On deterministic finite state machines in random environments*

Joel Ratsaby

Electrical and Electronics Engineering Department
Ariel University
Ariel 40700, ISRAEL
e-mail: ratsaby@ariel.ac.il

June 11, 2020

Abstract:

The general problem under investigation is to understand how complexity of a system which has been adapted to its random environment affects the level of randomness of its output (which is a function of its random input). In this paper we consider a specific instance of this problem in which a deterministic finite-state decision system operates in a random environment that is modeled by a binary Markov chain. The system interacts with it by trying to match states of inactivity (represented by 0). Matching means that the system selects the $(t+1)^{th}$ bit from the Markov chain whenever it predicts at time t that the environment will take a 0 value. The actual value at time $t+1$ may be 0 or 1 thus the selected sequence of bits (which forms the system's output) may have both binary values. To try to predict well, the system's decision function is inferred based on a sample of the random environment.

We are interested in assessing how non-random the output sequence may be. To do that, we apply the adapted system on a second random sample of the environment and derive an upper bound on the deviation between the average number of 1 bits in the output sequence and the probability of a 1. The bound shows that the complexity of the system has a direct effect on this deviation and hence on how non-random the output sequence may be. The bound takes the form of $O\left(\sqrt{(2^k/n)}\right)$ where 2^k is the complexity of the system and n is the length of the second sample.

Keywords and phrases: Prediction of random binary sequence, Markov chain, subsequence selection, frequency instability.

1. Introduction

This paper stems from our aim to understand the interaction between deterministic systems and their random environment. This is of interest in many fields, for instance, in reliability theory and failure analysis [15, 21], queuing theory [8], biology [5, 19], aerospace [20], learning automata [13]. Another field is theoretical computer science which deals with the interaction of deterministic

*This is the latest version. The published paper in *Probability in the Engineering and Informational Sciences*, Vol. 33(4), pp. 528-563, 2019, has the following error: throughout the paper, β should be replaced by β_d which is defined in the current version in (5.3) systems with highly complex input. Here, systems called *selection rules*, which are modeled as Turing machines, are used to test if a highly complex binary input sequence is random [4]. One consequence of this theory [1, 9] is that the complexity of a selection rule (system) affects the rate of convergence of the law of large numbers (LLN) for subsequences that are selected by it (this is referred to as *frequency instability*). In other words, such finite state machines select subsequences for which the average number of 1 bits deviates by a large amount from the probability of a 1. The deviation is proportional to the complexity of the input sequence and the complexity of the system.

We wish to investigate this fact, however, from a different angle. Instead of using selection rules for testing randomness of a highly complex input sequences, we use them to select subsequences from an input sequence (environment) which is, as a given fact, random. We are interested in the rate of the LLN of such subsequences. The systems that we consider are not chosen arbitrarily but are first adapted to their random environment and then are used to predict its future behavior. As byproduct of this prediction, the systems select subsequences from the environment. We aim to understand how the complexity of such systems affects the frequency stability of these subsequences.

It is less interesting to consider an environment which is represented as a sequence of i.i.d. random variables because the best prediction strategy in this case is the trivial one, namely, at any time instant predict the symbol with the largest *a priori* probability (which would need to be estimated). Instead, predicting an environment which is modeled by a probability model known as a Markov chain is more interesting and applicable because it assumes interdependences between consecutive segments of symbols. (Nonetheless, one can predict sequences without assuming an underlying probability model as done in universal prediction [11].)

We consider a basic random environment which is modeled by a binary Markov chain and consists of a sequence of dependent binary random variables that indicate if the environment is *active* (denoted by 1) or *inactive* (denoted by 0).

As a system, we consider an automaton, or a deterministic finite state machine (FSM).We define the complexity of the system to be its number of parameters which are estimated from a sample of the environment.

The FSM predicts the next bit of the environment based on its current state. By repeating this process of prediction multiple times, we let the FSM select a subsequence from the input environment as follows: the system decides at every time instant whether the next input bit is likely to be a 0 (that is, if the environment is likely to be inactive). If the system predicts 0 it selects the next input bit to be the next output bit. Otherwise, it does not select this bit and does not output any value. We refer to this selection behavior as trying to match inactivity in the environment, or *matching the environment* for short. (We could also define matching to be based on predicting a 1 instead of a 0, but a nice consequence of the current definition is that the selected subsequence can be interpreted as the subsequence of prediction-error indicators, as described further below.)

This matching behavior is interesting to analyze because it allows us to represent the prediction of an input sequence as selection of a subsequence from this input (the selected random subsequence is just a sequence of error indicators when predicting zeros).

In order to measure how nonrandom the output sequence may be, we compare the deviation between the average number of 1 bits in the output versus the probability of a symbol 1. Empirical investigations of such matching systems [17, 18] show that subsequences that are selected by them have empirical frequencies that deviate from the probabilities in proportion to the complexity of the system's FSM. The current paper confirms this. We obtain a mathematical relationship that shows the dependence of this deviation in terms of the length of the input sequence, the length of the sample from which the FSM is inferred, the order of the environment's Markov chain and the complexity of the FSM. Our analysis can be applied to other deterministic functions of the input (not just selection functions) as long as the functions are Lipschitz continuous.

The aim of the paper is not to study the accuracy of learning to predict but rather to analyze a predictor as a selection system and investigate its output frequency instability. The motivation comes from the general aim to understand how a complex system interacts in a random environment (see the references mentioned at the start of this section and [16]). For the FSM system considered here, its interaction with the environment is represented by the selected subsequence which is defined as its output. While we are not aware of practical applications of the results, the analysis may be useful in problems of learning to predict Markov stochastic environments.

Let us summarize the remaining parts of the paper: section 2 has the mathematical setup and notation, section 3 describes how the system adapts to the environment, section 4 states the aim of the paper, section 5 defines the system's decision function, section 6 states an assumption about the Markov environment that is needed for using a concentration bound, section 7 states the results and the proof is presented in section 8.

2. Preliminaries

We use capital letters X, S, Y, Ξ to represent random variables and lower case letters x, s, y, ξ to represent their values. Let $\mathbb{I}(E)$ denote the indicator function for a logical (Boolean) expression E.

Let

$$\{X_t : t \in \mathbb{Z}\} \tag{2.1}$$

be a sequence of binary random variables possessing the following Markov property,

$$\begin{aligned} & P\left(X_t = x_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots\right) \\ = \quad & P\left(X_t = x \mid X_{t-1} = x_{t-1}, \ldots, X_{t-k^*} = x_{t-k^*}\right) \end{aligned} \tag{2.2}$$

where $x_{t-k^*}, \ldots, x_{t-1}$, x_t take a binary value of 0 or 1. This sequence is known as a discrete-time Markov stochastic process, or Markov *chain*, of order k^*.

The Markov chain has a Markov probability model

$$\mathcal{M}_{k^*} := (\mathbb{S}_{k^*}, Q)$$

which consists of a state space $\mathbb{S}_{k^*} = \{0,1\}^{k^*}$ and a $2^{k^*} \times 2^{k^*}$ state-transition probability matrix Q. We denote the i^{th} state of $\mathcal{M}_{k^*}$ by $s^{*(i)}$, $i = 0, 1, \ldots, 2^{k^*}-1$, with $s^{*(0)} := [s^{*(0)}_{k^*-1}, \ldots, s^{*(0)}_0] = [0, \ldots, 0, 0]$ (the all-zero state), $s^{*(1)} := [s^{*(1)}_{k^*-1}, \ldots, s^{*(1)}_0] = [0, \ldots, 0, 1]$, $\ldots$,$s^{*(2^{k^*}-1)} := [1, \ldots, 1]$. The $(ij)^{th}$ entry of Q is denoted by

$$Q[i,j] := q\left(s^{*(j)} \middle| s^{*(i)}\right). \tag{2.3}$$

Henceforth, by *environment* we mean the Markov chain (2.1) and we assume that it is *stationary* (that there exists a stationary probability distribution is discussed further below).

We can express the Markov chain $\{X_t : t \in \mathbb{Z}\}$ as a random sequence of states $\{S^*_t : t \in \mathbb{Z}\}$ where

$$S^*_t \;:=\; (X_{t-k^*+1}, X_{t-k^*+2}, \ldots, X_t) \in \mathbb{S}_{k^*} \tag{2.4}$$

is the random state at time t. The transition matrix Q is fixed with respect to time hence the chain is homogeneous.

There are two possible outgoing transitions from a state S^*_t to the next state S^*_{t+1} since S^*_{t+1} can take only one of the two values $(X_{t-k^*+2}, \ldots, X_t, 0)$ or $(X_{t-k^*+2}, \ldots, X_t, 1)$. We call them *type-0* and *type-1* transitions. Thus the matrix Q is sparse and with each state $s^{*(i)} \in \mathbb{S}_{k^*}$ we associate two non-zero valued parameters which are the probability of transition to those two states $s^{*(j)}$ that are obtained by a type-0 or type-1 transition from state $s^{*(i)}$. We denote these parameters by

$$q(1|i) := q(1|s^{*(i)}),\; q(0|i) := q(0|s^{*(i)}) = 1 - q(1|i). \tag{2.5}$$

The set $\left\{q(1|i) : 0 \le i \le 2^{k^*}-1\right\}$ serves as parameters of $\mathcal{M}_{k^*}$. We assume that for all $0 \le i \le 2^{k^*}-1$,

$$0 < q(1|i) < 1. \tag{2.6}$$

The structure of $\mathcal{M}_{k^*}$ is isomorphic to a directed De Bruijn graph with vertices mapped to states of $\mathcal{M}_{k^*}$ and edges mapped to state transitions. The graph is connected because (2.6) for all i hence it follows that all the states of $\mathcal{M}_{k^*}$ communicate and thus the chain is irreducible. With the finiteness of the state space this implies that the Markov chain has a stationary distribution (Corollary 8.2, [14]).

Let the vector

$$\pi^* := \left[\pi^*_0, \ldots, \pi^*_{2^{k^*}-1}\right] \tag{2.7}$$

denote the state stationary probability distribution, where π^*_i is the probability that $S^*_t = s^{*(i)}$. We also write $\pi^*_{s^{*(i)}} := \pi^*_i$, $0 \le i \le 2^{k^*}-1$. We denote by $\mathbb{P}$ the stationary joint probability distribution of a state sequence $(S_1^*, \ldots, S_l^*)$ defined as follows: for any sequence $(s_1^*, \ldots, s_l^*) \in \mathbb{S}_{k^*}^l$,

$$\mathbb{P}\left((S_1^*, \ldots, S_l^*) = (s_1^*, \ldots, s_l^*)\right) := \pi^*_{s_1^*} \prod_{r=1}^{l-1} q\left(s^*_{r+1} | s^*_r\right).$$

Henceforth, by a *system* we mean an FSM based on the states and transitions of a Markov probability model $\mathcal{M}_k$ of order k (which may be different from k^*) with a state space $\mathbb{S}_k$. According to this model, denote by $S_t \in \mathbb{S}_k$ a random state at time t. (To distinguish between the automaton system's states from the states of the random environment, we place a star for variables that represent the environment, as for instance S_t^* in (2.4), and no star for variables that represent the system's state.) For every state of a system there are two possible outgoing transitions to the next state. We denote the two probability parameters that correspond to state $s^{(i)} \in \mathbb{S}_k$ by $p(1|i)$ and $p(0|i)$.

3. Adapting the system to the environment

In the previous section we defined a system to be a finite state machine whose states and transition probabilities are according to the probability model $\mathcal{M}_k$. In this section we start by showing that these transition probabilities are determined by the unknown transition probabilities of the environment's model $\mathcal{M}_{k^*}$ and then we can therefore define statistical estimates for these probabilities.

In case $k^* > k$ we denote by $\mathbb{P}(ji)$ the stationary probability of the environment's state $S_t^* = s^{*(ji)}$, where $s^{*(ji)} \in \mathbb{S}_{k^*}$ is a binary number whose $k^* - k$ leftmost bits amount to a binary number whose value $0 \le j \le 2^{k^*-k} - 1$ and whose k rightmost bits amount to a binary number whose value is $0 \le i \le 2^k - 1$. For $1 \le k \le k^*$ we denote by

$$\mathbb{P}(j|i) := \frac{\mathbb{P}(ji)}{\sum_{j=0}^{2^{k^*-k}-1} \mathbb{P}(ji)} \tag{3.1}$$

the probability that the $k^* - k$ leftmost bits amount to the decimal value j given that the k rightmost bits amount to i.

For $1 \le q \le r$, define the projection operator $< \cdot >_q : \mathbb{S}_r \to \mathbb{S}_q$ as a mapping that takes a state $s^{(i)} \in \mathbb{S}_r$ to a state

$$s = < s^{(i)} >_q = [s_{q-1}^{(i)}, \ldots, s_0^{(i)}] \in \mathbb{S}_q \tag{3.2}$$

whose q digits correspond to the q least significant (rightmost) bits of $s^{(i)}$.

From the environment, we sample $m + \max\{k, k^*\}$ consecutive values and obtain a stationary Markov chain

$$X^{(m)} \quad := \quad \{X_{T+t}\}_{t=-\max\{k,k^*\}+1}^{m} \tag{3.3}$$

where $T \in \mathbb{Z}$ is any fixed time index. (It is stationary because the environment is assumed to be a stationary Markov chain, see start of section 2).

According to the system's model $\mathcal{M}_k$, the true (not estimated) probability that the chain $X^{(m)}$ makes a type-1 transition from state $s^{(i)} \in \mathbb{S}_k$ is $p(1|i)$. If $k \geq k^*$ then

$$p(1|i) \quad = \quad q\left(1 \middle| < s^{(i)} >_{k^*}\right) \tag{3.4}$$

which is a parameter of the environment model $\mathcal{M}_{k^*}$. If $k < k^*$ then

$$p(1|i) \quad = \quad \sum_{j=0}^{2^{k^*-k}-1} q\left(1 \middle| S^* = s^{*(ji)}\right) \mathbb{P}(j|i) \tag{3.5}$$

where the state $s^{*(ji)}$ is defined above. In either case, $p(1|i)$ is determined directly from $\mathcal{M}_{k^*}$ since it is determined by the stationary probability distribution and the true transition probabilities $q\left(1 \mid S^* =< s^{(i)} >_{k^*}\right)$ or $q\left(1 \mid S^* = s^{*(ji)}\right)$ of the environment's model $\mathcal{M}_{k^*}$. Thus the true values of the system's model parameters $p(1|i)$ are unknown and are completely determined by the environment's model $\mathcal{M}_{k^*}$. Because they are unknown, next we define estimates of these parameters based on the sample $X^{(m)}$. Let

$$S^{*(m)} = \{S_t^*\}_{t=1}^m, \quad S_t^* = [X_{T+t-k^*+1}, \ldots, X_{T+t}], \; 1 \leq t \leq m \tag{3.6}$$

and

$$S^{(m)} = \{S_t\}_{t=1}^m, \quad S_t = [X_{T+t-k+1}, \ldots, X_{T+t}], \; 1 \leq t \leq m. \tag{3.7}$$

There is a one-to-one correspondence between these two sequences since each one can be obtained from the other by going through $X^{(m)}$. (We will choose to use them as needed depending on the context.)

For $1 \leq i \leq 2^k - 1$ let

$$\alpha_i := \frac{1}{m} \sum_{t=1}^m \mathbb{I}\left\{S_t = s^{(i)}\right\}$$

denote the empirical frequency that state $s^{(i)} \in \mathbb{S}_k$ appears in the state sequence $S^{(m)}$. We denote by

$$\alpha := \alpha(S^{(m)}) \quad = \quad [\alpha_0(S^{(m)}), \ldots, \alpha_{2^k-1}(S^{(m)})]$$

where α satisfies

$$\sum_{i=0}^{2^k-1} \alpha_i = 1 \tag{3.8}$$

and for brevity we also write

$$\alpha_i \quad := \quad \alpha_i(X^{(m)}) = \alpha_i(S^{(m)}).$$

Note that the initial state $S_0 = [X_{-k+1}, \ldots, X_0]$ at time $t = 0$ is not accounted for in α but only states S_t at times $1 \leq t \leq m$. For instance, if $T = 0$,

$m = 6$, $k = 3$, $\max\{k, k^*\} = 3$ and $i = 1$ then $s^{(i)} = [001]$ and for the sequence $X^{(m)} := \{X_{T+t}\}_{t=-\max\{k,k^*\}+1}^{m} = 010010001$ we have $\alpha_i = \frac{2}{6}$ since the states at time instants $t = 1,2, \ldots$, are $s^{(4)}, s^{(1)}, \ldots$, respectively.

We denote by

$$\pi_i = \mathbb{E}[\alpha_i] \tag{3.9}$$

the stationary probability that $S_t = s^{(i)}$, for $s^{(i)} \in \mathbb{S}_k$, $1 \leq t \leq m$.

If $\alpha_i > 0$ we can define the estimate of $p(1|i)$ as the frequency of type-1 transitions from state $s^{(i)}$ in the sequence $S^{(m)}$, which is denoted by

$$\hat{p}(1|i) \quad := \quad \frac{1}{\alpha_i m} \sum_{l:\, S_{t_l} = s^{(i)}} \mathbb{I}\{X_{t_l+1} = 1\}. \tag{3.10}$$

The index in the sum runs over those time instants $t_l \in \{0, \ldots, m-1\}$, $1 \leq l \leq \alpha_i m$, where $S_{t_l} = s^{(i)}$.

Henceforth we consider the *adapted system* to be an FSM with these parameter estimates (based on $X^{(m)}$) that are substituted for the unknown transition probabilities of the model $\mathcal{M}_k$. (For brevity, sometimes we omit the word 'adapted' and just refer to it as system.)

4. Statement of the problem

We are interested in how the complexity of the adapted system influences the frequency stability of subsequences of the environment which are selected by the system. To study this, we obtain another sample from the environment, an $(n + \max\{k, k^*\})$-bit stationary Markov chain, denoted by

$$X^{(n)} \quad := \quad \{X_{T'+t}\}_{t=-\max\{k,k^*\}+1}^{n}, \tag{4.1}$$

where $T' \in \mathbb{Z}$ is any fixed time index that satisfies

$$T' - \max\{k, k^*\} + 1 > T + m. \tag{4.2}$$

This sequence is used as input to the adapted system. The condition (4.2) ensures that $X^{(n)}$ is a sample that comes later in time after the sample $X^{(m)}$ ends. This ensures that it is possible to use $X^{(n)}$ to test the random subsequence which is selected by the system after the system has already adapted based on $X^{(m)}$. Obviously the two sequences are dependent since they are from the same Markov chain (2.1).

The FSM starts at the initial state

$$S_0 : \quad = \quad (X_{T'-k+1}, \ldots, X_{T'}) \tag{4.3}$$

where $X_{T'+t}$ is the t^{th} bit of $X^{(n)}$ and produces a sequence of *decisions*

$$Y^{(n)} \quad = \quad \{Y_t\}_{t=1}^{n} \tag{4.4}$$

where $Y_t \in \{0, 1, REJECT\}$ is the value taken by a decision function $\hat{d}$ (described in the next section) at time t. By definition of matching, it behaves as follows: if $Y_t = 0$, the system decides to select bit $X_{T'+t}$. Otherwise, if $Y_t = 1$ or $Y_t = REJECT$, it decides not to select $X_{T'+t}$, $1 \le t \le n$. We define the output of the system to be this selected subsequence of $X^{(n)}$ and denote it by

$$\Xi^{(\nu)} \quad = \quad \{\Xi_{t_l}\}_{l=1}^{\nu} \tag{4.5}$$

where the time instants t_l correspond to 0-decisions, $Y_{t_l} = 0$, for $1 \le l \le \nu \le n$.

Our aim is to assess the frequency instability of $\Xi^{(\nu)}$. The main result, Theorem 7.1, states a bound on the deviation between the frequency of 1 in $\Xi^{(\nu)}$ and the probability of 1 given that $Y_t = 0$, with dependence on the complexity of the adapted system.

5. Decision function

In this section we define the adapted system's decision which is a function of the probability estimates $\hat{p}(1|i)$, $0 \le i \le 2^k - 1$, defined in (3.10). Let the set

$$\Upsilon := \{0, 1, REJECT\}.$$

Let us denote a decision function

$$d \quad \in \quad \Upsilon^{2^k}$$

which is a vector that consists of the individual decisions at each of the states of the FSM,

$$d := [d(0), \ldots, d(2^k - 1)].$$

We denote by $\hat{d}$ a decision function which is obtained from $X^{(m)}$,

$$\hat{d} := \hat{d}(X^{(m)}) = [\hat{d}(0), \ldots, \hat{d}(2^k - 1)], \tag{5.1}$$

where

$$\hat{d}(i) := \hat{d}(i, \eta) = \begin{cases} 1 & \text{if } \hat{p}(1|i) > \frac{1}{2} + \mathcal{T}(\alpha_i, \eta) \\ 0 & \text{if } \hat{p}(1|i) < \frac{1}{2} - \mathcal{T}(\alpha_i, \eta) \\ REJECT & \text{otherwise} \end{cases} \tag{5.2}$$

for $0 \le i \le 2^k - 1$, and $\mathcal{T}$ defines a threshold that determines if a non-REJECT decision is made. The option of choosing a 'REJECT' action is a standard way to minimize decision errors by refraining from making a decision that has a small confidence (it had been in use at least as early as [7]). The value of $\mathcal{T}$ depends on a confidence parameter $\eta \in (0, 1]$ and on the state frequency $\alpha_i(X^{(m)})$. We henceforth write $\hat{d}$ without explicitly showing the dependence on η. Note that if $\alpha_i = 0$, then $X^{(m)}$ did not visit state $s^{(i)}$ and $\hat{d}(i) = REJECT$ because the threshold $\mathcal{T}$ equals infinity (this is described below) which means that the third line holds in (5.2).

Given a fixed η and a realization $x^{(m)}$ of the sample $X^{(m)}$, we measure $\alpha(x^{(m)})$ and evaluate the threshold $\mathcal{T}(\alpha_i, \eta)$ for each state $s^{(i)}$, $0 \le i \le 2^k - 1$. With these threshold values, we have a decision function $\hat{d}$ which is defined according to (5.2) and is used at every state of the FSM.

The function $\mathcal{T}$ equals infinity at $\alpha_i = 0$ and decreases with α_i and η. It is defined such that if $\hat{d}(i) = 1$ then with probability at most η the Bayes' optimal decision is $d^*(i) = 0$, or if $\hat{d}(i) = 0$ then with probability at most η the Bayes' decision is $d^*(i) = 1$, $0 \le i \le 2^k - 1$. That is, with confidence at least $1 - 2\eta$ the system's decision $\hat{d}$ agrees with the Bayes' optimal decision function d^*. We note that in Theorem 7.1 the choice for η is dictated by the choice of a confidence parameter δ.

The selected subsequence $\Xi^{(\nu)}$ is defined as follows: for any decision rule d, for time instants t_l that correspond to 0-decisions ($Y_{t_l} = 0$) by d, we have $\Xi_{t_l} = X_{t_l}$, $1 \le l \le \nu$. Define

$$\beta_d := P\left(X_t = 1 \middle| Y_t = 0\right) \tag{5.3}$$

which is fixed for every t due to the stationarity of the Markov chain (2.1). (Y_t obviously depends on the decision rule and this dependence is implicit in the notation).

We are interested in the event that the frequency of 1 in $\Xi^{(\nu)}$ deviates (in absolute value) from $\beta_{\hat{d}}$ by at least ϵ and that the subsequence is not too short (since we are interested in the rate of the LLN).

To bound the probability of this event we use a concentration bound for Markov chains. Such bounds involve mixing characteristics of the chain and are discussed in the next section.

6. Mixing

Concentration bounds for Markov chains depend on the rate in which chains mix [3]. Our analysis uses such a bound and we therefore need to state explicitly any assumption about the mixing properties of the environment's Markov chain. In the Appendix, we show that there exists a minimum integer $\mathfrak{l}_0$, such that for $l \ge \mathfrak{l}_0$, the environment's transition matrix Q in (2.3) satisfies $Q^l > 0$, that is, every entry of Q^l, denoted by $q^{(l)}(s^{(j)}|s^{(i)})$, is positive. We henceforth choose

$$\mathfrak{l}_0 := \min\{l : Q^l > 0\} \tag{6.1}$$

and in theory, if Q was known then $\mathfrak{l}_0$ can be evaluated by computing Q^l for a sequence of $l \ge 1$ until the first l is found such that $Q^l > 0$. Denote by μ_0 the minimum entry of $Q^{\mathfrak{l}_0}$,

$$\mu_0 := \min_{i,j} q^{(\mathfrak{l}_0)}(j|i) \tag{6.2}$$

then the fact that $Q^{\mathfrak{l}_0} > 0$ implies that $\mu_0 > 0$. We henceforth make the following assumption.

Assumption 1. *The environment's transition matrix Q satisfies one of the following conditions: (i) the minimum entry μ_0 of $Q^{\mathfrak{l}_0}$ satisfies $\mu_0 \neq 2^{-k^*}$ or (ii) $\mu_0 = 2^{-k^*}$ and for all $0 \le i \le 2^{k^*} - 1$, the transitions probabilities (2.5) are $\frac{1}{2}$.*

Remark 1. In both parts (i) and (ii) of the assumption, Q may have a uniform stationary distribution $\pi^* = \left[2^{-k^*}, \ldots, 2^{-k^*}\right]$, which means Q is doubly stochastic and $\lim_{l\to\infty} Q^l$ is a matrix U, of the same size as Q, with all its entries identical to 2^{-k^*}. Part (ii) treats the special case where this limit U is reached exactly at time $\mathfrak{l}_0$, that is, $Q^{\mathfrak{l}_0} = U$.

We use $\mathfrak{l}_0$ and μ_0 in the following definition. According to the cases of Assumption 1, define

$$\rho(k^*, \mathfrak{l}_0) := \begin{cases} \frac{1-2^{k^*}\mu_0}{2\mu_0} & \text{if case } (i) \text{ holds and } \mathfrak{l}_0 = 1 \\ \frac{2^{k^*-1}}{\left(1-2^{k^*}\mu_0\right)^{(\mathfrak{l}_0-1)/\mathfrak{l}_0}\left(1-\left(1-2^{k^*}\mu_0\right)^{1/\mathfrak{l}_0}\right)} & \text{if case } (i) \text{ holds and } \mathfrak{l}_0 \geq 2 \\ 2^{k^*-1} & \text{if case } (ii) \text{ holds.} \end{cases} \tag{6.3}$$

In the first condition of (6.3), (2.6) implies that $Q > 0$ and thus μ_0 is the minimum entry of Q. Define

$$r(k, k^*) := \begin{cases} 1 & \text{if } k^* \geq k+1 \\ k - k^* + 2 & \text{if } k^* \leq k. \end{cases} \tag{6.4}$$

7. Result

The next theorem states that if the adapted system's decision function $\hat{d}$ uses a threshold $\mathcal{T}$ (which is specified below) then the system's output, namely the selected subsequence $\Xi^{(\nu)}$ of $X^{(n)}$, can have an average value that deviates from the probability β of symbol 1 by as much as ϵ, where the expression of ϵ is stated. This establishes the rate of the LLN for the sequence $\Xi^{(\nu)}$ as aimed for in section 4.

The parameters k^*, $\mathfrak{l}_0$, μ_0 of the environment's Markov chain are assumed to be unknown. We assume that an upper bound on $\rho(k^*, \mathfrak{l}_0)$ is available for the system.

Theorem 7.1. *Let $m, n, k, k^*, \mathfrak{l}_0 \geq 1$ be positive integers and ϱ a positive constant. The environment is a stationary binary Markov chain $\{X_t\}_{t\in\mathbb{Z}}$ of unknown order k^* and $\rho(k^*, \mathfrak{l}_0) \leq \varrho$. Let $\mathcal{M}_k$ be a Markov model of order k used by a finite state machine (system). Let $X^{(m)} = \{X_{T+t}\}_{t=-\max\{k,k^*\}+1}^{m}$, $X^{(n)} = \{X_{T'+t}\}_{t=-\max\{k,k^*\}+1}^{n}$ be samples of the environment with $T + m < T' - \max\{k, k^*\} + 1$. Denote by $\alpha_i := \alpha_i(X^{(m)})$, $0 \leq i \leq 2^k - 1$, the random variables representing the empirical frequency that state $s^{(i)}$ of $\mathcal{M}_k$ appears in the state sequence $S^{(m)}$ that corresponds to $X^{(m)}$. For any,*

- $0 < \delta \leq 1$,
- $\ell \in \{1, \ldots, n\}$,
- $\omega \in (\frac{\ell}{n}, 1]$,

if the system uses a decision function $\hat{d}(X^{(m)})$ *defined according to (5.2) with*

$$\mathcal{T}(\alpha_i, \delta) := \varrho\sqrt{\frac{32}{m\alpha_i^2}\ln\left(\frac{2^{k+4}}{\delta\alpha_i}\right)}, \tag{7.1}$$

then with probability no more than δ *the subsequence* $\Xi^{(\nu)}$ *of* $X^{(n)}$ *that corresponds to time instants when the system's decision is* $Y_{t_l} = 0$, $t_l \in \{1, \ldots, n\}$, *satisfies both of the following conditions:*

- *the absolute value of the deviation between* $\beta_{\hat{d}}$ *and the frequency of* 1 *in* $\Xi^{(\nu)}$ *is more than*

$$\frac{4r(k,k^*)\rho(k^*,\mathfrak{l}_0)}{\omega}\sqrt{\frac{2}{n}\left((2^k+4)\ln 2 + \ln\left(\frac{1-\frac{\ell}{n}}{\left(\omega - \frac{\ell}{n}\right)\delta}\right)\right)}, \tag{7.2}$$

  where $\beta_{\hat{d}} := \mathbb{P}\left(X_t = 1 \middle| Y_t = 0\right)$ *is the stationary probability of symbol* 1 *given that the system predicts a* 0.
- *it has a length* $\nu \geq \omega n$.

The theorem implies that with probability at least $1-\delta$, either the absolute deviation between $\beta_{\hat{d}}$ and the frequency of 1 in $\Xi^{(\nu)}$ is no more than (7.2) or the length of the subsequence $\Xi^{(\nu)}$ satisfies $\nu < \omega n$, where in the latter case there is no guarantee about the size of the deviation. In the former case, if we take the upper bound (7.2) on the deviation to be a measure of non-randomness of the system's output sequence and view the number of states 2^k as the complexity of the finite state machine (which is based on model $\mathcal{M}_k$), then it follows that the *larger* the complexity the *less* random the output. This is evident in numerical simulations [17] (section 6.3). Also, the output becomes less random if the mismatch $k - k^*$ between the system's model order and the environment's Markov order grows or if the environment's Markov order k^* increases.

Note that the threshold $\mathcal{T}$ depends on the random variables α_i, $0 \leq i \leq 2^k - 1$, and hence the theorem allows for evaluating $\mathcal{T}$ after the random sample $X^{(m)}$ is drawn. Thus the result of the theorem applies to the output of a system which has adapted to its random environment, namely, it depends on the random sample $X^{(m)}$. As we describe in the next section, we use data-dependent concentration bounds to achieve this.

If k^* is known but yet the system is constrained to use some k which may differ from k^*, then we have the following corollary which has an improved (lower) threshold $\mathcal{T}$ in case $k \geq k^*$ (Q and hence ρ are still unknown as above). The advantage in this case is that the length m of the sample $X^{(m)}$ can be smaller while still maintaining the result of the theorem for the same given values of δ, n, ℓ.

Corollary 7.1. *With the same definitions and under the same conditions as in Theorem 7.1, except that* k^* *is known by the system, if the system uses a* *decision function $\hat{d}(X^{(m)})$ defined according to (5.2) with the following choice for $\mathcal{T}$,*

$$\mathcal{T}(\alpha_i, \delta) := \begin{cases} \sqrt{\frac{3}{2\alpha_i m} \ln\left(\frac{2^{k+3}}{\delta\alpha_i}\right)} & \text{if} \quad k \geq k^* \\ \varrho\sqrt{\frac{32}{m\alpha_i^2} \ln\left(\frac{2^{k+4}}{\delta\alpha_i}\right)} & \text{if} \quad k < k^*, \end{cases} \tag{7.3}$$

then the statement of Theorem 7.1 holds.

The proofs of Theorem 7.1 and Corollary 7.1 differ by a small amount which is described in the last paragraph of section 8.

8. Proof of Theorem 7.1

Considering all possible decision functions based on Markov models $\mathcal{M}_k$ of order k, the Bayes decision function d^* yields the minimal expected number of decision mistakes. Its decision at state $s^{(i)}$ is $d^*(i) = 1$ if the true unknown probability $p(1|i) \geq \frac{1}{2}$ and $d^*(i) = 0$ otherwise, $0 \leq i \leq 2^k - 1$. The decision function $\hat{d}$ in (5.1) is obtained based on the sample $X^{(m)}$ and is hence a random variable itself that may differ from the optimal Bayes decision function.

We need to bound the absolute deviation between the frequency of 1 in the system's output $\Xi^{(\nu)}$ which is selected by this random decision function $\hat{d}$. We define the relation $\asymp$ on the set $\Upsilon^{2^k} \times \Upsilon^{2^k}$ as follows: for d, $d' \in \Upsilon^{2^k}$, $d \asymp d'$ if for all $0 \leq i \leq 2^k - 1$ either $d(i) = d'(i)$ or one of the values $d(i)$, $d'(i)$, is equal to REJECT. That is,

$$\begin{aligned} d \asymp d' \iff & \forall\, 0 \leq i \leq 2^k - 1, \\ & (d(i) = d'(i)) \vee (d'(i) = REJECT) \vee (d(i) = REJECT) \end{aligned}$$

and

$$\begin{aligned} d \not\asymp d' \iff & \exists\, 0 \leq i \leq 2^k - 1, \\ & (d(i) \neq d'(i)) \wedge (d'(i) \neq REJECT) \wedge (d^*(i) \neq REJECT) \end{aligned}$$

When $d \asymp d'$ we say that d and d' are equal up to rejects. When $d \not\asymp d'$ we say that d and d' are 'strictly' different. The Bayes' decision function d^* has no rejects, that is $d^*(i) \in \{0, 1\}$ thus the event $\hat{d} \not\asymp d^*$ means that there exists a state $s^{(i)}$ such that $\hat{d}(i) \neq d^*(i)$ and $\hat{d}(i) \neq REJECT$ and $\alpha_i > 0$ (recall, from section 5, if $\alpha_i = 0$ then $\hat{d}(i) = REJECT$).

To obtain a bound on the probability that the absolute deviation is larger than ϵ we consider two cases, $\hat{d} \not\asymp d^*$ and $\hat{d} \asymp d^*$. In the former, we bound the probability that $\hat{d}$ strictly differs from d^*. In the latter case, we bound the probability that the deviation is larger than ϵ for an output sequence $\Xi^{(\nu)}$ of a system which is based on a decision function that is equal to the Bayes' optimal function d^* up to rejects.

The theorem and corollary state finite sample sizes m and n, thus the proof does not rely on any asymptotic approximations. As mentioned above, we derive concentration bounds that are data-dependent. These two aspects of the proof come at the expense of more work and a longer exposition. Therefore we summarize next the main steps of the proof.

8.1. Summary of main steps

Denote by $E_{\hat{d},\epsilon}^{(\ell)}$ the event that the absolute deviation between the frequency of 1 in the system's output sequence $\Xi^{(\nu)}$ which is selected by the system's decision function $\hat{d}$, and the expected value of 1 given that $\hat{d}$ decides 0, is larger than ϵ and its length ν satisfies $\nu \geq \ell$. Then

$$\begin{aligned}\mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}\right) &= \mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}\middle|\hat{d} \not\asymp d^*\right)\mathbb{P}\left(\hat{d} \not\asymp d^*\right) + \mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}\middle|\hat{d} \asymp d^*\right)\mathbb{P}\left(\hat{d} \asymp d^*\right) \\ &\leq \mathbb{P}\left(\hat{d} \not\asymp d^*\right) + \mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}\middle|\hat{d} \asymp d^*\right). \qquad (8.1)\end{aligned}$$

In order to bound (8.1) from above, in section 8.2 we bound from above the first term of (8.1) and in section 8.3 we bound the second term. In section 8.4 we combine these two bounds to yield the result.

Section 8.2 is divided into two subsections. The event $\hat{d} \not\asymp d^*$ implies that there is at least one state $s^{(i)}$ such that the average $\hat{p}(1|i)$ of the number of type-1 transitions from it deviates from its expectation by a large amount. In case of the Corollary, we take advantage of the known value of k^* by splitting into two cases. If $k \geq k^*$, then although the underlying sequence is Markov, this can be represented as a deviation of the average of a sequence of i.i.d. Bernoulli random variables from its mean. This is treated in section 8.2.1 where we exploit this independence and use a Chernoff bound. The second case where $k < k^*$ is more involved and is treated in section 8.2.2; it is split in two subsections 8.2.3 and 8.2.4. Here $\hat{p}(1|i)$ is represented in terms of a sum of a selector function $f_i(s)$ which takes the value of a bit X_{t+1} if the state at time t is $s^{(i)}$. We show that f_i is Lipschitz with a constant 1 and use it in a concentration bound for Markov chains. In section 8.3, there is no independence to be exploited. We use a different selection function f_d that takes the value X_{t+1} if the system's decision $Y_t = 0$. We show that its Lipschitz constant is $r(k, k^*)$ and use the same concentration inequality.

As mentioned in section 7, we derive data-dependent bounds. In section 8.2 this is done to handle the fact that the random variables α_i which appear in the statement of the theorem are dependent on $X^{(m)}$ and in section 8.3 this is done to handle the fact that the length ν of the system's output sequence $\Xi^{(\nu)}$ is dependent on $X^{(n)}$.

8.2. Bounding the probability of the event $\hat{d} \not\asymp d^*$

Consider a state $s^{(i)} \in \mathcal{M}_k$ and define the indicator random variable

$$v_{t_l}^{(i)} = \begin{cases} 1 & \text{if a type 1 transition occurs at state } S_{t_l} = s^{(i)} \\ 0 & \text{otherwise.} \end{cases}$$

We consider two cases,
(a) $k \geq k^*$
(b) $k < k^*$.

In case (a), the probability that $v_{t_l}^{(i)} = 1$ only depends on the state $s^{(i)}$ at time t_l, that is on $S_{t_l} = s^{(i)}$. Therefore from (3.4) we have

$$\mathbb{P}\left(v_{t_l}^{(i)} = 1\right) = q\left(1 \middle| < s^{(i)} >_{k^*}\right). \tag{8.2}$$

Since the probability $q\left(1 \,|< s^{(i)} >_{k^*}\right)$ is a parameter value of the model $\mathcal{M}_{k^*}$ then it is constant with respect to time t_l.

In case (b), from (3.5) we have

$$\mathbb{P}\left(v_{t_l}^{(i)} = 1\right) = \sum_{j=0}^{2^{k^*-k}-1} q\left(1 \middle| S_{t_l}^* = s^{*(ji)}\right) \mathbb{P}(j|i) \tag{8.3}$$

where both of the probabilities in the sum of (8.3) are completely determined by the environment's model $\mathcal{M}_{k^*}$.

We henceforth denote by

$$p(1|i) \quad := \quad \mathbb{P}\left(v_{t_l}^{(i)} = 1\right) \tag{8.4}$$

remembering that it is either (8.2) or (8.3) depending on whether case (a) or (b) holds, respectively. We also have

$$p(1|i) \quad = \quad \mathbb{E}\hat{p}(1|i) \tag{8.5}$$

in both cases, where expectation is taken with respect to the stationary probability distribution $\mathbb{P}$. This follows from the following: let $1 \leq t_1, \ldots, t_{\alpha_i m} \leq m$ be random variables that represent the time instances at which the state $S_{t_l} = s^{(i)}$. Denote by $\mathbb{E}_{\alpha_i}$ the expected value with respect to the probability distribution of α_i, then we have

$$\begin{aligned} \mathbb{E}\hat{p}(1|i) &= \mathbb{E}_{\alpha_i}\mathbb{E}\left[\sum_{l=1}^{\alpha_i m} \frac{1}{\alpha_i m} v_{t_l}^{(i)} \middle| \alpha_i\right] \\ &= \mathbb{E}_{\alpha_i}\mathbb{E}\left[\frac{1}{\alpha_i m} v_{t_1}^{(i)} + \cdots + \frac{1}{\alpha_i m} v_{t_{\alpha_i m}}^{(i)} \middle| \alpha_i\right]. \end{aligned} \tag{8.6}$$

The inner expectation above is the expected value of a sum of $\alpha_i m$ i.i.d. random variables $(1/\alpha_i m)v_{t_l}^{(i)}$, $1 \leq l \leq \alpha_i m$, each with expected value $p(1|i)/\alpha_i m$. Continuing from (8.6) we have

$$\begin{aligned}\mathbb{E}\hat{p}(1|i) &= \mathbb{E}_{\alpha_i}\left[\alpha_i m(1/\alpha_i m)p(1|i)\right] \\ &= p(1|i).\end{aligned}$$

Consider the event $\hat{d} \not\asymp d^*$. By definition of the relation $\asymp$, this event implies that there exists an $i \in \{0,\ldots,2^k-1\}$ such that $\hat{d}(i) \neq d^*(i)$ and $\hat{d}(i) \neq$ *REJECT* and $\alpha_i > 0$. If $p(1|i) \leq \frac{1}{2}$ then $d^*(i) = 0$ and therefore the event $\hat{d}(i) \neq d^*(i)$ implies that $\hat{d}(i) = 1$. According to the decision function (5.2) this implies that $\hat{p}(1|i) > \frac{1}{2} + \mathcal{T}(\alpha_i,\eta)$. If $p(1|i) > \frac{1}{2}$ then $d^*(i) = 1$ and in this case the event $\hat{d}(i) \neq d^*(i)$ implies that $\hat{p}(1|i) < \frac{1}{2} - \mathcal{T}(\alpha_i,\eta)$.

Thus the event that there exists an $i \in \{0,\ldots,2^k-1\}$ with $\alpha_i > 0$ such that $\hat{d}(i) \neq d^*(i)$ and $\hat{d}(i) \neq$ *REJECT* implies the existence of some $i \in \{0,\ldots,2^k-1\}$ such that.. $|\hat{p}(1|i) - p(1|i)| > \mathcal{T}(\alpha_i,\eta)$, $\alpha_i > 0$. (Here $\hat{p}(1|i)$ depends on α_i since it is the number of type-1 transitions from state $s^{(i)}$ in the sequence $s^{(m)}$ divided by $m\alpha_i$, see (3.10)).

Let

$$I: \quad = \quad \left\{\frac{1}{m},\ldots,1\right\} \tag{8.7}$$

and let

$$A_d := \left\{x^{(m)} \in \{0,1\}^m : \exists 0 \leq i \leq 2^k-1,\ \hat{d}\,(i) \neq d(i),\ \hat{d}\,(i) \neq REJECT\right\}. \tag{8.8}$$

The event $\hat{d} \not\asymp d^*$ is equivalent to the event that $X^{(m)}$ belongs to the set A_{d^*}. We wish to bound from above the probability of A_{d^*} hence we use the following fact,

$$\begin{aligned}A_{d^*} &\subseteq \left\{x^{(m)} \in \{0,1\}^m : \exists 0 \leq i \leq 2^k-1,\ |\hat{p}(1|i) - p(1|i)| > \mathcal{T}(\alpha_i,\eta),\ \alpha_i\left(x^{(m)}\right) > 0\right\} \\ &= \left\{x^{(m)} \in \{0,1\}^m : \exists 0 \leq i \leq 2^k-1,\ \exists\gamma_i \in I,\ |\hat{p}(1|i) - p(1|i)| > \mathcal{T}(\gamma_i,\eta),\ \alpha_i\left(x^{(m)}\right) = \gamma_i\right\} \\ &\subseteq \left\{x^{(m)} \in \{0,1\}^m : \exists 0 \leq i \leq 2^k-1,\ \exists\gamma_i \in I,\ |\hat{p}(1|i) - p(1|i)| > \mathcal{T}(\gamma_i,\eta),\ \alpha_i\left(x^{(m)}\right) \geq \gamma_i\right\}.\end{aligned}$$

Let $0 < \delta_i \leq 1$ and choose η to be a function $\eta(\gamma_i,\delta_i)$ such that

$$\mathbb{P}\left(\exists\gamma_i \in I,\ |\hat{p}(1|i) - p(1|i)| > \mathcal{T}\left(\gamma_i,\eta(\gamma_i,\delta_i)\right),\ \alpha_i \geq \gamma_i\right) \leq \delta_i. \tag{8.9}$$

Then for any $0 < \delta \leq 1$ if we let

$$\delta_i \quad = \quad \frac{\delta}{2^{k+1}} \tag{8.10}$$

we obtain

$$\begin{aligned}
\mathbb{P}\left(A_{d^*}\right) &\leq \sum_{i=0}^{2^k-1} \mathbb{P}\left(\exists \gamma_i \in I,\, |\hat{p}(1|i) - p(1|i)| > \mathcal{T}\left(\gamma_i, \eta(\gamma_i, \delta_i)\right),\, \alpha_i \geq \gamma_i\right) \\
&\leq \frac{\delta}{2} \sum_{i=0}^{2^k-1} \frac{1}{2^k} \\
&= \frac{\delta}{2}.
\end{aligned} \tag{8.11}$$

We henceforth hide the dependence on η and write simply A_{d^*} for $A_{d^*,\eta}$.

In the next sections we derive the functions $\mathcal{T}(\gamma_i, \eta)$ and $\eta(\gamma_i, \delta_i)$ such that (8.11) holds.

8.2.1. $\mathcal{T}$ in case (a)

In case (a), although the sequence is a Markov chain, the indicator random variables $v_{t_l}^{(i)}$ are i.i.d. Bernoulli with success probability $p(1|i) := q(1| < s^{(i)} >_{k^*})$. To see that, consider two instants t_l and t_r where the state is $s^{(i)}$ then we have

$$\begin{aligned}
\mathbb{P}\left(v_{t_l}^{(i)} = 1 \middle| v_{t_r}^{(i)}\right) &= \mathbb{P}\left(v_{t_l}^{(i)} = 1 \middle| S_{t_l} = s^{(i)},\, v_{t_r}^{(i)}\right) \\
&= \mathbb{P}\left(v_{t_l}^{(i)} = 1 \middle| S_{t_l} = s^{(i)}\right) \\
&\overset{(\mathtt{i})}{=} \mathbb{P}\left(v_{t_l}^{(i)} = 1\right) \\
&\overset{(\mathtt{ii})}{=} p(1|i)
\end{aligned}$$

where the equality (`i`) shows the independence of $v_{t_l}^{(i)}$ and $v_{t_r}^{(i)}$ and the equality (`ii`) which follows from (8.4) implies that they are identically distributed. Thus in case (a), $\hat{p}(1|i)$ is the average of a subsequence of $\alpha_i m$ i.i.d. Bernoulli random variables where $\alpha_i > 0$ is a random variable taking values in I. (Clearly, the random variables α_i, $0 \leq i \leq 2^k - 1$, are mutually dependent but this is irrelevant here.)

The term in (8.9) that we need to bound from above by δ_i is

$$\begin{aligned}
&\mathbb{P}\left(\exists \gamma_i \in I :\, |\hat{p}(1|i) - p(1|i)| > \mathcal{T}(\gamma_i, \eta(\gamma_i, \delta_i)),\, \alpha_i \geq \gamma_i\right) \\
&= \mathbb{P}\left(\bigcup_{\gamma_i \in I} \left\{|\hat{p}(1|i) - p(1|i)| > \mathcal{T}(\gamma_i, \eta(\gamma_i, \delta_i)),\, \alpha_i \geq \gamma_i\right\}\right).
\end{aligned} \tag{8.12}$$

We have

$$\begin{aligned}\mathbb{P}\left(\hat{p}(1|i) > p(1|i) + \mathcal{T},\, \alpha_i \geq \gamma_i\right) &= \sum_{r \geq \gamma_i} \mathbb{P}\left(\hat{p}(1|i) > p(1|i) + \mathcal{T},\, \alpha_i = r\right) \\ &= \sum_{r \geq \gamma_i} \mathbb{P}\left(\hat{p}(1|i) > p(1|i) + \mathcal{T} \middle| \alpha_i = r\right) \mathbb{P}\left(\alpha_i = r\right).\end{aligned} \tag{8.13}$$

Using Chernoff's bound [6] for the average of a sequence of i.i.d. Bernoulli random variables we obtain

$$\mathbb{P}\left(\hat{p}(1|i) > p(1|i) + \mathcal{T} \middle| \alpha_i = r\right) \quad \leq \quad \exp\left\{-2rm\mathcal{T}^2\right\} \tag{8.14}$$

thus (8.13) is bounded from above by

$$\begin{aligned}\sum_{r \geq \gamma_i} \exp\left\{-2rm\mathcal{T}^2\right\} \mathbb{P}\left(\alpha_i = r\right) &\leq \exp\left\{-2\gamma_i m\mathcal{T}^2\right\} \sum_{r \geq \gamma_i} \mathbb{P}\left(\alpha_i = r\right) \\ &\leq \exp\left\{-2\gamma_i m\mathcal{T}^2\right\}.\end{aligned}$$

In a similar way one obtains

$$\mathbb{P}\left(\hat{p}(1|i) \leq p(1|i) - \mathcal{T},\, \alpha_i \geq \gamma_i\right) \quad \leq \quad \exp\left\{-2\gamma_i m\mathcal{T}^2\right\} \tag{8.15}$$

therefore we have

$$\mathbb{P}\left(|\hat{p}(1|i) - p(1|i)| > \mathcal{T},\, \alpha_i \geq \gamma_i\right) \quad \leq \quad 2\exp\left\{-2\gamma_i m\mathcal{T}^2\right\}. \tag{8.16}$$

Let

$$\mathcal{T}^{(a)}\left(\gamma_i, \eta\right) = \sqrt{\frac{1}{2\gamma_i m} \ln\left(\frac{2}{\eta}\right)} \tag{8.17}$$

then we have for any $0 < \eta := \eta(\gamma_i, \delta_i) \leq 1$,

$$\mathbb{P}\left(|\hat{p}(1|i) - p(1|i)| > \mathcal{T}^{(a)}\left(\gamma_i, \eta(\gamma_i, \delta_i)\right), \alpha_i \geq \gamma_i\right) \leq \eta(\gamma_i, \delta_i). \tag{8.18}$$

Following a method of proof of Proposition 8 in [2] let us now define

$$J(\gamma^{(1)}, \gamma^{(2)}, \eta) \quad := \quad \left\{x^{(m)} : |\hat{p}(1|i) - p(1|i)| > \mathcal{T}^{(a)}\left(\gamma^{(1)}, \eta\right),\, \alpha_i(x^{(m)}) \geq \gamma^{(2)}\right\}.$$

Note that because $\mathcal{T}$ decreases with respect to increasing γ_i and η then for $\gamma^{(1)} \leq \gamma \leq \gamma^{(2)}$ we have $J(\gamma^{(1)}, \gamma^{(2)}, \eta) \subseteq J(\gamma, \gamma, \eta)$ and for $\eta_a \leq \eta_b$ we have $J(\gamma, \gamma, \eta_a) \subseteq J(\gamma, \gamma, \eta_b)$.

From (8.18) we have,

$$\begin{aligned}\mathbb{P}\left(J\left(\gamma_i,\gamma_i,\eta(\gamma_i,\delta_i)\right)\right) &= \mathbb{P}\left(|\hat{p}(1|i)-p(1|i)| > \mathcal{T}^{(a)}\left(\gamma_i,\eta(\gamma_i,\delta_i)\right), \alpha_i \geq \gamma_i\right) \\ &\leq \eta(\gamma_i,\delta_i). \end{aligned} \tag{8.19}$$

Define the set $\Gamma_j \subset [0,1]$ as follows,

$$\Gamma_j = \left[\left(\frac{1}{2}\right)^{j+1}, \left(\frac{1}{2}\right)^{j}\right]. \tag{8.20}$$

The set I defined in (8.7) clearly satisfies $I \subseteq \bigcup_{j=0}^{\infty}\Gamma_j$.

In (8.12) let us choose

$$\mathcal{T}(\gamma_i,\eta(\gamma_i,\delta_i)) = \mathcal{T}^{(a)}\left(\frac{\gamma_i}{3},\eta(\gamma_i,\delta_i)\right) \tag{8.21}$$

with

$$\eta(\gamma_i,\delta_i) := \frac{\delta_i\gamma_i}{2}. \tag{8.22}$$

Then (8.12) is now bounded from above as follows,

$$\mathbb{P}\left(\bigcup_{\gamma_i\in I}\{|\hat{p}(1|i)-p(1|i)| > \mathcal{T}(\gamma_i,\eta(\gamma_i,\delta_i)),\ \alpha_i\geq\gamma_i\}\right)$$

$$= \mathbb{P}\left(\bigcup_{\gamma_i\in I}\left\{|\hat{p}(1|i)-p(1|i)| > \mathcal{T}^{(a)}\left(\frac{\gamma_i}{3},\eta(\gamma_i,\delta_i)\right),\ \alpha_i\geq\gamma_i\right\}\right)$$

$$= \mathbb{P}\left(\bigcup_{\gamma_i\in I} J\left(\frac{\gamma_i}{3},\gamma_i,\frac{\delta_i\gamma_i}{2}\right)\right) \tag{8.23}$$

$$\leq \mathbb{P}\left(\bigcup_{j=0}^{\infty}\bigcup_{\gamma_i\in\Gamma_j} J\left(\frac{\gamma_i}{3},\gamma_i,\frac{\delta_i\gamma_i}{2}\right)\right) \tag{8.24}$$

$$\leq \sum_{j=0}^{\infty}\mathbb{P}\left(\bigcup_{\gamma_i\in\Gamma_j} J\left(\frac{\gamma_i}{3},\gamma_i,\frac{\delta_i\gamma_i}{2}\right)\right)$$

$$\leq \sum_{j=0}^{\infty}\mathbb{P}\left(J\left(\left(\frac{1}{2}\right)^{j+1},\left(\frac{1}{2}\right)^{j+1},\frac{\delta_i}{2}\left(\frac{1}{2}\right)^{j}\right)\right) \tag{8.25}$$

where the last inequality follows from the fact that for all $\gamma \in \Gamma_j$ we have $\gamma/3 \leq (\frac{1}{2})^{j+1} \leq \gamma$ and $\gamma \leq (\frac{1}{2})^j$. Using (8.19) it follows that (8.25) is bounded from above by

$$\frac{\delta_i}{2}\sum_{j=0}^{\infty}\left(\frac{1}{2}\right)^{j} \leq \delta_i$$

as required for (8.9) to hold. Thus in case (a) our choice for $\mathcal{T}$ in (8.11) is as defined in (8.21) and the choice for $\eta(\gamma_i, \delta_i)$ is as defined in (8.22). This gives $\mathcal{T}(\gamma_i, \eta(\gamma_i, \delta_i)) = \mathcal{T}^{(a)}\left(\frac{\gamma_i}{3}, \frac{\delta_i \gamma_i}{2}\right)$.

8.2.2. $\mathcal{T}$ in case (b)

In case (b) the indicator random variables $v_{t_l}^{(i)}$ are dependent. So we use a different bound for the convergence of $\hat{p}(1|i)$ to $p(1|i)$.

Let us fix $i \in \{0, \ldots, 2^k - 1\}$. For a state $s^* \in \mathbb{S}_{k^*}$ and $0 \leq r < q \leq k^* - 1$ we denote by $< s^* >_r^q$ the binary vector $[s_q^*, \ldots, s_r^*] \in \{0,1\}^{q-r+1}$. Define $f_i : \mathbb{S}_{k^*} \to \{0,1\}$ as follows: for a state $s^* \in \mathbb{S}_{k^*}$ and $s^{(i)} \in \mathbb{S}_k$ let

$$f_i(s^*) = \begin{cases} 0, & \text{if } < s^* >_1^k \neq s^{(i)} \\ < s^* >_0, & \text{if } < s^* >_1^k = s^{(i)}. \end{cases}$$

In words, given a state $s^* = [s_{k^*-1}^*, \ldots, s_k^*, \ldots, s_1^*, s_0^*] \in \mathbb{S}_{k^*}$ if $[s_k^*, \ldots, s_1^*] \neq s^{(i)}$ then $f_i(s^*) = 0$. Else, $[s_k^*, \ldots, s_1^*] = s^{(i)}$ and $f_i(s^*)$ equals $< s^* >_0 = s_0^*$. For a sequence $s^{*(m)} = \{s_t^*\}_{t=1}^m$, $s_t^* \in \mathbb{S}_{k^*}$, $1 \leq t \leq m$, define

$$F_i(s^{*(m)}) := \sum_{t=1}^m f_i(s_t^*).$$

Therefore $\hat{p}(1|i)$ can be represented as

$$\hat{p}(1|i) = \frac{F_i(S^{*(m)})}{m\alpha_i(X^{(m)})} \tag{8.26}$$

where $S^{*(m)}$ is defined in (3.6). To see this, consider a sequence $S^{(m)}$ in which state $s^{(i)}$ appears at least once, that is $\alpha_i m \geq 1$. Then for its corresponding sequence $S^{*(m)}$ we have

$$\frac{1}{\alpha_i m}\sum_{t=1}^m f_i(S_t^*) = \frac{1}{\alpha_i m}\sum_{l:<S_{t_l}^*>_1^k = s^{(i)}} f_i(S_{t_l}^*) \tag{8.27}$$

$$= \frac{1}{\alpha_i m}\sum_{l:<S_{t_l}^*>_1^k = s^{(i)}} v_{t_l-1}^{(i)}$$

$$= \frac{1}{\alpha_i m}\sum_{r=1}^{m\alpha_i} v_{t_r}^{(i)} \tag{8.28}$$

$$= \hat{p}(1|i)$$

where (8.28) holds by definition of the indicator variables $v_{t_r}^{(i)}$ whose index t_r runs over those time instants where the state $S_{t_r} = s^{(i)}$, $s^{(i)} \in \mathcal{M}_k$, $t_r \in$

$\{0, \ldots, m-1\}$. We have

$$\begin{aligned}
\mathbb{E}\left[F_i(S^{*(m)})\right] &= \mathbb{E}_{\alpha_i}\left[\mathbb{E}\left[F_i(S^{*(m)})\middle|\alpha_i\right]\right] \\
&= \mathbb{E}_{\alpha_i}\left[\mathbb{E}\left[\sum_{l=1}^{m\alpha_i} v_{t_l}^{(i)}\middle|\alpha_i\right]\right] \\
&= \mathbb{E}_{\alpha_i}[m\alpha_i p(1|i)] \\
&= mp(1|i)\mathbb{E}[\alpha_i] \\
&= mp(1|i)\pi_i. \qquad (8.29)
\end{aligned}$$

For two sequences $s^{*(m)}$ and $q^{*(m)} \in \mathbb{S}_{k^*}^m$ let us use the Hamming metric $d_H : \mathbb{S}_{k^*}^m \times \mathbb{S}_{k^*}^m \to \{0, 1, \ldots m\}$ which is defined by

$$d_H(s^{*(m)}, q^{*(m)}) := \sum_{t=1}^{m} \mathbb{I}\{s_t^* \neq q_t^*\}.$$

Consider some $r \geq 1$ and consider $s^{*(m)} = (s_1^*, \ldots, s_m^*)$ and $q^{*(m)} = (q_1^*, \ldots, q_m^*) \in \mathbb{S}_{k^*}^m$ such that $d_H(s^{*(m)}, q^{*(m)}) \leq r$. Define the following subsets of $\{1, \ldots, m\}$,

$$\begin{aligned}
I_1 &:= \left\{l :< s_{t_l}^* >_1^k \neq s^{(i)}, < q_{t_l}^* >_1^k \neq s^{(i)}\right\} \\
I_2 &:= \left\{l :< s_{t_l}^* >_1^k \neq s^{(i)}, < q_{t_l}^* >_1^k = s^{(i)}\right\} \\
I_3 &:= \left\{l :< s_{t_l}^* >_1^k = s^{(i)}, < q_{t_l}^* >_1^k \neq s^{(i)}\right\} \\
I_4 &:= \left\{l :< s_{t_l}^* >_1^k = s^{(i)}, < q_{t_l}^* >_1^k = s^{(i)}\right\} \qquad (8.30)
\end{aligned}$$

then we have

$$\begin{aligned}
\left|F_i(s^{*(m)}) - F_i(q^{*(m)})\right| &= \left|\sum_{t=1}^{m} f_i(s_t^*) - \sum_{t=1}^{m} f_i(q_t^*)\right| \\
&\le \sum_{t=1}^{m} |f_i(s_t^*) - f_i(q_t^*)| \\
&\overset{(\mathtt{i})}{=} \sum_{l\in I_1} \left|f_i(s_{t_l}^*) - f_i(q_{t_l}^*)\right| + \sum_{l\in I_2} \left|f_i(s_{t_l}^*) - f_i(q_{t_l}^*)\right| + \sum_{l\in I_3} \left|f_i(s_{t_l}^*) - f_i(q_{t_l}^*)\right| \\
&\quad + \sum_{l\in I_4} \left|f_i(s_{t_l}^*) - f_i(q_{t_l}^*)\right| \\
&= \sum_{l\in I_2} \left|0- < q_{t_l}^* >_0\right| + \sum_{l\in I_3} \left|< s_{t_l}^* >_0 -0\right| + \sum_{l\in I_4} \left|< s_{t_l}^* >_0 - < q_{t_l}^* >_0\right| \\
&\le \sum_{l\in I_2} \mathbb{I}\left\{s_{t_l}^* \neq q_{t_l}^*\right\} + \sum_{l\in I_3} \mathbb{I}\left\{s_{t_l}^* \neq q_{t_l}^*\right\} + \sum_{l\in I_4} \mathbb{I}\left\{s_{t_l}^* \neq q_{t_l}^*\right\} \\
&\le \sum_{t=1}^{m} \mathbb{I}\left\{s_t^* \neq q_t^*\right\} \\
&= d_H(s^{*(m)}, q^{*(m)}) \\
&\le r. \qquad (8.31)
\end{aligned}$$

Hence it follows that the function F_i is Lipschitz with constant 1.

In the Appendix, we show that the chains $S^{*(m)}$ and $S^{*(n)}$ satisfy the condition of a concentration bound (Lemma A.1) which holds for a Lipschitz function with constant 1. Hence it follows that we can apply the lemma for F_i. Before we do that we have for any non-negative numbers a,b, c, d,

$$\begin{aligned}
|ab - cd| &= |ab - ad + ad - cd| \\
&\le a\,|b - d| + d\,|a - c|. \qquad (8.32)
\end{aligned}$$

For any fixed $i \in \left\{0, \dots, 2^k - 1\right\}$ and positive α_i we have,

$$\begin{aligned}
|\hat{p}(1|i) - p(1|i)| &= \left|\frac{1}{\alpha_i}\alpha_i\hat{p}(1|i) - \frac{1}{\pi_i}\pi_i p(1|i)\right| \\
&\le \frac{1}{\alpha_i}\left|\alpha_i\hat{p}(1|i) - \pi_i p(1|i)\right| + \pi_i p(1|i)\left|\frac{1}{\alpha_i} - \frac{1}{\pi_i}\right| \\
&= \frac{1}{\alpha_i}\left|\alpha_i\hat{p}(1|i) - \pi_i p(1|i)\right| + \frac{p(1|i)}{\alpha_i}\left|\pi_i - \alpha_i\right|.
\end{aligned}$$

Hence,

$$
\begin{aligned}
&\mathbb{P}\left(\exists\gamma_i \in I,\, |\hat{p}(1|i) - p(1|i)| > \mathcal{T}\left(\gamma_i,\eta\right),\, \alpha_i \geq \gamma_i\right) \\
&\quad\leq \quad \mathbb{P}\left(\exists\gamma_i \in I:\, \frac{1}{\alpha_i}\left|\alpha_i\hat{p}(1|i) - \pi_i p(1|i)\right| > \frac{\mathcal{T}\left(\gamma_i,\eta\right)}{2},\, \alpha_i \geq \gamma_i\right) \\
&\qquad\quad + \mathbb{P}\left(\exists\gamma_i \in I:\, \frac{p(1|i)}{\alpha_i}\left|\pi_i - \alpha_i\right| > \frac{\mathcal{T}\left(\gamma_i,\eta\right)}{2},\, \alpha_i \geq \gamma_i\right) \\
&\quad= \quad \mathbb{P}\left(\exists\gamma_i \in I:\, \left|\alpha_i\hat{p}(1|i) - \pi_i p(1|i)\right| > \frac{\alpha_i\mathcal{T}\left(\gamma_i,\eta\right)}{2},\, \alpha_i \geq \gamma_i\right) \\
&\qquad\quad + \mathbb{P}\left(\exists\gamma_i \in I:\, \left|\pi_i - \alpha_i\right| > \frac{\alpha_i\mathcal{T}\left(\gamma_i,\eta\right)}{2p(1|i)},\, \alpha_i \geq \gamma_i\right). \qquad (8.33)
\end{aligned}
$$

8.2.3. Bounding the first term of (8.33)

We have

$$
\begin{aligned}
&\mathbb{P}\left(\exists\gamma_i \in I:\, \left|\alpha_i\hat{p}(1|i) - \pi_i p(1|i)\right| > \frac{\alpha_i\mathcal{T}\left(\gamma_i,\eta\right)}{2},\, \alpha_i \geq \gamma_i\right) \\
&\quad= \quad \mathbb{P}\left(\exists\gamma_i \in I:\, \left|\hat{p}(1|i) - \frac{\pi_i p(1|i)}{\alpha_i}\right| > \frac{\mathcal{T}\left(\gamma_i,\eta\right)}{2},\, \alpha_i \geq \gamma_i\right) \\
&\quad\leq \quad \mathbb{P}\left(\bigcup_{0<\gamma_i\leq 1}\left\{\left|\hat{p}(1|i) - \frac{\pi_i p(1|i)}{\alpha_i}\right| > \frac{\mathcal{T}\left(\gamma_i,\eta\right)}{2},\, \alpha_i \geq \gamma_i\right\}\right). \qquad (8.34)
\end{aligned}
$$

In the following, we avoid conditioning on the value of α_i in order not to violate the Markov property of the probability distribution (this property is assumed in Lemma A.1). Let us bound the probability of one term in the above union. We have,

$$
\begin{aligned}
&\mathbb{P}\left(\left|\hat{p}(1|i) - \frac{\pi_i p(1|i)}{\alpha_i}\right| > \frac{\mathcal{T}}{2},\, \alpha_i \geq \gamma_i\right) \\
&\quad= \quad \mathbb{P}\left(\left|F_i(S^{*(m)}) - \mathbb{E}\left[F_i(S^{*(m)})\right]\right| > \frac{m\alpha_i\mathcal{T}}{2},\, \alpha_i \geq \gamma_i\right) \\
&\quad= \quad \sum_{s^{*(m)}:\alpha_i(s^{*(m)})\geq\gamma_i} \mathbb{P}\left(s^{*(m)}\right)\mathbb{I}\left\{\left|F_i(s^{*(m)}) - \mathbb{E}F_i(S^{*(m)})\right| > \frac{m\alpha_i(s^{*(m)})\mathcal{T}}{2}\right\}
\end{aligned}
$$

(8.35)

and (8.35) is bounded from above by

$$\begin{aligned}
&\sum_{s^{*(m)}:\alpha_i(s^{*(m)})\geq\gamma_i} \mathbb{P}\left(s^{*(m)}\right)\mathbb{I}\left\{\left|F_i(s^{*(m)})-\mathbb{E}F_i(S^{*(m)})\right|>\frac{m\gamma_i\mathcal{T}}{2}\right\}\\
&\leq \sum_{s^{*(m)}\in\mathbb{S}^m_{k^*}} \mathbb{P}\left(s^{*(m)}\right)\mathbb{I}\left\{\left|F_i(s^{*(m)})-\mathbb{E}F_i(S^{*(m)})\right|>\frac{m\gamma_i\mathcal{T}}{2}\right\}\\
&= \mathbb{P}\left(\left|F_i(S^{*(m)})-\mathbb{E}F_i(S^{*(m)})\right|>\frac{m\gamma_i\mathcal{T}}{2}\right). \qquad (8.36)
\end{aligned}$$

Now, use Lemma A.1 for the Markov chain $S^{*(m)}$ and F_i. Applying (A.98) we obtain

$$\mathbb{P}\left(\left|F_i(S^{*(m)})-\mathbb{E}F_i(S^{*(m)})\right|>m\kappa\right)\leq 2\exp\left\{-\frac{m}{2}\left(\frac{\kappa}{\rho}\right)^2\right\} \qquad (8.37)$$

where ρ is defined in (6.3). It follows that

$$\mathbb{P}\left(\left|F_i(S^{*(m)})-\mathbb{E}F_i(S^{*(m)})\right|>\frac{m\gamma_i\mathcal{T}}{2}\right)\leq 2\exp\left\{-m\gamma_i^2\mathcal{T}^2/8\rho^2\right\}. \qquad (8.38)$$

Let

$$\mathcal{T}^{(b)}\left(\gamma_i,\eta\right):=\frac{2\rho}{\gamma_i}\sqrt{\frac{2}{m}\ln\left(\frac{2}{\eta}\right)} \qquad (8.39)$$

then for any $0<\eta(\gamma_i,\delta_i)\leq 1$ and from (8.36), (8.38), (8.39) we have

$$\mathbb{P}\left(\left|\hat{p}(1|i)-\frac{\pi_i p(1|i)}{\alpha_i}\right|>\frac{\mathcal{T}^{(b)}\left(\gamma_i,\eta(\gamma_i,\delta_i)\right)}{2},\ \alpha_i\geq\gamma_i\right)\leq\eta(\gamma_i,\delta_i). \qquad (8.40)$$

Now define

$$J(\gamma^{(1)},\gamma^{(2)},\eta) := \left\{x^{(m)}:\left|\hat{p}(1|i)-\frac{\pi_i p(1|i)}{\alpha_i(x^{(m)})}\right|>\frac{\mathcal{T}^{(b)}(\gamma^{(1)},\eta)}{2},\ \alpha_i(x^{(m)})\geq\gamma^{(2)}\right\}.$$

As in section 8.2.1, for $\gamma^{(1)}\leq\gamma\leq\gamma^{(2)}$ we have $J(\gamma^{(1)},\gamma^{(2)},\eta)\subseteq J(\gamma,\gamma,\eta)$ and for $\eta_a\leq\eta_b$ we have $J(\gamma^{(1)},\gamma^{(2)},\eta_a)\subseteq J(\gamma^{(1)},\gamma^{(2)},\eta_b)$.

From (8.40) we have,

$$\begin{aligned}
\mathbb{P}\left(J\left(\gamma_i,\gamma_i,\eta(\gamma_i,\delta_i)\right)\right) &= \mathbb{P}\left(\left|\hat{p}(1|i)-\frac{\pi_i p(1|i)}{\alpha_i}\right|>\frac{\mathcal{T}^{(b)}(\gamma_i,\eta(\gamma_i,\delta_i))}{2},\ \alpha_i\geq\gamma_i\right)\\
&\leq \eta(\gamma_i,\delta_i). \qquad (8.41)
\end{aligned}$$

Now choose

$$\mathcal{T}\left(\gamma_i,\eta(\gamma_i,\delta_i)\right):=\mathcal{T}^{(b)}\left(\frac{\gamma_i}{2},\eta(\gamma_i,\delta_i)\right) \qquad (8.42)$$

with

$$\eta(\gamma_i,\delta_i):=\frac{\gamma_i\delta_i}{4}. \qquad (8.43)$$

Using the definition of the sets in (8.20) then (8.34) is now bounded as follows,

$$
\begin{aligned}
& \mathbb{P}\left(\bigcup_{0<\gamma_i\le 1}\left\{\left|\hat{p}(1|i)-\frac{\pi_i p(1|i)}{\alpha_i}\right|>\frac{\mathcal{T}\left(\gamma_i,\eta(\gamma_i,\delta_i)\right)}{2},\ \alpha_i\ge\gamma_i\right\}\right) \\
&= \mathbb{P}\left(\bigcup_{0<\gamma_i\le 1}\left\{\left|\hat{p}(1|i)-\frac{\pi_i p(1|i)}{\alpha_i}\right|>\frac{\mathcal{T}^{(b)}\left(\frac{\gamma_i}{2},\frac{\gamma_i\delta_i}{4}\right)}{2},\ \alpha_i\ge\gamma_i\right\}\right) \\
&= \mathbb{P}\left(\bigcup_{j=0}^{\infty}\bigcup_{\gamma_i\in\Gamma_j} J\left(\frac{\gamma_i}{2},\gamma_i,\frac{\delta_i\gamma_i}{4}\right)\right) \qquad (8.44)\\
&\le \sum_{j=0}^{\infty}\mathbb{P}\left(J\left(\left(\frac{1}{2}\right)^{j+1},\left(\frac{1}{2}\right)^{j+1},\frac{\delta_i}{4}\left(\frac{1}{2}\right)^{j}\right)\right) \\
&\le \frac{\delta_i}{4}\sum_{j=0}^{\infty}\left(\frac{1}{2}\right)^{j} \\
&\le \frac{\delta_i}{2} \qquad (8.45)
\end{aligned}
$$

where the first inequality follows from the fact that for all $\gamma\in\Gamma_j$ we have $\gamma/2\le(\frac{1}{2})^{j+1}\le\gamma$ and $\gamma\le(\frac{1}{2})^{j}$ and the second inequality follows from (8.41).

It follows therefore that (8.34) is bounded from above by $\delta_i/2$.

8.2.4. *Bounding the second term of (8.33)*

We have

$$
\begin{aligned}
& \mathbb{P}\left(\exists\gamma_i\in I:\ |\pi_i-\alpha_i|>\frac{\alpha_i\mathcal{T}}{2p(1|i)},\ \alpha_i\ge\gamma_i\right) \\
&\le \mathbb{P}\left(\bigcup_{0<\gamma_i\le 1}\left\{|\pi_i-\alpha_i|>\frac{\alpha_i\mathcal{T}}{2p(1|i)},\ \alpha_i\ge\gamma_i\right\}\right). \qquad (8.46)
\end{aligned}
$$

Let us bound the probability of one term in the above union. Define $g_i:\mathbb{S}_{k^*}\to\{0,1\}$ as follows: for any state $s^*\in\mathbb{S}_{k^*}$ let

$$
g_i(s^*)=\begin{cases}0, & \text{if } <s^*>_1^k\neq s^{(i)}\\ 1, & \text{if } <s^*>_1^k= s^{(i)}.\end{cases}
$$

For a sequence $s^{*(m)}=\{s_t^*\}_{t=1}^m$, $s_t^*\in\mathbb{S}_{k^*}$, $1\le t\le m$, define

$$
G_i(s^{*(m)}):=\sum_{t=1}^{m} g_i(s_t^*).
$$

Using the regions in (8.30) one can show using the same argument as in (8.31) that for any two sequences $s^{*(m)}$ and $q^{*(m)}$ such that $d_H(s^{*(m)}, q^{*(m)}) \leq r$ then $|G_i(s^{*(m)}) - G_i(q^{*(m)})| \leq r$ and hence G_i is Lipschitz with constant 1.

We have

$$\frac{G_i(S^{*(m)})}{m} = \alpha_i$$

and

$$\mathbb{E}\left[G_i(S^{*(m)})\right] = m\pi_i.$$

So,

$$\begin{aligned}
&\mathbb{P}\left(|\alpha_i - \pi_i| > \frac{\alpha_i \mathcal{T}}{2p(1|i)},\, \alpha_i \geq \gamma_i\right) \\
&= \mathbb{P}\left(\left|G_i(S^{*(m)}) - \mathbb{E}\left[G_i(S^{*(m)})\right]\right| > \frac{m\alpha_i \mathcal{T}}{2p(1|i)},\, \alpha_i \geq \gamma_i\right) \\
&= \sum_{s^{*(m)}:\alpha_i(s^{*(m)})\geq\gamma_i} \mathbb{P}\left(s^{*(m)}\right) \mathbb{I}\left\{\left|G_i(s^{*(m)}) - \mathbb{E}G_i(S^{*(m)})\right| > \frac{m\alpha_i(s^{*(m)})\mathcal{T}}{2p(1|i)}\right\} \\
&\leq \sum_{s^{*(m)}:\alpha_i(s^{*(m)})\geq\gamma_i} \mathbb{P}\left(s^{*(m)}\right) \mathbb{I}\left\{\left|G_i(s^{*(m)}) - \mathbb{E}G_i(S^{*(m)})\right| > \frac{m\gamma_i\mathcal{T}}{2p(1|i)}\right\} \\
&\leq \sum_{s^{*(m)}\in\mathbb{S}^m_{k^*}} \mathbb{P}\left(s^{*(m)}\right) \mathbb{I}\left\{\left|G_i(s^{*(m)}) - \mathbb{E}G_i(S^{*(m)})\right| > \frac{m\gamma_i\mathcal{T}}{2p(1|i)}\right\} \\
&= \mathbb{P}\left(\left|G_i(S^{*(m)}) - \mathbb{E}G_i(S^{*(m)})\right| > \frac{m\gamma_i\mathcal{T}}{2p(1|i)}\right).
\end{aligned} \tag{8.47}$$

As above, we use the concentration bound (A.98) for the Markov chain $S^{*(m)}$ and function G_i. For any positive κ this yields

$$\mathbb{P}\left(\left|G_i(S^{*(m)}) - \mathbb{E}G_i(S^{*(m)})\right| > m\kappa\right) \leq 2\exp\left\{-\frac{m}{2}\left(\frac{\kappa}{\rho}\right)^2\right\}. \tag{8.48}$$

It follows that

$$\begin{aligned}
\mathbb{P}\left(\left|G_i(S^{*(m)}) - \mathbb{E}G_i(S^{*(m)})\right| > \frac{m\gamma_i\mathcal{T}}{2p(1|i)}\right) &\leq 2\exp\left\{-m\gamma_i^2\mathcal{T}^2/8p^2(1|i)\rho^2\right\} \\
&\leq 2\exp\left\{-m\gamma_i^2\mathcal{T}^2/8\rho^2\right\}
\end{aligned} \tag{8.49}$$

because $0 < p(1|i) \leq 1$. Using $\mathcal{T}^{(b)}$ defined in (8.39) and with $\eta(\gamma_i, \delta_i)$ defined in (8.43), we obtain

$$\begin{aligned}
&\mathbb{P}\left(|\alpha_i-\pi_i|>\frac{\alpha_i\mathcal{T}^{(b)}\left(\gamma_i,\eta(\gamma_i,\delta_i)\right)}{2p(1|i)},\,\alpha_i\geq\gamma_i\right)\\
&\leq\ \mathbb{P}\left(\left|G_i(S^{*(m)})-\mathbb{E}G_i(S^{*(m)})\right|>\frac{m\gamma_i\mathcal{T}^{(b)}\left(\gamma_i,\eta(\gamma_i,\delta_i)\right)}{2p(1|i)}\right)\\
&\leq\ \eta(\gamma_i,\delta_i). \qquad (8.50)
\end{aligned}$$

Let

$$J(\gamma^{(1)},\gamma^{(2)},\eta)\ :=\ \left\{x^{(m)}:|\alpha_i-\pi_i|>\frac{\alpha_i\mathcal{T}^{(b)}(\gamma^{(1)},\eta)}{2p(1|i)},\,\alpha_i(x^{(m)})>\gamma^{(2)}\right\}.$$

From (8.50) we have,

$$\begin{aligned}
\mathbb{P}\left(J\left(\gamma_i,\gamma_i,\eta(\gamma_i,\delta_i)\right)\right)\ &=\ \mathbb{P}\left(|\alpha_i-\pi_i|>\frac{\alpha_i\mathcal{T}^{(b)}\left(\gamma_i,\eta(\gamma_i,\delta_i)\right)}{2p(1|i)},\,\alpha_i\geq\gamma_i\right)\\
&\leq\ \eta(\gamma_i,\delta_i). \qquad (8.51)
\end{aligned}$$

Substituting the value of (8.42) for $\mathcal{T}$ in (8.46) and substituting for η the expression in (8.43) then we obtain the following upper bound on (8.46),

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{0<\gamma_i\leq 1}\left\{|\pi_i-\alpha_i|>\frac{\alpha_i\mathcal{T}}{2p(1|i)},\,\alpha_i\geq\gamma_i\right\}\right)\\
&=\ \mathbb{P}\left(\bigcup_{j=0}^{\infty}\bigcup_{\gamma_i\in\Gamma_j}J\left(\frac{\gamma_i}{2},\gamma_i,\frac{\delta_i\gamma_i}{4}\right)\right)\\
&\leq\ \frac{\delta_i}{2} \qquad (8.52)
\end{aligned}$$

where the last inequality follows from (8.51) and the same reasoning as in (8.45).

From (8.45) and (8.52) it follows that (8.33) is bounded from above by δ_i as required for (8.9) to hold. Thus for case (b) we choose the value of $\mathcal{T}$ as (8.42) and the choice of η as (8.43). This gives $\mathcal{T}(\gamma_i,\eta(\gamma_i,\delta_i))=\mathcal{T}^{(b)}\left(\frac{\gamma_i}{2},\frac{\delta_i\gamma_i}{4}\right)$.

8.3. Bounding the probability of large deviation

Recall that the adapted system decides Y_t, $1\leq t\leq n$, using the decision function $\hat{d}$. The first decision Y_1 is based on an initial given state (4.3). Let ν be the number of times that $Y_t=0$. Since $\Xi^{(\nu)}$ is a subsequence of $X^{(n)}$ then we can associate a selection rule

$$R_{\hat{d}}:\{0,1\}^n\rightarrow\{0,1\}^\nu$$

based on $\hat{d}$ as defined in (5.2) which selects $\Xi^{(\nu)}$ from $X^{(n)}$. Note that $\nu = \nu(S^{*(n)}, \hat{d})$ is a random variable that is dependent on the random decision function $\hat{d}$ and on the random sequence of states $S^{*(n)}$ that corresponds to $X^{(n)}$, that is, on

$$S^{*(n)} = \{S^*_t\}_{t=1}^n, \quad S^*_t = [X_{T'+t-k^*+1}, \ldots, X_t]$$

where $X_{T'+t}$ is the t^{th} bit of $X^{(n)}$, $1 \leq t \leq n$.

For every $\epsilon > 0$, $1 \leq \ell \leq n$ and $d \in \Upsilon^{2^k}$ we denote by $E^{(\ell)}_{d,\epsilon} \subseteq \{0,1\}^n$ the set of all sequences $x^{(n)}$ for which there exists $\omega \in \left(\frac{\ell}{n}, 1\right]$ such that the subsequence $\xi^{(\nu)}$ of $x^{(n)}$ selected by R_d is of length $\nu \geq \omega n$ and its frequency of 1 deviates from the expected value β_d (defined in (5.3)) by at least ϵ. Formally, this is defined as,

$$E^{(\ell)}_{d,\epsilon} = \left\{ x^{(n)} : \exists \omega \in (\ell/n, 1], \xi^{(\nu)} = R_d\left(x^{(n)}\right), \nu \geq \omega n, \left| \frac{1}{\nu} \sum_{l=1}^{\nu} \xi^{(\nu)}_l - \beta_d \right| > \epsilon \right\}. \tag{8.53}$$

To show the dependence on the decision function we sometimes explicitly write $\Xi^{(\nu)} = R_d\left(X^{(n)}\right)$ where $X^{(n)}$ is the second sample (4.1). We wish to bound the second term of (8.1), which is the probability of the event $E^{(\ell)}_{\hat{d},\epsilon}$ given that $\hat{d} \asymp d^*$. Below, we take an approach to bounding this which holds for any decision function $d \in \Upsilon^{2^k}$.

The technique here is similar to that of the previous section. There, the concentration bound was for a subsequence that consists of time instants at which the random state is $s^{(i)}$. We avoided conditioning on the random empirical frequency α_i since that could make the probability distribution violate the Markov property. In the present section, we apply the concentration result to a subsequence that consists of time instants where the states happen to be members of a set of states at which the selection rule selects a bit from $X^{(n)}$. While these states are determined solely by the selection rule $R_{\hat{d}}$, the length ν of the subsequence depends also on the sequence of states $S^{*(n)}$, namely, it is the number of times that $S^{*(n)}$ visits one of the states in this set. Here too, if we condition on the value of ν this may culminate in a conditional probability distribution of $S^{*(n)}$ which violates the Markov condition. Hence in the following analysis we avoid the conditional distribution as done in the previous section.

The current section differs from the previous one in that we do not need to consider two cases (a) and (b) since the sequence of interest here is the subsequence $\Xi^{(\nu)}$ which consists of dependent random variables, unlike the i.i.d. subsequence of section 8.2.1. To see that, consider first the case where $k \geq k^*$ and define by $L \subseteq \mathbb{S}_k$ the set of states $s^{(i)} \in \mathbb{S}_k$ where $\hat{p}(1|i) < \frac{1}{2} - \mathcal{T}$ (from (5.2) these are the states $S_{t_{t_l}}$ at which the system decides $Y_{t_l+1} = 0$). Let v_{t_l} be the indicator random variable whose value is the type of transition from state S_{t_l} at time t_l where $S_{t_l} \in L$. Clearly, since $Y_{t_l+1} = 0$ then the error at time $t_l + 1$ is $\Xi_{t_{l+1}} = v_{t_l}$. For any two elements Ξ_{t_l}, Ξ_{t_r} of the subsequence $\Xi^{(\nu)}$ where t_l,

$t_r \in L$, the following holds

$$\begin{aligned}\mathbb{P}\left(\Xi_{t_l}=1\left|\Xi_{t_r}\right.\right) &= \mathbb{P}\left(v_{t_l-1}=1\left|v_{t_r-1}\right.\right)\\ &= \sum_{i\in L}\mathbb{P}\left(v_{t_l-1}=1\left|S_{t_l-1}=s^{(i)},\,v_{t_r-1}\right.\right)\mathbb{P}\left(S_{t_l-1}=s^{(i)}\middle|v_{t_r-1}\right)\\ &= \sum_{i\in L}\mathbb{P}\left(v_{t_l-1}=1\left|S_{t_l-1}=s^{(i)}\right.\right)\mathbb{P}\left(S_{t_l-1}=s^{(i)}\middle|v_{t_r-1}\right) \qquad (8.54)\end{aligned}$$

where (8.54) follows since the type of transition v_{t_l-1} depends only on the state S_{t_l-1} because $k \ge k^*$.

In (8.54), the conditional probability $\mathbb{P}\left(S_{t_l-1}=s^{(i)}\left|v_{t_r-1}\right.\right)$ can in general be different from $\mathbb{P}\left(S_{t_l-1}=s^{(i)}\right)$ because the time t_r-1 may be in the set $\{t_l-k,\ldots,t_l-1\}$, that is, the bit v_{t_r-1} can be one of the bits of the state $S_{t_l-1}=(X_{t_l-k},\ldots,X_{t_l-1})$. Therefore knowledge of the value of v_{t_r-1} can limit the possible values of the state S_{t_l-1}. Thus one cannot remove the dependency of v_{t_l-1} on v_{t_r-1}, that is Ξ_{t_l} depends on Ξ_{t_r}. This conclusion also holds in case $k<k^*$. Hence, in general, we don't have $\mathbb{P}(\Xi_{t_l}=1|\Xi_{t_r})=\mathbb{P}(\Xi_{t_l}=1)$ and unlike the previous section where we split into two cases, here we just assume that $\Xi^{(\nu)}$ is a subsequence of dependent random variables.

We now continue the analysis and derive a bound on the probability of (8.53) that holds for any d. We assume that $X^{(m)}$ has been drawn and d is a *fixed* decision function which is the realization of the random variable $\hat{d}(X^{(m)})$ defined in (5.1). This means in particular that $\nu=\nu\left(S^{*(n)},d\right)$ is random only due to $S^{*(n)}$.

Consider any state-sequence $\theta \in \mathbb{S}_{k^*}^{r(k,k^*)}$ consisting of $r(k,k^*)$ states in $\mathbb{S}_{k^*}$ where r is defined in (6.4) and if $r(k,k^*)=1$ then θ is a single state of k^* bits. We extend the definition (3.2) and denote by $<\theta>_q$ the q least significant bits of the binary sequence that corresponds to the state sequence θ, and let $<\theta>_i^j$ denote the binary subsequence starting at the i^{th} bit and ending at the j^{th} bit from the right where the rightmost bit has index 0. For example, suppose $T'=0$, $k^*=2$, $k=5$ and for $n=4$ we have

$$\begin{aligned}x^{(4)} &= (x_{-4},x_{-3},x_{-2},x_{-1},x_0,x_1,x_2,x_3,x_4)\\ &= 111001010\end{aligned}$$

then at $t=2$ we have $\theta_t \in \mathbb{S}_2^5$ (since $k-k^*+2=5$) which equals

$$\begin{aligned}\theta_t &= (s^*_{t-4},s^*_{t-3},s^*_{t-2},s^*_{t-1},s^*_t)\\ &= (11,10,00,01,10)\\ &= (s^{*(3)},s^{*(2)},s^{*(0)},s^{*(1)},s^{*(2)})\end{aligned}$$

so $<\theta>_1^5=11001$ and $<\theta>_0=0$. Note that the binary sequence that corresponds to a state subsequence $\theta \in \mathbb{S}_{k^*}^{r(k,k^*)}$ has a length of $r(k,k^*)-1+k^*=k+1$ bits if $k^* \le k$ or k^* bits if $k^* \ge k+1$.

We denote by $\Theta_t^* = (S^*_{t-r(k,k^*)+1}, \ldots, S^*_{t-2}, S^*_{t-1}, S^*_t)$ a sequence of $r(k,k^*)$ random state variables where Θ_t^* takes values in $\mathbb{S}_{k^*}^{r(k,k^*)}$ according to the joint probability measure $\mathbb{P}$. Define $f_d : \mathbb{S}_{k^*}^{r(k,k^*)} \to \{0,1\}$ as follows,

$$f_d(\theta) = \begin{cases} <\theta>_0, & \text{if } \hat{p}(1| <\theta>_1^k) < \frac{1}{2} - \mathcal{T} \\ 0 & \text{otherwise} \end{cases}$$

where to avoid complicating the notation we just write $\mathcal{T}$, although it can in general depend on the value of θ (as is the case for $\hat{d}$ where the threshold (5.2) depends on the state i through α_i).

Note that if at time t the random state $\Theta_t^* = \theta$ where θ satisfies $\hat{p}\left(1| <\theta>_1^k\right) < \frac{1}{2} - \mathcal{T}$ then from (5.2) the decision function d decides $Y_t = 0$ in which case the value of the function f_d (which is the 'output' of the selection rule R_d) is the least significant bit of θ, that is, $f_d(\theta)$ equals the bit selected by the rule R_d. Otherwise, $\hat{p}(1| <\theta>_1^k) \geq \frac{1}{2} - \mathcal{T}$ and the decision function decides $Y_t = 1$ or REJECT, and R_d does not select it; this is represented by $f_d = 0$. Hence we represent the selection rule R_d by the function f_d.

For a sequence $s^{*(n)} = \{s_t^*\}_{t=1}^n$ where $s_t^* \in \mathbb{S}_{k^*}$, $1 \leq t \leq n$, define

$$F_d(s^{*(n)}) := \sum_{t=1}^n f_d(\theta_t)$$

where $\theta_t = (s^*_{t-r(k,k^*)+1}, \ldots, s^*_{t-1}, s^*_t)$. For convenience of notation, we sometimes write $F_d(x^{(n)})$ where $x^{(n)}$ corresponds to the state sequence $s^{*(n)}$.

Consider the random state sequence $S^{*(n)} = S_1^*, \ldots, S_n^*$ associated with $X^{(n)}$. We can express the average of the subsequence $\Xi^{(\nu)}$ as follows,

$$\frac{1}{\nu}\sum_{l=1}^{\nu} \Xi_l = \frac{1}{\nu} F_d(S^{*(n)})$$

which follows from

$$\begin{aligned} \frac{1}{\nu} F_d(S^{*(n)}) &= \frac{1}{\nu}\sum_{t=1}^n f_d(\Theta_t) && (8.55) \\ &= \frac{1}{\nu} \sum_{l:\hat{p}(1|<\Theta_{t_l}>_1^k)<\frac{1}{2}-\mathcal{T}} f_d(\Theta_{t_l}) && (8.56) \\ &= \frac{1}{\nu} \sum_{l:\hat{p}(1|<\Theta_{t_l}>_1^k)<\frac{1}{2}-\mathcal{T}} \Xi_l \\ &= \frac{1}{\nu}\sum_{l=1}^{\nu} \Xi_l && (8.57) \end{aligned}$$

where (8.57) holds since there are ν time instants t_l where the decision $Y_{t_l} = 0$ and hence $1 \leq l \leq \nu$. Define the random variable

$$\lambda := \lambda(\nu) = \frac{\nu}{n}$$

as the empirical frequency that a bit is selected. We henceforth denote

$$\begin{aligned}\hat{F}_d(S^{*(n)}) &:= \frac{1}{\nu}F_d\left(S^{*(n)}\right)\\ &= \frac{1}{\nu}\sum_{l=1}^{\nu}\Xi_l\\ &= \frac{1}{\lambda n}\sum_{l=1}^{\lambda n}\Xi_l, \qquad (8.58)\end{aligned}$$

where (8.58) follows from (8.57). We have

$$\begin{aligned}\mathbb{E}\left[F_d(S^{*(n)})\right] &= \mathbb{E}_\nu\left[\mathbb{E}\left[F_d(S^{*(n)})\middle|\nu\right]\right]\\ &= \mathbb{E}_\nu\left[\mathbb{E}\left[\sum_{l=1}^{\nu}X_{t_l}\middle|\nu\right]\right] \qquad (8.59)\\ &= \mathbb{E}_\nu\left[\mathbb{E}\left[\sum_{l:Y_{t_l}=0}X_{t_l}\middle|\nu\right]\right] \qquad (8.60)\\ &= \mathbb{E}_\nu[\nu\beta_d] \qquad (8.61)\\ &= \beta_d\mathbb{E}_\nu[\nu]\\ &= \beta_d n\mathbb{E}[\lambda], \qquad (8.62)\end{aligned}$$

where (8.59) follows because $\Xi^{(\nu)}$ is a subsequence of $X^{(n)}$ and (8.61) follows since for each l such that $Y_{t_l}=0$, $\mathbb{E}\left[X_{t_l}\right]=P\left(X_{t_l}=1\middle|Y_{t_l}=0\right)=\beta_d$ by (5.3). The expected value of λ depends on d therefore we denote by

$$\pi_d := \mathbb{E}[\lambda] \qquad (8.63)$$

and have

$$\mathbb{E}\left[F_d(S^{*(n)})\right] = \beta_d n\pi_d. \qquad (8.64)$$

For two state sequences $s^{*(n)}$ and $q^{*(n)}\in\mathbb{S}^n_{k^*}$ we use the Hamming metric $d_H:\mathbb{S}^n_{k^*}\times\mathbb{S}^n_{k^*}\to[0,\infty)$ defined by

$$d_H(s^{*(n)},q^{*(n)}) := \sum_{t=1}^{n}\mathbb{I}\left\{s^*_t\neq q^*_t\right\}.$$

Consider $s^{*(n)}=(s^*_1,\dots,s^*_n)$ and $q^{*(n)}=(q^*_1,\dots,q^*_n)\in\mathbb{S}^n_{k^*}$ such that

$$d_H(s^{*(n)},q^{*(n)}) \le \Delta$$

and define

$$\theta_t = (s^*_{t-r(k,k^*)+1},\dots,s^*_{t-1},s^*_t)$$

and

$$\psi_t = (q^*_{t-r(k,k^*)+1}, \ldots, q^*_{t-1}, q^*_t).$$

Define the following subsets of $\{1, \ldots, n\}$,

$$\begin{aligned}
I_1 &:= \left\{ l : \hat{p}(1| < \theta_{t_l} >_1^k) \geq \frac{1}{2} - \mathcal{T}, \hat{p}(1| < \psi_{t_l} >_1^k) \geq \frac{1}{2} - \mathcal{T} \right\} \\
I_2 &:= \left\{ l : \hat{p}(1| < \theta_{t_l} >_1^k) \geq \frac{1}{2} - \mathcal{T}, \hat{p}(1| < \psi_{t_l} >_1^k) < \frac{1}{2} - \mathcal{T} \right\} \\
I_3 &:= \left\{ l : \hat{p}(1| < \theta_{t_l} >_1^k) < \frac{1}{2} - \mathcal{T}, \hat{p}(1| < \psi_{t_l} >_1^k) \geq \frac{1}{2} - \mathcal{T} \right\} \\
I_4 &:= \left\{ l : \hat{p}(1| < \theta_{t_l} >_1^k) < \frac{1}{2} - \mathcal{T}, \hat{p}(1| < \psi_{t_l} >_1^k) < \frac{1}{2} - \mathcal{T} \right\}. \quad (8.65)
\end{aligned}$$

We have

$$\begin{aligned}
&\left| F_d(s^{*(n)}) - F_d(q^{*(n)}) \right| \leq \sum_{t=1}^n |f_d(\theta_t) - f_d(\psi_t)| \\
&= \sum_{l \in I_1} |f_d(\theta_{t_l}) - f_d(\psi_{t_l})| + \sum_{l \in I_2} |f_d(\theta_{t_l}) - f_d(\psi_t)| + \sum_{l \in I_3} |f_d(\theta_{t_l}) - f_d(\psi_{t_l})| \\
&\quad + \sum_{l \in I_4} |f_d(\theta_{t_l}) - f_d(\psi_{t_l})| \\
&= \sum_{l \in I_2} |0 - < \psi_{t_l} >_0| + \sum_{l \in I_3} |< \theta_{t_l} >_0 - 0| + \sum_{l \in I_4} |< \theta_{t_l} >_0 - < \psi_{t_l} >_0| \\
&\leq \sum_{l \in I_2} \mathbb{I}\{\theta_{t_l} \neq \psi_{t_l}\} + \sum_{l \in I_3} \mathbb{I}\{\theta_{t_l} \neq \psi_{t_l}\} + \sum_{l \in I_4} \mathbb{I}\{\theta_{t_l} \neq \psi_{t_l}\} \\
&\leq \sum_{t=1}^n \mathbb{I}\{\theta_t \neq \psi_t\} \\
&\leq r(k,k^*) \sum_{t=1}^n \mathbb{I}\{s^*_t \neq q^*_t\} \quad (8.66) \\
&= r(k,k^*) d_H(s^{*(n)}, q^{*(n)}) \\
&\leq r(k,k^*)\Delta \quad (8.67)
\end{aligned}$$

where (8.66) holds since for every t such that $s^*_t \neq q^*_t$ there are at most $r(k,k^*)$ time instants τ such that θ_τ contains state s^*_t, ψ_τ contains state q^*_t and $\theta_\tau \neq \psi_\tau$. For instance, let $T' = 0$, $n = 9$, $k = 3$, $k^* = 2$ and let the binary sequences that correspond to $s^{*(n)}$ and $q^{*(n)}$ be

$$\begin{aligned}
x^{(n)} &= (x_{-2}, x_{-1}, x_0, x_1, \ldots, x_9) = 011010011000 \\
\tilde{x}^{(n)} &= (\tilde{x}_{-2}, \tilde{x}_{-1}, \tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_9) = 011100001000
\end{aligned}$$

then $r(k, k^*) = 3$ and $\theta,\ \psi \in \mathbb{S}_2^3$. We have $\sum_{t=1}^{n} \mathbb{I}\{s_t^* \neq q_t^*\} = 5$. Looking at subsequences of length $k+1 = 4$ and comparing them across $x^{(n)}$ and $\tilde{x}^{(n)}$ over $1 \leq t \leq n$ we obtain $\sum_{t=1}^{n} \mathbb{I}\{\theta_t \neq \psi_t\} = 8$ and indeed $8 \leq r(3,2) \cdot 5 = 15$.

From (8.67) it follows that the function F_d is Lipschitz with constant $r(k, k^*)$.

Let the constant ω_0 be defined as

$$\omega_0 : \;=\; \frac{\ell}{n},$$

then from (8.53), we wish to bound the probability that there is some $\omega \in (\omega_0, 1]$ such that

$$\left|\hat{F}_d - \beta_d\right| \;>\; \epsilon \text{ and } \lambda \geq \omega$$

where from above $\hat{F}_d := \frac{1}{\nu} F_d$.

The probability of this event is bounded from above by the following probability,

$$\mathbb{P}\left(\exists \omega \in (\omega_0, 1] : \left|\hat{F}_d - \beta_d\right| > \epsilon,\ \lambda \geq \omega\right) \leq \mathbb{P}\left(\bigcup_{\omega_0 \leq \omega \leq 1} \left\{\left|\hat{F}_d - \beta_d\right| > \epsilon,\ \lambda \geq \omega\right\}\right). \tag{8.68}$$

From (8.32) it follows that,

$$\begin{aligned}
\left|\hat{F}_d - \beta_d\right| &= \left|\frac{1}{\lambda}\lambda \hat{F}_d - \frac{1}{\pi_d}\pi_d \beta_d\right| \\
&\leq \frac{1}{\lambda}\left|\lambda \hat{F}_d - \pi_d \beta_d\right| + \pi_d \beta_d \left|\frac{1}{\lambda} - \frac{1}{\pi_d}\right| \\
&= \frac{1}{\lambda}\left|\lambda \hat{F}_d - \pi_d \beta_d\right| + \frac{\beta_d}{\lambda}\left|\lambda - \pi_d\right|.
\end{aligned}$$

Hence the right side of (8.68) is bounded from above by

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\omega_0 \leq \omega \leq 1} \left\{\frac{1}{\lambda}\left|\lambda \hat{F}_d - \pi_d \beta_d\right| > \frac{\epsilon}{2},\ \lambda \geq \omega\right\}\right) + \mathbb{P}\left(\bigcup_{\omega_0 \leq \omega \leq 1} \left\{\frac{\beta_d}{\lambda}\left|\lambda - \pi_d\right| > \frac{\epsilon}{2},\ \lambda \geq \omega\right\}\right) \\
&\leq \mathbb{P}\left(\bigcup_{\omega_0 \leq \omega \leq 1} \left\{\left|\lambda \hat{F}_d - \pi_d \beta_d\right| > \frac{\lambda\epsilon}{2},\ \lambda \geq \omega\right\}\right) + \mathbb{P}\left(\bigcup_{\omega_0 \leq \omega \leq 1} \left\{\left|\lambda - \pi_d\right| > \frac{\lambda\epsilon}{2\beta_d},\ \lambda \geq \omega\right\}\right) \\
&\leq \mathbb{P}\left(\bigcup_{\omega_0 \leq \omega \leq 1} \left\{\left|\lambda \hat{F}_d - \pi_d \beta_d\right| > \frac{\lambda\epsilon}{2},\ \lambda \geq \omega\right\}\right) \qquad (8.69) \\
&\quad + \mathbb{P}\left(\bigcup_{\omega_0 \leq \omega \leq 1} \left\{\left|\lambda - \pi_d\right| > \frac{\lambda\epsilon}{2},\ \lambda \geq \omega\right\}\right) \qquad (8.70)
\end{aligned}$$

where the last inequality follows from the fact that $0 \leq \beta_d \leq 1$.

We now bound (8.69). Let us bound the probability of one term in the union of (8.69),

$$\begin{aligned}
\mathbb{P}\left(\left|\lambda \hat{F}_d - \pi_d\beta_d\right| > \frac{\lambda\epsilon}{2}, \lambda \geq \omega\right) &= \mathbb{P}\left(\left|\hat{F}_d - \frac{\pi_d\beta_d}{\lambda}\right| > \frac{\epsilon}{2}, \lambda \geq \omega\right) \\
&= \mathbb{P}\left(\left|\frac{F_d(S^{*(n)})}{\lambda n} - \frac{n\pi_d\beta_d}{\lambda n}\right| > \frac{\epsilon}{2}, \lambda \geq \omega\right) \\
&= \mathbb{P}\left(\left|F_d(S^{*(n)}) - n\pi_d\beta_d\right| > \frac{n\lambda\epsilon}{2}, \lambda \geq \omega\right) \\
&= \mathbb{P}\left(\left|F_d(S^{*(n)}) - \mathbb{E}\left[F_d(S^{*(n)})\right]\right| > \frac{n\lambda\epsilon}{2}, \lambda \geq \omega\right). \qquad (8.71)
\end{aligned}$$

Now, we can write (8.71) as follows,

$$\begin{aligned}
&\sum_{s^{*(n)}:\lambda(s^{*(n)})\geq\omega} \mathbb{P}\left(s^{*(n)}\right)\mathbb{I}\left\{\left|F_d(s^{*(n)}) - \mathbb{E}F_d(S^{*(n)})\right| > \frac{n\lambda(s^{*(n)})\epsilon}{2}\right\} \\
&\leq \sum_{s^{*(n)}:\lambda(s^{*(n)})\geq\omega} \mathbb{P}\left(s^{*(n)}\right)\mathbb{I}\left\{\left|F_d(s^{*(n)}) - \mathbb{E}F_d(S^{*(n)})\right| > \frac{n\omega\epsilon}{2}\right\} \\
&\leq \sum_{s^{*(n)}\in\mathbb{S}^n_{k^*}} \mathbb{P}\left(s^{*(n)}\right)\mathbb{I}\left\{\left|F_d(s^{*(n)}) - \mathbb{E}F_d(S^{*(n)})\right| > \frac{n\omega\epsilon}{2}\right\} \\
&= \mathbb{P}\left(\left|F_d(S^{*(n)}) - \mathbb{E}F_d(S^{*(n)})\right| > \frac{n\omega\epsilon}{2}\right). \qquad (8.72)
\end{aligned}$$

Since the Markov chain $S^{*(n)}$ has all the properties of the chain $S^{*(m)}$ in the previous section, we use Lemma A.1. We apply (A.98) to the function F_d/r which is Lipschitz with constant 1 and define $\kappa(\omega,\eta)$ as

$$\kappa(\omega,\eta) := \frac{2r\rho}{\omega}\sqrt{\frac{2}{n}\ln\left(\frac{2}{\eta}\right)} \qquad (8.73)$$

then we obtain

$$\mathbb{P}\left(\left|\frac{F_d(S^{*(n)})}{r} - \mathbb{E}\left(\frac{F_d(S^{*(n)})}{r}\right)\right| > \frac{n\omega\kappa(\omega,\eta)}{2r}\right) \leq \eta$$

from which it follows that

$$\mathbb{P}\left(\left|F_d(S^{*(n)}) - \mathbb{E}F_d(S^{*(n)})\right| > \frac{n\omega\kappa(\omega,\eta)}{2}\right) \leq \eta. \qquad (8.74)$$

Let

$$J(\omega^{(1)},\omega^{(2)},\eta) := \left\{x^{(n)} : \left|\hat{F}_d(x^{(n)}) - \frac{\pi_d\beta_d}{\lambda(x^{(n)})}\right| > \frac{\kappa(\omega^{(1)},\eta)}{2}, \lambda(x^{(n)}) > \omega^{(2)}\right\}$$

then

$$\begin{aligned}
\mathbb{P}\left(J(\omega,\omega,\eta)\right) &= \mathbb{P}\left(\left|\hat{F}_d(S^{*(n)}) - \frac{\pi_d\beta_d}{\lambda(X^{(n)})}\right| > \frac{\kappa(\omega,\eta)}{2},\, \lambda(X^{(n)}) > \omega\right) \\
&= \mathbb{P}\left(\left|F_d(S^{*(n)}) - n\pi_d\beta_d\right| > \frac{n\lambda(X^{(n)})\kappa(\omega,\eta)}{2},\, \lambda(X^{(n)}) > \omega\right) \\
&\leq \mathbb{P}\left(\left|F_d(S^{*(n)}) - \mathbb{E}F_d(S^{*(n)})\right| > \frac{n\omega\kappa\left(\omega,\eta\right)}{2}\right) \\
&\leq \eta. \qquad (8.75)
\end{aligned}$$

As above, for $\omega^{(1)} \leq \omega \leq \omega^{(2)}$ we have $J(\omega^{(1)},\omega^{(2)},\eta) \subseteq J(\omega,\omega,\eta)$ and for $\eta_a \leq \eta_b$, $J(\omega^{(1)},\omega^{(1)},\eta_a) \subseteq J(\omega^{(1)},\omega^{(1)},\eta_b)$. Define the set $\Delta_j \subset [0,1]$ as follows,

$$\Delta_j = \left[\omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{j+1}, \omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{j}\right]. \qquad (8.76)$$

We have $\bigcup_{j=0}^{\infty} \Delta_j = [\omega_0, 1]$. Now, substitute for

$$\epsilon \quad := \quad \kappa\left(\frac{\omega}{2}, \frac{(\omega-\omega_0)\eta}{2}\right) \qquad (8.77)$$

then the probability in (8.69) is bounded from above as follows,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\omega_0\leq\omega\leq 1}\left\{\left|\hat{F}_d - \frac{\pi_d\beta_d}{\lambda}\right| > \frac{\epsilon}{2},\, \lambda \geq \omega\right\}\right) = \mathbb{P}\left(\bigcup_{\omega_0\leq\omega\leq 1} J\left(\frac{\omega}{2}, \omega, \frac{(\omega-\omega_0)\eta}{2}\right)\right) \\
&= \mathbb{P}\left(\bigcup_{j=0}^{\infty}\bigcup_{\omega\in\Delta_j} J\left(\frac{\omega}{2}, \omega, \frac{(\omega-\omega_0)\eta}{2}\right)\right) \\
&\leq \sum_{j=0}^{\infty}\mathbb{P}\left(\bigcup_{\omega\in\Delta_j} J\left(\frac{\omega}{2}, \omega, \frac{(\omega-\omega_0)\eta}{2}\right)\right) \\
&\leq \sum_{j=0}^{\infty}\mathbb{P}\left(J\left(\omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{j+1}, \omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{j+1}, \frac{\eta}{2}(1-\omega_0)\left(\frac{1}{2}\right)^{j}\right)\right)
\end{aligned} \qquad (8.78)$$

where the last inequality follows from the fact that for all $\omega \in \Delta_j$ we have $\omega/2 \leq \omega_0 + (1-\omega_0)(\frac{1}{2})^{j+1} \leq \omega$ and $\omega \leq \omega_0 + (1-\omega_0)(\frac{1}{2})^j$. From (8.75) it follows that (8.78) is bounded from above by

$$\frac{\eta}{2}(1-\omega_0)\sum_{j=0}^{\infty}\left(\frac{1}{2}\right)^j \leq (1-\omega_0)\eta$$

and thus (8.69) is bounded by $(1-\omega_0)\eta$.

We now proceed to bound (8.70).

Define $g_d : \mathbb{S}_{k^*}^{r(k,k^*)} \to \{0,1\}$ as follows: for any state-subsequence $\theta \in \mathbb{S}_{k^*}^{r(k,k^*)}$ let

$$g_d(\theta) = \begin{cases} 1, & \text{if } \hat{p}(1| < \theta >_1^k) < \frac{1}{2} - \mathcal{T} \\ 0, & \text{otherwise.} \end{cases}$$

For a sequence $s^{*(n)} = \{s_t^*\}_{t=1}^n$, $s_t^* \in \mathbb{S}_{k^*}$ for $1 \le t \le n$, define

$$G_d(s^{*(n)}) := \sum_{t=1}^n g_d(\theta_t)$$

where $\theta_t = (s^*_{t-r(k,k^*)+1}, \dots, s^*_{t-1}, s^*_t)$.

Note that G_d/n is the average number of times that the sequence $s^{*(n)}$ visits a state at which the decision is 0. That is, we have

$$\frac{G_d(S^{*(n)})}{n} = \frac{\nu}{n} = \lambda$$

and so from (8.63) we have

$$\mathbb{E}\left[G_d\right] = n\mathbb{E}\left[\lambda\right] = n\pi_d.$$

Using the definition of the regions (8.65) one can show using the argument of (8.67) that G_d is Lipschitz with constant $r(k,k^*)$. Consider the probability of one term in the union in (8.70). We have,

$$\begin{aligned} \mathbb{P}\left(|\lambda - \pi_d| > \frac{\lambda\epsilon}{2}, \lambda > \omega\right) &= \mathbb{P}\left(\left|\frac{G_d(S^{*(n)})}{n} - \frac{\mathbb{E}G_d}{n}\right| > \frac{\lambda\epsilon}{2}, \lambda > \omega\right) \\ &= \mathbb{P}\left(\left|G_d(S^{*(n)}) - \mathbb{E}G_d\right| > \frac{n\lambda\epsilon}{2}, \lambda > \omega\right). \quad (8.79) \end{aligned}$$

Now, we can write (8.79) as follows,

$$\begin{aligned} &\sum_{s^{*(n)}:\lambda(s^{*(n)})>\omega} \mathbb{P}\left(s^{*(n)}\right) \mathbb{I}\left\{\left|G_d(s^{*(n)}) - \mathbb{E}G_d(S^{*(n)})\right| > \frac{n\lambda(s^{*(n)})\epsilon}{2}\right\} \\ &\le \sum_{s^{*(n)}:\lambda(s^{*(n)})>\omega} \mathbb{P}\left(s^{*(n)}\right) \mathbb{I}\left\{\left|G_d(s^{*(n)}) - \mathbb{E}G_d(S^{*(n)})\right| > \frac{n\omega\epsilon}{2}\right\} \\ &\le \sum_{s^{*(n)}\in\mathbb{S}_{k^*}^n} \mathbb{P}\left(s^{*(n)}\right) \mathbb{I}\left\{\left|G_d(s^{*(n)}) - \mathbb{E}G_d(S^{*(n)})\right| > \frac{n\omega\epsilon}{2}\right\} \\ &= \mathbb{P}\left(\left|G_d(S^{*(n)}) - \mathbb{E}G_d(S^{*(n)})\right| > \frac{n\omega\epsilon}{2}\right). \quad (8.80) \end{aligned}$$

We use Lemma A.1 since as is shown in the Appendix, the Markov chain $S^{*(n)}$ satisfies the necessary conditions. With the choice of (8.73) for κ, applying (A.98) to G_d/r and $S^{*(n)}$ yields the following bound (as in (8.74))

$$\mathbb{P}\left(\left|G_d(S^{*(n)}) - \mathbb{E}G_d(S^{*(n)})\right| > \frac{n\omega\kappa(\omega,\eta)}{2}\right) \le \eta. \quad (8.81)$$

Let

$$J(\omega^{(1)},\omega^{(2)},\eta) \quad := \quad \left\{ x^{(n)} : \left|\lambda(x^{(n)}) - \pi_d\right| > \frac{\lambda\left(x^{(n)}\right)}{2}\kappa(\omega^{(1)},\eta),\ \lambda(x^{(n)}) > \omega^{(2)} \right\}$$

then

$$\begin{aligned} \mathbb{P}\left(J(\omega,\omega,\eta)\right) &= \mathbb{P}\left(\left|\lambda(X^{(n)}) - \pi_d\right| > \frac{\lambda(X^{(n)})\kappa(\omega,\eta)}{2},\ \lambda(X^{(n)}) > \omega\right) \\ &\leq \mathbb{P}\left(\left|G_d(S^{*(n)}) - \mathbb{E}G_d(S^{*(n)})\right| > \frac{n\omega\kappa\left(\omega,\eta\right)}{2}\right) \\ &\leq \eta. \end{aligned} \tag{8.82}$$

Substituting for ϵ the value in (8.77) and using the sets of (8.76), with (8.78) and (8.82), implies that (8.70) is bounded from above by $(1-\omega_0)\eta$.

Hence for a fixed decision function d, with the choice of (8.77) we conclude from (8.68)–(8.82) that

$$\begin{aligned} \mathbb{P}\left(E_{d,\epsilon}^{(\ell)}\right) &\leq \mathbb{P}\left(\exists\omega \in [\omega_0, 1] : \left|\frac{1}{\lambda n}\sum_{t=1}^{\lambda n} \Xi_l - \beta_d\right| > \kappa\left(\frac{\omega}{2}, \frac{(\omega-\omega_0)\eta}{2}\right),\ \lambda \geq \omega\right) \\ &\leq 2\left(1-\omega_0\right)\eta. \end{aligned} \tag{8.83}$$

8.4. Combining

We now combine the results of sections 8.2 and 8.3 to obtain an upper bound on the probability that for the random decision function $\hat{d}$ the event $E_{\hat{d},\epsilon}^{(\ell)}$ occurs. We have

$$\begin{aligned} \mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}\right) &= \mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}, A_{d^*}\right) + \mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}, \overline{A}_{d^*}\right) \\ &= \mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}\middle| A_{d^*}\right)\mathbb{P}\left(A_{d^*}\right) + \mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}\middle|\overline{A}_{d^*}\right)\mathbb{P}\left(\overline{A}_{d^*}\right). \end{aligned} \tag{8.84}$$

From (8.8), we have $\overline{A}_{d^*} := \left\{x^{(m)} : \forall i,\ \hat{d}(i) = d^*(i) \text{ or } \hat{d}(i) = REJECT\right\}$. Hence given that $\overline{A}_{d^*}$ holds then $\hat{d} \asymp d^*$ and hence $\mathbb{P}\left(E_{\hat{d},\epsilon}^{(\ell)}\middle|\overline{A}_{d^*}\right) = \mathbb{P}\left(\bigcup_{d:d\asymp d^*} E_{d,\epsilon}^{(\ell)}\right)$ where the index d runs over all decision functions in Υ^{2^k} such that $d \asymp d^*$. Thus (8.84) is bounded from above by

$$\mathbb{P}\left(A_{d^*}\right) + \mathbb{P}\left(\bigcup_{d:d\asymp d^*} E_{d,\epsilon}^{(\ell)}\right). \tag{8.85}$$

For arbitrary $0 < \delta \leq 1$, recalling the choice of (8.10) then the first term of (8.85) is bounded from above by $\delta/2$ provided that the system's model threshold

$\mathcal{T}$ for state $s^{(i)}$ equals $\mathcal{T}^{(a)}\left(\frac{\gamma_i}{3}, \frac{\delta\gamma_i}{2^{k+2}}\right)$ or $\mathcal{T}^{(b)}\left(\frac{\gamma_i}{2}, \frac{\delta\gamma_i}{2^{k+3}}\right)$ subject to whether case (a) or (b) holds, respectively.

In the statement of the theorem, k^* is assumed to be unknown so the system cannot know which case holds. The threshold for case (a) is bounded from above by that of case (b), and the latter is bounded from above by $\mathcal{T}(\gamma_i, \delta)$ (which is defined in (7.1)). Letting (7.1) be the system's decision threshold $\mathcal{T}$ for state $s^{(i)}$ means that the first term of (8.85) is bounded from above by $\delta/2$ in either case (a) or (b).

The derivation leading to the bound of (8.83) holds for any d, in particular for d that satisfies $d \asymp d^*$. Let us choose for η the following value

$$\eta := \frac{\delta}{2^{2^k+2}(1-\omega_0)}$$

then, with the choice of (8.77) for ϵ we bound the second term of (8.85) using (8.83) as follows,

$$\begin{aligned}\mathbb{P}\left(\bigcup_{d:d\asymp d^*} E_{d,\epsilon}^{(\ell)}\right) &\leq \sum_{d:d\asymp d^*} \mathbb{P}\left(E_{d,\epsilon}^{(\ell)}\right) \\ &\leq 2^{2^k+1}(1-\omega_0)\eta \qquad (8.86) \\ &= \frac{\delta}{2}\end{aligned}$$

where (8.86) follows from the fact that for any fixed d^* there are $\sum_{j=0}^{2^k}\binom{2^k}{j} = 2^{2^k}$ possible decision functions d that satisfy $d \asymp d^*$ since for every $0 \leq i \leq 2^k - 1$ there are two choices: REJECT or not REJECT (in which case $d(i) = d^*(i)$). Thus (8.86) and (8.11) imply that (8.85) is bounded from above by δ.

The proof of Corollary 7.1 follows directly from the above, with the exception of section 8.4 where instead of bounding $\mathcal{T}^{(a)}$ and $\mathcal{T}^{(b)}$ we use them directly. □

References

[1] A. E. Asarin. On some properties of finite objects random in an algorithmic sense. *Soviet Mathematics Doklady*, 36(1):109–112, 1988.

[2] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[3] E. Behrends. *Introduction to Markov Chains: With Special Emphasis on Rapid Mixing*. Advanced Lectures in Mathematics. Springer Fachmedien Wiesbaden, 2000.

[4] L. Bienvenu. Kolmogorov-Loveland stochasticity and Kolmogorov complexity. In *Proc. of 24th Annual Symposium on Theoretical Aspects of Computer Science (STACS 2007)*, volume LNCS 4393, pages 260–271, 2007.

[5] P. C. Bressloff. *Stochastic processes in cell biology*. Springer, 2014.

[6] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.
[7] C. K. Chow. An optimum character recognition system using decision functions. *Electronic Computers, IRE Transactions on*, EC-6(4):247–254, 1957.
[8] G. Falin. A heterogeneous blocking system in a random environment. *Journal of Applied Probability*, 33(1):211–216, 1996.
[9] A. N. Kolmogorov. On tables of random numbers. *Theoretical Computer Science*, 207(2):387–395, 1998.
[10] A. Kontorovich and R. Weiss. Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. *Journal of Applied Probability*, 51(4):1100–1113, 2014.
[11] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, Oct 1998.
[12] C. D. Meyer, editor. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.
[13] K. Najim and A. S. Poznyak. *Learning Automata: Theory and Applications*. Pergamon Press, Inc., Elmsford, NY, USA, 1994.
[14] P. Nicolas. *Understanding Markov Chains, Examples and Applications*. Springer, 2013.
[15] K. Petakos and T. Tsapelas. Reliability analysis for systems in a random environment. *Journal of Applied Probability*, 34(4):1021–1031, 1997.
[16] J. Ratsaby. An algorithmic complexity interpretation of Lin's third law of information theory. *Entropy*, 10(1):6–14, 2008.
[17] J. Ratsaby. An empirical study of the complexity and randomness of prediction error sequences. *Communications in Nonlinear Science and Numerical Simulation*, 16:2832–2844, 2011.
[18] J. Ratsaby. On the sysratio and its critical point. *Mathematical and Computer Modelling*, 53:939–944, 2011.
[19] B. S. White. The effects of a rapidly-fluctuating random environment on systems of interacting species. *SIAM Journal on Applied Mathematics*, 32(3):666–693, 1977.
[20] J. J. Wijker. *Random Vibrations in Spacecraft Structures Design, Theory and Applications*. Springer, 2009.
[21] M. A. Wortman and G. A. Klutke. On maintained systems operating in a random environment. *Journal of Applied Probability*, 31(2):589–594, 1994.

Acknowledgements

The author thanks the referees and Aryeh Kontorovich for helpful comments.

Appendix

The total variation distance between two discrete probability distributions p and q on a domain $\mathcal{X}$ is defined as (see for instance, Definition 13.2, [3])

$$\|p-q\|_{TV} := \sup_{A\subset\mathcal{X}} |p(A)-q(A)|$$

and is related to the l_1-distance as follows (Lemma 13.3, [3]),

$$\|p-q\|_{TV} = \frac{1}{2}\|p-q\|_1 . \tag{A.87}$$

A normalized Hamming metric for sequences $s^{*(n)} = \{s_t^*\}_{t=1}^n \in \mathbb{S}_{k^*}^n$ is defined as follows: for any $s^{*(n)}, q^{*(n)} \in \mathbb{S}_{k^*}^n$,

$$d_H\left(s^{*(n)}, q^{*(n)}\right) := \sum_{t=1}^{n} \mathbb{I}\left\{s_t^* \neq q_t^*\right\}. \tag{A.88}$$

For a discrete time Markov chain with transition matrix Q defined in (2.3), denote its i^{th} row by $q(\cdot|i)$, that is, the conditional probability distribution given that the current state is $s^{*(i)}$. For discrete time l, we denote the entries of the matrix Q^l by $q^{(l)}(j|i)$, and $q^{(l)}(\cdot|i)$ denotes the i^{th} row of Q^l. Define

$$\tau_l := \max_{0\le i\le 2^{k^*}-1}\left\|q^{(l)}(\cdot|i)-\pi^*\right\|_{TV}$$

where π^* is defined in (2.7). The next lemma is Theorem 1.1 of [10] applied to our finite Markov chain. It establishes a concentration bound for $S^{*(n)}$ and functions that are Lipschitz with respect to the Hamming norm. (That we can apply this theorem follows from the fact that a Markov chain can be regarded as a hidden Markov chain by letting the emission alphabet be identical to the state space and emission probabilities to be delta-functions.)

Lemma A.1. *([10] Theorem 1.1) For $1 \le \alpha < \infty$ and $0 \le \beta < 1$, if $\tau_l \le \alpha\beta^{l-1}$ for $l = 1, 2, \ldots$, then for any $\varphi : \mathbb{S}_{k^*}^n \to \mathbb{R}$ with Lipschitz constant 1 with respect to the Hamming metric, the following holds:*

$$\mathbb{P}\left(\varphi\left(S^{*(n)}\right) - \mathbb{E}\varphi\left(S^{*(n)}\right) > n\kappa\right) \le \exp\left(-\frac{n(1-\beta)^2\kappa^2}{2\alpha^2}\right) \tag{A.89}$$

and

$$\mathbb{P}\left(\mathbb{E}\varphi\left(S^{*(n)}\right) > \varphi\left(S^{*(n)}\right) + n\kappa\right) \le \exp\left(-\frac{n(1-\beta)^2\kappa^2}{2\alpha^2}\right).$$

For this to be useful we need to ensure that we can apply this lemma to the samples $X^{(m)}$ and $X^{(n)}$ of the environment's Markov chain. Let us investigate this for $X^{(n)}$ (the case for $X^{(m)}$ would then follow directly).

The sequence $X^{(n)}$, and hence $S^{*(n)}$, is a sample of a homogeneous Markov chain with two types of transitions, a type-1 and type-0 transition (see section

2) which occur with a probability $q(1|s^{*(i)})$ and $q(0|s^{*(i)})$, respectively, $0 \le i \le 2^{k^*} - 1$. The transition matrix Q of (2.3) has the following form (recall the notation of (2.5)),

$$Q := \left(\begin{array}{cccccccccc}
q(0|0) & q(1|0) & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & q(0|1) & q(1|1) & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 0 & 0 & q(0|2) & q(1|2) & 0 & \cdots & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & q(0|2^{k^*-1}-1) & q(1|2^{k^*-1}-1) \\
q(0|2^{k^*-1}) & q(1|2^{k^*-1}) & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & q(0|2^{k^*-1}+1) & q(1|2^{k^*-1}+1) & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 0 & 0 & q(0|2^{k^*-1}+2) & q(1|2^{k^*-1}+2) & 0 & \cdots & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & q(0|2^{k^*}-1) & q(1|2^{k^*}-1)
\end{array} \right).$$

It is clear that for all $k^* \ge 2$, Q is not a (strictly) positive matrix. Thus we need some additional work to show that the condition of Lemma A.1 holds for the environment's Markov chain under all the cases of Assumption 1 with ρ as defined in (6.3).

A nonnegative irreducible matrix is regular (or primitive) if it has a single eigenvalue on the unit circle [12]. As mentioned in section 2, the Markov chain is irreducible with a nonnegative transition matrix. By assumption (2.6), Q has at least one positive entry on the diagonal, hence it follows that Q is primitive ([12] Example 8.3.3). From [12] p.693, $\lim_{l\to\infty} Q^l$ exists and is a matrix all of whose rows are identical to the stationary probability distribution of the Markov chain. More relevant for us, by the Frobenius's test for primitivity ([12], (8.3.16)), it follows that because Q is primitive then there exists a minimum integer $\mathfrak{l}_0$, such that for $l \ge \mathfrak{l}_0$, $Q^l > 0$, that is, all the elements of Q^l are positive ([12], Example 8.3.4 shows that $\mathfrak{l}_0 \le 2^{2k^*} - 2^{k^*+1} + 2$). We choose $\mathfrak{l}_0$ as in (6.1). That $Q^{\mathfrak{l}_0} > 0$ implies that the minimum value μ_0 of $Q^{\mathfrak{l}_0}$ satisfies

$$\mu_0 > 0. \tag{A.90}$$

Recall from (2.7) that $\pi^* := \left[\pi_0^*, \ldots, \pi_{2^{k^*}-1}^*\right]$ is the stationary probability distribution of the chain. Note that every row of $Q^{\mathfrak{l}_0}$ (which is a $2^{k^*} \times 2^{k^*}$ matrix) has an entry whose value is no larger than $1/2^{k^*}$ because $Q^{\mathfrak{l}_0}$ is row-stochastic and hence the sum of the entries in every row is 1. Hence

$$\mu_0 \le \frac{1}{2^{k^*}}. \tag{A.91}$$

Define the constant

$$c_0 = c_0(k^*, \mathfrak{l}_0) := 1 - 2^{k^*} \mu_0 \tag{A.92}$$

then, by (A.90) and (A.91), it follows that

$$0 \le c_0 < 1.$$

We need to consider the cases of Assumption 1. We start with case (i) where the environment has a Q such that $\mu_0 \ne 2^{-k^*}$ so that $c_0 > 0$. In this case, by

Proposition 10.5(ii) [3], if $\mathfrak{l}_0 \geq 2$, we have for every $0 \leq i,\, j \leq 2^{k^*}-1$, and every $l \geq \mathfrak{l}_0$,

$$\left|q^{(l)}(j|i) - \pi_j^*\right| \leq \left(\frac{1}{c_0}\right) c_0^{l/\mathfrak{l}_0}. \tag{A.93}$$

This means that the distance between the i^{th} row of Q^l and the stationary distribution is

$$\left\|q^{(l)}\left(\cdot|i\right) - \pi^*\right\|_1 = \sum_{j=0}^{2^{k^*}-1} \left|q^{(l)}(j|i) - \pi_j^*\right| \leq \left(\frac{2^{k^*}}{c_0}\right) c_0^{l/\mathfrak{l}_0} \tag{A.94}$$

where $\|\cdot\|_1$ denotes the l_1-norm. Therefore, from (A.87) and (A.94),

$$\begin{aligned}\left\|q^{(l)}\left(\cdot|i\right) - \pi^*\right\|_{TV} &\leq \left(\frac{2^{k^*}}{2c_0}\right) c_0^{l/\mathfrak{l}_0} \\ &= \left(\frac{2^{k^*} c_0^{1/\mathfrak{l}_0}}{2c_0}\right) \left(c_0^{1/\mathfrak{l}_0}\right)^{l-1}.\end{aligned} \tag{A.95}$$

Letting

$$\alpha = 2^{k^*-1} c_0^{-(\mathfrak{l}_0-1)/\mathfrak{l}_0} \tag{A.96}$$

and

$$\beta = c_0^{1/\mathfrak{l}_0} \tag{A.97}$$

means that we may use Lemma A.1 for $S^{*(n)}$ (whose transition matrix is (2.3)) together with any function φ that is Lipschitz with constant 1. Substituting (A.92) for c_0 and plugging (A.96) and (A.97) for α and β in the bound of Lemma A.1, then the concentration bound (A.89) becomes

$$\mathbb{P}\left(\varphi\left(S^{*(n)}\right) - \mathbb{E}\varphi\left(S^{*(n)}\right) > n\kappa\right) \leq \exp\left\{-\frac{n}{2}\left(\frac{\kappa}{\rho}\right)^2\right\}, \tag{A.98}$$

with

$$\rho := \frac{\alpha}{1-\beta} = \frac{2^{k^*-1}}{(1-2^{k^*}\mu_0)^{(\mathfrak{l}_0-1)/\mathfrak{l}_0}\left(1-(1-2^{k^*}\mu_0)^{1/\mathfrak{l}_0}\right)}. \tag{A.99}$$

If $\mathfrak{l}_0 = 1$ (and still under case (i) of Assumption 1) then $Q > 0$ and by Proposition 10.5(i) [3], $\left|q^{(l)}(j|i) - \pi_j^*\right| \leq c_0^l$. Following the above steps it suffices to choose $\alpha = 2^{k^*-1}c_0$ and $\beta = c_0$ to obtain

$$\rho := \frac{\alpha}{1-\beta} = \frac{1-2^{k^*}\mu_0}{2\mu_0}.$$

We now consider case (ii) of Assumption 1 where the environment's Q has $\mu_0 = 2^{-k^*}$ and therefore (A.99) cannot be used. The stationary distribution in this case is uniform so Q is doubly stochastic and $\lim_{l\to\infty} Q^l = U$ is reached exactly at time $\mathfrak{l}_0$, that is, $Q^{\mathfrak{l}_0} = U$. The matrix Q is as above with $q(1|i) = q(0|i) = \frac{1}{2}$, for all $0 \leq i \leq 2^{k^*} - 1$. We have $\mathfrak{l}_0 = k^*$ and the limit matrix U has all entries equal to 2^{-k^*} (if $k^* = 1$ then $Q = U$). For $l \geq \mathfrak{l}_0$ the left side of (A.93) equals zero because the limit is reached at $\mathfrak{l}_0$. But for $1 \leq l < \mathfrak{l}_0$ the left side of (A.93) is bounded from above by 2^{-l}. Thus for all $l \geq 1$, the left side of (A.95) is bounded from above by $2^{k^*-1}(1/2)^l$. We let $\alpha = 2^{k^*-2}$ and $\beta = 1/2$ to yield $\rho = 2^{k^*-1}$.

All the above holds also for the sample $S^{*(m)}$ of the Markov environment, with m replacing n.

In summary, we showed that for every case of Assumption 1, the necessary condition of Lemma A.1 that $\tau_l \leq \alpha\beta^{l-1}$ holds and therefore the lemma can be used as a concentration inequality for both samples $X^{(m)}$ and $X^{(n)}$ in the proof of Theorem 7.1.