1-\gamma)/2)\log_{|A|}(n)$ for some $\gamma \in (0,1)$. Then there exists a constant $C = C(\beta,\gamma,\eta,p^{\star}_{min},\Delta^{\star}_{min},\varepsilon,\mu) > 0$ such that $\mathbb{P}(\hat{\Lambda}_{2,n} \neq \Lambda^*) \to 0$ as $n \to \infty$, whenever

$$\delta^2_{min,n} \geq \frac{C\log(n)}{n^{(1+\gamma)/2}}. \tag{36}$$

The proof of Corollary 12 is given in Appendix A.2.4.

Remark 13 *(a) Under Assumption 4, one can check that $\ell_{*,n} \leq ((1-\gamma)/2)\log_{|A|}(n)$ whenever,*

$$\frac{\tilde{\delta}^2_{min}}{|\Lambda|} \geq \frac{16(1-\gamma)}{\Gamma_1^2(p^{\star}_{min}\min\{|b-c| : b \neq c\})^4\log_{|A|}(n)}.$$

(b) Comparing Corollaries 12 and 4, we observe that the consistency of both `FSC` and `PCP` estimators require the same lower bound on the decay of minimal oscillation $\delta_{min,n}$. Despite requiring additional assumptions (Assumption 3 and a condition on the growth of ℓ_*), `FSC` estimator do not need prior knowledge of a small subset S containing the set of relevant lags Λ as opposed to the `PCP` estimator, which is a significant advantage in practice.

(c) Let us mention that under the assumptions of Corollary 12, we have that the algorithmic complexity of the FSC is $O(|A|^3n^2\log_{|A|}(n))$. This follows immediately from item (c) of Remark 8.

3.3 Improving the efficiency for the binary case

In this section, we show that when the alphabet is binary, i.e., $A = \{0,1\}$, we can further improve the FSC algorithm if we consider Assumptions 4 and 5. Observe that when the alphabet is binary, Assumption 3 holds automatically (see Section 2.6). Moreover, we have that

$$\begin{aligned}\bar{\nu}_{k,S} &= 2\mathbb{E}\left(\mathbb{P}_{X_S}(X_k=1)\mathbb{P}_{X_S}(X_k=1)\left|\mathbb{P}_{X_S}(X_0=1|X_k=1) - \mathbb{P}_{X_S}(X_0=1|X_k=0)\right|\right)\\ &= 2\mathbb{E}\left(|\mathrm{Cov}_{X_S}(X_0,X_k)|\right),\end{aligned}$$

for any lag $k \in [\![-d,-1]\!]$, subset $S \subseteq [\![-d,-1]\!] \setminus \{k\}$ and configuration $x_S \in \{0,1\}^S$.

For a binary MTD, we have the following result.

Theorem 14 *Under Assumptions 4 and 5, it holds that*

$$\min_{S\subset\Lambda}\left(\max_{j\in\Lambda\setminus S}\bar{\nu}_{j,S} - \max_{j\in(\Lambda)^c}\bar{\nu}_{j,S}\right) \geq 2(\Gamma_1-\Gamma_2)p^2_{min}\delta_{min},$$

where δ_{min} and p_{min} are defined in (6) and (8) respectively. In particular, if $\Gamma_1 > \Gamma_2$ and $\hat{S}_m$ denotes the candidate set constructed at the end of the FS step of Algorithm 1 with parameter $\ell \geq |\Lambda|$, then $\Lambda \subseteq \hat{S}_m$ whenever the event $G_m(\ell,\xi)$ holds where

$$0 < \xi < (\Gamma_1-\Gamma_2)p^2_{min}\delta_{min}. \tag{37}$$

The proof of Theorem 14 is given in Appendix A.3.1.

Remark 15 *Notice that if the size of Λ is known, then Theorem 14 implies $\hat{S}_m = \Lambda$ on the event $G_m(|\Lambda|, \xi)$ with ξ satisfying (37). In particular, in this case, we neither need to execute the* **CUT** *step nor to split the data into two pieces.*

In the same spirit of the previous corollaries, we can show the following result.

Corollary 16 *For each n, consider a MTD model with set of relevant lags Λ_n and transition probabilities $p_n(a|x_{\Lambda_n})$ satisfying Assumptions 4 and 5 with $\Gamma_{1,n} = \Gamma_1 > \Gamma_2 = \Gamma_{2,n}$ and such that $|\Lambda_n| \leq L$ for some integer L, $p_{min,n} \geq p^{\star}_{min}$ and $\Delta_n \geq \Delta^{\star}_{min}$ for some positive constants $p^{\star}_{min}$ and $\Delta^{\star}_{min}$. Let $X_{1:n}$ be a sample from the MTD specified by Λ_n and $p_n(a|x_{\Lambda_n})$, and denote $\hat{\Lambda}_{2,n}$ the* **FSC** *estimator with parameters with parameters $m_n = n/2$, $\mu_n = \mu \in (0,3)$ such that $\mu > \psi(\mu)$, $\varepsilon_n = \varepsilon > 0$, $\alpha_n = (1+\eta)\log(n)$ with $\eta > 0$ and $\ell = L$. Suppose that $d_n = \beta n$ with $\beta \in (0,1)$. Then there exists a constant $C = C(\beta, L, \Delta^{*}_{min}, p^{*}_{min}, \Gamma_1, \Gamma_2, \eta, \mu, \varepsilon) > 0$ such that $\mathbb{P}(\hat{\Lambda}_{2,n} \neq \Lambda^{*}) \to 0$ as $n \to \infty$, as long as*

$$\delta_{min,n} \geq C\sqrt{\frac{\log(n)}{n}}, \tag{38}$$

The proof of Corollary 16 is given in Appendix A.3.2.

3.4 Post-selection transition probabilities estimation

Once the set of relevant lags have been estimated by applying the FSC estimator to the sample $X_{1:n}$, we reuse the entire sample to compute the estimator $\hat{p}_n(a|x_{\hat{\Lambda}_{2,n}})$ of the transition probability $p(a|x_{\Lambda})$. In the next result, we provide an estimate for rate of convergence of $\hat{p}_n(a|x_{\hat{\Lambda}_{2,n}})$ towards $p(a|x_{\Lambda})$, simultaneously for all pasts $x_{-d:-1} \in A^{[\![-d,-1]\!]}$.

Theorem 17 *Under assumptions and notation of Theorem 10,*

$$\mathbb{P}\left(\bigcup_{a\in A}\bigcup_{x_{-d:-1}\in A^{[\![-d,-1]\!]}}\left\{|\hat{p}_n(a|x_{\hat{\Lambda}_{2,n}}) - p(a|x_{\Lambda})| \geq \sqrt{\frac{2\alpha(1+\epsilon)\hat{V}_n(a,x_{\hat{\Lambda}_{2,n}})}{\bar{N}_n(x_{\hat{\Lambda}_{2,n}})}} + \frac{\alpha}{3\bar{N}_n(x_{\hat{\Lambda}_{2,n}})}\right\}\right) \leq 4|A|(n-d)\left\lceil\frac{\log(\mu(n-m-d)/\alpha+2)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha} + \mathbb{P}(\hat{\Lambda}_{2,n} \neq \Lambda), \tag{39}$$

where $\hat{V}_n(a, x_{\hat{\Lambda}_{2,n}})$ is given by

$$\hat{V}_n(a,x_{\hat{\Lambda}_{2,n}}) = \frac{\mu}{\mu-\psi(\mu)}\hat{p}_n(a|x_{\hat{\Lambda}_{2,n}}) + \frac{\alpha}{\mu-\psi(\mu)}\frac{1}{\bar{N}_n(x_{\hat{\Lambda}_{2,n}})}.$$

The proof of Theorem 17 is given in Appendix A.4.

3.5 A remark on the minimax rate for the lag selection

We take $A = \{0,1\}$ and consider the set of $\{p^{(j)}(\cdot|\cdot), j \in [\![-d,-1]\!]\}$ of transition probabilities of the following form:

$$p^{(j)}(1|x_{-d:-1}) = \frac{(1-\lambda)}{2} + \lambda p(1|x_j),\ j \in [\![-d,-1]\!], \lambda \in (0,1), \tag{40}$$

where $\lambda|p(1|1) - p(1|0)| := \delta > 0$. For each $j \in [\![-d, -1]\!]$, we denote $\mathbb{P}^{(j)}$ the probability measure under which $(X_t)_{t\in\mathbb{Z}}$ is a stationary MTD model having transition probability $p^{(j)}(\cdot|\cdot)$. For each $t \geq 1$, we denote $P_t^{(j)}$ the marginal distribution with respect to the variables $X_{1:t}$:

$$P_t^{(j)}(x_{1:t}) = \mathbb{P}^{(j)}(X_{1:t} = x_{1:t})$$

In what follows, $KL(P_t^{(j)}||P_t^{(k)})$ denotes the *Kullback-Leibler* divergence between the distributions $P_t^{(j)}$ and $P_t^{(k)}$. We denote $MTD_{d,\delta}$ the set all transition probabilities $p = \{p(a|x_\Lambda) : a \in A, x_\Lambda \in A^\Lambda\}$ of a MTD model of order d whose corresponding $\delta_{min} \geq \delta$. For a given $p \in MTD_{d,\delta}$, we denote $\mathbb{P}_p$ the probability distribution under which $(X_t)_{t\in\mathbb{Z}}$ is a stationary MTD model of order d with transition probabilities given by p. With this notation, we have the following result.

Proposition 18 *Let $n > d$. Then the following inequality holds: for $j, k \in [\![-d, -1]\!]$,*

$$KL(P_n^{(j)}||P_n^{(k)}) \leq \frac{2n\delta^2}{1-\lambda}. \tag{41}$$

In particular, if $\beta \in (0, 1)$, $d = n\beta$, and

$$\delta^2 \leq \frac{(1-\lambda)}{2n}\left(\frac{\log(n\beta)}{2} - \log(2)\right), \tag{42}$$

then

$$\inf_{\hat{\Lambda}_n} \sup_{p\in MTD_{d,\delta}} \mathbb{P}_p(\hat{\Lambda}_n \neq \Lambda) \geq 1/4, \tag{43}$$

where the infimum is over all lag estimators $\hat{\Lambda}_n$ based on a sample of size n.

The proof of (43) follows immediately from Fano's inequality and the upper bound (41). Combining (38) and (42), we deduce that the condition on the minimal oscillation required for the consistency of the FSC estimator in Corollary 16 is sharp. The proof of Proposition 18 is given in Appendix A.5

4. Simulations

Here, we investigated the performance of the proposed methods using simulations.

4.1 Experiment 1

We first used a MTD model on alphabet $A = \{0, 1\}$ with two relevant lags, denoted here as $-i$ and $-j$ for notational convenience. The choices for the order d and for the values of i and j are shown in the first three columns of Table 1. Let $p_0(1) = p_0(0) = 0.5$ and $\lambda_0 = 0.4$. Also, let $\lambda_{-i} = 0.2$, $\lambda_{-j} = 0.4$, $p_{-i}(0|0) = 0.3$, $p_{-i}(0|1) = 0.6$, $p_{-j}(0|0) = 0.5$, and $p_{-j}(0|1) = 0.9$. For all $x_{-d:-1} \in \{0, 1\}^{[\![-d,-1]\!]}$ and $a \in A$, the transition probability of the model was given by

$$p(a|x_{-d:-1}) = \lambda_0 p_0(a) + \lambda_{-i}p_{-i}(a|x_{-i}) + \lambda_{-j}p_{-j}(a|x_{-j}).$$

We simulated the above model using sample sizes

$$n \in \{10^2, 5.10^2, 10^3, 5.10^3, 10^4, 5.10^4, 10^5, 5.10^5\}.$$

For each choice of i, j, d, and n we simulated 100 realizations. We compared four different methods to select the relevant lags. FSC(ℓ) stands for the Forward Stepwise and Cut algorithm described in Algorithm 1 with parameter ℓ, $\epsilon = 0.1$, $\mu = 0.5$, and $\alpha = C\log(n)$, where the values of the constant C was chosen by optimizing the probability to select the relevant lags correctly only for sample size $n = 100$, for the given choice of d, i and j. We used the first $n/2$ samples for the Forward Stepwise and the last $n/2$ for Cut. Remember that ϵ, μ, α are used to define the random threshold for the Cut step. BSS(2) stands for the best subset selection algorithm, where we first estimated the parameters of the MTD model using n samples and the algorithm described in Berchtold (2001) with python implementation mtd-learn. This algorithm estimated for $k \in \{1, 2, \ldots, d\}$ the weight parameters λ_{-k}. We then choose the lags of the two largest λ_{-k} as the lags selected by BSS(2). We were not able to run the mtd-learn on models with order d larger than 15 in our computers because that algorithm did not converge. Finally, CTF(ℓ) stands for Conditional Tensor Factorization based Higher Order Markov Chain estimation together with the test for relevance of lags described in Sarkar and Dunson (2016), the parameter ℓ being the maximal number of relevant lags. We used the code available at https://github.com/david-dunson/bnphomc. The maximal possible order of the Markov chain was set to d and the number of simulation for the Gibbs sampler was set to 1000. The set of relevant lags chosen by CTF was given by the lags with non-null inclusion probability estimated using the Gibbs sampler. We were not able to run CTF(ℓ) when $j = n/5$ and $d = n/4$ because the algorithm did not converge when $n > 10^3$. We note that FS and BSS assume prior knowledge of the number of relevant sites, giving advantage over FSC and CTF. The results are indicated in Table 1.

Table 1: Estimated probability of correctly selecting only the relevant lags.

Model parameter			Method	Sample size (n)							
i	j	d		**100**	**500**	$\mathbf{10^3}$	$\mathbf{5.10^3}$	$\mathbf{10^4}$	$\mathbf{5.10^4}$	$\mathbf{10^5}$	$\mathbf{5.10^5}$
1	8	8	FSC(3)	0.05	0.08	0.13	0.53	0.81	0.86	0.93	1
1	8	8	CTF(3)	0	0	0.04	0.67	0.99	1	1	1
1	8	8	FS(2)	0.07	0.3	0.47	0.98	1	1	1	1
1	8	8	BSS(2)	0.05	0.14	0.23	0.41	0.79	0.78	0.84	0.87
1	15	15	FSC(5)	0.03	0.36	0.51	0.82	0.97	1	1	1
1	15	15	CTF(5)	0	0	0.01	0.62	0.99	1	1	1
1	15	15	FS(2)	0.02	0.2	0.66	0.92	1	1	1	1
1	15	15	BSS(2)	0.04	0.18	0.17	0.28	0.31	0.8	0.8	0.93
1	n/5	n/4	FSC(5)	0	0	0.04	0.19	0.46	1	1	1
1	n/5	n/4	CTF(5)	0	0	0	-	-	-	-	-
1	n/5	n/4	FS(2)	0.01	0.11	0.27	0.89	0.96	1	1	1
1	n/5	n/4	BSS(2)	-	-	-	-	-	-	-	-

4.2 Experiment 2

Here we used the following MTD model on alphabet $A = \{0, 1\}$. We considered different choices of order d and relevant lags $-i, -j \in [\![1, d]\!]$ (see Table 2). Let $p_0(1) = p_0(0) = 0.5$ and $\lambda_0 = 0.2$. Also, let $p_{-i}(0|0) = 1 - p_{-i}(0|1) = p_{-j}(0|1) = 1 - p_{-j}(0|1) = 0.7$ and $\lambda_{-i} = \lambda_{-j} = 0.4$. For all $x_{-d:-1} \in \{0, 1\}^{[\![-d,-1]\!]}$ and $a \in A$, the transition probability of the model was given by

$$p(a|x_{-d:-1}) = \lambda_0 p_0(a) + \lambda_{-i} p_{-i}(a|x_{-i}) + \lambda_{-j} p_{-j}(a|x_{-j}).$$

We simulated the above model using sample sizes $n \in \{2^8, 2^9, 2^{10}, 2^{11}, 2^{12}, 2^{13}\}$. For each choice of i, j, d, and n we simulated 100 realizations. For each realization, we estimated the transition probability $p(0|0^d)$. We used different estimators for the comparisons. `FSC`(ℓ) and `FS`(ℓ) are the same as described in Experiment 1. For transition probability estimation with `FSC`, we used $X_{1:n/2}$ for Forward Stepwise and $X_{n/2+1:n}$ for Cut step, obtaining the estimated relevant lag set $\hat{\Lambda}_n$. Then we used $X_{1:n}$ to calculate $\hat{p}_n(0|0_{\hat{\Lambda}_n})$. For transition probability estimation after `PCP`, we used $X_{1:n}$ to calculate $\hat{\Lambda}_n$ for the `PCP` relevant lag estimator with initial set $S = [\![-d, -1]\!]$. The parameters for the threshold were chosen as follows: $\epsilon = 0.1$, $\mu = 0.5$, and $\alpha = C \log(n)$, where we choose the values of the constant C by optimizing the probability to select the relevant lags correctly only for sample size $n = 100$, for the given choice of d, i and j. Then we used $X_{1:n}$ to calculate $\hat{p}_n(0|0_{\hat{\Lambda}_n})$. We also compared the performance of transition probability estimator $\hat{p}_n(0|0_{-d:-1})$, where we did not select the relevant lags (`Naive` estimator). In our simulations, when d was larger than 5, for both `PCP` and `Naive` estimators we did not obtain meaningful results because $\bar{N}_n(0^d) = 0$ with high probability. Therefore, we compared `PCP` and `Naive` estimators only for $d = 5$. In this case, `FSC` showed similar performance to `PCP` estimator and was in general better than `Naive` estimator. When $d > 5$, *e.g.* $d = n/8$, `FSC` still exhibited good performance. The results are indicated in Table 2.

Table 2: Empirical standard deviation of the estimator of $p(0|0^d)$. `FSC`, `FS`, `PCP`, and `Naive` are described in the main text.

Model parameter			Method	Sample size (n)					
i	*j*	*d*		**256**	**512**	**1024**	**2048**	**4096**	**8192**
1	5	5	`FS(2)`	0.0774	0.0682	0.0506	0.0286	0.0174	0.0133
1	5	5	`FSC(5)`	0.0745	0.0835	0.0602	0.0426	0.0222	0.0129
1	5	5	`PCP`	0.0965	0.0786	0.0577	0.0432	0.0242	0.0131
1	5	5	`Naive`	0.1518	0.0933	0.0624	0.0455	0.0340	0.0252
1	5	10	`FSC(5)`	0.0836	0.0842	0.0659	0.0425	0.0228	0.0141
1	10	15	`FSC(5)`	0.0864	0.0781	0.0641	0.0438	0.0249	0.0151
1	15	15	`FSC(5)`	0.0833	0.0834	0.0747	0.0488	0.0222	0.0167
11	100	120	`FSC(5)`	-	-	0.0838	0.0647	0.0312	0.0169
1	10	n/8	`FSC(5)`	0.0563	0.0543	0.0780	0.0698	0.0504	0.0105

4.3 Application

We applied the proposed method to study the relevant lags on a daily weather data registering the rainy and non-rainy days in Canberra Australia for a $n = 1000$ days. We obtained the data from kaggle (`https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package`). We used the first 1000 time points of the data. We used Forward Stepwise algorithm with $\ell = 3$ (`FS`(3)) and maximal order $d = 400$ to include the possibility of the annual cycle. We obtained as the three relevant lags $\{1, 62, 330\}$. The selected relevant lags were the same for $d = 365$ and $d = 450$, showing teh robustness of the result. The day before (lag 1) is clearly relevant and is often included in weather prediction models. Annual cycles (≈ 12 months) are also predictor of the weather, matching the 330 days lag that we found. Finally, the 62 days lag is consistent with the cycle of Madden-Julian oscillator (≈ 60 days), which is the largest inter-seasonal source of precipitation events in Australia (Wheeler et al. (2009)). We note that the estimated Markov chain is of order 330, which is around one-third of the sample size $n = 1000$, whereas using VLMC we do not expect to typically estimate Markov chains of order larger than $\log(10) \approx 7$. Indeed, using VLMC with BIC model selection criterion we selected a model with order 1. We set the upper limit of the model size as 400 for the VLMC order selection. As a further comparison, we applied the Conditional Tensor Factorization based Higher Order Markov Chain estimation together with the test for relevance of lags described in Sarkar and Dunson (2016). We again used the code available at `https://github.com/david-dunson/bnphomc`. The maximal possible order of the Markov chain was set to 400, the maximal number of relevant lags was 3, and the number of simulation for the Gibbs sampler was set to 1000. The inclusion probability calculated using Gibbs sampler for lags $(1, 2, 3, 4, 5, 6, 7)$ were $(100, 1.2, 0.2, 0.6, 0.2, 0.4, 0.2)$ percent, respectively. For all other orders the inclusion probability was zero. Therefore, no larger lags were detected by this method.

Acknowledgments

This research has been conducted as part of FAPESP project *Research, Innovation and Dissemination Center for Neuromathematics* (grant 2013/07699-0). GO thanks FAPERJ (grants E-26/201.397/2021 and E-26/211.343/2019) and CNPq (grant 303166/2022-3).

Appendix A. Proofs of Section 3

A.1 Proofs of Section 3.1

A.1.1 Proof of Theorem 1

Proof [Proof of Theorem 1] Since the set $S \subseteq [\![-d, -1]\!]$ containing the set Λ is fixed, we will write x instead of x_S to alleviate the notation. We start proving Item 1.

Proof of Item 1. For each $x \in A^S$, let us define the event

$$G_x = \bigcap_{a \in A} \left\{ |\hat{p}_n(a|x) - p(a|x)| < \sqrt{\frac{2\alpha(1+\varepsilon)\hat{V}_n(a,x)}{\bar{N}_n(x)}} + \frac{\alpha}{3\bar{N}_n(x)} \right\},$$

where $\hat{V}_n(a,x)$ is given by

$$\hat{V}_n(a,x) = \frac{\mu}{\mu-\psi(\mu)}\hat{p}_n(a|x) + \frac{\alpha}{(\mu-\psi(\mu))\bar{N}_n(x)}.$$

By using first the union bound and then by applying Proposition 34, we deduce that for each $x \in A^S$,

$$\mathbb{P}(G_x^c) \leq 4|A|\left\lceil \frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}\mathbb{P}\left(\bar{N}_n(x) > 0\right). \tag{44}$$

Now, observe that on G_x, we have that

$$d_{TV}(\hat{p}_n(\cdot|x), p(\cdot|x)) < \sum_{a\in A}\sqrt{\frac{\alpha(1+\varepsilon)\hat{V}_n(a,x)}{2\bar{N}_n(x)}} + \frac{\alpha|A|}{6\bar{N}_n(x)} = s_n(x),$$

which, together with (44), implies that

$$\mathbb{P}\left(d_{TV}(\hat{p}_n(\cdot|x), p(\cdot|x)) \geq s_n(x)\right) \leq 4|A|\left\lceil \frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}\mathbb{P}\left(\bar{N}_n(x) > 0\right). \tag{45}$$

Note that if $j \notin \Lambda$, then by the definition of the set Λ it follows that $p(a|x) = p(a|y)$ for all $x, y \in A^S$ which are $(S\backslash\{j\})$-compatible. Hence, by applying first the triangle inequality and then using that $t_n(x,y) = s_n(x) + s_n(y)$, we deduce that the event $\{d_{TV}(\hat{p}_n(\cdot|x), \hat{p}_n(\cdot|y)) \geq t_n(x,y)\}$ is contained in the event

$$\{d_{TV}(\hat{p}_n(\cdot|x), p(\cdot|x)) \geq s_n(x)\} \cup \{d_{TV}(\hat{p}_n(\cdot|y), p(\cdot|y)) \geq s_n(y)\},$$

so that

$$\begin{aligned}\mathbb{P}\left(j \in \hat{\Lambda}_{1,n}\right) &\leq 2\sum_{x\in A^S}\mathbb{P}(d_{TV}(\hat{p}_n(\cdot|x), p_n(\cdot|x)) \geq s_n(x)) \\ &\leq 8|A|\left\lceil \frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}\sum_{x\in A^S}\mathbb{P}(\bar{N}_n(x) > 0),\end{aligned}$$

where in the second inequality we have used (45).

Since $n - d = \sum_{x\in A^S}\bar{N}_n(x) \geq \sum_{x\in A^S} 1\{\bar{N}_n(x) > 0\}$ which implies that $n - d \geq \mathbb{E}\left[\sum_{x\in A^S} 1\{\bar{N}_n(x) > 0\}\right] = \sum_{x\in A^S}\mathbb{P}(\bar{N}_n(x) > 0)$, we obtain from the above inequality that,

$$\mathbb{P}\left(j \in \hat{\Lambda}_{1,n}\right) \leq 8|A|\left\lceil \frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}(n-d),$$

concluding the the proof of Item 1.

Proof of Item 2. Let $j \in \Lambda$, recall that $\delta_j = \lambda_j \max_{b,c\in A} d_{TV}(p_j(\cdot|b), p_j(\cdot|c))$ and consider the event $E = \{\delta_j \geq \gamma_{n,j}\}$. Take $b^\star, c^\star \in A$ such that $d_{TV}(p_j(\cdot|b^\star), p_j(\cdot|c^\star)) = \max_{b,c\in A} d_{TV}(p_j(\cdot|b), p_j(\cdot|c))$, and observe that with this choice,

$$\delta_j = \lambda_j d_{TV}(p_j(\cdot|b^\star), p_j(\cdot|c^\star)).$$

From this equality it follows that for any pair $(x,y)\in\mathcal{C}_j(b^\star,c^\star)$, we have

$$E=\{d_{TV}(p(\cdot|x),p(\cdot|y))\geq 2t_{n,j}\},$$

where we have used also that $\gamma_{n,j}=2t_{n,j}$. Now, take a pair $(x^\star,y^\star)\in\mathcal{C}_j(a^\star,b^\star)$ attaining the minimum in (14):

$$t_{n,j}(b^\star,c^\star)=t_n(x^\star,y^\star).$$

From the definition of $t_{n,j}$, it follows then that

$$t_{n,j}\geq t_{n,j}(b^\star,c^\star)=t_n(x^\star,y^\star).$$

Therefore, we conclude that

$$E\subseteq\{d_{TV}(p(\cdot|x^\star),p(\cdot|y^\star))\geq 2t_n(x^\star,y^\star)\},$$

so that by the triangle inequality, we obtain that on E,

$$\begin{aligned}2t_n(x^\star,y^\star)\leq d_{TV}(\hat{p}_n(\cdot|x^\star),p(\cdot|x^\star))&\\ &+d_{TV}(\hat{p}_n(\cdot|y^\star),p(\cdot|y^\star))+d_{TV}(\hat{p}_n(\cdot|x^\star),\hat{p}_n(\cdot|y^\star)).\end{aligned}$$

Hence, on $\{j\notin\hat{\Lambda}_{1,n}\}\cap E$, we have

$$t_n(x^\star,y^\star)\leq d_{TV}(\hat{p}_n(\cdot|x^\star),p(\cdot|x^\star))+d_{TV}(\hat{p}_n(\cdot|y^\star),p(\cdot|y^\star)),$$

implying that

$$\begin{aligned}\mathbb{P}\left(\{j\notin\hat{\Lambda}_{1,n}\}\cap E\right)\leq\mathbb{P}(d_{TV}(\hat{p}_n(\cdot|x^\star),p(\cdot|x^\star))\geq s_n(x^\star))&\\ &+\mathbb{P}(d_{TV}(\hat{p}_n(\cdot|y^\star),p(\cdot|y^\star))\geq s_n(y^\star)).\end{aligned}$$

From (45), it follows then that

$$\mathbb{P}\left(j\notin\hat{\Lambda}_{1,n},\gamma_{n,j}\leq\delta_j\right)=\mathbb{P}\left(\{j\notin\hat{\Lambda}_{1,n}\}\cap E\right)\leq 8|A|\left\lceil\frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha},$$

concluding the proof of Item 2.

Proof of Item 3. Observe that by combining Items 1 and 2 together with the union bound, we deduce that

$$\begin{aligned}\mathbb{P}\left(\hat{\Lambda}_{1,n}\neq\Lambda\right)\leq 8|A|\left((|S|-|\Lambda|)(n-d)+|\Lambda|\right)\left\lceil\frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha}&\\ &+\sum_{j\in\Lambda}\mathbb{P}\left(\gamma_{n,j}>\delta_j\right).\end{aligned}$$

Hence, to conclude the proof of Item 3, it suffices to show that

$$\mathbb{P}\left(\gamma_{n,j}>\delta_j\right)\leq 6|A|(|A|-1)\exp\left\{\frac{-\Delta^2(n-d)^2\mathbf{P}_S^2}{8n(|S|+1)^2}\right\},\tag{46}$$

for all $j \in \Lambda$, whenever the sample size n satisfies (16).

By the union bound, we have that

$$\mathbb{P}\left(\gamma_{n,j} > \delta_j\right) \le \sum_{b\in A}\sum_{c\in A: c\neq b} \mathbb{P}\left(t_{n,j}(b,c) > \delta_j/2\right), \tag{47}$$

and for each $b, c \in A$ with $b \neq c$,

$$\mathbb{P}\left(t_{n,j}(b,c) > \delta_j/2\right) \le \mathbb{P}\left(t_n(x,y) > \delta_j/2\right) \le \mathbb{P}\left(s_n(x) > \delta_j/4\right) + \mathbb{P}\left(s_n(y) > \delta_j/4\right), \tag{48}$$

for any $(x,y) \in \mathcal{C}_j(b,c)$.

By using again the union bound, we can deduce that for each in $x \in A^S$,

$$\mathbb{P}\left(s_n(x) > \delta_j/4\right) \le \mathbb{P}\left(\sum_{a\in A}\sqrt{\frac{\alpha(1+\varepsilon)\hat{V}_n(a,x)}{2\bar{N}_n(x)}} > \delta_j/8\right) + \mathbb{P}\left(\frac{\alpha}{6\bar{N}_n(x)} > \frac{\delta_j}{8|A|}\right).$$

and also that

$$\mathbb{P}\left(\sum_{a\in A}\sqrt{\frac{\alpha(1+\varepsilon)\hat{V}_n(a,x)}{2\bar{N}_n(x)}} > \delta_j/8\right) \le \mathbb{P}\left(\sum_{a\in A}\sqrt{\frac{\alpha(1+\varepsilon)\hat{p}_n(a|x)\mu}{2(\mu-\psi(\mu))\bar{N}_n(x)}} > \delta_j/16\right) + \mathbb{P}\left(\frac{|A|\alpha}{\bar{N}_n(x)}\sqrt{\frac{(1+\varepsilon)}{2(\mu-\psi(\mu))}} > \delta_j/16\right).$$

By applying Proposition 26 with $u_1 = \mathbf{P}(x) - (4|A|\alpha)/(3\delta_j(n-d))$ and $u_2 = \mathbf{P}(x) - (16|A|\alpha\sqrt{(1+\varepsilon)/2(\mu-\psi(\mu))})/(\delta_j(n-d))$, one can show that

$$\mathbb{P}\left(\frac{|A|\alpha}{6\bar{N}_n(x)} > \delta_j/8\right) + \mathbb{P}\left(\frac{|A|\alpha}{\bar{N}_n(x)}\sqrt{\frac{(1+\varepsilon)}{2(\mu-\psi(\mu))}} > \delta_j/16\right) \le 2\exp\left\{-\frac{\Delta^2(n-d)^2}{2n(|S|+1)^2}\left(\mathbf{P}(x) - \frac{16|A|\alpha}{\delta_j(n-d)}\sqrt{\frac{(1+\varepsilon)}{2(\mu-\psi(\mu))}}\right)^2\right\},$$

as long as $(n-d) > \frac{16|A|\alpha}{\delta_j\mathbf{P}(x)}\sqrt{\frac{(1+\varepsilon)}{2(\mu-\psi(\mu))}}$.

By using Jensen inequality, one can verify that

$$\mathbb{P}\left(\sum_{a\in A}\sqrt{\frac{\alpha(1+\varepsilon)\hat{p}_n(a|x)\mu}{2(\mu-\psi(\mu))\bar{N}_n(x)}} > \delta_j/16\right) \le \mathbb{P}\left(\bar{N}_n(x) < \frac{128\alpha(1+\varepsilon)\mu|A|}{\delta_j^2(\mu-\psi(\mu))}\right),$$

so that by Proposition 26 with $u_3 = \mathbf{P}(x) - (128\alpha(1+\varepsilon)\mu|A|)/(\delta_j^2(\mu-\psi(\mu)))$, it follows that

$$\mathbb{P}\left(\sum_{a\in A}\sqrt{\frac{2\alpha(1+\varepsilon)\hat{p}_n(a|x)\mu}{(\mu-\psi(\mu))\bar{N}_n(x)}} > \delta_j/16\right) \le \le \exp\left\{-\frac{\Delta^2(n-d)^2}{2n(|S|+1)^2}\left(\mathbf{P}(x) - \frac{128|A|\alpha\mu(1+\varepsilon)}{\delta_j^2(\mu-\psi(\mu))(n-d)}\right)^2\right\},$$

whenever $(n-d) > (128|A|\alpha\mu(1+\varepsilon))/(\delta_j^2\mathbf{P}(x)(\mu-\psi(\mu)))$.

Therefore, we have shown that for any $x \in A^S$,

$$\mathbb{P}\left(s_n(x) > \delta_j/4\right) \leq 3\exp\left\{-\frac{\Delta^2(n-d)^2\mathbf{P}^2(x)}{8n(|S|+1)^2}\right\}, \tag{49}$$

as long as

$$(n-d) \geq 2\left(16\frac{|A|\alpha}{\delta_j^2\mathbf{P}(x)}\left(\frac{8\mu(1+\varepsilon)}{(\mu-\psi(\mu))}+\sqrt{\frac{(1+\varepsilon)}{2(\mu-\psi(\mu))}}\right)\right).$$

Now, considering $b^{*,j}, c^{*,j} \in A$ with $b^* \neq c^*$ and $(x^{*,j}, y^{*,j}) \in \mathcal{C}_j(a^{*,j}, b^{*,j})$ such that

$$\min_{b,c\in A: b\neq c}\max_{(x,y)\in\mathcal{C}_j(a,b)}(\mathbf{P}(x)\wedge\mathbf{P}(y)) = (\mathbf{P}(x^{*,j})\wedge\mathbf{P}(y^{*,j})),$$

we can deduce from (48) and (49) that

$$\mathbb{P}\left(t_{n,j}(b,c) > \delta_j/2\right) \leq 6\exp\left\{-\frac{\Delta^2(n-d)^2(\mathbf{P}(x^{*,j})\wedge\mathbf{P}(y^{*,j}))^2}{8n(|S|+1)^2}\right\},$$

whenever

$$(n-d) \geq 2\left(16\frac{|A|\alpha}{\delta_j^2\mathbf{P}(x^{*,j})\wedge\mathbf{P}(y^{*,j})}\left(\frac{8\mu(1+\varepsilon)}{(\mu-\psi(\mu))}+\sqrt{\frac{(1+\varepsilon)}{2(\mu-\psi(\mu))}}\right)\right).$$

Since $\mathbf{P}(x^{*,j})\wedge\mathbf{P}(y^{*,j}) \geq \mathbf{P}_S$ for all $j \in \Lambda^*$, we can take

$$C = C(\mu,\varepsilon) = 32\left(\frac{8\mu(1+\varepsilon)}{(\mu-\psi(\mu))}+\sqrt{\frac{(1+\varepsilon)}{2(\mu-\psi(\mu))}}\right),$$

to deduce that (46) is indeed satisfied whenever

$$(n-d) \geq \frac{C|A|\alpha}{\delta_{min}^2\mathbf{P}_S},$$

concluding the proof. ■

A.1.2 Proof of Corollary 4

Proof [Proof of Corollary 4] Notice that Assumptions 1 and 2 are satisfied for all values of n, since $p_{min,n} \geq p^\star_{min}$ and $\Delta_n \geq \Delta^\star_{min}$ for positive constants $p^\star_{min}$ and $\Delta^\star_{min}$. Hence, the result follows immediately from Theorem 1-Item 3 and Remark 2-Item (c). ■

A.2 Proofs of Section 3.2

A.2.1 Proof of Proposition 6

In this section we prove Proposition 6. To that end, we need some auxiliary results. The first auxiliary result is the following. Recall that we write $\mathrm{Cov}_{x_S}(X_0, m_k(X_k))$ to denote the conditional covariance between the random variables X_0 and $m_k(X_k)$ given that $X_S = x_S$, where m_k is defined (3).

Lemma 19 *For each $S \subseteq [\![-d, -1]\!]$, $k \in S^c$ and $x_S \in A^S$, the following identity holds:*

$$Cov_{x_S}(X_0, m_k(X_k)) = \sum_{j \in \Lambda \setminus S} \lambda_j \, Cov_{x_S}(m_j(X_j), m_k(X_k)). \tag{50}$$

Remark 20 *In (50), we use the convention that $\sum_{j \in \emptyset} \lambda_j \, Cov_{x_S}(m_j(X_j), m_k(X_k)) = 0$.*

Proof [Proof of Lemma 19]

Observe that if $\Lambda \subseteq S$, then the both sides of (50) are 0, so that the result holds immediately in this case.

Now suppose that $\Lambda \not\subseteq S$. In this case, to shorten the notation, let us write

$$\mathbf{P}_{x_S}(x_{\Lambda \setminus S}) = \mathbb{P}_{x_S}(X_{\Lambda \setminus S} = x_{\Lambda \setminus S}), \text{ for } x_{\Lambda \setminus S} \in A^{\Lambda \setminus S}.$$

We want to compute

$$\mathrm{Cov}_{x_S}(X_0, m_k(X_k)) = \mathbb{E}_{x_S}(X_0 m_k(X_k)) - \mathbb{E}_{x_S}(X_0)\mathbb{E}_{x_S}(m_k(X_k)).$$

We first compute $\mathbb{E}_{x_S}(X_0)$. To that end, write

$$\mathbb{E}_{x_S}(X_0) = \sum_{a \in A} a \mathbb{P}_{x_S}(X_0 = a),$$

and observe that for each $a \in A$,

$$\begin{aligned}
\mathbb{P}_{x_S}(X_0 = a) &= \sum_{x_{\Lambda \setminus S} \in A^{\Lambda \setminus S}} \mathbf{P}_{x_S}(x_{\Lambda \setminus S}) p(a | x_S x_{\Lambda \setminus S}) \\
&= \lambda_0 p_0(a) + \sum_{j \in \Lambda \cap S} \lambda_j p_j(a | x_j) + \sum_{j \in \Lambda \setminus S} \lambda_j \sum_{x_{\Lambda \setminus S} \in A^{\Lambda \setminus S}} \mathbf{P}_{x_S}(x_{\Lambda \setminus S}) p_j(a | x_j) \\
&= \lambda_0 p_0(a) + \sum_{j \in \Lambda \cap S} \lambda_j p_j(a | x_j) + \sum_{j \in \Lambda \setminus S} \lambda_j \mathbb{E}_{x_S}(p_j(a | X_j)),
\end{aligned}$$

where in the second equality we have used the definition of the transition probabilities (2). Hence, we have that

$$\mathbb{E}_{x_S}(X_0) = \lambda_0 m_0 + \sum_{j \in \Lambda \cap S} \lambda_j m_j(x_j) + \sum_{j \in \Lambda \setminus S} \lambda_j \mathbb{E}_{x_S}(m_j(X_j)),$$

where $m_0 = \sum_{a \in A} a p_0(a)$.

As a consequence of the above equality, it follows that

$$\mathbb{E}_{x_S}(X_0)\mathbb{E}_{x_S}(m_k(X_k)) = \left[\lambda_0 m_0 + \sum_{j\in\Lambda\cap S}\lambda_j m_j(x_j)\right]\mathbb{E}_{x_S}(m_k(X_k)) \\ + \sum_{j\in\Lambda\setminus S}\lambda_j \mathbb{E}_{x_S}(m_j(X_j))\mathbb{E}_{x_S}(m_k(X_k)). \quad (51)$$

We now compute $\mathbb{E}_{x_S}(X_0 m_k(X_k))$. We consider only the case $k \in \Lambda$, the other is treated similarly. In this case, we first write

$$\mathbb{E}_{x_S}(X_0 m_k(X_k)) = \sum_{a\in A}\sum_{b\in A} a m_k(b)\mathbb{P}_{x_S}(X_0 = a, X_k = b), \quad (52)$$

and then we proceed similar as above to deduce that for each $a, b \in A$,

$$\mathbb{P}_{x_S}(X_0 = a, X_k = b) = \sum_{x_{\Lambda\setminus S}\in A^{\Lambda\setminus S}} \mathbf{P}_{x_S}(x_{\Lambda\setminus S})p(a|x_S x_{\Lambda\setminus S})1\{x_k = b\} \\ = \left[\lambda_0 p_0(a) + \sum_{j\in\Lambda\cap S}\lambda_j p_j(a|x_j) + \lambda_k p_k(a|b)\right]\mathbb{P}_{x_S}(X_k = b) \\ + \sum_{j\in\Lambda\setminus(S\cup\{k\})}\lambda_j \sum_{c\in A} p_j(a|c)\mathbb{P}_{x_S}(X_j = c, X_k = b). \quad (53)$$

where in the second equality we have used the definition of the transition probabilities (2). Combining (52) and (53), we deduce that

$$\mathbb{E}_{x_S}(X_0 m_k(X_k)) = \left[\lambda_0 m_0 + \sum_{j\in\Lambda\cap S}\lambda_j m_j(x_j)\right]\mathbb{E}_{x_S}(m_k(X_k)) + \lambda_k \mathbb{E}_{x_S}(m_k^2(X_k)) \\ + \sum_{j\in\Lambda\setminus(S\cup\{k\})}\lambda_j \mathbb{E}_{x_S}(m_k(X_k)m_j(X_j)) \\ = \left[\lambda_0 m_0 + \sum_{j\in\Lambda\cap S}\lambda_j m_j(x_j)\right]\mathbb{E}_{x_S}(m_k(X_k)) \\ + \sum_{j\in\Lambda\setminus S}\lambda_j \mathbb{E}_{x_S}(m_k(X_k)m_j(X_j)). \quad (54)$$

Putting together the identities (51) and (54), we then conclude that

$$\mathrm{Cov}_{x_S}(X_0 m_k(X_k)) = \sum_{j\in\Lambda\setminus S}\lambda_j\left(\mathbb{E}_{x_S}(m_j(X_j)m_k(X_k)) \right. \\ \left. - \mathbb{E}_{x_S}(m_j(X_j))\mathbb{E}_{x_S}(m_k(X_k))\right), \quad (55)$$

and the result follows. ■

The next auxiliary result is the following.

Lemma 21 *Suppose Assumptions 1 and 3 hold. Then there exists a constant $\kappa' > 0$ such that the following property holds: for any $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \not\subseteq S$, we have*

$$\sum_{j\in\Lambda\setminus S}\sum_{k\in\Lambda\setminus S}\lambda_j\lambda_k\mathbb{E}(Cov_{X_S}(m_j(X_j),m_k(X_k))) \geq \kappa'.$$

Proof [Proof of Lemma 21] It suffices to show that for $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \not\subseteq S$, we have

$$\sum_{j\in\Lambda\setminus S}\sum_{k\in\Lambda\setminus S}\lambda_j\lambda_k\mathbb{E}(\mathrm{Cov}_{x_S}(m_j(X_j),m_k(X_k))) > 0.$$

Suppose that this is not the case. Then,

$$\begin{aligned}
0 &= \sum_{j\in\Lambda\setminus S}\sum_{k\in\Lambda\setminus S}\lambda_j\lambda_k\mathbb{E}(\mathrm{Cov}_{x_S}(m_j(X_j),m_k(X_k)))\\
&= \sum_{j\in\Lambda\setminus S}\sum_{k\in\Lambda\setminus S}\lambda_j\lambda_k\mathrm{Cov}\left(m_j(X_j)-\mathbb{E}_{X_S}(m_j(X_j)), m_k(X_k)-\mathbb{E}_{X_S}(m_k(X_k))\right)\\
&= \mathrm{Var}\left(\sum_{j\in\Lambda\setminus S}\lambda_j(m_j(X_j)-\mathbb{E}_{X_S}(m_j(X_j)))\right),
\end{aligned}$$

so that $\mathbb{P}$-almost surely,

$$\sum_{j\in\Lambda\setminus S}\lambda_j(m_j(X_j)-\mathbb{E}_{X_S}(m_j(X_j))) = 0.$$

This implies that $\mathbb{P}$-almost surely,

$$\sum_{j\in\Lambda\setminus S}\lambda_j m_j(X_j) = \mathbb{E}_{X_S}\left(\sum_{j\in\Lambda\setminus S}\lambda_j m_j(X_j)\right),$$

or equivalently,

$$\sum_{j\in\Lambda\setminus S}\lambda_j m_j(X_j) = f(X_S),\ \mathbb{P}\text{-a.s.},$$

for some function $f : A^S \to \mathbb{R}$.

Now take any configuration $x_S \in A^S$ and consider the event $A = \{X_S = x_S\}$. From the above identity, it follows that $\mathbb{P}$-a.s.,

$$1_A\sum_{j\in\Lambda\setminus S}\lambda_j m_j(X_j) = 1_A f(x_S).$$

Finally, take any configuration $x_{\Lambda\setminus S} \in A^{\Lambda\setminus S}$ such that

$$\sum_{j\in\Lambda\setminus S}\lambda_j m_j(x_j) \neq f(x_S),$$

and let $B = \{X_{\Lambda\setminus S} = x_{\Lambda\setminus S}\}$. Such a configuration must exist by Assumption 3. As a consequence, we have that $\mathbb{P}$-a.s.,

$$1_{A\cap B}\sum_{j\in\Lambda\setminus S}\lambda_j m_j(x_j) = 1_{A\cap B}f(x_S),$$

implying that

$$\mathbb{P}(A\cap B)\sum_{j\in\Lambda\setminus S}\lambda_j m_j(x_j) = \mathbb{P}(A\cap B)f(x_S).$$

By Assumption 1, we have $\mathbb{P}(A\cap B) = \mathbf{P}(x_\Lambda) > 0$ so that the identify above would imply that

$$\sum_{j\in\Lambda\setminus S}\lambda_j m_j(x_j) = f(x_S),$$

which is a contradiction. Therefore, we must have

$$\mathrm{Var}\left(\sum_{j\in\Lambda\setminus S}\lambda_j(m_j(X_j) - \mathbb{E}_{X_S}(m_j(X_j)))\right) > 0,$$

and the result follows. ■

We also need the following result.

Lemma 22 *For each $S\subseteq [\![-d,-1]\!]$, $k\in S^c$ and $x_S\in A^S$, the following identity holds:*

$$|Cov_{x_S}(X_0, m_k(X_k))| \le \|m_k\|_{Lip}|Cov_{x_S}(X_0, X_k)|, \tag{56}$$

where m_k and $\|m_k\|_{Lip}$ are defined (3) *and* (6) *respectively.*

Proof [Proof of Lemma 22] First observe that $\mathrm{Cov}_{x_S}(X_0, m_k(X_k)) = \mathrm{Cov}_{x_S}(X_0, m_k(X_k) - m_k(c))$, for any $c\in A$. Since,

$$\mathrm{Cov}_{x_S}(X_0, m_k(X_k) - m_k(c)) = \sum_{b\in A}(m_k(b) - m_k(c))\mathrm{Cov}_{x_S}(X_0, 1\{X_k = b\})$$

and $|m_k(b) - m_k(c)| \le \|m_k\|_{Lip}|b-c|$, it follows then that

$$\mathrm{Cov}_{x_S}(X_0, m_k(X_k)) \le \|m_k\|_{Lip}\sum_{b\in A}|b-c|\mathrm{Cov}_{x_S}(X_0, 1\{X_k = b\}).$$

By taking $c = \min(A)$, we have that $|b-c| = (b-c)$ for any $b\in A$ and we deduce from the above inequality that

$$\mathrm{Cov}_{x_S}(X_0, m_k(X_k)) \le \|m_k\|_{Lip}\mathrm{Cov}_{x_S}(X_0, X_k) \le \|m_k\|_{Lip}|\mathrm{Cov}_{x_S}(X_0, X_k)|.$$

A similar argument shows that $\mathrm{Cov}_{x_S}(X_0, m_k(X_k)) \ge -\|m_k\|_{Lip}|\mathrm{Cov}_{x_S}(X_0, X_k)|$, concluding the proof. ■

Our last auxiliary result is the following.

Lemma 23 *For each $S \subseteq [\![-d,-1]\!]$ such $\Lambda \not\subset S$, $x_S \in A^S$ and $j,k \in \Lambda \setminus S$, the following identity holds:*

$$Cov_{x_S}(m_j(X_j), m_k(X_k)) = \frac{1}{2}\sum_{b\in A}\sum_{c\in A}\mathbb{P}_{x_S}(X_k=b)\mathbb{P}_{x_S}(X_k=c)(m_k(b)-m_k(c))\times \\ (\mathbb{E}_{x_S}(m_j(X_j)|X_k=b)-\mathbb{E}_{x_S}(m_j(X_j)|X_k=c)) \,. \quad (57)$$

where m_j is defined (3).

Proof [Proof of Lemma 23] First notice that

$$\text{Cov}_{x_S}(m_j(X_j), m_k(X_k)) = \sum_{a,b\in A} m_j(a)m_k(b)\text{Cov}_{x_S}(1\{X_j=a\}, 1\{X_k=b\}).$$

Now, for any $a, b \in A$, one can check that

$$\text{Cov}_{x_S}(1\{X_j=a\}, 1\{X_k=b\}) = \sum_{c\in A}\mathbb{P}_{x_S}(X_k=b)\mathbb{P}_{x_S}(X_k=c) \\ \times (\mathbb{P}_{x_S}(X_j=a|X_k=b)-\mathbb{P}_{x_S}(X_j=a|X_k=c))$$

Hence, we deduce from the above equalities that

$$\text{Cov}_{x_S}(m_j(X_j), m_k(X_k)) = \sum_{b\in A}\sum_{c\in A} m_k(b)\mathbb{P}_{x_S}(X_k=b)\mathbb{P}_{x_S}(X_k=c)) \\ \times (\mathbb{E}_{x_S}(m_j(X_j)|X_k=b)-\mathbb{E}_{x_S}(m_j(X_j)|X_k=c)).$$

Interchanging the role of the symbols b and c in the equality above, we obtain that

$$\text{Cov}_{x_S}(m_j(X_j), m_k(X_k)) = -\sum_{c\in A}\sum_{b\in A} m_k(c)\mathbb{P}_{x_S}(X_k=b)\mathbb{P}_{x_S}(X_k=c)) \\ \times (\mathbb{E}_{x_S}(m_j(X_j)|X_k=b)-\mathbb{E}_{x_S}(m_j(X_j)|X_k=c)).$$

The result follows by combing the last two equalities above. ■

We now prove Proposition 6.

Proof [Proof of Proposition 6]

We first prove inequality (26). Let us denote $D_{k,S}(a,b,c,x_S) = \mathbb{P}_{x_S}(X_0=a|X_k=b) - \mathbb{P}_{x_S}(X_0=a|X_k=c)$, for each $a,b,c \in A$, $x_S \in A^S$ and $k \notin S$. With this notation, one can check that for any $x_S \in A^S$ and $k \notin S$, we have that

$$\text{Cov}_{x_S}(X_0, X_k) = \frac{1}{2}\sum_{b\in A}\sum_{c\in A}(b-c)w_{k,S}(b,c,x_S)\sum_{a\in A} aD_{k,S}(a,b,c,x_S). \quad (58)$$

Now, observe that the triangle inequality and the equality

$$\frac{1}{2}\sum_{a\in A}|D_{k,S}(a,b,c,x_S)| = d_{k,S}(b,c,x_S),$$

imply that

$$|\mathrm{Cov}_{x_S}(X_0, X_k)| \leq \mathit{Diam}(A)\|A\|_\infty \sum_{b\in A}\sum_{c\in A} w_{k,S}(b,c,x_S)d_{k,S}(b,c,x_S),$$

so that

$$\mathbb{E}\left(|\mathrm{Cov}_{X_S}(X_0, X_k)|\right) \leq \mathit{Diam}(A)\|A\|_\infty \bar{\nu}_{k,S},$$

proving inequality (26).

We now prove (27). This is done as follows. In the sequel, we shall write $\lambda 1_{\Lambda\setminus S}$ to denote the vector $\lambda = (\lambda_j)_{j\in\Lambda}$ restricted to the coordinates in $\Lambda \setminus S$: $\lambda 1_{\Lambda\setminus S} = (\lambda_j)_{j\in\Lambda\setminus S}$. With this notation, it follows from Lemma 19 and Lemma 21 that for any $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \not\subseteq S$,

$$\sum_{k\in\Lambda\setminus S} \lambda_k \mathbb{E}\left(\mathrm{Cov}_{X_S}(X_0, m_k(X_k))\right) \geq \kappa'.$$

By the triangle inequality, it then follows that

$$\sum_{k\in\Lambda\setminus S} \lambda_k \left|\mathbb{E}\left(\mathrm{Cov}_{X_S}(X_0, m_k(X_k))\right|\right) \geq \kappa'.$$

Now using that

$$\max_{k\in\Lambda\setminus S} \left|\mathbb{E}\left(\mathrm{Cov}_{X_S}(X_0, m_k(X_k))\right)\right| \|\lambda 1_{\Lambda\setminus S}\|_1 \geq \sum_{k\in\Lambda\setminus S} \lambda_k \left|\mathbb{E}\left(\mathrm{Cov}_{X_S}(X_0, m_k(X_k))\right)\right|,$$

we conclude that

$$\max_{k\in\Lambda\setminus S} \left|\mathbb{E}\left(\mathrm{Cov}_{X_S}(X_0, m_k(X_k))\right)\right| \|\lambda 1_{\Lambda\setminus S}\|_1 \geq \kappa'.$$

By observing that $1-\lambda_0 = \sum_{k\in\Lambda}\lambda_k \geq \|\lambda 1_{\Lambda\setminus S}\|_1$, we conclude from the above inequality that

$$\max_{k\in\Lambda\setminus S} \left|\mathbb{E}\left(\mathrm{Cov}_{X_S}(X_0, m_k(X_k))\right)\right| \geq \kappa'/(1-\lambda_0) > 0,$$

and the result follows from Lemma 22.

Therefore, it remains to prove (28). To that end, we first use Lemma 19, Lemma (22) and Lemma 23 to obtain that

$$\sum_{k\in\Lambda\setminus S} \lambda_k \|m_k\|_{Lip} |\mathrm{Cov}_{x_S}(X_0, X_k)| \geq \frac{1}{2}\sum_{k\in\Lambda\setminus S}\sum_{j\in\Lambda\setminus S}\sum_{b\in A}\sum_{c\in A} \lambda_k\lambda_j \mathbb{P}_{x_S}(X_k = b)\mathbb{P}_{x_S}(X_k = c)$$
$$\times \left(m_k(b) - m_k(c)\right)\left(\mathbb{E}_{x_S}(m_j(X_j)|X_k = b) - \mathbb{E}_{x_S}(m_j(X_j)|X_k = c)\right).$$

Next, we observe that Assumption 4 implies that

$$\left(1 - \sum_{j\in\Lambda\setminus S: j\neq k} \frac{\lambda_j\left(|\mathbb{E}_{x_S}(m_j(X_j)|X_k = b) - \mathbb{E}_{x_S}(m_j(X_j)|X_k = c)|\right)}{\lambda_k |m_k(b) - m_k(c)|}\right) \geq \Gamma_1,$$

so that

$$\sum_{k\in\Lambda\setminus S} \lambda_k \|m_k\|_{Lip} \mathbb{E}\left(|\mathrm{Cov}_{x_S}(X_0, X_k)|\right) \geq \frac{p_{min}^2 \Gamma_1}{2} \sum_{k\in\Lambda\setminus S} \lambda_k^2 \sum_{b\in A}\sum_{c\in A}(m_k(b) - m_k(c))^2,$$

where we have also used that $\mathbb{P}_{x_S}(X_k = b) \geq p_{min}$. Finally, note that $|m_k(b) - m_k(c)| \geq \min\{|b-c|^2 : b \neq c\}\|m_k\|^2_{Lip}$ to obtain that

$$\max_{k \in \Lambda \setminus S} \mathbb{E}\left(|\mathrm{Cov}_{x_S}(X_0, X_k)|\right) \sum_{k \in \Lambda \setminus S} \lambda_k \|m_k\|_{Lip} \geq \frac{p^2_{min}\Gamma_1}{2\min\{|b-c|^2 : b \neq c\}} \sum_{k \in \Lambda \setminus S} \lambda_k^2 \|m_k\|^2_{Lip}.$$

Then, by using Cauchy-Schwartz inequality, we deduce that

$$\sum_{k \in \Lambda \setminus S} \lambda_k \|m_k\|_{Lip} \leq \sqrt{\sum_{k \in \Lambda \setminus S} \lambda_k^2 \|m_k\|^2_{Lip}} \sqrt{|\Lambda \setminus S|} \leq \sqrt{\sum_{k \in \Lambda \setminus S} \lambda_k^2 \|m_k\|^2_{Lip}} \sqrt{|\Lambda|}.$$

The result follows by combining the last two inequalities. ■

A.2.2 Proof of Theorem 9

Before starting the proof of Theorem 9, we recall some definitions from Information Theory. In what follows, for $S \in [\![-d, -1]\!]$ and $j \in [\![-d, -1]\!]$, we write $I(X_0; X_j | X_S)$ to denote the conditional mutual information between X_0 and X_j given X_S, defined as

$$I(X_0; X_j | X_S) = \sum_{x_S \in A^S} \mathbf{P}(x_S) I(X_0; X_j | X_S = x_S), \tag{59}$$

where $I(X_0; X_j | X_S = x_S) := I_j(x_S)$ denotes the conditional mutual information between X_0 and X_j given $X_S = x_S$, defined as

$$I_j(x_S) = \sum_{a,b \in A} \mathbb{P}_{x_S}(X_0 = a, X_j = b) \log\left(\frac{\mathbb{P}_{x_S}(X_0 = a, X_j = b)}{\mathbb{P}_{x_S}(X_0 = a)\mathbb{P}_{x_S}(X_j = b)}\right). \tag{60}$$

We use the convention that when $S = \emptyset$, the conditional probability $\mathbb{P}_{x_s}$ is the unconditional probability $\mathbb{P}$. Hence, in this case, the conditional mutual information between X_0 and X_j is the mutual information between these random variables, denoted $I(X_0; X_j) := I_j$.

The entropy $H(X_0)$ of X_0 is defined as

$$H(X_0) = -\sum_{a \in A} \mathbb{P}(X_0 = a) \log(\mathbb{P}(X_0 = a)). \tag{61}$$

To prove Theorem 9 we proceed similarly to Bresler (2015). During the proof we will need the the following lemma.

Lemma 24 *Suppose that the event $G_m(\xi, \ell)$ defined in* (29) *holds and let $S \subseteq [\![-d, -1]\!]$ with $|S| \leq \ell$. If $\hat{\nu}_{m,k,S} \geq \tau$ with $k \in S^c$, then $I(X_0; X_k | X_S) \geq 2(\tau - \xi)^2$.*

Proof Definition (59) together with Jensen inequality implies that for any $j \in S^c$,

$$\begin{aligned} \sqrt{\frac{1}{2} I(X_0; X_j | X_S)} &= \sqrt{\frac{1}{2} \sum_{x_S \in A^S} \mathbf{P}(x_S) I_j(x_S)} \\ &\geq \sum_{x_S \in A^S} \mathbf{P}(x_S) \sqrt{\frac{1}{2} I_j(x_S)}. \end{aligned}$$

By Pinsker inequality, it then follows that

$$\begin{aligned}\sqrt{\frac{1}{2}I_j(x_S)} &\geq \frac{1}{2}\sum_{a,b\in A}|\mathbb{P}_{x_S}(X_0=a,X_j=b)-\mathbb{P}_{x_S}(X_0=a)\mathbb{P}_{x_S}(X_j=b)| \\ &= \sum_{b\in A}\mathbb{P}_{x_S}(X_j=b)\frac{1}{2}\sum_{a\in A}|\mathbb{P}_{x_S}(X_0=a|X_j=b)-\mathbb{P}_{x_S}(X_0=a)| \\ &= \sum_{b\in A}\sum_{c\in A}w_{j,S}(b,c,x_S)d_{TV}(\mathbb{P}_{x_S}(X_0\in\cdot|X_j=b),\mathbb{P}_{x_S}(X_0\in\cdot|X_j=c)) \\ &= \nu_{j,S}(x_S),\end{aligned}$$

where in the second equality we have used that for any $a,b\in A$,

$$\mathbb{P}_{x_S}(X_0=a|X_j=b)=\sum_{c\in A}\mathbb{P}_{x_S}(X_j=c)\mathbb{P}_{x_S}(X_0=a|X_j=b),$$

and also that

$$\mathbb{P}_{x_S}(X_0=a)=\sum_{c\in A}\mathbb{P}_{x_S}(X_j=c)\mathbb{P}_{x_S}(X_0=a|X_j=c)).$$

As a consequence, we deduce that

$$\sqrt{\frac{1}{2}I(X_0;X_j|X_S)}\geq\sum_{x_S\in A^S}\mathbf{P}(x_S)\nu_{j,S}(x_S)=\bar{\nu}_{j,S}.$$

Now, on the event $G_m(\xi,\ell)$, we have that $\bar{\nu}_{k,S}\geq\hat{\nu}_{m,k,S}-\xi$ so that

$$\sqrt{\frac{1}{2}I(X_0;X_k|X_S)}\geq\hat{\nu}_{m,k,S}-\xi\geq\tau-\xi,$$

where in the rightmost inequality we have used that $\hat{\nu}_{m,k,S}\geq\tau$. Hence,

$$I(X_0;X_j|X_S)\geq 2(\tau-\xi)^2,$$

and the result follows. ■

We now prove Theorem 9.

Proof Suppose the event $G_m(\xi_*,\ell_*)$ holds and let $\hat{S}_m$ be the set obtained at the end of FS step of Algorithm 1 with parameter ℓ_*, where the parameters ξ_* and ℓ_* are defined as in (30). In the sequel, let $S_0=\emptyset$ and $S_k=S_{k-1}\cup\{j_k\}$, where $j_k\in\arg\max_{j\in S^c_{k-1}}\hat{\nu}_{m,j,S_{k-1}}$ for $1\leq k\leq d$, and observe that by construction $\hat{S}_m=S_{\ell_*}$. We want to show that $\Lambda\subseteq\hat{S}_m$. We argue by contraction. Suppose that Λ is not contained in $\hat{S}_m$. In this case, it follows that $\Lambda\not\subseteq S_k$ for all $1\leq k\leq\ell_*$, and Proposition 6 implies that for all $1\leq k\leq\ell_*$,

$$\max_{j\in S^c_k}\bar{\nu}_{j,S_k}\geq\frac{\kappa}{\|A\|_\infty Diam(A)}=4\xi_*,$$

where the equality holds by the choice of ξ_*. Since the event $G_m(\xi_*, \ell_*)$ holds and $|S_k| \le |S_{\ell_*}| = \ell_*$, it follows from the above inequality that

$$\hat{\nu}_{m,j_k,S_{k-1}} = \max_{j \in S_{k-1}^c} \hat{\nu}_{m,j,S_{k-1}} \ge 3\xi_*,$$

for all $1 \le k \le \ell_* + 1$. By Lemma 24, we then deduce that $I(X_0, X_{j_k}|X_{S_{k-1}}) \ge 8\xi_*^2$ for all $1 \le k \le \ell_* + 1$.

Now, notice that

$$\log_2(|A|) \ge H(X_0) \ge I(X_0; X_{\hat{S}_m \cup \{j_{\ell_*+1}\}}) = \sum_{k=1}^{\ell_*+1} I(X_0, X_{j_k}|X_{S_{k-1}}), \tag{62}$$

where we have used Gibbs inequality in the first passage, the fact that the entropy is always larger than the mutual information in the second passage and the Chain Rule in the last passage. The proof of these facts can be found, for instance, in (Cover and Thomas, 2006).

By the choice of $\ell_* = \lfloor \log_2(|A|)/8\xi_*^2 \rfloor$, we have that $\ell_* + 1 > \log_2(|A|)/8\xi_*^2$ so that it follows from (62) that

$$\log_2(|A|) \ge (\ell_* + 1)8\xi_*^2 > \log_2(|A|),$$

a contradiction. Thus, we must have $\Lambda \subseteq \hat{S}_m$ and the result follows. ■

A.2.3 Proof of Theorem 10

To prove Theorem 10 we shall need the following result.

Proposition 25 *Suppose Assumptions 1 and 2 hold, and let $\Delta^\star > 0$ the quantity defined in Assumption 2. Then, for any $\xi > 0$ and $m > d \ge 2\ell \ge 0$,*

$$\mathbb{P}(G_m^c(\ell, \xi)) \le 2d(\ell+1)\binom{d}{\ell}|A|^{\ell+2} \exp\left\{-\frac{\xi^2(m-d)^2(\Delta_*)^2}{18|A|^{2(\ell+2)}m(\ell+2)^2}\right\}. \tag{63}$$

During the proof of Proposition 25 we will make use of the following proposition. For any function $f : A^{[\![1,m]\!]} \to \mathbb{R}$, define for each $1 \le j \le m$,

$$\delta_j(f) = \sup\left\{|f(x_{1:j-1}ax_{j+1:m}) - f(x_{1:j-1}bx_{j+1:m})| : a, b \in A, x_{1:m} \in A^{[\![1,m]\!]}\right\}, \tag{64}$$

with the convention that $x_{1:0} = x_{m+1:m} = \emptyset$, $\emptyset ax_{2:m} = ax_{2:m}$ and $x_{1:m-1}a = x_{1:m-1}a\emptyset$. Let $\underline{\delta}(f) = (\delta_1(f), \ldots, \delta_m(f))$ and denote $\|\underline{\delta}(f)\|_2^2 = \sum_{j=1}^m \delta_j^2(f)$. In what follows, we write $\mathbb{E}_{1:m}[f] = \sum_{x_{1:m} \in A^{[\![1,m]\!]}} \mathbb{P}(X_{1:m} = x_{1:m})f(x_{1:m})$.

Proposition 26 (Theorem 3.4. of (Chazottes et al., 2020)) *Suppose Assumption 2 holds, that is, $\Delta > 0$.*

1. *For any $u > 0$ and $f : A^{[\![1,m]\!]} \to \mathbb{R}$,*

$$\mathbb{P}\left(|f(X_{1:m}) - \mathbb{E}_{1:m}[f]| > u\right) \le 2\exp\left\{-\frac{2u^2\Delta^2}{\|\underline{\delta}(f)\|_2^2}\right\}.$$

2. For $m > d$, any $g : A^S \times A \to \mathbb{R}$ with $S \subseteq [\![-d,-1]\!]$ and $u > 0$,

$$\mathbb{P}\left(\left|\frac{1}{(m-d)}\sum_{t=d+1}^{m} g(X_{t+S}, X_t) - \mathbb{E}_S[g]\right| > u\right) \leq 2\exp\left\{-\frac{u^2(m-d)^2\Delta^2}{2m(|S|+1)^2\|g\|_\infty^2}\right\},$$

where $\mathbb{E}_S[g] = \sum_{x_S \in A^S}\sum_{a\in A}\mathbb{P}(X_S = x_S, X_0 = a)g(x_S, a)$ and $X_{t+S} = (X_{t+j})_{j\in S}$.

Before starting the proof Proposition 25, we need to introduce some additional notation. For each $x \in A^S$ with $S \subseteq [\![-d,-1]\!]$, we write

$$\hat{\mathbf{P}}_m(x) = \frac{\bar{N}_m(x)}{m-d}. \tag{65}$$

In what follows, we write $xa_{V\cup\{0\}}$, with $a \in A$ and $V \subseteq [\![-d,-1]\!]$, to denote the configuration $((xa)_j)_{j\in V\cup\{0\}}$, defined as

$$(xa)_j = \begin{cases} x_j, & \text{for } j \in V \\ a, & \text{for } j = 0 \end{cases}.$$

When $V = S \cup \{k\}$ and $x_k = b \in A$, we shall write $xba_{S\cup\{k,0\}}$ instead of $xa_{V\cup\{0\}}$.

With this notation, the empirical version of $\bar{\nu}_{k,S}$ is defined as follows:

$$\hat{\nu}_{m,k,S} = \sum_{x_S\in A^S}\hat{\mathbf{P}}_m(x_S)\hat{\nu}_{m,k,S}(x_S) \tag{66}$$

where for $x_S \in A^S$, we define

$$\hat{\nu}_{m,k,S}(x_S) = \sum_{b\in A}\sum_{c\in A}\hat{w}_{m,k,S}(b,c,x_S)\hat{d}_{m,k,S}(b,c,x_S), \tag{67}$$

and for $b, c \in A$,

$$\hat{w}_{m,k,S}(b,c,x_S) = \frac{\hat{\mathbf{P}}_m(xb_{S\cup\{k\}})}{\hat{\mathbf{P}}_m(x_S)}\frac{\hat{\mathbf{P}}_m(xc_{S\cup\{k\}})}{\hat{\mathbf{P}}_m(x_S)}, \tag{68}$$

and

$$\hat{d}_{m,k,S}(b,c,x_S) = \frac{1}{2}\sum_{a\in A}\left|\hat{\mathbb{P}}_{x_S}(X_0 = a|X_k = b) - \hat{\mathbb{P}}_{x_S}(X_0 = a|X_k = c)\right|,$$

where for each $b \in A$,

$$\hat{\mathbb{P}}_{x_S}(X_0 = a|X_k = b) = \frac{\hat{\mathbf{P}}_m(xba_{S\cup\{k,0\}})}{\hat{\mathbf{P}}_m(xb_{S\cup\{k\}})}.$$

Hereafter, we omit the dependence on S and on m, whenever there is no risk of confusion. We now prove Proposition 25.

Proof [Proof of Proposition 25]

Claim 1. Let $S \subseteq [\![-d,-1]\!]$ with $|S| \leq \ell < d/2$ and take $j \in S^c$. Then,

$$|\bar{\nu}_{j,S} - \hat{\nu}_{j,S}| \leq 3\sum_{x\in A^S}\sum_{a\in A}\sum_{b\in A}\left|\hat{\mathbb{P}}(X_S = x, X_j = b, X_0 = a) - \mathbb{P}(X_S = x, X_j = b, X_0 = a)\right|.$$

Proof of the Claim 1. By applying the triangle inequality twice, one can check that

$$|\bar{\nu}_{j,S}-\hat{\nu}_{j,S}|\leq \frac{1}{2}\sum_{x\in A^S}\sum_{a,b,c\in A}\Big|\mathbf{P}(x)w_{j,S}(b,c,x)\left(\mathbb{P}_x(X_0=a|X_j=b)-\mathbb{P}_x(X_0=a|X_j=c)\right)$$
$$-\hat{\mathbf{P}}(x)\hat{w}_{j,S}(b,c,x)\left(\hat{\mathbb{P}}_x(X_0=a|X_j=b)-\hat{\mathbb{P}}_x(X_0=a|X_j=c)\right)\Big|. \quad (69)$$

Now observe that for fixed $x\in A^S$ and $a,b,c\in A$,

$$\mathbf{P}(x)w_{j,S}(b,c,x)\mathbb{P}_x(X_0=a|X_j=b)=\mathbb{P}_x(X_j=c)\mathbb{P}(X_S=x,X_j=b,X_0=a)$$

and similarly,

$$\hat{\mathbf{P}}(x)\hat{w}_{j,S}(b,c,x)\hat{\mathbb{P}}_x(X_0=a|X_j=b)=\hat{\mathbb{P}}_x(X_j=c)\hat{\mathbb{P}}(X_S=x,X_j=b,X_0=a).$$

By using these identities in (69) and then by applying the triangle inequality, one can deduce that

$$|\bar{\nu}_{j,S}-\hat{\nu}_{j,S}|\leq \sum_{x\in A^S}\sum_{a,b,c\in A}\Big|\mathbb{P}_x(X_j=c)\mathbb{P}(X_S=x,X_j=b,X_0=a)$$
$$-\hat{\mathbb{P}}_x(X_j=c)\hat{\mathbb{P}}(X_S=x,X_j=b,X_0=a)\Big|. \quad (70)$$

By adding and subtracting the term $\mathbb{P}_x(X_j=c)\hat{\mathbb{P}}(X_S=x,X_j=b,X_0=a)$ in the right-hand side of the above inequality and using again the triangle inequality, it follows that

$$\sum_{x\in A^S}\sum_{a,b,c\in A}\Big|\mathbb{P}_x(X_j=c)\mathbb{P}(X_S=x,X_j=b,X_0=a)$$
$$-\hat{\mathbb{P}}_x(X_j=c)\hat{\mathbb{P}}(X_S=x,X_j=b,X_0=a)\Big|$$
$$\leq \sum_{x\in A^S}\sum_{a,b\in A}|\mathbb{P}(X_S=x,X_j=b,X_0=a)-\hat{\mathbb{P}}(X_S=x,X_j=b,X_0=a)|$$
$$+\sum_{x\in A^S}\sum_{a,c\in A}\hat{\mathbb{P}}(X_S=x,X_0=a)|\mathbb{P}_x(X_j=c)-\hat{\mathbb{P}}_x(X_j=c)|. \quad (71)$$

By adding and subtracting the term $\mathbb{P}(X_S=x)\mathbb{P}(X_S=x,X_j=c)$, we can then check that

$$|\mathbb{P}_x(X_j=c)-\hat{\mathbb{P}}_x(X_j=c)|\leq \frac{\mathbb{P}(X_j=c)}{\hat{\mathbb{P}}(X_S=x)}|\hat{\mathbb{P}}(X_S=x)-\mathbb{P}(X_S=x)|$$
$$+\frac{1}{\hat{\mathbb{P}}(X_S=x)}|\hat{\mathbb{P}}(X_S=x,X_j=c)-\mathbb{P}(X_S=x,X_j=c)|. \quad (72)$$

From (71) and (72), one deduces that

$$\begin{aligned}|\bar{\nu}_{j,S}-\hat{\nu}_{j,S}| \le \sum_{x\in A^S}\sum_{a,b\in A} |\mathbb{P}(X_S=x,X_j=b,X_0=a)-\hat{\mathbb{P}}(X_S=x,X_j=b,X_0=a)| \\ + \sum_{x\in A^S}\sum_{c\in A}|\mathbb{P}(X_S=x,X_j=c)-\hat{\mathbb{P}}(X_S=x,X_j=c)| \\ + \sum_{x\in A^S}|\mathbb{P}(X_S=x)-\hat{\mathbb{P}}(X_S=x)|. \quad (73)\end{aligned}$$

Since

$$\begin{aligned}|\mathbb{P}(X_S=x,X_j=c)-\hat{\mathbb{P}}(X_S=x,X_j=c)| \\ \le \sum_{a\in A}|\mathbb{P}(X_S=x,X_j=c,X_0=a)-\hat{\mathbb{P}}(X_S=x,X_j=c,X_0=a)|\end{aligned}$$

and

$$\begin{aligned}|\mathbb{P}(X_S=x)-\hat{\mathbb{P}}(X_S=x)| \\ \le \sum_{a,c\in A}|\mathbb{P}(X_S=x,X_j=c,X_0=a)-\hat{\mathbb{P}}(X_S=x,X_j=c,X_0=a)|,\end{aligned}$$

the proof of Claim 1 follows from (73).

Claim 2. For any $u>0$,

$$\begin{aligned}\mathbb{P}\left(3\sum_{x\in A^S}\sum_{a\in A}\sum_{b\in A}\left|\hat{\mathbb{P}}(X_S=x,X_j=b,X_0=a)-\mathbb{P}(X_S=x,X_j=b,X_0=a)\right|>u\right) \\ \le 2|A|^{|S|+2}\exp\left\{-\frac{u^2(m-d)^2\Delta^2}{18|A|^{2(|S|+2)}m(|S|+2)^2}\right\}.\end{aligned}$$

Proof of Claim 2. It follows from the union bound and Proposition 26.

We now will conclude the proof. Let $\mathcal{S}_k=\{S\subseteq [\![-d,-1]\!] : |S|=k\}$ and observe that by the union bound

$$\mathbb{P}(G_m^c(\xi,\ell)) \le \sum_{k=0}^{\ell}\sum_{S\in\mathcal{S}_k}\sum_{j\in S^c}\mathbb{P}\left(|\bar{\nu}_{j,S}-\hat{\nu}_{j,S}|>\xi\right).$$

Combining Claims 1 and 2, it follows that

$$\mathbb{P}\left(|\bar{\nu}_{j,S}-\hat{\nu}_{j,S}|>\xi\right) \le 2|A|^{|S|+2}\exp\left\{-\frac{\xi^2(m-d)(\Delta_*)^2}{18|A|^{2(|S|+2)}m(|S|+2)^2}\right\},$$

which implies that

$$\mathbb{P}(G_m^c(\xi,\ell)) \le 2\sum_{k=0}^{\ell}\binom{d}{k}(d-k)|A|^{k+2}\exp\left\{-\frac{\xi^2(m-d)^2\Delta^2}{18|A|^{2(k+2)}m(k+2)^2}\right\}.$$

Since $\ell \leq d/2$, we can use that $\binom{d}{k} \leq \binom{d}{\ell}$ for all $0 \leq k \leq \ell$ to obtain that

$$\mathbb{P}(G_m^c(\xi, \ell)) \leq 2d(\ell+1)\binom{d}{\ell}|A|^{\ell+2} \exp\left\{-\frac{\xi^2(m-d)(\Delta_*)^2}{18|A|^{2(\ell+2)}m(\ell+2)^2}\right\},$$

and the result follows. ■

We now prove Theorem 10.

Proof [Proof of Theorem 10]

First, observe that by Theorem 9,

$$\mathbb{P}(\hat{\Lambda}_{2,n} \neq \Lambda) \leq \mathbb{P}(G_m^c(\xi_*, \ell_*)) + \mathbb{P}(\Lambda \subseteq \hat{S}_m, \hat{\Lambda}_{2,n} \neq \Lambda) \tag{74}$$

Next, notice that the second term on the right hand side of (74) can be written as

$$\mathbb{P}(\Lambda \subseteq \hat{S}_m, \hat{\Lambda}_{2,n} \neq \Lambda) = \sum_{S \subseteq [\![-d,-1]\!] : \Lambda \subseteq S, |S| \leq \ell_*} \mathbb{P}(\hat{S}_m = S, \hat{\Lambda}_{2,n} \neq \Lambda).$$

Now for any $S \in [\![-d,-1]\!]$ such that $\Lambda \subseteq S, |S| \leq \ell_*$, it follows from the union bound that

$$\mathbb{P}(\hat{S}_m = S, \hat{\Lambda}_{2,n} \neq \Lambda) \leq \sum_{j \in \Lambda} \mathbb{P}(\hat{S}_m = S, j \notin \hat{\Lambda}_{2,n}) + \sum_{j \in S \setminus \Lambda} \mathbb{P}(\hat{S}_m = S, j \in \hat{\Lambda}_{2,n}).$$

By proceeding similarly as in the proof of Item 1 of Theorem 1, one can deduce that for any $j \in S \setminus \Lambda$,

$$\mathbb{P}(\hat{S}_m = S, j \in \hat{\Lambda}_{2,n}) \leq 8|A| \left\lceil \frac{\log(\mu(n-m-d)/\alpha + 2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha} \sum_{x \in A^S} \mathbb{P}(\hat{S}_m = S, \bar{N}_{m,n}(x) > 0),$$

so that

$$\sum_{j \in S \setminus \Lambda} \mathbb{P}(\hat{S}_m = S, j \in \hat{\Lambda}_{2,n}) \leq 8(\ell_* - |\Lambda|)|A| \left\lceil \frac{\log(\mu(n-m-d)/\alpha + 2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha} \times \sum_{x \in A^S} \mathbb{P}(\hat{S}_m = S, \bar{N}_{m,n}(x) > 0).$$

Since

$$\sum_{S \subseteq [\![-d,-1]\!] : \Lambda \subseteq S, |S| \leq \ell_*} \sum_{x \in A^S} \mathbb{P}(\hat{S}_m = S, \bar{N}_{m,n}(x) > 0) \leq (n-m-d)\mathbb{P}(\Lambda \subseteq \hat{S}_m)$$

we then deduce that

$$\sum_{S \subseteq [\![-d,-1]\!] : \Lambda \subseteq S, |S| \leq \ell_*} \sum_{j \in S \setminus \Lambda} \mathbb{P}(\hat{S}_m = S, j \in \hat{\Lambda}_{2,n}) \leq 8(\ell_* - |\Lambda|)|A| \times \left\lceil \frac{\log(\mu(n-m-d)/\alpha + 2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}(n-m-d).$$

Following the steps of the proof of Item 2 of of Theorem 1, we can also show that

$$\sum_{S\subseteq[\![-d,-1]\!]:\Lambda\subseteq S,|S|\leq \ell_*}\sum_{j\in\Lambda}\mathbb{P}(\hat{S}_m=S, j\notin\hat{\Lambda}_{2,n},\delta_j\geq\gamma^S_{m,n,j})\leq 8|\Lambda||A| \times\left\lceil\frac{\log(\mu(n-m-d)/\alpha+2)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha}\sum_{S\subseteq[\![-d,-1]\!]:\Lambda\subseteq S,|S|\leq \ell_*}\mathbb{P}(\hat{S}_m=S)$$

$$\leq 8|\Lambda||A|\left\lceil\frac{\log(\mu(n-m-d)/\alpha+2)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha},$$

where $\gamma^S_{m,n,j}$ is defined as in (14) with $t_{m,n,j}=\max_{b,c\in A:b\neq c}\min_{(x_S,y_S)\in\mathcal{C}_j(b,c)}t_{m,n}(x_S,y_S)$ in the place of $t_{n,j}$.

Hence, it remains to estimate

$$\sum_{S\subseteq[\![-d,-1]\!]:\Lambda\subseteq S,|S|\leq \ell_*}\sum_{j\in\Lambda}\mathbb{P}(\hat{S}_m=S, j\notin\hat{\Lambda}_{2,n},\delta_j<\gamma^S_{m,n,j}).$$

By proceeding similarly to the proof of Item 3 of Theorem 1, one can show that for each $S\subseteq[\![-d,-1]\!]$ such that $|S|\leq\ell_*$,

$$\mathbb{P}\left(\gamma^S_{m,n,j}>\delta_j\right)\leq 6|A|(|A|-1)\exp\left\{\frac{-(\Delta p_{min})^2(n-m-d)^2}{2(n-m)|A|^{2(\ell_*-1)}(\ell_*+1)^2}\left(1-\frac{n_{min}}{n-m-d}\right)^2\right\},$$

for all $j\in\Lambda$ as long as n satisfies $n>m+d+n_{min}$. By using this upper bound and by recalling that $\sum_{S\subseteq[\![-d,-1]\!]:\Lambda\subseteq S,|S|\leq\ell_*}\leq\sum_{k=0}^{\ell_*}\binom{d}{k}\leq(\ell_*+1)\binom{d}{\ell_*}$ (since $2\ell_*\leq d$), we deduce that

$$\sum_{S\subseteq[\![-d,-1]\!]:\Lambda\subseteq S,|S|\leq \ell_*}\sum_{j\in\Lambda}\mathbb{P}(\hat{S}_m=S, j\notin\hat{\Lambda}_{2,n},\delta_j<\gamma^S_{m,n,j})\leq$$

$$6|A|(|A|-1)(\ell_*+1)\binom{d}{\ell_*}\exp\left\{\frac{-(\Delta p_{min})^2(n-m-d)^2}{2(n-m)|A|^{2(\ell_*-1)}(\ell_*+1)^2}\left(1-\frac{n_{min}}{n-m-d}\right)^2\right\},$$

for all $j\in\Lambda$ whenever $n>m+d+n_{min}$, implying the result. ■

A.2.4 Proof of Corollary 12

Proof [Proof of Corollary 12] Assumptions 1 and 2 are satisfied for all values of n, since $p^\star_n\geq p^\star_{min}$ and $\Delta^\star_n\geq\Delta^\star_{min}$ for positive constants $p^\star_{min}$ and $\Delta^\star_{min}$ for all n. Since the sequence of MTD models also satisfy Assumption 3, the result follows from Theorem 10 and Remark 11-Item (b). ■

A.3 Proofs of Section 3.3

A.3.1 Proof of Theorem 14

Proof [Proof of Theorem 14]

Notice that we can write for each $j \in [\![-d,-1]\!]$,

$$m_j(X_j) = (p_j(1|1) - p_j(1|0)X_{-j} + p_j(1|0),$$

so that equality (50) can be rewritten for any $j \in [\![-d,-1]\!]$ satisfying $(p_j(1|1) - p_j(1|0)) \neq 0$, $S \subseteq [\![-d,-1]\!] \setminus \{j\}$ and $x_S \in \{0,1\}^S$, as

$$\mathrm{Cov}_{x_S}(X_0, X_j) = \sum_{\ell \in \Lambda \setminus S} \Delta_\ell \mathrm{Cov}_{x_S}(X_\ell, X_j),$$

where $\Delta_\ell = \lambda_\ell(p_\ell(1|1) - p_\ell(1|0))$ for $\ell \in \Lambda$. Recalling that $\bar{\nu}_{j,S} = 2\mathbb{E}\left(|\mathrm{Cov}_{X_S}(X_0, X_j)|\right)$ in the binary case, we can deduce that for any $S \subseteq [\![-d,-1]\!]$, $x_S \in \{0,1\}^S$ and $j \in [\![-d,-1]\!] \setminus S$,

$$\bar{\nu}_{j,S} = 2\mathbb{E}\left(\left|\sum_{\ell \in \Lambda \setminus S} \Delta_\ell \mathrm{Cov}_{X_S}(X_\ell, X_j)\right|\right).$$

As a consequence, it follows that for $S \subset \Lambda$ and $j \in \Lambda \setminus S$,

$$\begin{aligned}\bar{\nu}_{j,S} \geq 2\mathbb{E}&\left(\mathrm{Var}_{X_S}(X_j)|\Delta_j| - \sum_{\ell \in \Lambda \setminus S : \ell \neq j} |\Delta_\ell||\mathrm{Cov}_{X_S}(X_\ell, X_j)|\right) \\ &= 2\mathbb{E}\left(|\Delta_j|\mathrm{Var}_{X_S}(X_j) \times\right. \\ &\left.\left(1 - \sum_{\ell \in \Lambda \setminus S : \ell \neq j} \frac{|\Delta_\ell|}{|\Delta_j|} |\mathbb{P}_{x_S}(X_\ell = 1 | X_j = 1) - \mathbb{P}_{x_S}(X_\ell = 1 | X_j = 0)|\right)\right),\end{aligned}$$

where in the second inequality we have used that

$$|\mathrm{Cov}_{x_S}(X_\ell, X_j)| = \mathrm{Var}_{x_S}(X_j)|\mathbb{P}_{x_S}(X_\ell = 1 | X_j = 1) - \mathbb{P}_{x_S}(X_\ell = 1 | X_j = 0)|.$$

By Assumption 4, we then deduce that

$$\bar{\nu}_{j,S} \geq 2\Gamma_1 |\Delta_j| \mathbb{E}(\mathrm{Var}_{X_S}(X_j)). \tag{75}$$

Now, take $S \subset \Lambda$ and let $j_S \in \arg\min_{\ell \in \Lambda \setminus S} |\Delta_j| \mathbb{E}(\mathrm{Var}_{X_S}(X_\ell))$. For any $j \in (\Lambda)^c$, use the triangle inequality, equality (75) and Assumption 5 to deduce that

$$\begin{aligned}\bar{\nu}_{j,S} \leq 2\mathbb{E}&\left(\sum_{\ell \in \Lambda \setminus S} |\Delta_\ell||\mathrm{Cov}_{X_S}(X_\ell, X_j)|\right) \\ &\leq 2|\Delta_{j_S}| \mathbb{E}\left(\mathrm{Var}_{X_S}(X_{j_S})\right)\Gamma_2\end{aligned} \tag{76}$$

Using that $\delta_j = |\Delta_j|$ and combing inequalities (75) and (76), it follows then that

$$\max_{j \in \Lambda \setminus S} \bar{\nu}_{j,S} - \max_{j \in (\Lambda)^c} \bar{\nu}_{j,S} \geq 2(\Gamma_1 - \Gamma_2)|\Delta_{j_S}| \mathbb{E}\left(\mathrm{Var}_{X_S}(X_{j_S})\right), \tag{77}$$

where we have used also that $\max_{j\in\Lambda\setminus S}\bar{\nu}_{j,S} \geq \bar{\nu}_{j_S,S}$. Using that $\mathbb{E}\left(\mathrm{Var}_{X_S}(X_{j_S})\right) \geq (p^\star)^2$ and $|\Delta_{j_S}| \geq \min_{j\in\Lambda}|\Delta_j|$ we obtain that

$$\min_{S\subset\Lambda}\left(\max_{j\in\Lambda\setminus S}\bar{\nu}_{j,S} - \max_{j\in(\Lambda)^c}\bar{\nu}_{j,S}\right) \geq 2(\Gamma_1-\Gamma_2)p_{min}^2\min_{j\in\Lambda}|\Delta_j|.$$

concluding the first half of the proof.

To show the second assertion of the theorem, take $S\subset\Lambda$, let $j_S^* \in \arg\max_{j\in\Lambda\setminus S}\bar{\nu}_{j,S}$ and note that on $G_n(\ell,\xi)$,

$$\max_{j\in\Lambda\setminus S}\hat{\nu}_{n,j,S} \geq \hat{\nu}_{n,j_S^*,S} \geq \bar{\nu}_{j_S^*,S} - \xi = \max_{j\in\Lambda\setminus S}\bar{\nu}_{j,S} - \xi.$$

Similarly, one can show that on $G_n(\ell,\xi)$,

$$\max_{j\in(\Lambda)^c}\hat{\nu}_{n,j,S} \leq \max_{j\in(\Lambda)^c}\bar{\nu}_{j,S} + \xi.$$

As a consequence, it follows that

$$\max_{j\in\Lambda\setminus S}\hat{\nu}_{n,j,S} - \max_{j\in(\Lambda)^c}\hat{\nu}_{n,j,S} \geq \left(\max_{j\in\Lambda\setminus S}\bar{\nu}_{j,S} - \max_{j\in(\Lambda)^c}\bar{\nu}_{j,S}\right) - 2\xi,$$

whenever $G_n(\ell,\xi)$. By taking ξ as in (37), we have that

$$\max_{j\in\Lambda\setminus S}\hat{\nu}_{n,j,S} - \max_{j\in(\Lambda)^c}\hat{\nu}_{n,j,S} > 0,$$

implying that $\arg\max_{j\in S^c}\hat{\nu}_{n,j,S} \in \Lambda$ for all $S\subset\Lambda$, and the result follows. ■

A.3.2 Proof of Corollary 16

Proof [Proof of Corollary 16] By Theorem 14, we have that

$$\mathbb{P}(\hat{\Lambda}_{2,n} \neq \Lambda) \leq \mathbb{P}(G_m^c(\xi,L)) + \mathbb{P}(\Lambda \subseteq \hat{S}_m, \hat{\Lambda}_{2,n} \neq \Lambda),$$

for any $\xi < (\Gamma_1-\Gamma_2)p_{min}^\star\min_{j\in\Lambda}\delta_j$.

By Proposition 25, we have that

$$\mathbb{P}(G_m^c(\xi,L)) \leq 2d(L+1)\binom{d}{L}|A|^{L+2}\exp\left\{-\frac{\xi^2(m-d)^2\Delta^2}{18|A|^{2(L+2)}m(L+2)^2}\right\}.$$

By taking $\xi = (\Gamma_1-\Gamma_2)p_{min}^\star\min_{j\in\Lambda}\delta_j$, one can check that if

$$\min_{j\in\Lambda}\delta_j \geq C_1\frac{\log(n)}{n},$$

for some constant $C_1 = C_1(\beta, L, \Delta_{min}^*, p_{min}^*, \Gamma_1, \Gamma_2)$, then $\mathbb{P}(G_m^c(\xi,L)) \to 0$ as $n\to\infty$.

By proceeding exactly as in the proof of Theorem 10 and using 11-item (b), we can show that

$$\begin{aligned}\mathbb{P}(\Lambda \subseteq \hat{S}_m, \hat{\Lambda}_{2,n} \neq \Lambda) \leq \\ 6|A|(L+1)\binom{d}{L}\left[d|A|^{L+1}(|A|-1)|\Lambda| \exp\left\{-\frac{(\Delta p_{min})^2(n-d)^2}{8n(L+1)^2|A|^{2(L-1)}}\right\}\right] \\ + 8|A|\left((L-|\Lambda|)(n-m-d)+|\Lambda|\right)\left\lceil\frac{\log(\mu(n-m-d)/\alpha+2)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha},\end{aligned}$$

as long as

$$n \geq d + \frac{C|A|^L\alpha}{p_{min}\delta^2_{min}},$$

where $C = C(\mu, \varepsilon)$.

Therefore, using that $\Delta \geq \Delta^\star_{min}$, $p_{min} \geq p^\star_{min}$, $d = n\beta$, $\alpha = (1+\eta)\log(n)$, we also see that if $\delta^2_{min} \geq C_2\frac{\log(n)}{n}$ for some $C_2 = C_2(\mu, \varepsilon, \eta, \Delta^\star_{min}, p^\star_{min}, \beta, L)$ then $\mathbb{P}(\Lambda \subseteq \hat{S}_m, \hat{\Lambda}_{2,n} \neq \Lambda) \to 0$ as $n \to \infty$.

By taking $C = C_1 \vee C_2$, we deduce that $\mathbb{P}(\hat{\Lambda}_{2,n} \neq \Lambda) \to 0$ as $n \to \infty$ as long as $\delta^2_{min} \geq C\frac{\log(n)}{n}$, and the result follows.

■

A.4 Proofs of Section 3.4

Proof [Proof of Theorem 17]

By the union bound, we have that

$$\begin{aligned}\mathbb{P}\left(\bigcup_{a\in A}\bigcup_{x_{-d:-1}\in A^{[\![-d,-1]\!]}}\left\{|\hat{p}_n(a|x_{\hat{\Lambda}_{2,n}}) - p(a|x_\Lambda)| \geq \sqrt{\frac{2\alpha(1+\epsilon)\hat{V}_n(a, x_{\hat{\Lambda}_{2,n}})}{\bar{N}_n(x_{\hat{\Lambda}_{2,n}})}}\right.\right. \\ \left.\left.+\frac{\alpha}{3\bar{N}_n(x_{\hat{\Lambda}_{2,n}})}\right\}\right) \leq \mathbb{P}(\Lambda \neq \hat{\Lambda}_{2,n}) \\ + \sum_{a\in A}\sum_{x_\Lambda\in A^\Lambda}\mathbb{P}\left(|\hat{p}_n(a|x_\Lambda) - p(a|x_\Lambda)| \geq \sqrt{\frac{2\alpha(1+\epsilon)\hat{V}_n(a,x_\Lambda)}{\bar{N}_n(x_\Lambda)}} + \frac{\alpha}{3\bar{N}_n(x_\Lambda)}\right).\end{aligned}$$

Now, Proposition 34 implies that for any $a \in A$ and $x_\Lambda \in A^\Lambda$,

$$\begin{aligned}\mathbb{P}\left(|\hat{p}_n(a|x_\Lambda) - p(a|x_\Lambda)| \geq \sqrt{\frac{2\alpha(1+\epsilon)\hat{V}_n(a,x_\Lambda)}{\bar{N}_n(x_\Lambda)}} + \frac{\alpha}{3\bar{N}_n(x_\Lambda)}\right) \\ \leq 4\left\lceil\frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha}\mathbb{P}\left(\bar{N}_n(x_\Lambda) > 0\right),\end{aligned}$$

so that

$$\sum_{a\in A}\sum_{x_\Lambda\in A^\Lambda}\mathbb{P}\left(|\hat{p}_n(a|x_\Lambda)-p(a|x_\Lambda)|\geq\sqrt{\frac{2\alpha(1+\epsilon)\hat{V}_n(a,x_\Lambda)}{\bar{N}_n(x_\Lambda)}}+\frac{\alpha}{3\bar{N}_n(x_\Lambda)}\right)$$
$$\leq 4|A|\left\lceil\frac{\log(\mu(n-d)/\alpha+2)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha}\times\sum_{x_\Lambda\in A^\Lambda}\mathbb{E}\left[1\{\bar{N}_n(x_\Lambda)>0\}\right].$$

Since

$$\sum_{x_\Lambda\in A^\Lambda}\mathbb{E}[1\{\bar{N}_n(x_\Lambda)>0\}]\leq(n-d),$$

the result follows. ■

A.5 Proof of Section 3.5

A.5.1 Proof of Proposition 18

Proof [Proof of Proposition 18]

First observe that since all MTDs are stationary Markov chains of order at most d, we can use the Markov property to show that

$$KL(P_n^{(j)}||P_n^{(k)})=KL(P_d^{(j)}||P_d^{(k)})+(n-d)\mathbb{E}^{(j)}(KL(p^{(j)}(\cdot|X_{-d:-1})||p^{(k)}(\cdot|X_{-d:-1}))).$$

where $KL(p^{(j)}(\cdot|x_{-d:-1})||p^{(k)}(\cdot|x_{-d:-1}))$ denotes the Kullback-Leibler divergence between $p^{(j)}(\cdot|x_{-d:-1})$ and $p^{(k)}(\cdot|x_{-d:-1})$.

Now note that for each fixed $x_{-:d_1}\in A^{[\![-d,-1]\!]}$, we can use the definition of the transition probabilities $p^{(j)}(\cdot|\cdot)$ together with Lemma 6 of Csizar and Talata (2005) to deduce that

$$KL(p^{(j)}(\cdot|x_{-d:-1})||p^{(k)}(\cdot|x_{-d:-1}))\leq\lambda^2|p(1|1)-p(1|0)|^2(p_{min})^{-1}1_{\{x_j\neq x_k\}}.$$

Since $p_{min}\geq(1-\lambda)/2$ and $\delta=\lambda|p(1|1)-p(1|0)|$, it follows from the above inequality that

$$\mathbb{E}^{(j)}(KL(p^{(j)}(\cdot|X_{-d:-1})||p^{(k)}(\cdot|X_{-d:-1})))\leq\frac{2\delta^2}{1-\lambda}.$$

By using similar arguments, one can also show that

$$KL(P_d^{(j)}||P_d^{(k)})\leq\frac{1}{p_{min}}\left(\max_{x_{-d:-1},y_{-d:-1}}|p^{(j)}(1|x_{-d:-1})-p^{(k)}(1|y_{-d:-1})|^2\right.$$
$$\left.+\sum_{i=1}^{d-1}\max_{x_{-d:-1},y_{-d:-1}:x_{-i:-1}=y_{-i:-1}}|p^{(j)}(1|x_{-d:-1})-p^{(k)}(1|y_{-d:-1})|^2\right)\leq\frac{2d\delta^2}{1-\lambda}.$$

Therefore, it follows that

$$KL(P_n^{(j)}||P_n^{(k)})\leq\frac{2d\delta^2}{1-\lambda}+(n-d)\frac{2\delta^2}{1-\lambda}=\frac{2n\delta^2}{1-\lambda},$$

and the result follows. ■

A.6 Computation of PCP and FSC estimators

We will first show that one can compute the PCP estimator with at most $O(|A|^2|S|(n-d))$ computations, as claimed in item (c) of Remark 2.

Proof of item (c) of Remark 2. One way to compute the PCP estimator is the following. First, we compute $N_n(x_S, a)$ simultaneously for all pasts x_S and symbols $a \in A$, and build the set $E_S = \{x_S : \bar{N}_n(x_S) > 0\}$. This can be done with $O(n-d)$ computations. Indeed, we set initially $N_n(x_S, a) = 0$ for all past x_S and symbol $a \in A$. Then at each time $d+1 \leq t \leq n$, we increment by 1 the count of $N_n(x_S, a)$ for which $X_{t+S} = x_S$ and $X_t = a$, leaving all the other counts unchanged. Moreover, at the first time that $N_n(x_S, a) > 0$, we include x_S in the set E. Note that the cardinality of the set E_S is at most $(n-d)$. Next, we need to compute $s_n(x_S)$ and $\hat{p}_n(\cdot|x_S)$ for each $x_S \in E$, which can be done with at most $O(|A|)$ additional computations. Once all these quantities are determined, we then need to test whether a given lag $j \in S$ has to be removed or not, by evaluating inequality (13) for all pairs of $(S \setminus \{j\})$-compatible pasts in E_S. This can be done with at most $O(|A|^2(n-d))$ more computations because 1) the number of different pasts in E_S is at most $(n-d)$; 2) there are at most $|A|$ pasts in E which are compatible with a fixed past x_S in E_S; and 3) one can evaluate whether inequality (13) holds or not to a given pair of compatible past with $O(|A|)$ additional computations. Finally, since the number of lags to be tested is $|S|$, it follows that we can implement the PCP estimator with at most $O(|A|^2|S|(n-d))$ computations, concluding the proof.

We now show that we can compute the FSC estimator by using at most $O(|A|^3\ell(m-d)(d-(\ell-1)/2) + |A|^2(n-m-d)\ell)$ computations, as stated in item (c) of Remark 8.

Proof of item (c) of Remark 8.

By the item (c) of Remark 2, the CUT step can be computed with at most $O(\ell|A|^2(n-m-d))$ computations since the FS step outputs a subset of size ℓ and the size of the second half of the sample is $n-m$. Hence, the proof will be concluded if we show that the FS step can be computed with at most $O(|A|^3(m-d)(\ell d - (\ell-1)\ell/2)$ computations. To see that, let us fix $S \subseteq [\![-d, -1]\!]$ and $j \notin S$. Proceeding as in the proof item (c) of Remark 2, one can check that we compute $N_m(xba_{S\cup\{0,j\}})$ simultaneously for all configurations $xba_{S\cup\{j,0\}}$ and build the set $E_S = \{x_S : N_m(x_S) > 0\}$ with $O(m-d)$ computations. Notice that the size of the set E_S is most $(m-d)$. Since for each $x_S \in E_S$, we need to perform at most $O(|A|^3)$ additional operations to compute $\hat{\mathbb{P}}_m(x_S)$ and $\hat{\nu}_{j,m}(x_S)$, it follows that with at most $O(|A|^3(m-d))$ computations we can determine $\hat{\nu}_{m,j,S}$. Therefore, the step 3 of the FS step (where we need to compute $\hat{\nu}_{m,j,S}$ for $j \in S^c$) can be implemented with $O(|A|^3(m-d)(d-|S|))$ calculations. Since we need to repeat step 3 of the FS step for ℓ different sets, we conclude that with at most

$$O(|A|^3(m-d)\sum_{|S|=0}^{\ell-1}(d-|S|)) = O(|A|^3(m-d)(\ell d - (\ell-1)\ell/2)$$

computations, we can implement the FS step. This concludes the proof.

Appendix B. Martingale concentration inequalities

In the sequel, $\mathbb{N}$ denotes the set of non-negative integers $\{0, 1, \ldots\}$ Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We assume that this probability space is rich enough so that the following stochastic processes may be defined on it. In what follows, let $(X_t)_{t\in\mathbb{Z}}$ be a Markov chain of order $d \in \mathbb{Z}_+$, taking values on a finite alphabet A, with family of transition probabilities $\{p(\cdot|x_{-d:-1}) : x_{-d:-1} \in \text{supp}(\mathbf{P})\}$. Denote $\mathcal{F}_t = \sigma(X_{-d:t})$ for $t \in \mathbb{N}$. For each $a \in A$, consider the stochastic process $M^a = ((M^a_t))_{t\in\mathbb{N}}$ defined as,

$$M^a_t = 1\{X_t = a\} - p(a|X_{(t-d):(t-1)}),\ t \in \mathbb{N}.$$

Let $H = (H_t)_{t\in\mathbb{N}}$ be a stochastic process taking values on a finite alphabet $B \subset \mathbb{R}$, satisfying $H_0 = 0$ and $H_t \in \mathcal{F}_{t-1}$ for all $t \in \mathbb{Z}_+$, and consider $H \bullet M^a = (H \bullet M^a_t)_{t\in\mathbb{N}}$ defined as,

$$H \bullet M^a_t = \sum_{s=0}^{t} H_s M^a_s,\ t \in \mathbb{N}. \tag{78}$$

Notice that $H \bullet M^a$ is adapted to the fitration $\mathbb{F} := (\mathcal{F}_t)_{t\in\mathbb{N}}$, that is $H \bullet M^a_t \in \mathcal{F}_t$ for all $t \in \mathbb{N}$. Also $H \bullet M^a_0 = 0$. Recall the notation $\|B\|_\infty = \max_{b\in B} |b|$.

Lemma 27 *Let $H \bullet M^a = (H \bullet M^a_t)_{t\in\mathbb{N}}$ be the stochastic process defined in* (78). *Then $H \bullet M^a$ is a square integrable Martingale w.r.t. $\mathbb{F}$ starting from $H \bullet M^a_0 = 0$. Moreover, the predictable quadratic variation of $H \bullet M^a$, denoted by $\langle H \bullet M^a\rangle = (\langle H \bullet M^a\rangle_t)_{t\in\mathbb{N}}$, is given by*

$$\langle H \bullet M^a\rangle_t = \sum_{s=0}^{t} H^2_s p\left(a|X_{(s-d):(s-1)}\right)\left(1 - p\left(a|X_{(s-d):(s-1)}\right)\right),\ t \in \mathbb{N}. \tag{79}$$

Furthermore, for any $\lambda > 0$ and $b > 0$ such that $\|B\|_\infty \le b$, the stochastic process

$$\exp\left(\lambda H \bullet M^a - \frac{e^{\lambda b} - \lambda b - 1}{b^2}\langle H \bullet M^a\rangle\right) = \left(\exp\left(\lambda H \bullet M^a_t - \frac{e^{\lambda b} - \lambda b - 1}{b^2}\langle H \bullet M^a\rangle_t\right)\right)_{t\in\mathbb{N}}$$

is a supermartingale w.r.t. $\mathbb{F}$ starting from 1.

Proof For each $t \in \mathbb{Z}_+$, we have that $H_t \in \mathcal{F}_{t-1}$ and also that $\mathbb{E}\left[1\{X_t = a\}|\mathcal{F}_{t-1}\right] = p(a|X_{(t-d):(t-1)})$. These two facts imply that for any $t \in \mathbb{Z}_+$,

$$\mathbb{E}\left[H_t M^a_t|\mathcal{F}_{t-1}\right] = H_t \mathbb{E}\left[M^a_t|\mathcal{F}_{t-1}\right] = 0,$$

which, in turn, implies that $E[H \bullet M^a_t|\mathcal{F}_{t-1}] = H \bullet M^a_{t-1}$. Hence, $H \bullet M^a$ is a martingale w.r.t. to $\mathbb{F}$. Since $|H \bullet M^a_t| \le \|B\|_\infty t$ for $t \ge 1$, it follows that $H \bullet M^a$ is also square integrable.

The predictable quadratic variation of $H \bullet M^a$ is defined as

$$\langle H \bullet M^a\rangle_t = \sum_{s=1}^{t} \mathbb{E}\left((M_s - M_{s-1})^2 |\mathcal{F}_{s-1}\right),$$

for $t \in \mathbb{Z}_+$ with $\langle H \bullet M^a\rangle_0 = 0$. For any $t \in \mathbb{Z}_+$, one can check that

$$\left(H \bullet M^a_t - H \bullet M^a_{t-1}\right)^2 = H^2_t(1\{X_t = a\} - 2p(a|X_{(t-d):(t-1)})1\{X_t = a\} - p^2(a|X_{(t-d):(t-1)})).$$

Using again that $H_t \in \mathcal{F}_{t-1}$ and also that $\mathbb{E}\left[1\{X_t = a\}|\mathcal{F}_{t-1}\right] = p(a|X_{(t-d):(t-1)})$, one then deduces that for any $t \in \mathbb{Z}_+$,

$$\mathbb{E}\left(\left(H \bullet M^a_t - H \bullet M^a_{t-1}\right)^2 |\mathcal{F}_{t-1}\right) = H_t^2 p(a|X_{(t-d):(t-1)})(1 - p(a|X_{(t-d):(t-1)})),$$

which establishes (79). The proof that $\exp\left(\lambda H \bullet M^a - \frac{e^{\lambda b} - \lambda b - 1}{b^2}\langle H \bullet M^a\rangle\right)$ is a supermartingale w.r.t $\mathbb{F}$ can be found in (Raginsky and Sason, 2014). ■

We will use Lemma 27 to prove the following concentration inequality.

Proposition 28 *Let $H \bullet M^a = (H \bullet M^a_t)_{t \in \mathbb{N}}$ be the stochastic process defined in (78). Suppose that $\|B\|_\infty \leq b$ for some $b > 0$. For any fixed $\alpha > 0$ and $v > 0$, we have for $t \in \mathbb{N}$,*

$$\mathbb{P}\left(H \bullet M^a_t \geq \sqrt{2v\alpha} + \frac{\alpha b}{3}, \langle H \bullet M^a\rangle_t \leq v\right) \leq \exp(-\alpha)\,\mathbb{P}(\langle H \bullet M^a\rangle_t > 0).$$

Remark 29 *This is basically Lemma 5 of (Oliveira, 2015) (see the Economical Freedman's inequality provided in Inequality (41)) applied to the square integrable martingale $H \bullet M^a$. The only difference is the factor 2 in front of the linear term $\frac{\alpha b}{3}$ which is not present here. Notice that for $t \in \mathbb{Z}_+$, the concentration inequality above can be rewritten in the following form:*

$$\mathbb{P}\left(H \bullet M^a_t \geq \sqrt{2v\alpha} + \frac{\alpha b}{3}, \langle H \bullet M^a\rangle_t \leq v | \langle H \bullet M^a\rangle_t > 0\right) \leq \exp(-\alpha).$$

The conditioning on event $\{\langle H \bullet M^a\rangle_t > 0\}$ reflects the fact that if $\langle H \bullet M^a\rangle_t = 0$ almost surely, then $H \bullet M^a_t = H \bullet M^a_0 = 0$ almost surely as well.

Proof For $t = 0$ the result holds trivially. Now, suppose $t \in \mathbb{Z}_+$. By considering the set $B/b = \{c/b : c \in B\}$ instead of B, it suffices to prove the case $b = 1$. To shorten the notation, we denote $M_t = H \bullet M^a_t$ in the sequel. By the Markov property, we have that for any $\lambda > 0$,

$$\mathbb{P}(\lambda M_t - \psi(\lambda)\langle M\rangle_t \geq \alpha) \leq \exp(-\alpha)\mathbb{E}\left[\exp\left(\lambda M_t - \psi(\lambda)\langle M\rangle_t\right) 1\{\langle M\rangle_t > 0\}\right],$$

where $\psi(\lambda) = e^\lambda - \lambda - 1$ and we have used that if $\langle M\rangle_t = 0$ almost surely, then $M_t = M_0 = 0$ almost surely. By using the fact that $(\exp(\lambda M_t - \psi(\lambda)\langle M\rangle_t))_{t \in \mathbb{N}}$ is a supermartingale (Lemma 27 with $b = 1$) together with the decomposition

$$\{\langle M\rangle_t > 0\} = \bigcup_{k=1}^{t}\{\langle M\rangle_k > 0 \text{ and } \langle M\rangle_j = 0 \text{ for all } j < k\},$$

as in (Oliveira, 2015), we can deduce that

$$\mathbb{E}\left[\exp\left(\lambda M_t - \psi(\lambda)\langle M\rangle_t\right) 1\{\langle M\rangle_t > 0\}\right] \leq \mathbb{P}(\langle M\rangle_t > 0),$$

which implies not only that for any $\lambda > 0$,

$$\mathbb{P}\left(M_t \geq \frac{\psi(\lambda)}{\lambda}\langle M\rangle_t + \frac{\alpha}{\lambda}\right) \leq \exp(-\alpha)\mathbb{P}(\langle M\rangle_t > 0), \tag{80}$$

but also that

$$\mathbb{P}\left(M_t \geq \frac{\psi(\lambda)}{\lambda}v + \frac{\alpha}{\lambda}, \langle M\rangle_t \leq v\right) \leq \exp(-\alpha)\mathbb{P}(\langle M\rangle_t > 0).$$

Now, we use that for $\lambda \in (0,3)$ it holds that $\psi(\lambda) \leq \lambda^2(1-\lambda/3)^{-1}/2$. Hence, from the above inequalities we deduce that for any $\lambda \in (0,3)$,

$$\mathbb{P}\left(M_t \geq \frac{\lambda}{2(1-\lambda/3)}\langle M\rangle_t + \frac{\alpha}{\lambda}\right) \leq \exp(-\alpha)\mathbb{P}(\langle M\rangle_t > 0), \tag{81}$$

and also that

$$\mathbb{P}\left(M_t \geq \frac{\lambda}{2(1-\lambda/3)}v + \frac{\alpha}{\lambda}, \langle M\rangle_t \leq v\right) \leq \exp(-\alpha)\mathbb{P}(\langle M\rangle_t > 0).$$

Minimizing $\lambda \in (0,3) \mapsto \frac{\lambda}{(1-\lambda/3)}v + \frac{\alpha}{\lambda}$, the result follows. ■

By using a peeling argument as in (Hansen et al., 2015), we deduce from the above result the following.

Proposition 30 *Let $H \bullet M^a = (H \bullet M^a_t)_{t\in\mathbb{N}}$ be the stochastic process defined in* (78). *Suppose that $\|B\|_\infty \leq b$ for some $b > 0$. For $\epsilon > 0$, $v > w > 0$ and $\alpha > 0$, we have for $t \in \mathbb{N}$,*

$$\mathbb{P}\left(H \bullet M^a_t \geq \sqrt{2\alpha(1+\epsilon)\langle H \bullet M^a\rangle_t} + \frac{\alpha b}{3}, w \leq \langle H \bullet M^a\rangle_t \leq v\right) \leq \left\lceil \frac{\log(v/w+1)}{\log(1+\varepsilon)} \right\rceil \exp\left(-\alpha\right)\mathbb{P}(\langle H \bullet M^a\rangle_t > 0).$$

Proof It suffices to prove the case $b = 1$. The general case follows from this one by first replacing B by $B/b = \{c/b : c \in B\}$ and then rearranging the terms properly. Let us denote $v_0 = w$ and $v_k = (1+\varepsilon)v_{k-1}$ for $1 \leq k \leq K := \left\lceil \frac{\log(v/w+1)}{\log(1+\varepsilon)} \right\rceil$. Notice that $v_K \geq v$, by the definition of K. To shorten the notation, we denote $H \bullet M^a_t = M_t$ in what follows.

Starting from (81), one can deduce that for any $0 \leq k < K$ and $\lambda \in (0,3)$, we have

$$\mathbb{P}\left(M_t \geq \frac{\lambda}{2(1-\lambda/3)}\langle M\rangle_t + \frac{\alpha}{\lambda}, v_k \leq \langle M\rangle_t \leq v_{k+1}\right) \leq \exp(-\alpha)\mathbb{P}(\langle M\rangle_t > 0),$$

which, in turn, implies that

$$\mathbb{P}\left(M_t \geq \frac{\lambda}{2(1-\lambda/3)}v_{k+1} + \frac{\alpha}{\lambda}, v_k \leq \langle M\rangle_t \leq v_{k+1}\right) \leq \exp(-\alpha)\mathbb{P}(\langle M\rangle_t > 0).$$

Minimizing w.r.t. to $\lambda \in (0,3)$ as in Proposition 28, it then follows that

$$\mathbb{P}\left(M_t \geq \sqrt{2v_{k+1}\alpha} + \frac{\alpha}{3}, v_k \leq \langle M\rangle_t \leq v_{k+1}\right) \leq \exp\left(-\alpha\right)\mathbb{P}(\langle M\rangle_t > 0).$$

Now, on the event $\{v_k \leq \langle M\rangle_t\}$, we have that $\langle M\rangle_t(1+\varepsilon) \geq (1+\varepsilon)v_k = v_{k+1}$, so that the inequality above implies

$$\mathbb{P}\left(M_t \geq \sqrt{2(1+\varepsilon)\langle M\rangle_t\alpha} + \frac{\alpha}{3}, v_k \leq \langle M\rangle_t \leq v_{k+1}\right) \leq \exp\left(-\alpha\right)\mathbb{P}(\langle M\rangle_t > 0).$$

Summing over k the result follows (recall that $v \leq v_K$ by the choice of K). ■

Hereafter, let $m \in \mathbb{N}$ and consider a function $\varphi : A^{[\![1,m]\!]} \times A^{[\![-d,-1]\!]}$ such that its supremum norm $\|\varphi\|_\infty = \max_{(x_{1:m},x_{-d:-1})\in A^{[\![1,m]\!]}\times A^{[\![-d,-1]\!]}} |\varphi(x_{1:m}, x_{-d:-1})| \leq b$. Here we use the convention that φ is a function defined only on $A^{[\![-d,-1]\!]}$ when $m = 0$. Given such a function φ, let us denote $H^\varphi = (H^\varphi_t)_{t\geq 0}$ the stochastic process defined as $H^\varphi_0 = \ldots = H^\varphi_{m+d} = 0$ and $H^\varphi_t = \varphi(X_{1:m}, X_{(t-d):(t-1)})$ for $t \geq d+m+1$.

Clearly, $H^\varphi_t \in \mathcal{F}_{t-1}$ for all $t \in \mathbb{Z}_+$. From (79), one can check that the predictable quadratic variation $\langle H^\varphi \bullet M^a\rangle$ of the martingale $H^\varphi \bullet M^a$ is given by $\langle H^\varphi \bullet M^a\rangle_0 = \ldots = \langle H^\varphi \bullet M^a\rangle_{m+d} = 0$ and for $t \geq m+d+1$,

$$\langle H^\varphi \bullet M^a\rangle_t = \sum_{s=m+d+1}^{t} \varphi^2(X_{1:m}, X_{(s-d):(s-1)})p\left(a|X_{(s-d):(s-1)}\right)\left(1 - p\left(a|X_{(s-d):(s-1)}\right)\right). \tag{82}$$

As a direct consequence of Proposition 30, we derive the following result.

Corollary 31 *Let $X_{1:n}$ be a sample from a MTD model of order d with set of relevant lags Λ. Let $\hat{\Lambda}_m$ be an estimator of Λ computed from $X_{1:m}$, where $n > m$. For each $x \in A^{[\![-d,-1]\!]}$, $a \in A$ and $S \subseteq [\![-d,-1]\!]$, let $\hat{p}_{m,n}(a|x_S)$ be the empirical transition probability defined in (7) computed from $X_{m+1:n}$. Then for any $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \subseteq S$, $\varepsilon > 0$, $\alpha > 0$ and $n \geq m+d+1$, we have*

$$\mathbb{P}\left(\hat{\Lambda}_m = S, |\hat{p}_{m,n}(a|x_S) - p(a|x_\Lambda)| \geq \sqrt{\frac{2\alpha(1+\varepsilon)p(a|x_\Lambda)(1-p(a|x_\Lambda))}{\bar{N}_{m,n}(x_S)}} + \frac{\alpha}{3\bar{N}_{m,n}(x_S)}\right) \leq 2\left\lceil\frac{\log(n-m-d+1)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha}\mathbb{P}\left(\bar{N}_{m,n}(x_S) > 0, \hat{\Lambda}_m = S\right). \tag{83}$$

In particular,

$$\mathbb{P}\left(\Lambda \subseteq \hat{\Lambda}_m, |\hat{p}_{m,n}(a|x_{\hat{\Lambda}_m}) - p(a|x_\Lambda)| \geq \sqrt{\frac{2\alpha(1+\varepsilon)p(a|x_\Lambda)(1-p(a|x_\Lambda))}{\bar{N}_{m,n}(x_{\hat{\Lambda}_m})}} + \frac{\alpha}{3\bar{N}_{m,n}(x_{\hat{\Lambda}_m})}\right) \leq 2\left\lceil\frac{\log(n-m-d+1)}{\log(1+\varepsilon)}\right\rceil e^{-\alpha}\mathbb{P}\left(\bar{N}_{m,n}(x_{\hat{\Lambda}_m}) > 0, \Lambda \subseteq \hat{\Lambda}_m\right). \tag{84}$$

Proof Summing in both sides of (83) over $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \subseteq S$, we obtain inequality (84). Thus, it remains to prove (83). To that end, take $\varphi(X_{1:m}, X_{(t-d):(t-1)}) = 1\{\hat{\Lambda}_m = S\}1\{X_{t+j} = x_j, j \in S\}$ and notice that in this case

$$\langle H^{\varphi} \bullet M^a \rangle_n = 1\{\hat{\Lambda}_m = S\}\bar{N}_{m,n}(x_S)p(a|x_\Lambda)(1-p(a|x_\Lambda)).$$

So, if either $p(a|x_\Lambda) = 0$ or $p(a|x_\Lambda) = 1$, then we have necessarily $\langle H^{\varphi} \bullet M^a \rangle_n = 0$ for all $n \geq m+d+1$, which implies that almost surely for all $n \geq m+d+1$,

$$H^{\varphi} \bullet M^a_n = 1\{\hat{\Lambda}_m = S\}(\bar{N}_{m,n}(x_S, a) - \bar{N}_{m,n}(x_S)p(a|x_\Lambda)) = 0.$$

By noticing that $H^{\varphi} \bullet M^a_n = 1\{\hat{\Lambda}_m = S\}\bar{N}_{m,n}(x_S)(\hat{p}_{m,n}(a|x_S) - p(a|x_\Lambda))$, it follows that, on the event $\{\hat{\Lambda}_m = S, \bar{N}_{m,n}(x_S) > 0\}$, we must have $\hat{p}_{m,n}(a|x_S) = p(a|x_\Lambda)$ almost surely and so the left-hand side of (83) is 0 and the result holds trivially.

Let us now suppose $0 < p(a|x_\Lambda) < 1$. In this case, we apply Proposition 30 with $w = p(a|x_\Lambda)(1-p(a|x_\Lambda))$, $v = (n-m-d)w$ and $b = 1$ to deduce that,

$$\begin{aligned}\mathbb{P}\Big(\hat{\Lambda}_m = S, N^*_{n,m}(x_S)(\hat{p}_{m,n}(a|x_S) - p(a|x_\Lambda)) \geq \\ \sqrt{2\alpha(1+\varepsilon)p(a|x_\Lambda)(1-p(a|x_\Lambda))\bar{N}_{m,n}(x_S)} + \frac{\alpha}{3}, \bar{N}_{m,n}(x_S) > 0\Big) \\ \leq \left\lceil \frac{\log(n-m-d+1)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}\mathbb{P}(1\{\hat{\Lambda}_m = S\}\bar{N}_{m,n}(x_S)p(a|x_\Lambda)(1-p(a|x_\Lambda)) > 0). \quad (85)\end{aligned}$$

To conclude the proof, observe that in this case

$$\{1\{\hat{\Lambda}_m = S\}\bar{N}_{m,n}(x_S)p(a|x_\Lambda)(1-p(a|x_\Lambda)) > 0\} = \{\hat{\Lambda}_m = S, \bar{N}_{m,n}(x_S) > 0\},$$

and use again Proposition 30 with $H^{-\varphi} \bullet M^a$ in the place of $H^{\varphi} \bullet M^a$ (by noting also that $\langle H^{\varphi} \bullet M^a \rangle = \langle H^{-\varphi} \bullet M^a \rangle$). ■

Remark 32 *Let us briefly comment on the results of Corollary 31. Suppose $x \in A^{[\![-d,-1]\!]}$ and $a \in A$ are such that $0 < p(a|x_\Lambda) < 1$ and also that $\hat{\Lambda}_m$ is a consistent estimator of Λ. By the CLT for aperiodic and irreducible Markov Chains it follows that $\sqrt{\bar{N}_{m,n}(x_\Lambda)}(\hat{p}_{m,n}(a|x_\Lambda) - p(a|x_\Lambda))1\{\hat{\Lambda}_m = \Lambda, \bar{N}_{m,n}(x_\Lambda) > 0\}$ converges in distribution (as $\min\{m,n\} \to \infty$) to a centered Gaussian random variable with variance $p(a|x_\Lambda)(1-p(a|x_\Lambda))$. This implies that for sufficiently large n,*

$$\begin{aligned}\mathbb{P}\Big(\hat{p}_{m,n}(a|x_\Lambda) - p(a|x_\Lambda) \\ \geq \sqrt{\frac{2\alpha p(a|x_\Lambda)(1-p(a|x_\Lambda))}{\bar{N}_{m,n}(x_\Lambda)}} | \hat{\Lambda}_m = \Lambda, \bar{N}_{m,n}(x_\Lambda) > 0\Big) \leq e^{-\alpha}.\end{aligned}$$

Let us compare this heuristic argument with Corollary 31 applied to $S = \Lambda$. In this case, in Inequality (83), the variance term $\sqrt{\frac{2\alpha(1+\varepsilon)p(a|x_\Lambda)(1-p(a|x_\Lambda))}{\bar{N}_{m,n}(x_\Lambda)}}$ can be made arbitrarily close to *optimal value* $\sqrt{\frac{2\alpha p(a|x_\Lambda)(1-p(a|x_\Lambda))}{\bar{N}_{m,n}(x_\Lambda)}}$*, at the cost* $\frac{1}{\log(1+\varepsilon)}$*. Both the linear term* $\frac{\alpha}{\bar{N}_{m,n}(x_\Lambda)}$ *and the* $\log(n-m-d+1)$ *factor are the price to pay to achieve the result which holds every* $n \geq m+d+1$ *and reflect the fact that* $\hat{p}_{m,n}(a|x_\Lambda) - p(a|x_\Lambda)$ *is Gaussian only asymptotically.*

In particular, Corollary 31 improves the Economical Freedman's Inequality (as stated in (Oliveira, 2015) - Lemma 5, Inequality (42)) when restricted to the martingale $H^\varphi \bullet M^a$*.*

In the sequel, let us denote $P^a = (P^a_t)_{t\in\mathbb{N}}$, for each $a \in A$, the stochastic process defined as $P^a_t = p(a|X_{(t-d:t-1)})$ for each $t \in \mathbb{N}$. With this notation, notice that

$$H^2 \bullet P^a_t = \sum_{s=0}^{t} H^2_s p(a|X_{(s-1):(s-d)}),\ t \in \mathbb{N}, \tag{86}$$

is such that $\langle H \bullet M^a \rangle_t \leq H^2 \bullet P^a_t$ for all $t \in \mathbb{N}$. In particular, Proposition 28 holds if we replace $\langle H \bullet M^a \rangle_t$ by $H^2 \bullet P^a_t$. A closer inspection of the proof of Proposition 30 reveals that this proposition also holds with $H^2 \bullet P^a_t$ in the place of $\langle H \bullet M^a \rangle_t$. In the next theorem, we show that we can replace $H^2 \bullet P^a_t$ by a linear transformation of its empirical version which is crucial for our analysis.

Theorem 33 *Let* $H \bullet M^a = (H \bullet M^a_t)_{t\in\mathbb{N}}$ *be the stochastic process defined in* (78)*. Suppose that* $\|B\|_\infty \leq b$ *for some* $b > 0$*. For any fixed* $\mu \in (0,3)$ *satisfying* $\mu > \psi(\mu) = \exp(\mu) - \mu - 1$ *and* $\alpha > 0$*, define for* $t \in \mathbb{N}$*,*

$$H^2 \bullet \hat{P}^a_t = \frac{\mu}{\mu - \psi(\mu)} \sum_{s=0}^{t} H^2_s \hat{P}^a_s + \frac{b^2\alpha}{\mu - \psi(\mu)},$$

where $\hat{P}^a_s = 1\{X_s = a\}$ *for all* $s \in \mathbb{N}$*. Then, for any fixed* $\epsilon > 0$ *and* $v > w > 0$*, we have for any* $t \in \mathbb{N}$*,*

$$\mathbb{P}\left(H \bullet M^a_t \geq \sqrt{2(1+\epsilon)\alpha H^2 \bullet \hat{P}^a_t} + \frac{\alpha b}{3}, w \leq H^2 \bullet \hat{P}^a_t \leq v\right) \\ \leq 2\left\lceil \frac{\log(v/w+1)}{\log(1+\varepsilon)} \right\rceil \exp(-\alpha)\, \mathbb{P}(\langle H \bullet M^a \rangle_t > 0).$$

Proof We prove only the case $b = 1$. Note that $-H^2 \bullet M^a_t = H^2 \bullet P^a_t - H^2 \bullet \hat{P}^a_t$. Also recall that $\langle H \bullet M^a \rangle_t \leq H^2 \bullet P^a_t$.

We now proceed to the proof. We first use Inequality (80) for the martingale $-H^2 \bullet M^a$ together with that fact that $\langle -H^2 \bullet M^a \rangle_t \leq H^4 \bullet P^a_t \leq H^2 \bullet P^a_t$ (the last inequality holds because $b = 1$) to deduce that for any $\mu > 0$,

$$\mathbb{P}\left(H^2 \bullet P^a_t \geq H^2 \bullet \hat{P}^a_t + \frac{\psi(\mu)}{\mu} H^2 \bullet P^a_t + \frac{\alpha}{\mu}\right) \leq \exp(-\alpha)\mathbb{P}(\langle H \bullet M^a \rangle_t > 0),$$

which implies that for any $\mu \in (0,3)$ satisfying $\mu > \psi(\mu)$, it holds

$$\mathbb{P}\left(H^2 \bullet P^a_t \geq H^2 \bullet \hat{P}^a_t\right) \leq \exp(-\alpha)\mathbb{P}(\langle H \bullet M^a \rangle_t > 0).$$

Hence, combining this inequality with (81), we conclude that for any $\lambda \in (0,3)$ and any $\mu \in (0,3)$ satisfying $\mu > \psi(\mu)$,

$$\mathbb{P}\left(H \bullet M_t^a \geq \frac{\lambda}{2(1-\lambda/3)} H^2 \bullet \hat{P}_t^a + \frac{\alpha}{\lambda}\right) \leq \\ \mathbb{P}\left(H \bullet M_t^a \geq \frac{\lambda}{2(1-\lambda/3)} H^2 \bullet \hat{P}_t^a + \frac{\alpha}{\mu}, H^2 \bullet P_t^a \leq H^2 \bullet \hat{P}_t^a\right) \\ + \mathbb{P}\left(H^2 \bullet P_t^a \geq H^2 \bullet \hat{P}_t^a\right) \leq 2\exp(-\alpha)\mathbb{P}(\langle H \bullet M^a \rangle_t > 0),$$

where we also used in the last inequality the fact that $\langle H \bullet M^a \rangle_t \leq H^2 \bullet P_t^a$.

To conclude the proof, we need to follow the same steps as Proposition 28 and then use the peeling argument as in the proof of Proposition 30 with $H^2 \bullet \hat{P}_t^a$ in the place of $\langle H \bullet M^a \rangle_t$. ■

As consequence of Theorem 33, we obtain the following result.

Proposition 34 *Let $X_{1:n}$ be a sample from a MTD model of order d with set of relevant lags Λ. Let $\hat{\Lambda}_m$ be an estimator of Λ computed from $X_{1:m}$ where $n > m$. For any $x \in A^{[\![-d,-1]\!]}$, $a \in A$ and $S \subseteq [\![-d,-1]\!]$, let $\hat{p}_{m,n}(a|x_S)$ be the empirical transition probability defined in (7) computed from $X_{m+1:n}$, and consider for $\alpha > 0$ and $\mu \in (0,3)$ satisfying $\mu > \psi(\mu) = \exp(\mu) - \mu - 1$,*

$$\hat{V}_{m,n}(a,x,S) = \frac{\mu}{\mu - \psi(\mu)} \hat{p}_{m,n}(a|x_S) + \frac{\alpha}{\mu - \psi(\mu)} \frac{1}{\bar{N}_{m,n}(x_S)}.$$

Then for any $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \subseteq S$ and $n \geq m+d+1$, we have

$$\mathbb{P}\left(\hat{\Lambda}_m = S, |\hat{p}_{m,n}(a|x_S) - p(a|x_\Lambda)| \geq \sqrt{\frac{2\alpha(1+\epsilon)\hat{V}_{m,n}(a,x,S)}{\bar{N}_{m,n}(x_S)}} + \frac{\alpha}{3\bar{N}_{m,n}(x_S)}\right) \\ \leq 4\left\lceil \frac{\log(\mu(n-m-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}\mathbb{P}\left(\hat{\Lambda}_m = S, \bar{N}_{m,n}(x_S) > 0\right). \quad (87)$$

In particular,

$$\mathbb{P}\left(\Lambda \subseteq \hat{\Lambda}_m, |\hat{p}_{m,n}(a|x_{\hat{\Lambda}_m}) - p(a|x_\Lambda)| \geq \sqrt{\frac{2\alpha(1+\epsilon)\hat{V}_{m,n}(a,x,\hat{\Lambda}_m)}{\bar{N}_{m,n}(x_{\hat{\Lambda}_m})}} + \frac{\alpha}{3\bar{N}_{m,n}(x_{\hat{\Lambda}_m})}\right) \\ \leq 4\left\lceil \frac{\log(\mu(n-m-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}\mathbb{P}\left(\Lambda \subseteq \hat{\Lambda}_n, \bar{N}_{m,n}(x_{\hat{\Lambda}_m}) > 0\right). \quad (88)$$

Proof Summing in both sides of (87) over $S \subseteq [\![-d,-1]\!]$ such that $\Lambda \subseteq S$, we obtain inequality (88). Hence, it remains to show (87). Arguing as in Corollary 31 we need to consider only the case $0 < p(a|x) < 1$.

By applying Theorem 33 with $H = H^{\pm\varphi}$ where φ is as in the proof of Corollary 31, $v = \frac{\mu}{\mu-\psi(\mu)}(n-m-d) + \frac{\alpha}{\mu-\psi(\mu)}$ and $w = \frac{\alpha}{\mu-\psi(\mu)}$, we obtain that

$$\begin{aligned}&\mathbb{P}\left(\hat{\Lambda}_m = S, \bar{N}_{m,n}(x_S)|\hat{p}_{m,n}(a|x_S) - p(a|x_\Lambda)| \geq \sqrt{2(1+\epsilon)\alpha\tilde{V}_{m,n}(a,x,S)} + \frac{\alpha}{3}\right)\\ &\quad\leq 4\left\lceil \frac{\log(\mu(n-m-d)/\alpha+2)}{\log(1+\varepsilon)} \right\rceil e^{-\alpha}\mathbb{P}(1\{\hat{\Lambda}_m = S\}\bar{N}_{m,n}(x_S)p(a|x_\Lambda)(1-p(a|x_\Lambda)) > 0),\end{aligned}$$

where $\tilde{V}_{m,n}(a,x,S) = \bar{N}_{m,n}(x_S)\hat{V}_{m,n}(a,x,S)$.

By using that when $0 < p(a|x) < 1$,

$$\{1\{\hat{\Lambda}_m = S\}\bar{N}_{m,n}(x_S)p(a|x_\Lambda)(1-p(a|x_\Lambda)) > 0\} = \{\hat{\Lambda}_m = S, \bar{N}_{m,n}(x_S) > 0\},$$

and the fact that $\tilde{V}_{m,n}(a,x,S) = \bar{N}_{m,n}(x_S)\hat{V}_{m,n}(a,x,S)$, we deduce (87) from the above inequality. ■

References

Andre Berchtold. Estimation in the mixture transition distribution model. *Journal of Time Series Analysis*, 22(4):379–397, 2001.

André Berchtold and Adrian Raftery. The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. *Statistical Science*, 17(3):328 – 356, 2002.

Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 771–782. Association for Computing Machinery, 2015.

Peter Bühlmann and Abraham J Wyner. Variable length markov chains. *The Annals of Statistics*, 27(2):480–513, 1999.

Gyorgy Buzsaki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304(5679):1926–1929, 2004.

J.R. Chazottes, S. Gallo, and D.Y. Takahashi. Optimal gaussian concentration bounds for stochastic chains of unbounded memory. *ArXiv*, 2020.

Yanqing Chen, Mingzhou Ding, and JA Scott Kelso. Long memory processes (1/f α type) in human coordination. *Physical Review Letters*, 79(22):4501, 1997.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Imre Csiszár and Zsolt Talata. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information theory*, 52(3):1007–1016, 2006.

Antonio Galves, Charlotte Galves, Jesus E Garcia, Nancy L Garcia, and Florencia Leonardi. Context tree selection and linguistic rhythm retrieval from written texts. *The Annals of Applied Statistics*, 6(1):186–209, 2012.

Jesús E Garcıa, 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.

David L Gilden, Thomas Thornton, and Mark W Mallon. 1/f noise in human cognition. *Science*, 267(5205):1837–1839, 1995.

Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83 – 143, 2015.

Matthew Heiner and Athanasios Kottas. Estimation and selection for high-order markov chains with bayesian mixture transition distribution models. *Journal of Computational and Graphical Statistics*, pages 1–13, 2021.

Väinö Jääskinen, Jie Xiong, Jukka Corander, and Timo Koski. Sparse markov chains for sequence data. *Scandinavian Journal of Statistics*, 41(3):639–655, 2014.

Andrea Király, Imre Bartos, and Imre M Jánosi. Correlation properties of daily temperature anomalies over land. *Tellus A: Dynamic Meteorology and Oceanography*, 58(5):593–600, 2006.

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.

Roberto Imbuzeiro Oliveira. Stochastic processes with random contexts: A characterization and adaptive estimators for the transition probabilities. *IEEE Transactions on Information Theory*, 61(12):6910–6925, 2015.

Adrian E. Raftery. A model for high-order markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(3):528–539, 1985.

Maxim Raginsky and Igal Sason. *Concentration of Measure Inequalities in Information Theory, Communications, and Coding: Second Edition.* 2014.

Jorma Rissanen. A universal data compression system. *IEEE Transactions on information theory*, 29(5):656–664, 1983.

Abhra Sarkar and David B Dunson. Bayesian nonparametric modeling of higher order markov chains. *Journal of the American Statistical Association*, 111(516):1791–1803, 2016.

M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Matthew C Wheeler, Harry H Hendon, Sam Cleland, Holger Meinke, and Alexis Donald. Impacts of the madden–julian oscillation on australian rainfall and circulation. *Journal of Climate*, 22(6):1482–1498, 2009.

Naiming Yuan, Zuntao Fu, and Shida Liu. Long-term memory in climate variability: A new look based on fractional integral techniques. *Journal of Geophysical Research: Atmospheres*, 118(23):12–962, 2013.