Finite Sample Analysis for Structured Discrete System Identification

Xiaotian Xie, Dimitrios Katselis, Carolyn L. Beck, *Fellow, IEEE*, and R. Srikant, *Fellow, IEEE*

***Abstract*— We consider a discrete-time dynamical system over a discrete state-space, which evolves according to a structured Markov model called Bernoulli Autoregressive (BAR) model. Our goal is to obtain sample complexity bounds for the problem of estimating the parameters of this model using an indirect Maximum Likelihood Estimator. Our sample complexity bounds exploit the structure of the BAR model and are established using concentration inequalities for random matrices and Lipschitz functions.**

***Index Terms*— Discrete state-space dynamical systems, Identification, Markov chains, Sample complexity**

I. Introduction

System identification aims to build mathematical models of dynamical systems from measurements. Extensive prior work has focused on the asymptotic properties of different identification methods [1]–[4]. The necessity of understanding how many observations in a single trajectory are sufficient to estimate model parameters within a prescribed level of confidence motivates more recent trends in non-asymptotic analysis [5]–[10], most of which focus on Linear Time Invariant (LTI) systems. To properly deal with the dependencies in observed data, various techniques have been considered in existing works, e.g., mixing-time arguments [8], [11], Mendelson's small ball method [6], [7], [10], [12], [13], and concentration inequalities for random matrices [14]–[16]. Relying on mixing-time arguments, one can treat dependent data as almost independent. Naturally, the resulting bounds degrade if the considered process mixes slowly, e.g., in linear dynamical systems with state matrix spectral radius close to one. In [6], the authors avoid mixing-time arguments by introducing the Block Martingale Small Ball (BMSB) condition, which corresponds to an adaptation of Mendelson's small ball method. Their analysis shows that in the marginal stability regime the statistical performance of the ordinary least-squares estimator depends on the minimum eigenvalue of the finite-time controllability Gramian.

Moreover, network (or graph) modeling has recently received increasing attention in the machine learning literature with various applications, e.g., in system biology [17], economics [18], epidemiology [19] and social sciences [20]. System identification can be used as an approach to perform topology inference and edge weight estimation.

In this note, we investigate the problem of identifying a discrete-time, discrete-state dynamical system, whose state evolves based on a particular Markov chain model, called Bernoulli Autoregressive (BAR) model [4], [8]. For a directed graph with p nodes and $\forall k \geq 0$, the BAR model describes the dynamics of each node $i \in [p]$ via

$$X_i(k+1) \sim \mathrm{Ber}\left(\mathbf{a}_i^\top \mathbf{X}(k) + b_i W_i(k+1)\right).$$

Here, $\mathbf{X}(k) \in \{0,1\}^p$ denotes the state vector at time k, $\{W_i(k+1) \sim \mathrm{Ber}(\rho_{w_i}), i \in [p]\}$ correspond to independent Bernoulli random variables and $\mathbf{X}(0) \sim \mu$, where μ is a probability measure on $\{0,1\}^p$. The tuples $(\mathbf{a}_i, b_i, \rho_{w_i}) \in \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}, i \in [p]$, to be explicitly defined in Section II, are the BAR model parameters. The statement of the BAR model identification problem relies on these parameter tuples.

Binary-state models of similar nature to our BAR model include the voter model and related variants [21]–[24], and the ALARM model [25]. Studies of such models on networks originate from [21]. The authors in [21] propose a linear voter model to describe interacting systems, where individuals holding one of two opinions update their stances under the influence of their 'friends'. It is worth noting that such systems will reach consensus in finite time, which corresponds to an absorbing state in the context of Markov chains. The voter model is also applicable in situations such as competition for territory between two distinct populations [26], spread of diseases (or viruses) in a population [19], [27], and particle interactions in statistical mechanics [28]. In the simplest case, the linear voter model is of the same form as the BAR model without the noise term. However, in many real-life cases, such as in social networks, it is uncommon that people can eventually reach (and remain in) a consensus. Also, people's opinions can be influenced by unexpected factors other than their friends. Compared to the voter model, the presence of Bernoulli noise in the BAR model eliminates the absorbing states. Moreover, the parameterized Bernoulli noise allows for capturing diffusion of opinions [23], herding behaviors in financial markets [29], and regulatory interactions in cellular systems [30], [31]. The BAR model can be reduced to the linear voter model by simply setting $b_i = 0, i \in [p]$. In addition, the original voter model in [21] only considers positive influence relationships, while negative influence relationships, commonly existing in real-life networks, are not captured. For example, in gene regulatory networks [17], [30], [31], products of a fraction of genes may have inductive or prohibitory influence on the expression of other gene fractions. The ALARM model [25], some variants of the voter model [32], and the generic BAR model [4], [8] take both relationships into account and therefore, they have broader range of applications.

Our BAR model is also intimately related to a generalized autoregressive linear model framework [33], [34], which covers a class of discrete-valued processes including Poisson and Bernoulli autoregressive processes. The definitions of the Bernoulli autoregressive model in this line of literature are significantly different from ours. In [33], the model is structurally similar to the aforementioned ALARM model, where the Bernoulli parameter relies on the logistic function. Although the model in [34] can stochastically capture ours and those with higher-order time lags by appropriate choices of the link functions f_i, the BAR model in our setting has a more natural form of a dynamical system driven by noise. Moreover, the proposed estimation approaches in [25], [33], [34] are primarily focused on inferring the unknown parameters with sparsity constraints; the techniques developed therein for the finite-time analysis follow different well-established ideas such as Restricted Strong Convexity.

Let $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]^\top$ be the BAR system matrix with entries reflecting the underlying connectivity of the network. Define $\mathbf{\Theta} = [\mathbf{A}, \mathbf{c}] \in \mathbb{R}^{p \times (p+1)}$, where $\mathbf{c} = [c_1, c_2, \cdots, c_p]^\top$ and $c_i = b_i \rho_{w_i}$. $\mathbf{\Theta}$ corresponds to a reparameterization of the BAR model, which is

This work was supported in part by NSF Grant CCF 17-04970 and ONR Grant N00014-19-1-2566.

Xiaotian Xie is with the School of Automation, Central South University, Changsha, 410083, China. Email: xxt19911031@gmail.com.

Carolyn L. Beck and R. Srikant are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. Emails: {beck3|rsrikant}@illinois.edu.

Dimitrios Katselis is with the ECE Department, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. Email: katselis@illinois.edu.

aligned with the underlying stochastic dynamics. In this note, our focus is on deriving a finite-sample bound for the Frobenius norm of the identification error $\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$ of an indirect Maximum Likelihood Estimator $\hat{\boldsymbol{\Theta}}$ for the BAR model. The major challenge in this derivation is two-fold. From the perspective of system identification, standard techniques for the finite-sample analysis of linear systems cannot be used due to the significant dependence on the linearity of these systems and the Gaussian properties that the laws of iterates enjoy [5]–[7], [9], [10], [35], as opposed to the discreteness of the BAR model. From the point view of discrete-time Markov chains, the difficulty lies in the exponential growth of the state space, and therefore, of the size of the transition probability matrix, which is $2^p \times 2^p$ for a system with p nodes. Our main result shows that given $T = \Omega\left(\frac{p^3 \log p}{(1-r)^2\epsilon^2 \min\{\lambda_{min}^2(\boldsymbol{\Pi}), \lambda_{min}(\boldsymbol{\Pi})\}}\right)$ observations from a single trajectory of the stationary BAR chain, our estimator is ϵ-close in the Frobenius-norm induced metric to the true $\boldsymbol{\Theta}$ with high probability. Here, $\epsilon > 0$ is the magnitude of the identification error, $r \in [0,1)$ is the underlying Dobrushin coefficient, and $\lambda_{min}(\boldsymbol{\Pi})$ is the minimum eigenvalue of an augmented version of the steady-state correlation matrix. This sample complexity bound can be extended to account for any initial measure of the underlying BAR chain. Moreover, by including appropriate bounds on the model parameters similar to those in [8] in the description of the BAR parameter space, we show that $\lambda_{min}(\boldsymbol{\Pi}) \gtrsim \frac{1}{p}$. This results in sample complexity $T = \Omega\left(\frac{p^5 \log p}{(1-r)^2\epsilon^2}\right)$ or $T = \Omega\left(\frac{p^5 \log p}{\epsilon^2}\right)$ for $\bar{r}$-contractive BAR chains.

II. Preliminaries

The BAR model, first introduced in [8], corresponds to a Markov chain defined over a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = p$ nodes, each node associated with a binary-valued state element. More specifically, $\forall i \in [p]$ and $\forall k \geq 0$,

$$X_i(k+1) \sim \text{Ber}\left(\mathbf{a}_i^\top \mathbf{X}(k) + b_i W_i(k+1)\right). \tag{1}$$

Here, $\mathbf{X}(k) \in \{0,1\}^p$ denotes the state vector at time k and $\{W_i(k+1) \sim \text{Ber}(\rho_{w_i})\}_{i=1}^p$ are independent Bernoulli noise random variables, independent of $\mathbf{X}(t)$ for any $t < k+1$, with $\rho_{w_i} \in (0,1)$. Suppose that $\mathbf{X}(0) \sim \mu$, where μ is a probability measure on $\{0,1\}^p$. Conditioned on $\mathbf{X}(k)$, the entries of $\mathbf{X}(k+1)$ are mutually independent. Moreover, we assume that $\sum_{j=1}^p a_{ij} + b_i \leq 1, a_{ij} \in [0,1), b_i \in (0,1], \forall i \in [p]$. Here, the b_i's are assumed to be nonzero for persistent excitation.

Let $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_p]^\top$. The entries of $\mathbf{A}$ intrinsically reflect the underlying connectivity of the graph, specifically note that

$$(j,i) \in \mathcal{E} \iff a_{ij} > 0, \quad \forall i,j \in [p], \tag{2}$$

where the ordered pair (j,i) denotes a directed edge from node j to node i. Moreover, by noting that $P(X_i(k+1) = 1|\mathbf{X}(k)) = \mathbf{a}_i^\top \mathbf{X}(k) + b_i \rho_{w_i}$ for every $i \in [p]$, we define $c_i = b_i \rho_{w_i}$ and let $\mathbf{c} = [c_1, c_2, \cdots, c_p]^\top$. With these definitions, the effective parameter set of the BAR model is given by

$$\Theta = \left\{\boldsymbol{\Theta} = [\mathbf{A}, \mathbf{c}] \,\middle|\, \sum_{j=1}^p a_{ij} + b_i \leq 1,\; a_{ij} \in [0,1), \forall i,j \in [p],\right.$$
$$\left. b_i \in (0,1], c_i = b_i \rho_{w_i} > 0, \forall i \in [p]\right\}, \tag{3}$$

which is also stochastically equivalent to

$$\Theta = \left\{\boldsymbol{\Theta} = [\mathbf{A}, \mathbf{c}] \,\middle|\, \sum_{j=1}^p a_{ij} + c_i < 1,\; a_{ij} \in [0,1), \forall i,j \in [p],\right.$$
$$\left. c_i > 0, \forall i \in [p]\right\} \tag{4}$$

for $X_i(k+1) \sim \text{Ber}\left(\mathbf{a}_i^\top \mathbf{X}(k) + c_i\right), \forall i \in [p]$, conditionally independent given $\mathbf{X}(k)$ for any $k \geq 0$. Note that in (3) we do not include in $\boldsymbol{\Theta}$ the auxiliary parameters $b_i, \forall i \in [p]$, since these only affect the transition probabilities via the products $c_i = b_i \rho_{w_i}, \forall i \in [p]$. In this note, we focus on estimating $\boldsymbol{\Theta} = [\mathbf{A}, \mathbf{c}] \in \mathbb{R}^{p \times (p+1)}$, which corresponds to identifying the stochastic dynamics and the connectivity of the underlying network.

Clearly, $\{\mathbf{X}(k)\}_{k \geq 0}$ is an irreducible and aperiodic Markov chain with finite state space $\{0,1\}^p$. We denote by π the associated equilibrium measure. Our goal is to estimate the parameters of the model from $T+1$ observations of the BAR sequence, i.e., $\{\mathbf{X}(k)\}_{k=0}^T$. Denote by $\vartheta_{\mathbf{u},r,l} = P\left((\cdot)_r = l|\mathbf{u}\right)$ the probability of transitioning from state $\mathbf{u} \in \{0,1\}^p$ to a state with r-th element equal to $l \in \{0,1\}$. For any states $\mathbf{u}, \mathbf{v} \in \{0,1\}^p$, let $N_{\mathbf{uv}} = \sum_{k=0}^{T-1} \mathbb{1}\left(\mathbf{X}(k) = \mathbf{u}, \mathbf{X}(k+1) = \mathbf{v}\right)$ and $N_{\mathbf{u},r,1} = \sum_{k=0}^{T-1} \mathbb{1}\left(\mathbf{X}(k) = \mathbf{u}, X_r(k+1) = 1\right)$. Moreover, let $N_{\mathbf{u}} = \sum_{\mathbf{v}} N_{\mathbf{uv}} = \sum_{k=0}^{T-1} \mathbb{1}\left(\mathbf{X}(k) = \mathbf{u}\right)$ be the amount of time spent in state $\mathbf{u}$. Define $\mathbf{y}_{T,r} = \left[\vartheta_{\mathbf{X}(0),r,1}, \ldots, \vartheta_{\mathbf{X}(T-1),r,1}\right]^\top$ and $\hat{\mathbf{y}}_{T,r} = \left[N_{\mathbf{X}(0),r,1}/N_{\mathbf{X}(0)}, \ldots, N_{\mathbf{X}(T-1),r,1}/N_{\mathbf{X}(T-1)}\right]^\top$, where the latter contains plug-in estimators of the entries of the former corresponding to solutions of a Maximum Likelihood estimation problem [4]. Here, we use the convention that $\frac{N_{\mathbf{u},r,1}}{N_{\mathbf{u}}} = 0$ if $N_{\mathbf{u}} = 0$, $\forall r \in [p]$. Also, let $\mathbf{Z}_T \in \mathbb{R}^{T \times (p+1)}$ be a matrix with k-th row $\left[\mathbf{X}(k)^\top, 1\right]$ for $k = 0, 1, \ldots, T-1$. With the above definitions, for any $r \in [p]$ we can write

$$\begin{bmatrix} \frac{N_{\mathbf{X}(0),r,1}}{N_{\mathbf{X}(0)}} \\ \vdots \\ \frac{N_{\mathbf{X}(T-1),r,1}}{N_{\mathbf{X}(T-1)}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}(0)^\top & 1 \\ \vdots & \vdots \\ \mathbf{X}(T-1)^\top & 1 \end{bmatrix} \cdot \begin{bmatrix} \hat{\mathbf{a}}_r \\ \hat{c}_r \end{bmatrix}$$
$$\text{or} \quad \hat{\mathbf{y}}_{T,r} = \mathbf{Z}_T \hat{\boldsymbol{\theta}}_r,$$

where $\hat{\mathbf{a}}_r$ and $\hat{c}_r$ are estimators of $\mathbf{a}_r$ and c_r, respectively.

Assuming $\mathbf{Z}_T^\top \mathbf{Z}_T$ is invertible, we get the following estimator:

$$\hat{\boldsymbol{\theta}}_r = \widehat{[\boldsymbol{\Theta}]}_{r,:}^\top = \left(\mathbf{Z}_T^\top \mathbf{Z}_T\right)^{-1} \mathbf{Z}_T^\top \hat{\mathbf{y}}_{T,r}, \quad \forall r \in [p].$$

Therefore,

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} \hat{\boldsymbol{\theta}}_1^\top \\ \vdots \\ \hat{\boldsymbol{\theta}}_p^\top \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{y}}_{T,1}^\top \mathbf{Z}_T \left(\mathbf{Z}_T^\top \mathbf{Z}_T\right)^{-1} \\ \vdots \\ \hat{\mathbf{y}}_{T,p}^\top \mathbf{Z}_T \left(\mathbf{Z}_T^\top \mathbf{Z}_T\right)^{-1} \end{bmatrix}. \tag{5}$$

In the rest of the paper, we derive a finite-sample bound for the Frobenius norm of the identification error $\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$, i.e., we bound the probability

$$P(\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F < \epsilon).$$

For convenience, we will assume that $\boldsymbol{\Theta}$ lies in the interior of the set Θ and also that ϵ is sufficiently small so that $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F < \epsilon$ implies that $\hat{\boldsymbol{\Theta}} \in \Theta$. The results can be extended to the case where $\boldsymbol{\Theta}$ is not an interior point of Θ by projecting the estimator onto the closure of Θ.

III. Main Results

In this section, we study the finite-sample properties of the estimator $\hat{\mathbf{\Theta}}$. Notice that

$$\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\right\|_F \le \sqrt{p}\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\right\|_2. \tag{6}$$

In the sequel, we will focus on deriving a high-probability bound on the event $\{\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 < \bar{\epsilon}\}$, $\forall \bar{\epsilon} > 0$.

Let $\mathbf{Y}_T = \left[\mathbf{y}_{T,1}, \cdots, \mathbf{y}_{T,p}\right]^\top$ and $\hat{\mathbf{Y}}_T = \left[\hat{\mathbf{y}}_{T,1}, \cdots, \hat{\mathbf{y}}_{T,p}\right]^\top$. By (5) we can write

$$\hat{\mathbf{\Theta}} = \hat{\mathbf{Y}}_T \mathbf{Z}_T \left(\mathbf{Z}_T^\top \mathbf{Z}_T\right)^{-1} \text{ and } \mathbf{\Theta} = \mathbf{Y}_T \mathbf{Z}_T \left(\mathbf{Z}_T^\top \mathbf{Z}_T\right)^{-1}$$

whenever $\mathbf{Z}_T^\top \mathbf{Z}_T$ is invertible. Note that a necessary condition for this invertibility is that $T \ge p + 1$. We can bound the error as

$$\begin{aligned} \|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 &= \left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right)^{-1}\right\|_2 \\ &\le \left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 \cdot \left\|\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right)^{-1}\right\|_2 \\ &= \frac{\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2}{\lambda_{min}\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right)}, \end{aligned} \tag{7}$$

where $\lambda_{min}(\cdot)$ denotes the smallest eigenvalue of a matrix. Then, for any $\bar{\epsilon} > 0, \xi > 0$ we have that

$$\begin{aligned} &P_\mu\left(\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 \ge \bar{\epsilon}\right) = P_\mu\left(\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 \ge \bar{\epsilon}, \lambda_{min}\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right) < \xi\right) \\ &+ P_\mu\left(\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 \ge \bar{\epsilon}, \lambda_{min}\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right) \ge \xi\right) \\ &\le P_\mu\left(\lambda_{min}\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right) < \xi\right) \\ &+ P_\mu\left(\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 \ge \bar{\epsilon}, \lambda_{min}\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right) \ge \xi\right) \\ &\le P_\mu\left(\lambda_{min}\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right) < \xi\right) + P_\mu\left(\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 \ge \bar{\epsilon}\xi\right). \end{aligned} \tag{8}$$

The following analysis provides bounds on the two terms on the right-hand side of (8).

Common techniques to show that $\lambda_{min}\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right)$ is not too small with high probability include anti-concentration results [36] and Mendelson's small-ball method [6], [7]. Here, we employ the following lemma from [9] together with covering arguments:

Lemma 1: (Lemma 1, [9]) Consider $\mathbf{X} \in \mathbb{R}^{m\times n}$, $n \le m$ and let $\mathbf{M} \in \mathbb{R}^{n\times n}$ be a full rank matrix. Let also $\tau > 0$ and assume that

$$\left\|(\mathbf{X}\mathbf{M})^\top \mathbf{X}\mathbf{M} - \mathbf{I}_n\right\|_2 \le \max\left(\tau, \tau^2\right). \tag{9}$$

Then,

$$\frac{1-\tau}{s_1(\mathbf{M})} \le s_n(\mathbf{X}) \le \cdots \le s_1(\mathbf{X}) \le \frac{1+\tau}{s_n(\mathbf{M})},$$

where $s_1(\cdot) \ge \cdots \ge s_n(\cdot)$ denote the singular values of the involved matrices arranged in non-increasing order.

Note that for $\tau \in (0, 1)$, the right-hand side of (9) becomes $\max\left(\tau, \tau^2\right) = \tau$. By this lemma, it suffices to show that for a sufficiently large T, the event

$$\left\|(\mathbf{Z}_T\mathbf{M})^\top (\mathbf{Z}_T\mathbf{M}) - \mathbf{I}_{p+1}\right\|_2 \le \tau \tag{10}$$

occurs with high probability for an appropriate choice of $\mathbf{M}$ and $\tau \in (0, 1)$. Further, the term $\left\|(\mathbf{Z}_T\mathbf{M})^\top (\mathbf{Z}_T\mathbf{M}) - \mathbf{I}_{p+1}\right\|_2$ can be viewed as the supremum of the deviation of a Lipschitz function from its expectation. Then, we can use the concentration results by Marton [37] and Bobkov and Götze [38] which together show that Lipschitz functions of a finite-state Markov chain with Dobrushin coefficient less than one concentrate around their expectations with high probability. This leads to the following result:

Proposition 1: Suppose that the BAR sequence $\{\mathbf{X}(k)\}_{k=0}^T$ is initialized with the stationary measure π, i.e., $\mathbf{X}(0) \sim \pi$. Let $r < 1$ be the Dobrushin coefficient. Then for any $\varepsilon \in [0, 1/2)$ and $\tau \in (0, 1)$,

$$\begin{aligned} &P_\pi\left(\left\|(\mathbf{Z}_T\mathbf{M})^\top (\mathbf{Z}_T\mathbf{M}) - \mathbf{I}_{p+1}\right\|_2 \ge \tau\right) \\ &\le 2\left(1 + \frac{2}{\varepsilon}\right)^{p+1} \exp\left(-\frac{T\lambda_{min}^2(\mathbf{\Pi})(1-r)^2(1-2\varepsilon)^2\tau^2}{2p(p+1)}\right), \end{aligned}$$

where $\mathbf{M} = \left(E_\pi\left[\sum_{k=0}^{T-1}\tilde{\mathbf{X}}(k)\tilde{\mathbf{X}}(k)^\top\right]\right)^{-1/2} = (T\mathbf{\Pi})^{-1/2}$, $\mathbf{\Pi} = E_\pi[\tilde{\mathbf{X}}(k)\tilde{\mathbf{X}}(k)^\top]$ and $\tilde{\mathbf{X}}(k) = \left[\mathbf{X}(k)^\top, 1\right]^\top$ for $k = 0, 1, \ldots, T-1$.

A proof of Proposition 1 can be found in the Appendix.

By Lemma 1, the following inequality holds with high probability as a direct consequence of the previous proposition:

$$\lambda_{min}(\mathbf{Z}_T^\top\mathbf{Z}_T) \ge \frac{(1-\tau)^2}{\lambda_{max}^2(\mathbf{M})} = T(1-\tau)^2\lambda_{min}\left(\mathbf{\Pi}\right).$$

Let μ be any measure on $\{0, 1\}^p$. Define $\left\|\frac{\mu}{\pi}\right\|_{2,\pi}^2 = \sum_{i=1}^{2^p}\frac{\mu_i^2}{\pi_i} \in [1, \infty]$, which is a measure of nonstationarity and satisfies $\left\|\frac{\mu}{\pi}\right\|_{2,\pi} \le 1/\sqrt{\min_i \pi_i}$ [39], [40]. We consider the following proposition from [39] to extend Proposition 1 for any initial distribution.

Proposition 2: (Proposition 3.10, [39]) Let $\{X(k)\}_{k=0}^{n-1}$ be a time-homogeneous Markov chain with state space $\mathcal{X}$, initial distribution μ, and stationary distribution π. Suppose that $g\left(X(0), \ldots, X(n-1)\right)$ is a real-valued measurable function. Then,

$$\begin{aligned} &P_\mu\left(g\left(X(0), \ldots, X(n-1)\right) \ge t\right) \\ &\le \left\|\frac{\mu}{\pi}\right\|_{2,\pi}\sqrt{P_\pi\left(g\left(X(0), \ldots, X(n-1)\right) \ge t\right)}. \end{aligned} \tag{11}$$

With the previous two propositions and Lemma 1, we have the following conclusion:

Proposition 3: Suppose that the BAR sequence $\{\mathbf{X}(k)\}_{k=0}^T$ is initialized with measure μ, i.e., $\mathbf{X}(0) \sim \mu$. For any $\varepsilon \in [0, 1/2)$, $\tau \in (0, 1)$ and $\delta \in (0, 1)$,

$$P_\mu\left(\lambda_{min}\left(\mathbf{Z}_T^\top\mathbf{Z}_T\right) \ge T(1-\tau)^2\lambda_{min}\left(\mathbf{\Pi}\right)\right) \ge 1 - \delta$$

when

$$T \ge \frac{4p(p+1)\log\left(\frac{\sqrt{2}\|\frac{\mu}{\pi}\|_{2,\pi}(1+\frac{2}{\varepsilon})^{\frac{p+1}{2}}}{\delta}\right)}{\lambda_{min}^2\left(\mathbf{\Pi}\right)(1-r)^2(1-2\varepsilon)^2\tau^2}.$$

Now let us consider $P_\mu\left(\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 \ge \bar{\epsilon}\xi\right)$ in (8) for $\xi = T(1-\tau)^2\lambda_{min}\left(\mathbf{\Pi}\right)$. A high probability bound on the event $\mathcal{E} = \left\{\left\|\frac{1}{T}\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 < \tilde{\epsilon}\right\}$ for any $\tilde{\epsilon} > 0$ is stated in the following result:

Proposition 4: Suppose that the BAR sequence $\{\mathbf{X}(k)\}_{k=0}^T$ is initialized with measure μ, i.e., $\mathbf{X}(0) \sim \mu$. Then for any $\tilde{\epsilon} > 0$ we have that

$$\begin{aligned} &P_\mu\left(\frac{1}{T}\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 \ge \tilde{\epsilon}\right) \\ &\le (2p+1)\exp\left(-\frac{\tilde{\epsilon}^2 T}{\frac{p(p+1)}{2} + \frac{2\sqrt{p(p+1)}\tilde{\epsilon}}{3}}\right). \end{aligned}$$

The proof is provided in the Appendix and relies on the matrix Freedman inequality [41].

The main result of this note can be stated now.

Theorem 1: Let the Dobrushin coefficient of the BAR model be denoted by r. Suppose that the BAR sequence $\{\mathbf{X}(k)\}_{k=0}^{T}$ is initialized with measure μ, i.e., $\mathbf{X}(0) \sim \mu$. Then

$$P_\mu\left(\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\right\|_F < \epsilon\right) \geq 1 - 2\delta$$

when

$$T \geq \max\left\{\frac{4p(p+1)\log\left(\frac{\sqrt{2}\|\frac{\mu}{\pi}\|_{2,\pi}(1+\frac{2}{\varepsilon})^{\frac{p+1}{2}}}{\delta}\right)}{\lambda_{min}^2\left(\mathbf{\Pi}\right)(1-r)^2(1-2\varepsilon)^2\tau^2},\right.$$
$$\left.\log\left(\frac{1+2p}{\delta}\right)\frac{\frac{p^2(p+1)}{2} + \frac{2}{3}p\sqrt{(p+1)}(1-\tau)^2\lambda_{min}\left(\mathbf{\Pi}\right)\epsilon}{(1-\tau)^4\lambda_{min}^2\left(\mathbf{\Pi}\right)\epsilon^2}\right\}$$

for any $\epsilon > 0$, $\tau \in (0, 1)$, $\varepsilon \in [0, 1/2)$ and $\delta \in (0, 1/2)$.

Proof: We begin by noting that $r < 1$ for the BAR model since the Markov chain can move from any state to any other state in one step with non-zero probability. By Proposition 4,

$$P_\mu\left(\frac{1}{T}\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 \geq \bar{\epsilon}(1-\tau)^2\lambda_{min}\left(\mathbf{\Pi}\right)\right)$$
$$\leq (1+2p)\exp\left(-\frac{(1-\tau)^4\lambda_{min}^2\left(\mathbf{\Pi}\right)\bar{\epsilon}^2 T}{\frac{p(p+1)}{2} + \frac{2\sqrt{p(p+1)}(1-\tau)^2\lambda_{min}(\mathbf{\Pi})\bar{\epsilon}}{3}}\right). \tag{12}$$

Then by (6) and (8),

$$P_\mu\left(\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\right\|_F \geq \epsilon\right) \leq P_\mu\left(\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 \geq \frac{\epsilon}{\sqrt{p}}\right)$$
$$\leq P_\mu\left(\lambda_{min}\left(\mathbf{Z}_T^\top \mathbf{Z}_T\right) < T(1-\tau)^2\lambda_{min}\left(\mathbf{\Pi}\right)\right)$$
$$+ P_\mu\left(\frac{1}{T}\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 \geq \frac{\epsilon}{\sqrt{p}}(1-\tau)^2\lambda_{min}\left(\mathbf{\Pi}\right)\right). \tag{13}$$

Combining (12), (13) and Proposition 3, the conclusion follows. ■

We now comment on how our proof techniques relate to prior work. We first bound $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F$ by $\sqrt{p}\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2$. Then, similarly to related works on LTI systems, we write the estimation error $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2$ as a ratio of two terms, and treat the resulting numerator and denominator separately [6], [7], [9], [36]. More specifically, inspired by an argument in [42], we derive a concentration result for an appropriate martingale sequence in the numerator term based on the matrix Freedman inequality [41]. For the denominator term, the key idea is to obtain an anti-concentration result for the minimum eigenvalue of an appropriate matrix by showing that a sufficient condition holds with high probability using Lemma 1 in [9], ε-net arguments [14] and transportation-cost inequalities [37], [38].

Finally, we provide the following lower bound for $\lambda_{min}(\mathbf{\Pi})$:

Lemma 2: Suppose that $c_i \in [\underline{c}, \bar{c}] \subset (0, 1), \|\mathbf{a}_i\|_1 + c_i \leq \bar{\alpha} < 1, \forall i \in [p]$ and $\forall p \geq 2$. Then, there exists $\bar{C} > 0$ independent of p such that $\lambda_{min}(\mathbf{\Pi}) \geq \frac{\bar{C}}{p}, \forall p \geq 2$.

The proof is provided in the Appendix. With this result, Theorem 1 holds when

$$T \geq \max\left\{\frac{4p^3(p+1)\log\left(\frac{\sqrt{2}\|\frac{\mu}{\pi}\|_{2,\pi}(1+\frac{2}{\varepsilon})^{\frac{p+1}{2}}}{\delta}\right)}{C^2(1-r)^2(1-2\varepsilon)^2\tau^2},\right.$$
$$\left.\log\left(\frac{1+2p}{\delta}\right)\left[\frac{\frac{p^4(p+1)}{2}}{(1-\tau)^4C^2\epsilon^2} + \frac{\frac{2}{3}p^2\sqrt{(p+1)}}{(1-\tau)^2C\epsilon}\right]\right\}.$$

Remark: Note that if we further constrain the parameter space to $\bar{r}$-contractive BAR chains, where $\bar{r} < 1$, then $T = \Omega(p^5 \log p)$ [37].

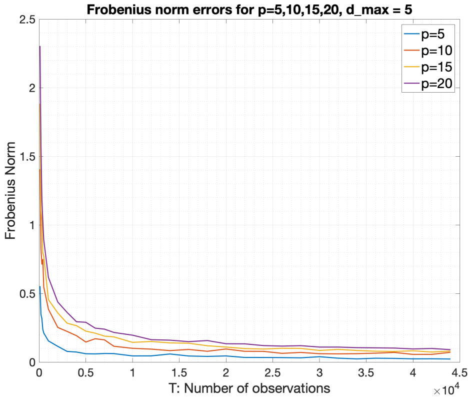

Fig. 1: Frobenius norm errors for $\hat{\mathbf{\Theta}}$

IV. Experimental results

We performed simulations for several synthetic networks with $p = 5, 10, 15, 20$ nodes and the same maximum in-degree $d_{max} = 5$. Fig. 1 illustrates the corresponding Frobenius norm errors of the estimator $\hat{\mathbf{\Theta}}$ for numbers of observations ranging from 100 to 44000. We observe that the estimator requires $T = 2500, 10000, 22000, 44000$ observations to achieve an error of 0.1 for systems with $p = 5, 10, 15, 20$ nodes, respectively. The results show that the estimator converges to arbitrary precision with a polynomial in the system size p number of observations (ignoring logarithmic terms). The polynomial order tends to be between $p^2 \log p$ and $p^3 \log p$, which is approximately a $\lambda_{min}^2(\mathbf{\Pi})$-factor away from the sample complexity obtained by our theoretical analysis.

V. Conclusion

In this note, we have established a sample complexity bound for an indirect Maximum Likelihood Estimator of the BAR model. The bound has polynomial dependence on the system dimension, and further depends on the Dobrushin coefficient and the minimum eigenvalue of an augmented version of the steady-state correlation matrix.

References

[1] T. Söderström, "Convergence properties of the generalised least squares identitication method," *Automatica*, vol. 10, no. 6, pp. 617–626, 1974.

[2] L. Ljung, "Consistency of the least-squares identification method," *IEEE Transactions on Automatic Control*, vol. 21, no. 5, pp. 779–781, 1976.

[3] L. Ljung, "On the consistency of prediction error identification methods," in *Mathematics in Science and Engineering*, vol. 126, pp. 121–164, Elsevier, 1976.

[4] X. Xie, D. Katselis, C. L. Beck, and R. Srikant, "On the consistency of maximum likelihood estimators for causal network identification," *IEEE Control Systems Letters*, 2021.

[5] M. C. Campi and E. Weyer, "Finite sample properties of system identification methods," *IEEE Transactions on Automatic Control*, vol. 47, no. 8, pp. 1329–1334, 2002.

[6] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conference On Learning Theory*, pp. 439–473, 2018.

[7] N. Matni and S. Tu, "A tutorial on concentration bounds for system identification," *arXiv preprint arXiv:1906.11395*, 2019.

[8] D. Katselis, C. L. Beck, and R. Srikant, "Mixing times and structural inference for Bernoulli Autoregressive Processes," *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 3, pp. 364–378, 2018.

[9] Y. Jedra and A. Proutiere, "Finite-time identification of stable linear systems: Optimality of the least-squares estimator," in *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 996–1001, IEEE, 2020.

[10] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite time LTI system identification," *Journal of Machine Learning Research*, vol. 22, no. 26, pp. 1–61, 2021.

[11] M. Hardt, T. Ma, and B. Recht, "Gradient descent learns linear dynamical systems," *Journal of Machine Learning Research*, vol. 19, pp. 1–44, 2018.

[12] V. H. Peña, T. L. Lai, and Q.-M. Shao, *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

[13] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Online least squares estimation with self-normalized processes: An application to bandit problems," *arXiv preprint arXiv:1102.2670*, 2011.

[14] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, p. 210–268. Cambridge University Press, 2012.

[15] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[16] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Finite-time system identification and adaptive control in autoregressive exogenous systems," in *Learning for Dynamics and Control*, pp. 967–979, PMLR, 2021.

[17] R.-S. Wang, A. Saadatpour, and R. Albert, "Boolean modeling in systems biology: an overview of methodology and applications," *Physical biology*, vol. 9, no. 5, p. 055001, 2012.

[18] M. J. Neely, "Stock market trading via stochastic network optimization," in *49th IEEE Conference on Decision and Control (CDC)*, pp. 2777–2784, IEEE, 2010.

[19] C. Nowzari, V. Preciado, and G. J. Pappas, "Analysis and control of epidemics: A survey of spreading processes on complex networks," *IEEE Control Systems Magazine*, vol. 36, pp. 26–46, 2016.

[20] D. Acemoglu, M. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *The Review of Economic Studies*, vol. 78, pp. 1201–1236, 2011.

[21] R. A. Holley and T. M. Liggett, "Ergodic theorems for weakly interacting infinite systems and the voter model," *The annals of probability*, pp. 643–663, 1975.

[22] Y. Hassin and D. Peleg, "Distributed probabilistic polling and applications to proportionate agreement," *Information and Computation*, vol. 171, p. 248–268, Dec 2001.

[23] A. Carro, R. Toral, and M. San Miguel, "The noisy voter model on complex networks," *Scientific reports*, vol. 6, no. 1, pp. 1–14, 2016.

[24] J. Pouget-Abadie and T. Horel, "Inferring graphs from cascades: A sparse recovery framework," in *International Conference on Machine Learning*, pp. 977–986, PMLR, 2015.

[25] A. Agaskar and Y. M. Lu, "Alarm: A logistic auto-regressive model for binary processes on networks," in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 305–308, IEEE, 2013.

[26] P. Clifford and A. Sudbury, "A model for spatial conflict," *Biometrika*, vol. 60, no. 3, pp. 581–588, 1973.

[27] O. A. Pinto and M. A. Munoz, "Quasi-neutral theory of epidemic outbreaks," *PloS one*, vol. 6, no. 7, p. e21946, 2011.

[28] T. M. Liggett and T. M. Liggett, *Interacting particle systems*, vol. 2. Springer, 1985.

[29] A. Kirman, "Ants, rationality, and recruitment," *The Quarterly Journal of Economics*, vol. 108, no. 1, pp. 137–156, 1993.

[30] W.-P. Lee and W.-S. Tzou, "Computational methods for discovering gene networks from expression data," *Briefings in bioinformatics*, vol. 10, no. 4, pp. 408–423, 2009.

[31] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the boolean network model," in *Biocomputing'99*, pp. 17–28, World Scientific, 1999.

[32] Y. Li, W. Chen, Y. Wang, and Z.-L. Zhang, "Voter model on signed social networks," *Internet Mathematics*, vol. 11, p. 93–133, Mar 2015.

[33] E. C. Hall, G. Raskutti, and R. Willett, "Inference of high-dimensional autoregressive generalized linear models," *arXiv preprint arXiv:1605.02693*, 2016.

[34] P. Pandit, M. Sahraee-Ardakan, A. A. Amini, S. Rangan, and A. K. Fletcher, "Generalized autoregressive linear models for discrete high-dimensional data," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 884–896, 2020.

[35] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Nonparametric finite time LTI system identification," *arXiv preprint arXiv:1902.01848*, 2019.

[36] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *International Conference on Machine Learning*, pp. 5610–5618, PMLR, 2019.

[37] K. Marton, "Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration," *The Annals of Probability*, vol. 24, no. 2, pp. 857–866, 1996.

[38] S. G. Bobkov and F. Götze, "Exponential integrability and transportation cost related to logarithmic Sobolev inequalities," *Journal of Functional Analysis*, vol. 163, no. 1, pp. 1–28, 1999.

[39] D. Paulin, "Concentration inequalities for Markov chains by Marton couplings and spectral methods," *Electronic Journal of Probability*, vol. 20, 2015.

[40] G. Wolfer and A. Kontorovich, "Statistical estimation of ergodic Markov chain kernel over discrete state space," *Bernoulli*, vol. 27, no. 1, pp. 532–553, 2021.

[41] J. A. Tropp, "Freedman's inequality for matrix martingales," *Electronic Communications in Probability*, vol. 16, pp. 262–270, 2011.

[42] G. Wolfer and A. Kontorovich, "Statistical estimation of ergodic Markov chain kernel over discrete state space," *arXiv preprint arXiv:1809.05014v3*, 2019.

[43] J. A. Tropp, "Integer factorization of a positive-definite matrix," *SIAM Journal on Discrete Mathematics*, vol. 29, no. 4, pp. 1783–1791, 2015.

[44] S. Puntanen, G. P. Styan, and J. Isotalo, "Block-diagonalization and the Schur complement," in *Matrix Tricks for Linear Statistical Models*, pp. 291–304, Springer, 2011.

APPENDIX

A. Proof of Proposition 1

Before specifying the matrix $\mathbf{M}$, we start with the following observation: for any full-rank matrix $\mathbf{M} \in \mathbb{R}^{(p+1)\times(p+1)}$,

$$\begin{aligned} &\left\| \left(\mathbf{Z}_T\mathbf{M}\right)^\top \left(\mathbf{Z}_T\mathbf{M}\right) - \mathbf{I}_{p+1} \right\|_2 \\ &= \sup_{\mathbf{v}\in S^p} \left| \mathbf{v}^\top \left[\left(\mathbf{Z}_T\mathbf{M}\right)^\top \left(\mathbf{Z}_T\mathbf{M}\right) - \mathbf{I}_{p+1} \right] \mathbf{v} \right| \\ &= \sup_{\mathbf{v}\in S^p} \left| \left\|\mathbf{Z}_T\mathbf{M}\mathbf{v}\right\|_2^2 - \left\|\mathbf{v}\right\|_2^2 \right|, \end{aligned} \tag{14}$$

where S^p is the unit Euclidean sphere in $\mathbb{R}^{p+1}$. The matrix $\mathbf{M}$ will be chosen so that we can make use of the following corollary, which is a combination of [37, Proposition 1] and [38, Theorem 1.3].

Corollary 1: Consider a Markov chain $\{X(k)\}_{k\geq 0}$ with a finite state space $\mathcal{X}$ and Dobrushin coefficient $r < 1$. Suppose that $\mathcal{X}^n$ is equipped with the Hamming metric $d_1(x_{0:n-1}, x'_{0:n-1}) = \sum_{k=0}^{n-1} \mathbb{1}(x(k) \neq x'(k))$ for any two elements $x_{0:n-1} = (x(0), \ldots, x(n-1))$ and $x'_{0:n-1} = \left(x'(0), \ldots, x'(n-1)\right) \in \mathcal{X}^n$. Then for any Lipschitz function $f : \mathcal{X}^n \to \mathbb{R}$ with Lipschitz constant L and $\forall \gamma > 0$,

$$\begin{aligned} &P_\mu\left(\left|f(X(0),\ldots,X(n-1)) - E_\mu\left[f(X(0),\ldots,X(n-1))\right]\right| > \gamma\right) \\ &\quad \leq 2\exp\left(-\frac{2(1-r)^2\gamma^2}{nL^2}\right), \end{aligned}$$

where P_μ is the law of the Markov chain when the initial measure is μ.

In our case, for some fixed $\mathbf{v} \in S^p$, we let

$$f\left(\mathbf{X}(0),\ldots,\mathbf{X}(T-1)\right) = \left\|\mathbf{Z}_T\mathbf{M}\mathbf{v}\right\|_2^2 = \sum_{k=0}^{T-1} \left[\mathbf{v}^\top\mathbf{M}^\top\tilde{\mathbf{X}}(k)\right]^2,$$

where $\tilde{\mathbf{X}}(k) = \left[\mathbf{X}(k)^\top, 1\right]^\top$. It can be shown that f is Lipschitz with respect to the Hamming metric d_1. Consider two T-tuples,

$$\begin{aligned} &\left|f(x_{0:T-1}) - f(x'_{0:T-1})\right| \\ &= \left| \left[\mathbf{v}^\top\mathbf{M}^\top \begin{bmatrix}\mathbf{x}(i)\\ 1\end{bmatrix}\right]^2 - \left[\mathbf{v}^\top\mathbf{M}^\top \begin{bmatrix}\mathbf{x}'(i)\\ 1\end{bmatrix}\right]^2 \right| \\ &= \left| \left[\mathbf{v}^\top\mathbf{M}^\top \begin{bmatrix}\left(\mathbf{x}(i) - \mathbf{x}'(i)\right)\\ 0\end{bmatrix}\right] \left[\mathbf{v}^\top\mathbf{M}^\top \left(\begin{bmatrix}\mathbf{x}(i)\\ 1\end{bmatrix} + \begin{bmatrix}\mathbf{x}'(i)\\ 1\end{bmatrix}\right)\right] \right| \\ &\leq \left\|\mathbf{M}\mathbf{v}\right\|_2^2 \left\|\mathbf{x}(i) - \mathbf{x}'(i)\right\|_2 \left(\left\|\begin{bmatrix}\mathbf{x}(i)\\ 1\end{bmatrix}\right\|_2 + \left\|\begin{bmatrix}\mathbf{x}'(i)\\ 1\end{bmatrix}\right\|_2 \right) \end{aligned}$$

$$\le \|\mathbf{M}\|_2^2 \left\|\mathbf{x}(i)-\mathbf{x}'(i)\right\|_2 \cdot 2\sqrt{p+1}$$
$$\le 2\sqrt{p(p+1)}\,\|\mathbf{M}\|_2^2\,\mathbb{1}\{\mathbf{x}(i)\neq\mathbf{x}'(i)\},$$

where the first inequality is due to the triangle and Cauchy–Schwarz inequalities; the second inequality follows from the fact that $\max_{\mathbf{x}\in\{0,1\}^p}\|\left[\mathbf{x}^\top,1\right]\|_2=\sqrt{p+1}$; and the last inequality holds since $\|\mathbf{x}-\mathbf{x}'\|\le\sqrt{p}$ for $\mathbf{x}\neq\mathbf{x}'$, $\mathbf{x},\mathbf{x}'\in\{0,1\}^p$.

Now notice that by choosing

$$\mathbf{M}=\left(E_\pi\left[\sum_{k=0}^{T-1}\tilde{\mathbf{X}}(k)\tilde{\mathbf{X}}(k)^\top\right]\right)^{-1/2}$$
$$=\left(E_\pi\left[\mathbf{Z}_T^\top\mathbf{Z}_T\right]\right)^{-1/2}=(T\mathbf{\Pi})^{-1/2},$$

where $\mathbf{\Pi}=E_\pi\left[\tilde{\mathbf{X}}(k)\tilde{\mathbf{X}}(k)^\top\right]$, we can write

$$\left|\|\mathbf{Z}_T\mathbf{M}\mathbf{v}\|_2^2-\|\mathbf{v}\|_2^2\right|=\left|\|\mathbf{Z}_T\mathbf{M}\mathbf{v}\|_2^2-E_\pi\left[\|\mathbf{Z}_T\mathbf{M}\mathbf{v}\|_2^2\right]\right|.$$

A direct application of Corollary 1 gives

$$P_\pi\left(\left|\|\mathbf{Z}_T\mathbf{M}\mathbf{v}\|_2^2-\|\mathbf{v}\|_2^2\right|>\gamma\right)$$
$$\le 2\exp\left(-\frac{T(1-r)^2\gamma^2\lambda_{min}^2(\mathbf{\Pi})}{2p(p+1)}\right), \qquad (15)$$

for any $\gamma>0$. The quantity $\left\|(\mathbf{Z}_T\mathbf{M})^\top(\mathbf{Z}_T\mathbf{M})-\mathbf{I}_{p+1}\right\|_2$, as the supremum of $\left|\|\mathbf{Z}_T\mathbf{M}\mathbf{v}\|_2^2-\|\mathbf{v}\|_2^2\right|$ with respect to $\mathbf{v}$ over S^p, will not be large with high probability due to the following lemma.

Lemma 3: ([9, Lemma 4]) Let $\varepsilon\in[0,1/2)$ and $\mathcal{N}(\varepsilon)$ be an ε-net of S^{n-1} with minimal cardinality. Then for symmetric $W\in\mathbb{R}^{n\times n}$ and any $\tau>0$, we have that

$$P\left(\|W\|_2>\tau\right)\le\left(1+\frac{2}{\varepsilon}\right)^n\max_{\mathbf{v}\in\mathcal{N}(\varepsilon)}P\left(\left|\mathbf{v}^\top W\mathbf{v}\right|>(1-2\varepsilon)\tau\right).$$

An application of Lemma 3 with $\gamma=(1-2\varepsilon)\tau$ in (15) and $n=p+1$ gives the desired result.

B. Proof of Proposition 4

Let $\{\mathbf{u}_1,\mathbf{u}_2,\dots,\mathbf{u}_{2^p}\}$ be an enumeration of the states in $\{0,1\}^p$. Recall that for the considered horizon $k=0,1,\dots,T$, $N_{\mathbf{u},r,1}$ is the number of one-step transitions from state $\mathbf{u}$ to some state with r-th entry equal to 1 and $N_{\mathbf{u}}$ denotes the total amount of time the chain spends in state $\mathbf{u}$. We can then rewrite $\left[\left(\hat{\mathbf{Y}}_T-\mathbf{Y}_T\right)\mathbf{Z}_T\right]_{r,:}$ as follows:

$$\left[\left(\hat{\mathbf{Y}}_T-\mathbf{Y}_T\right)\mathbf{Z}_T\right]_{r,:}=\left(\hat{\mathbf{y}}_{T,r}-\mathbf{y}_{T,r}\right)^\top\mathbf{Z}_T$$
$$=\sum_{k=0}^{T-1}\left(\frac{N_{\mathbf{X}(k),r,1}}{N_{\mathbf{X}(k)}}-\vartheta_{\mathbf{X}(k),r,1}\right)\begin{bmatrix}\mathbf{X}(k)\\1\end{bmatrix}^\top$$
$$=\sum_{j=1}^{2^p}N_{\mathbf{u}_j}\left(\frac{N_{\mathbf{u}_j,r,1}}{N_{\mathbf{u}_j}}-\vartheta_{\mathbf{u}_j,r,1}\right)\begin{bmatrix}\mathbf{u}_j\\1\end{bmatrix}^\top$$
$$=\sum_{j=1}^{2^p}\left(N_{\mathbf{u}_j,r,1}-\vartheta_{\mathbf{u}_j,r,1}N_{\mathbf{u}_j}\right)\begin{bmatrix}\mathbf{u}_j\\1\end{bmatrix}^\top. \qquad (16)$$

To upper bound the probability $P_\mu\left(\frac{1}{T}\left\|\left(\hat{\mathbf{Y}}_T-\mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2\ge\tilde{\epsilon}\right)$, we will rely on (16) and the following theorem.

Theorem 2: (Matrix Freedman inequality, [41]) Consider a matrix martingale $\{\mathbf{S}(k)\}_{k\ge0}$ whose values are matrices with dimension $d_1\times d_2$ and $\mathbf{S}(0)=\mathbf{0}_{d_1\times d_2}$. Let $\{\mathbf{D}(k)=\mathbf{S}(k)-\mathbf{S}(k-1)\}_{k\ge1}$ be the martingale difference sequence. Assume that $\{\mathbf{D}(k)\}_{k\ge1}$ is uniformly bounded, i.e.,

$$\|\mathbf{D}(k)\|_2\le R\quad\text{almost surely (a.s.) for } k=1,2,\dots$$

Define two predictable quadratic variation processes for this martingale:

$$\mathbf{W}_{\mathrm{col}}(k):=\sum_{j=1}^{k}E\left[\mathbf{D}(j)\mathbf{D}(j)^\top\middle|\mathcal{F}_{j-1}\right]\quad\text{and}$$
$$\mathbf{W}_{\mathrm{row}}(k):=\sum_{j=1}^{k}E\left[\mathbf{D}(j)^\top\mathbf{D}(j)\middle|\mathcal{F}_{j-1}\right],\quad\text{for}\quad k=1,2\dots$$

Here, $\mathcal{F}_0\subset\mathcal{F}_1\subset\cdots\subset\mathcal{F}$ is a filtration of the σ-algebra $\mathcal{F}$ corresponding to the underlying probability space and $\{\mathbf{S}(k)\}_{k\ge0}$ is adapted to this filtration. Let $\Sigma_k^2=\max\left\{\|\mathbf{W}_{\mathrm{col}}(k)\|_2,\|\mathbf{W}_{\mathrm{row}}(k)\|_2\right\}$. Then for all $\gamma\ge0$ and $\sigma^2>0$,

$$P\left(\exists k\ge0:\left\|\mathbf{S}(k)=\sum_{j=1}^{k}\mathbf{D}(j)\right\|_2\ge\gamma\ \text{ and }\ \Sigma_k^2\le\sigma^2\right)$$
$$\le(d_1+d_2)\exp\left(-\frac{\gamma^2/2}{\sigma^2+R\gamma/3}\right).$$

We now use Theorem 2 with $(d_1,d_2)=(p,p+1)$ and we define an appropriate martingale difference sequence corresponding to the $p\times(p+1)$ martingale $\left(\hat{\mathbf{Y}}_T-\mathbf{Y}_T\right)\mathbf{Z}_T$. For any $r\in[p]$ and for $k=1,2,\dots,T$ we let

$$\mathbf{D}_r(k)=\sum_{j=1}^{2^p}\mathbb{1}\left(\mathbf{X}(k-1)=\mathbf{u}_j\right)$$
$$\cdot\left[\mathbb{1}\left(X_r(k)=1\right)-\vartheta_{\mathbf{u}_j,r,1}\right]\cdot\begin{bmatrix}\mathbf{u}_j\\1\end{bmatrix}^\top$$

and we define $\bar{\mathbf{D}}(k)=\left[\mathbf{D}_1^\top(k),\cdots,\mathbf{D}_p^\top(k)\right]^\top$ as the matrix martingale difference sequence. We now observe that

$$\sum_{k=1}^{T}\mathbf{D}_r(k)=\sum_{j=1}^{2^p}\sum_{k=1}^{T}\mathbb{1}\left(\mathbf{X}(k-1)=\mathbf{u}_j\right)\cdot$$
$$\left(\mathbb{1}\left(X_r(k)=1\right)-\vartheta_{\mathbf{u}_j,r,1}\right)\begin{bmatrix}\mathbf{u}_j\\1\end{bmatrix}^\top$$
$$=\sum_{j=1}^{2^p}\left(N_{\mathbf{u}_j,r,1}-\vartheta_{\mathbf{u}_j,r,1}N_{\mathbf{u}_j}\right)\begin{bmatrix}\mathbf{u}_j\\1\end{bmatrix}^\top$$
$$=\left[\left(\hat{\mathbf{Y}}_T-\mathbf{Y}_T\right)\mathbf{Z}_T\right]_{r,:} \qquad (17)$$

and therefore,

$$\sum_{k=1}^{T}\bar{\mathbf{D}}(k)=\left[\left(\hat{\mathbf{Y}}_T-\mathbf{Y}_T\right)\mathbf{Z}_T\right]. \qquad (18)$$

Moreover, let $\{\mathcal{F}_k\}_{k=0}^{\infty}$ be the sequence of the canonical nested σ-fields $\mathcal{F}_k=\sigma\left(\mathbf{X}(0),\dots,\mathbf{X}(k)\right)$, which corresponds to a filtration and note that $\left[\left(\hat{\mathbf{Y}}_T-\mathbf{Y}_T\right)\mathbf{Z}_T\right]_{r,:}$ is $\mathcal{F}_T$-measurable $\forall r\in[p]$. By the Markov property it follows that

$$E\left[\mathbf{D}_r(k)\middle|\mathcal{F}_{k-1}\right]=E\left[\mathbf{D}_r(k)\middle|\mathbf{X}(k-1)\right]$$
$$=\left(E\left[\mathbb{1}\left(X_r(k)=1\right)\middle|\mathbf{X}(k-1)\right]-\vartheta_{\mathbf{X}(k-1),r,1}\right)$$
$$\cdot\begin{bmatrix}\mathbf{X}(k-1)\\1\end{bmatrix}^\top=\mathbf{0}_{p+1}^\top. \qquad (19)$$

Hence, $E\left[\bar{\mathbf{D}}(k)\middle|\mathcal{F}_{k-1}\right]=\mathbf{0}_{p\times(p+1)}$. Additionally, we observe that

$\forall k \geq 1,$

$$\begin{aligned}\left\|\bar{\mathbf{D}}(k)\right\|_2 &\leq \left\|\bar{\mathbf{D}}(k)\right\|_F \\ &\leq \left(p \cdot \max_{\mathbf{u}_j} \left\|\begin{bmatrix}\mathbf{u}_j \\ 1\end{bmatrix}\right\|_2^2\right)^{1/2} = \sqrt{p(p+1)} \quad \text{a.s.}\end{aligned} \tag{20}$$

Before continuing further, we note that in the light of the above discussion, $\{\bar{\mathbf{D}}(k)\}_{k\geq 1}$ is indeed a martingale difference sequence associated with the desired matrix martingale satisfying the integrability condition stated in [41]:

$$E\left[\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2\right] \leq T\sqrt{p(p+1)} < \infty, \quad \forall T \geq 1.$$

Furthermore, we have that $\forall r, m \in [p]$, $r \neq m$,

$$\begin{aligned}\mathbf{D}_r(k)\mathbf{D}_m(k)^\top = \sum_{j,l=1}^{2^p} &\mathbb{1}\left(\mathbf{X}(k-1) = \mathbf{u}_j\right) \cdot \\ &\mathbb{1}\left(\mathbf{X}(k-1) = \mathbf{u}_l\right)\left(\mathbb{1}\left(X_r(k) = 1\right) - \vartheta_{\mathbf{u}_j,r,1}\right) \cdot \\ &\left(\mathbb{1}\left(X_m(k) = 1\right) - \vartheta_{\mathbf{u}_l,m,1}\right)\begin{bmatrix}\mathbf{u}_j \\ 1\end{bmatrix}^\top \begin{bmatrix}\mathbf{u}_l \\ 1\end{bmatrix} \\ = \sum_{j=1}^{2^p} &\mathbb{1}\left(\mathbf{X}(k-1) = \mathbf{u}_j\right)\left(\mathbb{1}\left(X_r(k) = 1\right) - \vartheta_{\mathbf{u}_j,r,1}\right) \cdot \\ &\left(\mathbb{1}\left(X_m(k) = 1\right) - \vartheta_{\mathbf{u}_j,m,1}\right)\left\|\begin{bmatrix}\mathbf{u}_j \\ 1\end{bmatrix}\right\|_2^2.\end{aligned}$$

Taking the conditional expectation with respect to the sub-σ-algebra $\mathcal{F}_{k-1}$ we obtain

$$\begin{aligned}&E\left[\mathbf{D}_r(k)\mathbf{D}_m(k)^\top \middle| \mathcal{F}_{k-1}\right] = E\left[\mathbf{D}_r(k)\mathbf{D}_m(k)^\top \middle| \mathbf{X}(k-1)\right] \\ &= E\left[\left(\mathbb{1}\left(X_r(k) = 1\right) - \vartheta_{\mathbf{X}(k-1),r,1}\right) \cdot \left\|\tilde{\mathbf{X}}(k-1)\right\|_2^2 \cdot \right. \\ &\left.\left(\mathbb{1}\left(X_m(k) = 1\right) - \vartheta_{\mathbf{X}(k-1),m,1}\right) \middle| \mathbf{X}(k-1)\right] \\ &= 0 \quad \text{a.s.}, \qquad \forall r, m \in [p],\ r \neq m,\end{aligned}$$

and when $r = m$

$$\begin{aligned}&E\left[\mathbf{D}_r(k)\mathbf{D}_r(k)^\top \middle| \mathcal{F}_{k-1}\right] = E\left[\mathbf{D}_r(k)\mathbf{D}_r(k)^\top \middle| \mathbf{X}(k-1)\right] \\ &= E\left[\left(\mathbb{1}\left(X_r(k) = 1\right) - \vartheta_{\mathbf{X}(k-1),r,1}\right)^2 \middle| \mathbf{X}(k-1)\right]\left\|\tilde{\mathbf{X}}(k-1)\right\|_2^2 \\ &\leq \vartheta_{\mathbf{X}(k-1),r,1}(1 - \vartheta_{\mathbf{X}(k-1),r,1})(p+1) \quad \text{a.s.}\end{aligned}$$

Therefore,

$$\begin{aligned}&E\left[\bar{\mathbf{D}}(k)\bar{\mathbf{D}}(k)^\top \middle| \mathcal{F}_{k-1}\right] \\ &= E\left[\begin{bmatrix}\mathbf{D}_1(k) \\ \vdots \\ \mathbf{D}_p(k)\end{bmatrix}\begin{bmatrix}\mathbf{D}_1(k)^\top & \cdots & \mathbf{D}_p(k)^\top\end{bmatrix} \middle| \mathcal{F}_{k-1}\right] \\ &= \left\|\tilde{\mathbf{X}}(k-1)\right\|_2^2 \cdot \operatorname{diag}\left(\bar{\boldsymbol{\nu}}(k-1)\right),\end{aligned} \tag{21}$$

where $\operatorname{diag}\left(\bar{\boldsymbol{\nu}}(k-1)\right)$ denotes the diagonal matrix with diagonal elements contained in the vector $\bar{\boldsymbol{\nu}}(k-1) = [\bar{\nu}_1(k-1), \ldots, \bar{\nu}_p(k-1)]$, $\bar{\nu}_r(k-1) = \vartheta_{\mathbf{X}(k-1),r,1}(1 - \vartheta_{\mathbf{X}(k-1),r,1})$ for $r = 1, 2, \ldots, p$. Furthermore,

$$\begin{aligned}\|\mathbf{W}_{\text{col}}(k)\|_2 &= \left\|\sum_{j=1}^{k} E\left[\bar{\mathbf{D}}(j)\bar{\mathbf{D}}(j)^\top \middle| \mathcal{F}_{j-1}\right]\right\|_2 \\ &= \max_{r\in[p]} \sum_{j=1}^{k} \vartheta_{\mathbf{X}(j-1),r,1}(1 - \vartheta_{\mathbf{X}(j-1),r,1})\left\|\tilde{\mathbf{X}}(j-1)\right\|_2^2 \\ &\leq \frac{k(p+1)}{4} \quad \text{a.s.}, \quad \forall k \geq 1.\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbf{D}_r(k)^\top\mathbf{D}_r(k) = \sum_{j,l=1}^{2^p} &\mathbb{1}\left(\mathbf{X}(k-1) = \mathbf{u}_j\right) \cdot \\ &\mathbb{1}\left(\mathbf{X}(k-1) = \mathbf{u}_l\right)\left(\mathbb{1}\left(X_r(k) = 1\right) - \vartheta_{\mathbf{u}_j,r,1}\right) \cdot \\ &\left(\mathbb{1}\left(X_r(k) = 1\right) - \vartheta_{\mathbf{u}_l,r,1}\right)\begin{bmatrix}\mathbf{u}_j \\ 1\end{bmatrix}\begin{bmatrix}\mathbf{u}_l \\ 1\end{bmatrix}^\top \\ = \sum_{j=1}^{2^p} &\mathbb{1}\left(\mathbf{X}(k-1) = \mathbf{u}_j\right) \cdot \\ &\left(\mathbb{1}\left(X_r(k) = 1\right) - \vartheta_{\mathbf{u}_j,r,1}\right)^2\begin{bmatrix}\mathbf{u}_j \\ 1\end{bmatrix}\begin{bmatrix}\mathbf{u}_j \\ 1\end{bmatrix}^\top.\end{aligned}$$

Therefore,

$$\begin{aligned}&E\left[\bar{\mathbf{D}}(k)^\top\bar{\mathbf{D}}(k) \middle| \mathcal{F}_{k-1}\right] = E\left[\sum_{r=1}^{p}\mathbf{D}_r(k)^\top\mathbf{D}_r(k) \middle| \mathcal{F}_{k-1}\right] \\ &= \sum_{r=1}^{p}\vartheta_{\mathbf{X}(k-1),r,1}\left(1 - \vartheta_{\mathbf{X}(k-1),r,1}\right)\tilde{\mathbf{X}}(k-1)\tilde{\mathbf{X}}(k-1)^\top\end{aligned}$$

and

$$\begin{aligned}&\|\mathbf{W}_{\text{row}}(k)\|_2 \\ &= \left\|\sum_{j=1}^{k}\sum_{r=1}^{p}\vartheta_{\mathbf{X}(j-1),r,1}\left(1 - \vartheta_{\mathbf{X}(j-1),r,1}\right) \cdot \right. \\ &\qquad \left.\tilde{\mathbf{X}}(j-1)\tilde{\mathbf{X}}(j-1)^\top\right\|_2 \\ &\leq \frac{kp(p+1)}{4} \quad \text{a.s.}, \quad \forall k \geq 1.\end{aligned}$$

Summarizing the derived bounds, we conclude that

$$\begin{aligned}\Sigma_k^2 &= \max\left\{\|\mathbf{W}_{\text{col}}(k)\|_2, \|\mathbf{W}_{\text{row}}(k)\|_2\right\} \\ &\leq \frac{kp(p+1)}{4} \quad \text{a.s.}, \quad \forall k \geq 1.\end{aligned}$$

Combining the derived bounds, we can see that $\forall\tilde{\epsilon} > 0$

$$\begin{aligned}&P_\mu\left(\frac{1}{T}\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 \geq \tilde{\epsilon}\right) \\ &= P_\mu\left(\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 \geq T\tilde{\epsilon}\right) \\ &= P_\mu\left(\left\{\left\|\left(\hat{\mathbf{Y}}_T - \mathbf{Y}_T\right)\mathbf{Z}_T\right\|_2 \geq T\tilde{\epsilon}\right\} \cap \right. \\ &\qquad \left.\left\{\Sigma_T^2 \leq \frac{Tp(p+1)}{4}\right\}\right) \\ &\leq (2p+1)\exp\left(-\frac{\tilde{\epsilon}^2 T}{\frac{p(p+1)}{2} + \frac{2\tilde{\epsilon}\sqrt{p(p+1)}}{3}}\right),\end{aligned}$$

where the last inequality follows from applying Theorem 2 with $\gamma = \tilde{\epsilon}T$, $R = \sqrt{p(p+1)}$ due to (20) and $\sigma^2 = \frac{Tp(p+1)}{4}$.

C. Proof of Lemma 2

We first note that

$$\begin{aligned}\mathbf{\Pi} = E_\pi\left[\tilde{\mathbf{X}}(k)\tilde{\mathbf{X}}(k)^\top\right] &= \begin{bmatrix} E_\pi\left[\mathbf{X}(k)\mathbf{X}(k)^\top\right] & E_\pi\left[\mathbf{X}(k)\right] \\ E_\pi\left[\mathbf{X}(k)\right]^\top & 1\end{bmatrix} \\ &= E_\pi \underbrace{\begin{bmatrix} E\left[\mathbf{X}(k)\mathbf{X}(k)^\top|\mathbf{X}(k-1)\right] & E\left[\mathbf{X}(k)|\mathbf{X}(k-1)\right] \\ E\left[\mathbf{X}(k)|\mathbf{X}(k-1)\right]^\top & 1\end{bmatrix}}_{\mathbf{\Pi}(k-1)}.\end{aligned} \tag{22}$$

Moreover, it can be easily seen that

$$\begin{aligned} &E\left[\mathbf{X}(k)\mathbf{X}(k)^\top|\mathbf{X}(k-1)\right] = \\ &E\left[\mathbf{X}(k)|\mathbf{X}(k-1)\right]E\left[\mathbf{X}(k)|\mathbf{X}(k-1)\right]^\top + \mathrm{Cov}(\mathbf{X}(k)|\mathbf{X}(k-1)),\end{aligned} \tag{23}$$

where the conditional covariance matrix $\mathrm{Cov}(\mathbf{X}(k)|\mathbf{X}(k-1))$ is given by

$$\begin{aligned}\mathrm{Cov}(\mathbf{X}(k)|\mathbf{X}(k-1)) = E\left[(\mathbf{X}(k) - E\left[\mathbf{X}(k)|\mathbf{X}(k-1)\right])\cdot\right.\\ \left.(\mathbf{X}(k) - E\left[\mathbf{X}(k)|\mathbf{X}(k-1)\right])^\top|\mathbf{X}(k-1)\right].\end{aligned}$$

Additionally, we note that for $i \neq j$, the (i,j)-th entry of the conditional covariance matrix is

$$\begin{aligned}&E[(X_i(k) - E[X_i(k)|\mathbf{X}(k-1)])\cdot \\ &(X_j(k) - E[X_j(k)|\mathbf{X}(k-1)])|\mathbf{X}(k-1)] = 0,\end{aligned}$$

due to the conditional independence of X_i, X_j given $\mathbf{X}(k-1)$. Furthermore, the (i,i)-th entry of the conditional covariance matrix satisfies

$$\begin{aligned}&E[(X_i(k) - E[X_i(k)|\mathbf{X}(k-1)])^2|\mathbf{X}(k-1)] = \\ &(\mathbf{a}_i^\top\mathbf{X}(k-1) + c_i)(1 - \mathbf{a}_i^T\mathbf{X}(k-1) - c_i) \\ &= \vartheta_{\mathbf{X}(k-1),i,1}(1 - \vartheta_{\mathbf{X}(k-1),i,1}) = \bar{\nu}_i(k-1), \quad i \in [p].\end{aligned}$$

Note that $\bar{\nu}_i(k-1) \leq \frac{1}{4}, \forall i \in [p], \forall k \geq 1$ almost surely. By combining the above results and by following the notation of (21), it follows that

$$\mathrm{Cov}(\mathbf{X}(k)|\mathbf{X}(k-1)) = \mathrm{diag}\left(\bar{\boldsymbol{\nu}}(k-1)\right). \tag{24}$$

We now turn to bounding $\lambda_{min}(\mathbf{\Pi})$. Employing the concavity of the minimum eigenvalue on the space of symmetric matrices [43], we have that

$$\lambda_{min}(\mathbf{\Pi}) \geq E_\pi\left[\lambda_{min}\left(\mathbf{\Pi}(k-1)\right)\right]. \tag{25}$$

We now focus on the matrix $\mathbf{\Pi}(k-1)$. Let $\bar{\boldsymbol{\mu}}(k-1) = E\left[\mathbf{X}(k)|\mathbf{X}(k-1)\right]$. Using the Aitken block diagonalization of this matrix, which relies on the Schur complement of $\mathbf{\Pi}(k-1)$ with respect to 1 (lower diagonal block) [44], we can write:

$$\begin{aligned}&\mathbf{\Pi}(k-1) = \\ &\underbrace{\begin{bmatrix}\mathbf{I}_p & \bar{\boldsymbol{\mu}}(k-1) \\ \mathbf{0}_{p\times 1}^\top & 1\end{bmatrix}}_{\mathbf{K}(k-1)}\underbrace{\begin{bmatrix}\mathbf{F}(k-1) & \mathbf{0}_{p\times 1} \\ \mathbf{0}_{p\times 1}^\top & 1\end{bmatrix}}_{\mathbf{L}(k-1)}\underbrace{\begin{bmatrix}\mathbf{I}_p & \mathbf{0}_{p\times 1} \\ \bar{\boldsymbol{\mu}}(k-1)^\top & 1\end{bmatrix}}_{\mathbf{K}(k-1)^\top},\end{aligned}$$

where

$$\begin{aligned}\mathbf{F}(k-1) &= E\left[\mathbf{X}(k)\mathbf{X}(k)^\top|\mathbf{X}(k-1)\right] - \bar{\boldsymbol{\mu}}(k-1)\bar{\boldsymbol{\mu}}(k-1)^\top \\ &= \mathrm{Cov}(\mathbf{X}(k)|\mathbf{X}(k-1)) = \mathrm{diag}\left(\bar{\boldsymbol{\nu}}(k-1)\right)\end{aligned}$$

is the aforementioned Schur complement. Here, (23) has been used. Moreover, let $\mathbf{w}^*(k-1)$ be the vector in S^p corresponding to the minimum eigenvalue of $\mathbf{\Pi}(k-1)$ and set $\mathbf{q}^*(k-1) = \mathbf{K}(k-1)^\top\mathbf{w}_*(k-1)$. Then,

$$\begin{aligned}&\lambda_{min}(\mathbf{\Pi}(k-1)) = \mathbf{q}^*(k-1)^\top\mathbf{L}(k-1)\mathbf{q}^*(k-1) \geq \\ &\lambda_{min}(\mathbf{L}(k-1))\|\mathbf{q}^*(k-1)\|_2^2 \geq \lambda_{min}(\mathbf{F}(k-1))s_{p+1}^2(\mathbf{K}(k-1)^\top) \\ &= \min_{i\in[p]}\bar{\nu}_i(k-1)s_{p+1}^2(\mathbf{K}(k-1)).\end{aligned}$$

Here, $s_{p+1}(\mathbf{K}(k-1)) = \sqrt{\lambda_{min}(\mathbf{K}(k-1)^\top\mathbf{K}(k-1))}$ is the smallest singular value of $\mathbf{K}(k-1)$. We now note that the eigenvalues $\rho_1(k-1) \geq \ldots \geq \rho_{p+1}(k-1) > 0$ of

$$\begin{aligned}\mathbf{K}(k-1)^\top\mathbf{K}(k-1) &= \begin{bmatrix}\mathbf{I}_p & \bar{\boldsymbol{\mu}}(k-1) \\ \bar{\boldsymbol{\mu}}(k-1)^\top & \|\bar{\boldsymbol{\mu}}(k-1)\|_2^2 + 1\end{bmatrix} \\ &= \mathbf{I}_{p+1} + \begin{bmatrix}\mathbf{0}_{p\times p} & \bar{\boldsymbol{\mu}}(k-1) \\ \bar{\boldsymbol{\mu}}(k-1)^\top & \|\bar{\boldsymbol{\mu}}(k-1)\|_2^2\end{bmatrix}\end{aligned}$$

are 1 with multiplicity $p-1$ ($\rho_2(k-1) = \cdots = \rho_p(k-1)$) and

$$\begin{aligned}&1 + \frac{\|\bar{\boldsymbol{\mu}}(k-1)\|_2^2 \pm \sqrt{\|\bar{\boldsymbol{\mu}}(k-1)\|_2^2(\|\bar{\boldsymbol{\mu}}(k-1)\|_2^2 + 4)}}{2} \\ &= \frac{2 + \|\bar{\boldsymbol{\mu}}(k-1)\|_2^2 \pm \sqrt{\left(2 + \|\bar{\boldsymbol{\mu}}(k-1)\|_2^2\right)^2 - 4}}{2}.\end{aligned}$$

Based on the previous analysis we conclude that

$$\begin{aligned}\lambda_{min}(\mathbf{\Pi}) &\geq E_\pi\left[\min_{i\in[p]}\bar{\nu}_i(k-1)\rho_{p+1}(k-1)\right] \\ &\geq \min_{i\in[p]} c_i(1 - \|\mathbf{a}_i\|_1 - c_i)\cdot \\ \min_{\mathbf{x}(k-1)\in\{0,1\}^p} &\frac{2 + \|\bar{\boldsymbol{\mu}}(k-1)\|_2^2 - \sqrt{\left(2 + \|\bar{\boldsymbol{\mu}}(k-1)\|_2^2\right)^2 - 4}}{2},\end{aligned}$$

where in the last line we assume that $\mathbf{X}(k-1) = \mathbf{x}(k-1)$. Finally, the function $f(x) = \frac{2+x-\sqrt{(2+x)^2-4}}{2}$ is strictly decreasing for $x > 0$. Using the bounds in the statement of Lemma 2 and this observation, we have that

$$\lambda_{min}(\mathbf{\Pi}) \geq \underline{c}(1-\bar{\alpha})\underbrace{\frac{2 + p\bar{\alpha}^2 - \sqrt{\left(2 + p\bar{\alpha}^2\right)^2 - 4}}{2}}_{s_{min}}, \tag{26}$$

where

$$\max_{i\in[p],\mathbf{x}(k-1)\in\{0,1\}^p}\bar{\mu}_i(k-1) = \max_{i\in[p]}\|\mathbf{a}_i\|_1 + c_i \leq \bar{\alpha}$$

is employed. It can be verified that for any constant $\tilde{c}$ such that $\tilde{c}\bar{\alpha}^2 < 1$ and $2\tilde{c}\bar{\alpha}^2 + \tilde{c}\sqrt{\tilde{c}\bar{\alpha}^2} + \tilde{c} < 2$, we have that $s_{min} \geq \frac{\tilde{c}}{p}, \forall p \geq 2$. This implies that $\lambda_{min}(\mathbf{\Pi}) \geq \frac{C}{p}, \forall p \geq 2$ for any $C = \underline{c}(1-\bar{\alpha})\tilde{c}$ in an appropriate interval $(0, \bar{C})$, where $\bar{C}$ is independent of p.