# On the Consistency of Maximum Likelihood Estimators for Causal Network Identification

Xiaotian Xie, *Graduate Student Member, IEEE*, Dimitrios Katselis, Carolyn L. Beck, *Senior Member, IEEE*, and R. Srikant, *Fellow, IEEE*

***Abstract*—We consider the problem of identifying parameters of a particular class of Markov chains, called Bernoulli Autoregressive (BAR) processes. The structure of any BAR model is encoded by a directed graph. Incoming edges to a node in the graph indicate that the state of the node at a particular time instant is influenced by the states of the corresponding parental nodes in the previous time instant. The associated edge weights determine the corresponding level of influence from each parental node. In the simplest setup, the Bernoulli parameter of a particular node's state variable is a convex combination of the parental node states in the previous time instant and an additional Bernoulli noise random variable. This letter focuses on the problem of edge weight identification using Maximum Likelihood (ML) estimation and proves that the ML estimator is strongly consistent for two variants of the BAR model. We additionally derive closed-form estimators for the aforementioned two variants and prove their strong consistency.**

***Index Terms*—Identification, Markov chains.**

## I. Introduction

THE SPREADING of ideas and information, the propagation of viruses and diseases, and the fluctuation of stock prices are examples of processes evolving over social, information or other types of networks [1]–[8]. Identifying the underlying network structure in these systems motivates the so-called *network inference problem*, which aims at recovering the underlying connectivity between entities or nodes in the system based on observed data. The dependencies, correlations or causal relationships between network entities can be modeled as undirected or directed edges in a graph. The associated dependency strengths can be described as edge weights. Many algorithms have been proposed to identify the network structure and edge weights from time series data for various processes. Clearly, efficient algorithms in terms of sample complexity are desired.

The network inference problem for various dynamic processes has been recently studied in both the machine learning literature and the system identification literature. The so-called continuous-time independent cascade model (CICE) considered in [5], [9], [10] presents a typical model for capturing the dynamics of virus or information spreading in networks. A discrete-time version of CICE is studied in [11] and [12]. In [7], the Generalized Linear Model is formulated, which is a class of diffusion models encompassing both the discrete and continuous-time CICE models, and the Linear Voter model [13].

System identification focuses on estimating system parameters from measured input and output data [14]–[16]. Recently system identification has been used to study the problem of network inference. Broadly speaking, system identification methods can be classified as belonging to one of two types: methods that consider continuous state-spaces [17]–[21] and those that consider discrete state-spaces [22]. Our paper falls into the second category.

In this letter, we consider the Bernoulli Autoregressive (BAR) model, which is a parameterized discrete-time Markov chain initially introduced in [23]. In this model, the state of each node is a Bernoulli random variable with probability of success equal to a convex combination of the parental node states (or their flipped states) in the previous time step and an additional binary noise term ensuring persistence of excitation. The BAR model can be used to approximate opinion dynamics, biological and financial times series, and similar processes [23]–[25]. Another relevant discrete-time binary process is the ALARM model proposed in [26]. In contrast to the BAR model, the ALARM model defines the transition probabilities via a logistic function.

Relying on well-established statistical principles, we first formulate and study the consistency properties of the Maximum Likelihood (ML) parameter estimator for the BAR model in which every parental node causally influences each descendant node positively; the notion of positive correlations

Manuscript received October 29, 2020; revised December 28, 2020; accepted January 14, 2021. Date of publication January 22, 2021; date of current version June 24, 2021. This work was supported in part by NSF under Grant NeTS 1718203, Grant CPS ECCS 1739189, Grant ECCS 16-09370, Grant ECCS 2032321, and Grant CCF 1934986; in part by NSF/USDA under Grant AG 2018-67007-28379; in part by ARO under Grant W911NF-19-1-0379; and in part by ONR under Grant N00014-19-1-2566. Recommended by Senior Editor R. S. Smith. *(Corresponding author: Xiaotian Xie.)*

Xiaotian Xie, Carolyn L. Beck, and R. Srikant are with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: xx5@illinois.edu; beck3@illinois.edu; rsrikant@illinois.edu).

Dimitrios Katselis is with the Department of Electronics and Communication Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: katselis@illinois.edu).

Digital Object Identifier 10.1109/LCSYS.2021.3053610

is formalized in [23]. The consistency of ML estimators in the case of independent and identically distributed (i.i.d.) random variables has been studied extensively; see, e.g., [27], [28] and references therein. The consistency of ML estimators for Markov chains appears to be less well studied, see [29] for a reference.

To establish the (strong) consistency of the ML estimator for the BAR model, we prove that the vectorized transition probability matrix is an injective mapping of the model parameters. In the rest of this letter, we call the injectivity of this mapping *identifiability* of the BAR model. The strong consistency of the ML estimator is then shown by leveraging the injectivity and the continuity of the transition probabilities with respect to the parameters, as well as the compactness of the parameter set. By relying on the ML principle, a closed-form estimator is subsequently provided. Strong consistency is also shown to hold for this estimator. The identifiability proof is then extended to the *generic* BAR model with both positive and negative correlations, where the notion of negative correlations is also formalized in [23]. This identifiability extension establishes the strong consistency of the ML estimator for the general BAR model class. The closed-form estimator and its consistency are also extended to the *generic* BAR model. These analytical results provide a complement to the prior work [23].

*Notation:* Matrices and vectors are denoted by bold upper and lowercase letters, respectively. Probability distributions in vector form may be either denoted by bold upper or lowercase letters. Random vectors are also denoted by uppercase bold letters, while their corresponding realizations are denoted by lowercase bold letters. Scalar random variables are denoted by uppercase letters. The $i$-th entry of a vector $\mathbf{x}$ is denoted by $x_i$. For a matrix $\mathbf{A}$, $a_{ij}$ corresponds to its $(i,j)$-th entry. Depending on the context, vector and matrix entries may be indexed more generally, e.g., by state elements. $\mathbf{1}_m$ and $\mathbf{0}_m$ are the $m \times 1$ all-ones and all-zeros vectors, respectively, and $\mathbf{0}_{m\times n}$ is the all-zeros $m \times n$ matrix. $\mathbf{I}_m$ is the $m \times m$ identity matrix. Moreover, $\mathbf{e}_{m,i}$ is the $i$-th column of $\mathbf{I}_m$. For $m \in \mathbb{N}$, $[m] = \{1, 2, \ldots, m\}$. Finally, $\mathbb{I}(\cdot)$ stands for the indicator function.

## II. The Bar Model With Positive Correlations

The BAR model is a special form of a Markov chain defined on a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = p$ nodes. Let $X_i(k) \in \{0, 1\}$ be the state of node $i \in [p]$ at time instant $k$ and let $\mathbf{X}(k) \in \{0, 1\}^p$ be the associated BAR process state vector at the same time instant. The BAR model with positive correlations only is described by

$$X_i(k+1) \sim \text{Ber}\Big(\mathbf{a}_i^\top \mathbf{X}(k) + b_i W_i(k+1)\Big), \quad i = 1, \ldots, p, \quad (1)$$

where $\mathbf{a}_i \in [0, 1]^p$, $b_i \in [0, 1]$, $i = 1, \ldots, p$ are parameters of the BAR model and $\text{Ber}(\rho)$ represents the Bernoulli distribution with parameter $\rho$. Additionally, $\{W_i(k+1) \sim \text{Ber}(\rho_{w_i})\}_{i=1}^p$ are independent noise random variables, also independent of $\mathbf{X}(t)$ for any $t < k+1$, where $\rho_{w_i} \in [\rho_{min}, \rho_{max}]$ for all $i \in [p]$ with $0 < \rho_{min} < \rho_{max} < 1$. Moreover, the initial distribution is $P_{\mathbf{X}(0)}$, i.e., $\mathbf{X}(0) \sim P_{\mathbf{X}(0)}$. The interpretation here is that the entries of $\mathbf{X}(k+1)$ are conditionally independent Bernoulli random variables given $\mathbf{X}(k)$.

To ensure that the Bernoulli random variables in (1) are well-defined, we require that

$$\sum_{j=1}^{p} a_{ij} + b_i = 1, \quad \forall i \in [p]. \quad (2)$$

*Remark:* This can be relaxed to $\sum_{j=1}^{p} a_{ij} + b_i = \beta_i$, where $\beta_i \in (0, 1]$, $\forall i \in [p]$.

For persistent excitation, we further assume that $b_i \geq b_{min}, \forall i \in [p]$, where $b_{min} \in (0, 1)$ is a constant. Notice that if $b_i = 0$ for all $i \in [p]$, the BAR Markov chain will get absorbed in $\mathbf{0}_p$ or $\mathbf{1}_p$ upon visiting the state $\mathbf{0}_p$ or $\mathbf{1}_p$, respectively.

Furthermore, we assume that $\mathbf{a}_i$ encodes a part of the graph structure through the equivalence

$$(j, i) \in \mathcal{E} \iff a_{ij} > 0, \quad \forall i, j \in [p], \quad (3)$$

where the ordered pair $(j, i)$ denotes a directed edge from node $j$ to node $i$. The notion of positive correlations in (1) relies on the fact that $a_{ij} > 0$ increases the probability of the event $\{X_i(k+1) = 1\}$ when $X_j(k) = 1$. A more general form of the BAR model with both positive and negative correlations is introduced in Section V.

We now let $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p]^T$, i.e., $\mathbf{a}_r^T$ corresponds to the $r$-th row of $\mathbf{A}$, $\mathbf{b} = [b_1, b_2, \ldots, b_p]^T$, $\mathbf{W} = [W_1, W_2, \ldots, W_p]^T$ and $\rho_w = [\rho_{w_1}, \rho_{w_2}, \ldots, \rho_{w_p}]^T$. We note that $\{\mathbf{X}(k)\}_{k\geq 0}$ is an irreducible and aperiodic Markov chain with finite state space $\{0, 1\}^p$. Moreover, for any vectors $\mathbf{u}, \mathbf{v} \in \{0, 1\}^p$

$$\begin{aligned} p_{\mathbf{uv}} &= E_{\mathbf{W}}[P(\mathbf{X}(k+1) = \mathbf{v} | \mathbf{X}(k) = \mathbf{u}, \mathbf{W})] \\ &= \prod_{i=1}^{p} \Big[\mathbf{a}_i^\top \mathbf{u} + \rho_{w_i} b_i\Big]^{v_i} \Big[1 - \mathbf{a}_i^\top \mathbf{u} - \rho_{w_i} b_i\Big]^{1-v_i} \end{aligned} \quad (4)$$

specifies the transition probability from state $\mathbf{u}$ to state $\mathbf{v}$. We denote by $\pi \in \mathbb{R}^{2^p}$ the associated stationary distribution with component $\pi_{\mathbf{u}}$ corresponding to the state $\mathbf{u} \in \{0, 1\}^p$ and by $\mathbf{P} = (p_{\mathbf{uv}}) \in \mathbb{R}^{2^p \times 2^p}$ the BAR transition probability matrix.

The goal is to recover the model parameters from an observed sequence $\{\mathbf{X}(k) = \mathbf{x}(k)\}_{k=0}^T$. Clearly, by inferring $\mathbf{A}$, estimates of $\mathbf{b}$ and the underlying network structure are direct per (2) and (3), respectively. Moreover, by the subsequent analysis it will become apparent that the results in this letter can be extended to the case where $\sum_{j=1}^{p} a_{ij} + b_i \leq 1$ for every $i \in [p]$ when the Bernoulli noise parameters $\rho_{w_i}$ are assumed to be known.

## III. Maximum Likelihood Estimation

In this section, we consider recovering the BAR model parameters via ML estimation and we establish the strong consistency of the ML estimator. Suppose that $\{\mathbf{x}(k)\}_{k=0}^T$ is a sequence of observations generated by the BAR model (1). Let $\theta = (\mathbf{A}, \mathbf{b}, \rho_w)$ with the implicit relationship $\mathbf{b} = \mathbf{1}_p - \mathbf{A}\mathbf{1}_p$. Clearly, $\mathbf{b}$ is a redundant parameter, but it is preserved here to facilitate the subsequent analysis. From (4) the rescaled

log-likelihood function is given by

$$\begin{aligned} L_T(\theta) &= \frac{1}{T}\sum_{k=0}^{T-1} \log P(\mathbf{x}(k+1)|\mathbf{x}(k);\theta) + \frac{1}{T}\log P_{\mathbf{X}(0)}(\mathbf{x}(0);\theta) \\ &= \frac{1}{T}\sum_{k=0}^{T-1}\sum_{i=1}^{p}\Big[x_i(k+1)\log\Big(\mathbf{a}_i^\top \mathbf{x}(k) + \rho_{w_i} b_i\Big) \\ &\quad + (1-x_i(k+1))\log\Big(1-\mathbf{a}_i^\top \mathbf{x}(k) - \rho_{w_i} b_i\Big)\Big] \\ &\quad + \frac{1}{T}\log P_{\mathbf{X}(0)}(\mathbf{x}(0);\theta), \end{aligned} \tag{5}$$

Henceforth we assume that $P_{\mathbf{X}(0)}$ is independent of the model parameters. For this reason, the term containing the initial measure in (5) and in any subsequent log-likelihood function in the rest of the letter will be omitted. Such functions will still be called log-likelihood functions.

For any states $\mathbf{u}, \mathbf{v} \in \{0,1\}^p$, we denote by $N_{\mathbf{uv}}$ the number of one-step transitions from state $\mathbf{u}$ to state $\mathbf{v}$ in the observed sequence and we let $N_{\mathbf{u}} = \sum_{\mathbf{v}} N_{\mathbf{uv}}$ be the amount of time spent in state $\mathbf{u}$ over a horizon of $T$ time steps. Then (5) can be also written as

$$L_T(\theta) = \sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T}\log p_{\mathbf{uv}}(\theta). \tag{6}$$

*Remark:* The likelihood function in (6) is concave if $\rho_w$ is assumed known. A natural assumption for this value could be $\rho_w = (1/2)\mathbf{1}_p$. However, we do not make such an assumption here, in which case the likelihood function is no longer concave, in general. In numerical results shown in the longer version of the paper [30], we use standard numerical algorithms available in most software packages to find the maximum of the likelihood function.

Let $\theta_0 \in \Theta$ be the true parameter tuple $(\mathbf{A}, \mathbf{b}, \rho_w)$, where

$$\Theta = \Big\{(\mathbf{A},\mathbf{b},\rho_w)\Big|\sum_{j=1}^{p} a_{ij} + b_i = 1,\ a_{ij} \geq 0, \forall i,j \in [p],$$
$$b_i \geq b_{min},\ \text{and}\ \rho_{w_i} \in [\rho_{min},\ \rho_{max}], \forall i \in [p]\Big\}$$

is a compact set. An application of the Ergodic Theorem [31] for Markov chains reveals that

$$L_T(\theta_0) \xrightarrow[T\to\infty]{\text{a.s.}} \sum_{\mathbf{u},\mathbf{v}} \pi_{\mathbf{u}} p_{\mathbf{uv}}(\theta_0)\log p_{\mathbf{uv}}(\theta_0),$$

which is the negative of the entropy rate of the corresponding BAR chain with parameter tuple $\theta_0$ and is always finite since the BAR model has a finite state space.

The ML estimator $\hat{\theta}_T$ of $\theta_0$ satisfies

$$\hat{\theta}_T \in \arg\max_{\theta\in\Theta}\ T\cdot L_T(\theta) = \arg\max_{\theta\in\Theta}\ L_T(\theta). \tag{7}$$

In the rest of this section, we will show the strong consistency of $\hat{\theta}_T$, that is,

$$\hat{\theta}_T \xrightarrow[T\to\infty]{\text{a.s.}} \theta_0.$$

The proof is provided in Theorem 1 in the following by verifying that the conditions of in [29, Th. 2.1] for general discrete-time Markov chains hold for our BAR model and by combining these conditions with a proper and complete self-contained derivation. To summarize, we first prove that $\mathbf{P}(\hat{\theta}_T) \xrightarrow[T\to\infty]{\text{a.s.}} \mathbf{P}(\theta_0)$. To establish strong consistency, we then show that the vector-valued mapping $\mathbf{p} : \Theta \to \mathbb{R}^{2^{2p}}$ defined as $\mathbf{p}(\theta) = \text{vec}(\mathbf{P}(\theta))$ is injective, i.e.,

$$\forall \theta, \theta' \in \Theta, \quad \theta \neq \theta' \implies \mathbf{p}(\theta) \neq \mathbf{p}(\theta'). \tag{8}$$

Here, vec$(\cdot)$ denotes the vectorization of a matrix. Finally, we complete the proof by leveraging the compactness of the parameter set $\Theta$ and the continuity of the components of $\mathbf{p}(\theta) = \text{vec}(\mathbf{P}(\theta))$ or equivalently, of the transition probabilities with respect to the model parameters.

*Remark:* In the following, we say that the BAR model is *identifiable* when (8) holds.

*Theorem 1:* The ML estimator $\hat{\theta}_T$ of $\theta_0$, defined in (7), for the BAR model in (1) is strongly consistent.

*Proof:* We present the proof in three parts.

*Part I:* We first show that $\mathbf{P}(\hat{\theta}_T) \xrightarrow[T\to\infty]{\text{a.s.}} \mathbf{P}(\theta_0)$. The proof is a simplified, self-contained version of the proof of [29, Th. 2.1].

For each $\mathbf{u} \in \{0,1\}^p$, we define the (row) vector $\mathbf{Q}_{\mathbf{u}} = (N_{\mathbf{uv}}/N_{\mathbf{u}})_{\mathbf{v}\in\{0,1\}^p} \in \mathbb{R}^{2^p}$ with the convention $\mathbf{Q}_{\mathbf{u}} = 2^{-p}\mathbf{1}_{2^p}^\top$ for $N_{\mathbf{u}} = 0$ and we let $\mathbf{P}_{\mathbf{u}} = (P_{\mathbf{uv}})_{\mathbf{v}\in\{0,1\}^p} \in \mathbb{R}^{2^p}$ denote the transition distribution out of state $\mathbf{u}$, which is also a row vector in the transition matrix $\mathbf{P}$. In particular, it is well-known that the set $\{\mathbf{Q}_{\mathbf{u}}\}_{\mathbf{u}\in\{0,1\}^p}$ is the *ML estimator* of the transition matrix $\mathbf{P}$, assuming no further parameterization of the transition probabilities. By the non-negativity of the Kullback–Leibler divergence, we have

$$D_{\text{KL}}\Big(\mathbf{Q}_{\mathbf{u}}\Big\|\mathbf{P}_{\mathbf{u}}(\hat{\theta}_T)\Big) = -\sum_{\mathbf{v}} \frac{N_{\mathbf{uv}}}{N_{\mathbf{u}}}\log\frac{p_{\mathbf{uv}}(\hat{\theta}_T)}{N_{\mathbf{uv}}/N_{\mathbf{u}}} \geq 0$$

or equivalently,

$$\sum_{\mathbf{v}} \frac{N_{\mathbf{uv}}}{N_{\mathbf{u}}}\log\frac{N_{\mathbf{uv}}}{N_{\mathbf{u}}} \geq \sum_{\mathbf{v}} \frac{N_{\mathbf{uv}}}{N_{\mathbf{u}}}\log p_{\mathbf{uv}}(\hat{\theta}_T).$$

Multiply both sides of the above inequality by $\frac{N_{\mathbf{u}}}{T}$ and sum over $\mathbf{u}$ to obtain

$$\sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T}\log\frac{N_{\mathbf{uv}}}{N_{\mathbf{u}}} \geq \sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T}\log p_{\mathbf{uv}}(\hat{\theta}_T) \geq \sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T}\log p_{\mathbf{uv}}(\theta_0) \tag{9}$$

where the last inequality is due to (6) and the definition of the ML estimator. From (9), we can further obtain

$$0 \geq \sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T}\log\frac{p_{\mathbf{uv}}(\hat{\theta}_T)}{N_{\mathbf{uv}}/N_{\mathbf{u}}} \geq \sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T}\log\frac{p_{\mathbf{uv}}(\theta_0)}{N_{\mathbf{uv}}/N_{\mathbf{u}}}. \tag{10}$$

By the Ergodic Theorem for Markov chains,

$$\sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T}\log p_{\mathbf{uv}}(\theta_0) \xrightarrow[T\to\infty]{\text{a.s.}} \sum_{\mathbf{u},\mathbf{v}} \pi_{\mathbf{u}}(\theta_0) p_{\mathbf{uv}}(\theta_0)\log p_{\mathbf{uv}}(\theta_0) \tag{11}$$

and also

$$\sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T}\log\frac{N_{\mathbf{uv}}}{N_{\mathbf{u}}} \xrightarrow[T\to\infty]{\text{a.s.}} \sum_{\mathbf{u},\mathbf{v}} \pi_{\mathbf{u}}(\theta_0) p_{\mathbf{uv}}(\theta_0)\log p_{\mathbf{uv}}(\theta_0). \tag{12}$$

By (11) and (12) we have that

$$\sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T}\log\frac{p_{\mathbf{uv}}(\theta_0)}{N_{\mathbf{uv}}/N_{\mathbf{u}}} \xrightarrow[T\to\infty]{\text{a.s.}} 0.$$

This together with (10) yields

$$\sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T} \log \frac{p_{\mathbf{uv}}(\hat{\theta}_T)}{N_{\mathbf{uv}}/N_{\mathbf{u}}} \xrightarrow[T\to\infty]{\text{a.s.}} 0. \tag{13}$$

Employing Pinsker's inequality [32] and the fact that the total variation distance between two discrete measures $q, r$ with vector forms $\mathbf{q}, \mathbf{r}$, respectively, is $\|q-r\|_{\text{TV}} = (1/2)\|\mathbf{q}-\mathbf{r}\|_1$, we have that for each $\mathbf{u} \in \{0,1\}^p$

$$\frac{1}{2} D_{\text{KL}}\Big(\mathbf{Q}_{\mathbf{u}} \Big\| \mathbf{P}_{\mathbf{u}}(\hat{\theta}_T)\Big) \geq \Big\|\mathbf{Q}_{\mathbf{u}} - \mathbf{P}_{\mathbf{u}}(\hat{\theta}_T)\Big\|_{\text{TV}}^2 \geq \frac{1}{4}\Big\|\mathbf{Q}_{\mathbf{u}} - \mathbf{P}_{\mathbf{u}}(\hat{\theta}_T)\Big\|_2^2.$$

Multiplying again by $\frac{N_{\mathbf{u}}}{T}$ and summing over $\mathbf{u}$ gives

$$-2\sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{uv}}}{T} \log \frac{p_{\mathbf{uv}}(\hat{\theta}_T)}{N_{\mathbf{uv}}/N_{\mathbf{u}}} \geq \sum_{\mathbf{u},\mathbf{v}} \frac{N_{\mathbf{u}}}{T}\left(p_{\mathbf{uv}}(\hat{\theta}_T) - \frac{N_{\mathbf{uv}}}{N_{\mathbf{u}}}\right)^2 \geq 0. \tag{14}$$

Now employing again the Ergodic Theorem for Markov chains, i.e., $N_{\mathbf{u}}/T \xrightarrow[T\to\infty]{\text{a.s.}} \pi_{\mathbf{u}} > 0, \forall \mathbf{u} \in \{0,1\}^p$, and combining (13) and (14) yields

$$\left|p_{\mathbf{uv}}(\hat{\theta}_T) - \frac{N_{\mathbf{uv}}}{N_{\mathbf{u}}}\right| \xrightarrow[T\to\infty]{\text{a.s.}} 0, \quad \forall (\mathbf{u},\mathbf{v}) \in (\{0,1\}^p)^2.$$

We then end up with

$$\left|p_{\mathbf{uv}}(\hat{\theta}_T) - p_{\mathbf{uv}}(\theta_0)\right| \xrightarrow[T\to\infty]{\text{a.s.}} 0, \quad \forall (\mathbf{u},\mathbf{v}) \in (\{0,1\}^p)^2.$$

*Part II (Proof of Identifiability):* We reparameterize the BAR model as $\theta = (\mathbf{A}, \mathbf{c})$ with $\mathbf{c} = \text{diag}(\mathbf{b})\rho_w$. Clearly, there is a one-to-one correspondence between a given set of parameters $(\mathbf{A}, \mathbf{c})$ and $(\mathbf{A}, \mathbf{b}, \rho_w)$ via the relations $\mathbf{b} = \mathbf{1}_p - \mathbf{A}\mathbf{1}_p$ and $\rho_w = (\text{diag}(\mathbf{1}_p - \mathbf{A}\mathbf{1}_p))^{-1}\mathbf{c}$, where $(\cdot)^{-1}$ denotes matrix inversion. Suppose that two different sets of parameters $\theta = (\mathbf{A}, \mathbf{c})$ and $\theta' = (\mathbf{A}', \mathbf{c}')$ lead to the same transition probability matrix, i.e., $\mathbf{P}(\theta) = \mathbf{P}(\theta')$ or equivalently, $\mathbf{p}(\theta) = \mathbf{p}(\theta')$. First, consider the case of $\mathbf{c} \neq \mathbf{c}'$. The following argument is valid for both the cases of $\mathbf{A} = \mathbf{A}'$ and $\mathbf{A} \neq \mathbf{A}'$. Without loss of generality, assume that $c_1 \neq c_1'$. Let $\mathbf{u} = \mathbf{0}$ and $\mathbf{v}$ be some vector in $\{0,1\}^p$ with $v_1 = 0$. Since $p_{\mathbf{uv}}(\theta) = p_{\mathbf{uv}}(\theta')$ and $c_i, c_i' \neq 0$ $\forall i$, (4) implies that

$$1 - c_1' = (1-c_1)\prod_{i=2}^{p}\left(\frac{c_i}{c_i'}\right)^{v_i}\left(\frac{1-c_i}{1-c_i'}\right)^{1-v_i}. \tag{15}$$

Consider now the transition probability from $\mathbf{u} = \mathbf{0}$ to $\mathbf{v}'$, where $v_1' = 1$ and $v_j' = v_j$ for $j = 2, \ldots, p$. Since $p_{\mathbf{uv}'}(\theta) = p_{\mathbf{uv}'}(\theta')$,

$$c_1' = c_1\prod_{i=2}^{p}\left(\frac{c_i}{c_i'}\right)^{v_i}\left(\frac{1-c_i}{1-c_i'}\right)^{1-v_i}. \tag{16}$$

Combining (15) and (16), it is easy to see that $c_1 = c_1'$, which is a contradiction. Thus, $\mathbf{c} \neq \mathbf{c}' \implies \mathbf{P}(\theta) \neq \mathbf{P}(\theta')$ or equivalently, $\mathbf{p}(\theta) \neq \mathbf{p}(\theta')$.

Now we consider the second case where $\mathbf{c} = \mathbf{c}'$ and $\mathbf{A} \neq \mathbf{A}'$. Without loss of generality, let $a_{11} \neq a_{11}'$. Consider $\mathbf{u}' = \mathbf{e}_{p,1}$ and the same $\mathbf{v}, \mathbf{v}'$ as before. By our assumption that $p_{\mathbf{u}'\mathbf{v}}(\theta) = p_{\mathbf{u}'\mathbf{v}}(\theta')$ and $p_{\mathbf{u}'\mathbf{v}'}(\theta) = p_{\mathbf{u}'\mathbf{v}'}(\theta')$, the contradiction $a_{11} = a_{11}'$ arises. Thus, $\mathbf{c} = \mathbf{c}', \mathbf{A} \neq \mathbf{A}' \implies \mathbf{P}(\theta) \neq \mathbf{P}(\theta')$ or equivalently, $\mathbf{p}(\theta) \neq \mathbf{p}(\theta')$.

Finally, it is easy to see that $\mathbf{c} = \mathbf{c}'$ and $\mathbf{A} = \mathbf{A}'$ imply that $\mathbf{b} = \mathbf{b}'$ and $\rho_w = \rho_w'$ due to the aforementioned one-to-one correspondence between $(\mathbf{A}, \mathbf{c})$ and $(\mathbf{A}, \mathbf{b}, \rho_w)$.

*Part III (Completing the Proof):* Let $\Omega$ be a set of sample paths such that $P(\hat{\theta}_T)$ converges to $P(\theta_0)$. Suppose $\hat{\theta}_T$ does not converge to $\theta_0$ on one of these sample paths. Since $\{\hat{\theta}_T\}$ lies in a compact set, there exists a subsequence $\{\hat{\theta}_{T_n}\}$ that converges. By the continuity of $p_{\mathbf{uv}}$ $\forall \mathbf{u}, \mathbf{v}$, $\lim_{n\to\infty} p_{\mathbf{uv}}(\hat{\theta}_{T_n}) = p_{\mathbf{uv}}(\lim_{n\to\infty}\hat{\theta}_{T_n})$. Since this limit is equal to $p_{\mathbf{uv}}(\theta_0)$, and from the identifiability result in Part II, we can conclude that $\hat{\theta}_{T_n}$ converges to $\theta_0$. Since this argument applies to every convergent subsequence of $\{\hat{\theta}_T\}$ and $\Theta$ is a compact set, the original sequence also converges to $\theta_0$. ■

## IV. A Closed-Form Estimator

In this section, we provide a closed-form estimator for the BAR model parameters in (1), i.e., for $(\mathbf{A}, \mathbf{c} = \text{diag}(\mathbf{b})\rho_w)$; from an estimate for $(\mathbf{A}, \mathbf{c})$ we can recover an estimate for $(\mathbf{b}, \rho_w)$. Considering the log-likelihood function for $\theta = (\mathbf{A}, \mathbf{c})$, we have

$$\begin{aligned} L(\theta) &= \sum_{k=0}^{T-1}\sum_{\mathbf{u}}\sum_{\mathbf{v}} \mathbb{I}(\mathbf{x}(k) = \mathbf{u},\ \mathbf{x}(k+1) = \mathbf{v}) \log p_{\mathbf{uv}}(\theta) \\ &= \sum_{k=0}^{T-1}\sum_{\mathbf{u}}\sum_{\mathbf{v}} \mathbb{I}(\mathbf{x}(k) = \mathbf{u},\ \mathbf{x}(k+1) = \mathbf{v}) \\ &\quad \times \sum_{r=1}^{p}\big[v_r \log P(v_r = 1|\mathbf{u}) + (1-v_r)\log P(v_r = 0|\mathbf{u})\big]. \end{aligned} \tag{17}$$

Observe that $P(v_r = 1|\mathbf{u})$ and $P(v_r = 0|\mathbf{u})$ are independent of $\mathbf{v}$. We can therefore define $\vartheta_{\mathbf{u},r,l} = P((\cdot)_r = l|\mathbf{u})$, for $l \in \{0,1\}$. Furthermore, we define $N_{\mathbf{u},r,l} = \sum_{k=0}^{T-1}\mathbb{I}(\mathbf{x}(k) = \mathbf{u}, x_r(k+1) = l)$, which is the number of times the BAR chain transitions from state $\mathbf{u}$ to a state with $r$-th entry equal to $l$. Moreover, $N_{\mathbf{u},r,0} + N_{\mathbf{u},r,1} = N_{\mathbf{u}} = \sum_{\mathbf{v}} N_{\mathbf{uv}}$, $\forall \mathbf{u} \in \{0,1\}^p$ and $r \in [p]$. Suppose that there are $m$ distinct states $\mathbf{u}_1$, $\mathbf{u}_2, \ldots,$ $\mathbf{u}_m$ in the subsequence $\{\mathbf{x}(k)\}_{k=0}^{T-1}$ of the observed sequence $\{\mathbf{x}(k)\}_{k=0}^{T}$. Define $\mathbf{U}_m = [\mathbf{u}_1, \ldots, \mathbf{u}_m]^\top \in \mathbb{R}^{m\times p}$ and $\mathbf{y}_{m,r} = [N_{\mathbf{u}_1,r,1}/N_{\mathbf{u}_1}, \ldots, N_{\mathbf{u}_m,r,1}/N_{\mathbf{u}_m}]^\top$.

Now for $i \in [p]$, define $\hat{c}_i = \sum_{k=0}^{T-1}\mathbb{I}(\mathbf{x}(k) = \mathbf{0}_p, x_i(k+1) = 1)/\sum_{k=0}^{T-1}\mathbb{I}(\mathbf{x}(k) = \mathbf{0}_p)$ as an estimator of the entry $c_i$ when the state $\mathbf{0}_p$ is visited at least once in $\{\mathbf{x}(k)\}_{k=0}^{T-1}$ or let $\hat{\mathbf{c}} = \gamma\mathbf{1}_p$ for some $\gamma \in [b_{min}\rho_{min}, \rho_{max}]$ otherwise by convention.

Let us rewrite the log-likelihood function in (17) as

$$\mathscr{L} = L(\theta) = \sum_{\mathbf{u}}\sum_{r=1}^{p}\big(N_{\mathbf{u},r,0}\log\vartheta_{\mathbf{u},r,0} + N_{\mathbf{u},r,1}\log\vartheta_{\mathbf{u},r,1}\big).$$

Instead of maximizing this function with respect to $\theta = (\mathbf{A}, \mathbf{c})$, we maximize it with respect to the choice of the marginal conditional probabilities $\{\vartheta_{\mathbf{u},r,0}, \vartheta_{\mathbf{u},r,1}\}_{\mathbf{u},r}$. Consider the constrained ML estimation problem

$$\begin{aligned} \max_{\{\vartheta_{\mathbf{u},r,0}\geq 0, \vartheta_{\mathbf{u},r,1}\geq 0\}_{\mathbf{u},r}} \quad & \mathscr{L} \\ \text{s.t.} \quad & \vartheta_{\mathbf{u},r,0} + \vartheta_{\mathbf{u},r,1} = 1,\ \forall \mathbf{u} \in \{0,1\}^p,\ \forall r \in [p]. \end{aligned}$$

Forming the Lagrangian and setting the gradient (with respect to $\{\vartheta_{\mathbf{u},r,0}, \vartheta_{\mathbf{u},r,1}\}_{\mathbf{u},r}$) to zero, we obtain

$$\hat{\vartheta}_{\mathbf{u},r,i} = \frac{N_{\mathbf{u},r,i}}{N_{\mathbf{u}}}, \quad \forall \mathbf{u} \in \{0,1\}^p, \ \forall r \in [p], \ \forall i \in \{0,1\}.$$

Recall that $\vartheta_{\mathbf{u},r,1}$ is defined as the probability of transitioning from state $\mathbf{u}$ to some state with $r$-th component equal to 1. We can therefore require that

$$\begin{bmatrix} \frac{N_{\mathbf{u}_1,r,1}}{N_{\mathbf{u}_1}} \\ \vdots \\ \frac{N_{\mathbf{u}_m,r,1}}{N_{\mathbf{u}_m}} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{u}_1^\top \\ \vdots \\ \mathbf{u}_m^\top \end{bmatrix}}_{\mathbf{U}_m} \cdot \hat{\mathbf{a}}_r + \hat{c}_r \mathbf{1}_m \tag{18}$$

$$\text{or} \quad \mathbf{y}_{m,r} - \hat{c}_r \mathbf{1}_m = \mathbf{U}_m \hat{\mathbf{a}}_r. \tag{19}$$

Note that (18) is reminiscent of the *invariance property* for ML estimation [27]. Then, whenever $\mathbf{U}_m$ is full-column rank, $\hat{\mathbf{A}}$ is an estimator of $\mathbf{A}$, where

$$\hat{\mathbf{a}}_r = \left(\mathbf{U}_m^\top \mathbf{U}_m\right)^{-1} \mathbf{U}_m^\top \left(\mathbf{y}_{m,r} - \hat{c}_r \cdot \mathbf{1}_m\right), \quad \forall r \in [p].$$

Finally, a valid estimate of the parameter tuple for any $T \geq 1$ is

$$\hat{\theta} = \left[\left(\hat{\mathbf{A}}, \hat{\mathbf{b}} = \mathbf{1}_p - \hat{\mathbf{A}}\mathbf{1}_p, \hat{\rho}_w = \left(\text{diag}(\mathbf{1}_p - \hat{\mathbf{A}}\mathbf{1}_p)\right)^{-1} \hat{\mathbf{c}}\right)\right]^+, \tag{20}$$

where $[\,\cdot\,]^+$ corresponds to a projection onto the parameter set $\Theta$.

*Theorem 2:* The estimator in (20) is strongly consistent.

*Proof:* It is sufficient to show that $(\hat{\mathbf{A}}, \hat{\mathbf{c}})$ is strongly consistent. Since the BAR chain is finite-state, the stationary probabilities satisfy $\pi_{\mathbf{u}} > 0, \forall \mathbf{u} \in \{0,1\}^p$. Moreover, for any initial measure, the Ergodic Theorem for Markov chains implies that

- $\frac{N_{\mathbf{u}}}{T} \xrightarrow[T\to\infty]{\text{a.s.}} \pi_{\mathbf{u}}$. This further implies that $\mathbf{U}_m \xrightarrow[T\to\infty]{\text{a.s.}} \mathbf{U}_{2^p}$ in the sense that $[\mathbf{U}_m^\top \ \ \mathbf{0}_{2^p-m\times p}^\top]^\top \xrightarrow[T\to\infty]{\text{a.s.}} \mathbf{U}_{2^p}$.
- $\frac{N_{\mathbf{u},r,1}}{N_{\mathbf{u}}} \xrightarrow[T\to\infty]{\text{a.s.}} P((\cdot)_r = 1|\mathbf{u}), \forall \mathbf{u} \in \{0,1\}^p, \forall r \in [p]$ or equivalently, $\frac{N_{\mathbf{u},r,1}}{N_{\mathbf{u}}} \xrightarrow[T\to\infty]{\text{a.s.}} \mathbf{u}^\top \mathbf{a}_r + c_r, \forall \mathbf{u} \in \{0,1\}^p, \forall r \in [p]$. This implies that $\mathbf{y}_{m,r} \xrightarrow[T\to\infty]{\text{a.s.}} [P((\cdot)_r = 1|\mathbf{u})]_{\mathbf{u}\in\{0,1\}^p}, \forall r \in [p]$, which is a $2^p \times 1$ column vector, in the sense that $[\mathbf{y}_{m,r}^\top \ \ \mathbf{0}_{2^p-m}^\top]^\top \xrightarrow[T\to\infty]{\text{a.s.}} [P((\cdot)_r = 1|\mathbf{u})]_{\mathbf{u}\in\{0,1\}^p}, \forall r \in [p]$. As a consequence, $\hat{\mathbf{c}} \xrightarrow[T\to\infty]{\text{a.s.}} \mathbf{c}$.

Combining these observations with (19) we obtain

$$\lim_{T\to\infty} [\mathbf{U}_m^\top \ \ \mathbf{0}_{2^p-m\times p}^\top]^\top \hat{\mathbf{a}}_r = \lim_{T\to\infty} \mathbf{U}_{2^p} \hat{\mathbf{a}}_r = \mathbf{U}_{2^p} \mathbf{a}_r \ \text{ a.s.}$$

and the strong consistency of the closed-form estimator in (20) follows if $\mathbf{U}_{2^p}$ is full-column rank or equivalently if $\mathbf{U}_{2^p}^\top \mathbf{U}_{2^p}$ is nonsingular. It is easy to see that $\mathbf{U}_{2^p}^\top \mathbf{U}_{2^p}$ has diagonal entries equal to $2^{p-1}$ and off-diagonal entries equal to $2^{p-2}$. Thus, we can write $\mathbf{U}_{2^p}^\top \mathbf{U}_{2^p} = 2^{p-2}\mathbf{1}_p\mathbf{1}_p^\top + 2^{p-2}\mathbf{I}_p$. This matrix is invertible for every $p < \infty$, since $1 + 2^{p-2}\mathbf{1}_p^\top (2^{p-2}\mathbf{I}_p)^{-1}\mathbf{1}_p = 1 + p > 0$ as the condition in the Sherman—Morrison formula [33] dictates. ∎

*Remark:* The closed-form estimator in (20) is obtained from the ML estimators of the marginal probabilities $\vartheta_{\mathbf{u},r,l}$ under appropriate conditions and a projection operation. These features lead to a consistency proof different from the key ideas in the proof of Theorem 1.

## V. The Generic Bar Model

Motivated by modeling positive and negative influences from parental nodes, an extension of the BAR model in (1) has been introduced in [23]. We first reformulate this generic BAR model.

Denote by $\mathscr{S}_i = \mathscr{S}_i^+ \cup \mathscr{S}_i^-$ the parental set of node $i$, where $\mathscr{S}_i^+ \cap \mathscr{S}_i^- = \emptyset$. The nodes in $\mathscr{S}_i^+$ and $\mathscr{S}_i^-$ are said to have positive and negative influence on $i$, respectively. The generic BAR model, parameterized by $\tilde{\theta} = (\mathbf{A}, \tilde{\mathbf{A}}, \mathbf{b}, \rho_w)$, is defined as

$$X_i(k+1) \sim \text{Ber}\Big(\mathbf{a}_i^\top \mathbf{X}(k) + \tilde{\mathbf{a}}_i^\top (\mathbf{1} - \mathbf{X}(k)) + b_i W_i(k+1)\Big), \tag{21}$$

for all $i \in [p]$, where $\mathbf{a}_i^\top$ and $\tilde{\mathbf{a}}_i^\top$ are the $i$-th rows of $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\tilde{\mathbf{A}} \in \mathbb{R}^{p\times p}$, respectively. Furthermore, we assume that $\mathscr{S}_i^+ = \text{supp}(\mathbf{a}_i)$ and $\mathscr{S}_i^- = \text{supp}(\tilde{\mathbf{a}}_i)$. Here, supp$(\cdot)$ denotes the support of a vector. As in the previous case, the constraints $\sum_{j=1}^p (a_{ij} + \tilde{a}_{ij}) + b_i = 1, \forall i \in [p]$ are also required in this case. Similarly, we assume that $a_{ij}, \tilde{a}_{ij} \geq 0, \forall i,j \in [p]$, $b \geq b_{min}$ and $\rho_{w_i} \in [\rho_{min}, \rho_{max}], \forall i \in [p]$. Therefore, the parameter set is defined as

$$\tilde{\Theta} = \Bigg\{ \Big(\mathbf{A}, \tilde{\mathbf{A}}, \mathbf{b}, \rho_w\Big) \Big| \sum_{j=1}^p \big(a_{ij} + \tilde{a}_{ij}\big) + b_i = 1, \ b_i \geq b_{min}, \ \rho_{w_i} \in [\rho_{min}, \rho_{max}], \forall i \in [p], \text{ and } a_{ij}, \tilde{a}_{ij} \geq 0, \ a_{ij}\tilde{a}_{ij} = 0, \ \forall i,j \in [p] \Bigg\}.$$

*Remark:* The parameter set $\tilde{\Theta}$ is compact. A brief justification of this is provided in [30].

The ML estimator is a maximizer of the rescaled log-likelihood function, that is,

$$\tilde{\theta}_T \in \arg\max_{\tilde{\theta}\in\tilde{\Theta}} \tilde{L}_T(\tilde{\theta}),$$

where

$$\tilde{L}_T(\tilde{\theta}) = \frac{1}{T}\sum_{k=0}^{T-1}\sum_{i=1}^{p}\Big[x_i(k+1)\log\Big(\mathbf{a}_i^\top \mathbf{x}(k) + \tilde{\mathbf{a}}_i^\top (\mathbf{1} - \mathbf{x}(k)) + \rho_{w_i} b_i\Big) + (1 - x_i(k+1))\log\Big(1 - \mathbf{a}_i^\top \mathbf{x}(k) - \tilde{\mathbf{a}}_i^\top (\mathbf{1} - \mathbf{x}(k)) - \rho_{w_i} b_i\Big)\Big].$$

The ML estimator for the generic BAR model can be shown to be strongly consistent via a direct extension of the analysis in Section III. More precisely, it is sufficient to establish identifiability.

*Theorem 3:* For the generic BAR model in (21), $\tilde{\theta} \neq \tilde{\theta}' \implies \mathbf{P}(\tilde{\theta}) \neq \mathbf{P}(\tilde{\theta}'), \forall (\tilde{\theta}, \tilde{\theta}') \in \tilde{\Theta} \times \tilde{\Theta}$ with $\tilde{\theta} \neq \tilde{\theta}'$.

The key idea of the proof is similar to the identifiability proof in Theorem 1. For this reason, the complete proof is presented in [30].

We can also derive a closed-form estimator for the generic BAR model by first rewriting $P(X_i(k+1) = 1|\mathbf{X}(k))$ as

$$P(X_i(k+1) = 1|\mathbf{X}(k) = \mathbf{x}(k)) = (\mathbf{a}_i - \tilde{\mathbf{a}}_i)^\top \mathbf{x}(k) + \tilde{\mathbf{a}}_i^\top \mathbf{1} + b_i \rho_{w_i} = \bar{\mathbf{a}}_i^\top \mathbf{x}(k) + \bar{c}_i.$$

Here, $\bar{\mathbf{a}}_i = \mathbf{a}_i - \tilde{\mathbf{a}}_i$ and $\bar{c}_i = \tilde{\mathbf{a}}_i^\top \mathbf{1} + b_i \rho_{w_i}$ for $i \in [p]$. We further note that due to the nonoverlapping supports of $\mathbf{a}_i$ and $\tilde{\mathbf{a}}_i$ for every $i \in [p]$, the vector $\bar{\mathbf{a}}_i$ contains the entries of $\mathbf{a}_i$ and the entries of $\tilde{\mathbf{a}}_i$ with flipped signs, each at a different location.

Similarly, we can reparameterize the generic BAR model as $\bar{\theta} = (\bar{\mathbf{A}}, \bar{\mathbf{c}})$, where $\bar{\mathbf{A}} = [\bar{\mathbf{a}}_1, \ldots, \bar{\mathbf{a}}_p]^T$ and $\bar{\mathbf{c}} = [\bar{c}_1, \ldots, \bar{c}_p]^T$. With the same defintions on $\mathbf{U}_m$ and $\mathbf{y}_{m,r}$ as before, an extension of the estimator in (20) can be obtained.

First, $\forall i \in [p]$, define $\hat{\bar{c}}_i = \sum_{k=0}^{T-1} \mathbb{I}(\mathbf{x}(k) = \mathbf{0}_p, x_i(k+1) = 1) / \sum_{k=0}^{T-1} \mathbb{I}(\mathbf{x}(k) = \mathbf{0}_p)$ as an estimator of the entry $\bar{c}_i$ when the state $\mathbf{0}_p$ is visited at least once in $\{\mathbf{x}(k)\}_{k=0}^{T-1}$ or let $\hat{\bar{\mathbf{c}}} = \gamma \mathbf{1}_p$ for some $\gamma \in [b_{min}\rho_{min}, 1 - b_{min} + b_{min}\rho_{max}]$ otherwise by convention. Then we claim that whenever $\mathbf{U}_m$ is full-column rank, $\hat{\bar{\mathbf{A}}}$ is an estimator of $\bar{\mathbf{A}}$, where

$$\hat{\bar{\mathbf{a}}}_r = \left(\mathbf{U}_m^\top \mathbf{U}_m\right)^{-1} \mathbf{U}_m^\top \left(\mathbf{y}_{m,r} - \hat{\bar{c}}_r \cdot \mathbf{1}_m\right), \ \forall r \in [p].$$

Note that $\hat{\mathbf{A}}$ and $\hat{\bar{\mathbf{A}}}$ can be separated based on the signs of the entries in $\hat{\bar{\mathbf{A}}}$. Moreover, $\hat{\mathbf{b}} = \mathbf{1}_p - (\hat{\mathbf{A}} + \hat{\bar{\mathbf{A}}})\mathbf{1}_p$, $\hat{\mathbf{c}} = \hat{\bar{\mathbf{c}}} - \hat{\bar{\mathbf{A}}}\mathbf{1}_p$ and $\hat{\rho}_w = (\mathrm{diag}(\hat{\mathbf{b}}))^{-1}\hat{\mathbf{c}}$. Finally, $\hat{\bar{\theta}} = [(\hat{\mathbf{A}}, \hat{\bar{\mathbf{A}}}, \hat{\mathbf{b}}, \hat{\rho}_w)]^+$ is a valid estimate of the parameter tuple for any $T \geq 1$, where $[\cdot]^+$ corresponds to a projection onto the parameter set $\tilde{\Theta}$ by preserving the supports of $\hat{\mathbf{A}}$ and $\hat{\bar{\mathbf{A}}}$.

The derivation of this closed-form estimator and the proof of its strong consistency are straightforward based on the derivation of (20) and the proof of Theorem 2, respectively.

## VI. Conclusion

In this letter, we studied the problem of estimating the parameters of a class of Markov chains called BAR models. ML estimation for BAR chains was shown to be strongly consistent. Strong consistency was also established for certain closed-form estimators of the parameters of these BAR models.

## References


[1] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[2] M. J. Neely, "Stock market trading via stochastic network optimization," in *Proc. IEEE Conf. Decis. Control*, 2010, pp. 2777–2784.

[3] D. Acemoglu, M. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *Rev. Econ. Stud.*, vol. 78, no. 4, pp. 1201–1236, 2011.

[4] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games Econ. Behav.*, vol. 76, no. 1, pp. 210–225, 2012.

[5] M. Gomez-Rodriguez, L. Song, H. Daneshmand, and B. Schölkopf, "Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 793–801.

[6] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Directed information graphs," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6887–6909, Dec. 2015.

[7] J. Pouget-Abadie and T. Horel, "Inferring graphs from cascades: A sparse recovery framework," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 977–986.

[8] C. Nowzari, V. Preciado, and G. J. Pappas, "Analysis and control of epidemics: A survey of spreading processes on complex networks," *IEEE Control Syst. Mag.*, vol. 36, no. 1, pp. 26–46, Feb. 2016.

[9] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Trans. Knowl. Disc. Data*, vol. 5, no. 4, pp. 1–37, 2012.

[10] B. Abrahao, F. Chierichetti, R. Kleinberg, and A. Panconesi, "Trace complexity of network inference," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2013, pp. 491–499.

[11] P. Netrapalli and S. Sanghavi, "Learning the graph of epidemic cascades," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, pp. 211–222, 2012.

[12] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2003, pp. 137–146.

[13] T. M. Liggett, *Interacting Particle Systems*, vol. 276. New York, NY, USA: Springer, 2012.

[14] T. Söderström and P. Stoica, *System Identification*. New York, NY, USA: Prentice-Hall, 1989.

[15] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ, USA: Pearson Educ., 1998.

[16] M. Verhaegan and M. Verdult, *Filtering and System Identification*. Cambridge, U.K.: Cambridge University Press, 2007.

[17] D. Materassi and G. Innocenti, "Topological identification in networks of dynamical systems," *IEEE Trans. Autom. Control*, vol. 55, no. 8, pp. 1860–1871, Aug. 2010.

[18] D. Materassi, G. Innocenti, L. Giarré, and M. Salapaka, "Model identification of a network as compressing sensing," *Syst. Control Lett.*, vol. 62, no. 8, pp. 664–672, 2013.

[19] A. Chiuso and G. Pillonetto, "A Bayesian approach to sparse dynamic network identification," *Automatica*, vol. 48, no. 8, pp. 1553–1565, 2012.

[20] A. J. Seneviratne and V. Solo, "Topology identification of a sparse dynamic network," in *Proc. IEEE 51st IEEE Conf. Decis. Control (CDC)*, 2012, pp. 1518–1523.

[21] P. M. J. Van den Hof, A. G. Dankers, P. S. C. Heuberger, and X. Bombois, "Identification of dynamic models in complex networks with prediction error methods—Basic methods for consistent module estimates," *Automatica*, vol. 49, no. 10, pp. 2994–3006, 2013.

[22] R. Mattila, V. Krishnamurthy, and B. Wahlberg, "Recursive identification of chain dynamics in hidden Markov models using non-negative matrix factorization," in *Proc. 54th IEEE Conf. Decis. Control (CDC)*, 2015, pp. 4011–4016.

[23] D. Katselis, C. L. Beck, and R. Srikant, "Mixing times and structural inference for Bernoulli Autoregressive Processes," *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 3, pp. 364–378, Jul./Sep. 2019.

[24] K. Sznajd-Weron and J. Sznajd, "Opinion evolution in closed community," *Int. J. Mod. Phys. C*, vol. 11, no. 6, pp. 1157–1165, 2000.

[25] R.-S. Wang, A. Saadatpour, and R. Albert, "Boolean modeling in systems biology: An overview of methodology and applications," *Phys. Biol.*, vol. 9, no. 5, 2012, Art. no. 055001.

[26] A. Agaskar and Y. M. Lu, "ALARM: A logistic auto-regressive model for binary processes on networks," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2013, pp. 305–308.

[27] P. Moulin and V. V. Veeravalli, *Statistical Inference for Engineers and Data Scientists*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[28] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY, USA: Springer, 1994.

[29] B. Ranneby, "On necessary and sufficient conditions for consistency of MLE's in Markov chain models," *Scandinavian J. Stat.*, vol. 5, no. 2, pp. 99–105, 1978.

[30] X. Xie, D. Katselis, C. L. Beck, and R. Srikant, "On the consistency of maximum likelihood estimators for causal network identification," 2020. [Online]. Available: http://arxiv.org/abs/2010.08870.

[31] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, vol. 31. New York, NY, USA: Springer, 2013.

[32] S. Kullback, "A lower bound for discrimination information in terms of variation (corresp.)," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 126–127, Jan. 1967.

[33] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge, U.K.: Cambridge Univ. Press, 2007.