DiVA
http://www.diva-portal.org

This is the published version of a paper presented at *55th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2017, Allerton HouseMonticello, United States, 3 October 2017 through 6 October 2017.*

Citation for the original published paper:

Molavipour, S., Bassi, G., Skoglund, M. (2017)
Testing for Directed Information Graphs
In: *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (pp. 212-219). IEEE
Annual Allerton Conference on Communication Control and Computing
https://doi.org/10.1109/ALLERTON.2017.8262740

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-226276

Testing for Directed Information Graphs

Sina Molavipour, Germán Bassi, and Mikael Skoglund

***Abstract*—In this paper, we study a hypothesis test to determine the underlying directed graph structure of nodes in a network, where the nodes represent random processes and the direction of the links indicate a causal relationship between said processes. Specifically, a k-th order Markov structure is considered for them, and the chosen metric to determine a connection between nodes is the directed information. The hypothesis test is based on the empirically calculated transition probabilities which are used to estimate the directed information. For a single edge, it is proven that the detection probability can be chosen arbitrarily close to one, while the false alarm probability remains negligible. When the test is performed on the whole graph, we derive bounds for the false alarm and detection probabilities, which show that the test is asymptotically optimal by properly setting the threshold test and using a large number of samples. Furthermore, we study how the convergence of the measures relies on the existence of links in the true graph.**

I. Introduction

Causality is a concept that expresses the joint behavior in time of a group of components in a system. In general, it denotes the effect of one component to itself and others in the system during a time period. Consider a network of nodes, each producing a signal in time. These processes can behave independently, or there might be an underlying connection, by nature, between them. Inferring this structure is of great interest in many applications. In [1], for instance, neurons are taken as components while the time series of produced spikes is used to derive the underlying structure. Dynamical models are also a well-known tool to understand functionals of expressed neurons [2]. Additionally, in social networks, there is an increasing interest to estimate influences among users [3], while further applications exist in biology [4], economics [5], and many other fields.

Granger [6] defined the notion of causality between two time series by using a linear autoregressive model and comparing the estimation errors for two scenarios: when history of the second node is accounted for and when it is not. With this definition, however, we can poorly estimate models which operate non-linearly. Directed information was first introduced to address the flow of information in a communication set-up, and suggested by Massey [7] as a measure of causality since it is not limited to linear models. There exist other methods which may qualify for different applications. Several of these definitions are compared in [1], where directed information is argued as a robust measure for causality. There are also symmetric measures like correlation or mutual information, but they can only represent a mutual relationship between nodes and not a directed one.

The underlying causal structure of a network of processes can be properly visualized by a directed graph. In particular, in a Directed Information Graph (DIG) –introduced simultaneously by Amblard and Michel [8] and Quinn *et al.* [1]– the existence of an edge is determined by the value of the directed information between two nodes considering the history of the rest of the network. There are different approaches to tackle the problem of detecting and estimating such type of graphs. Directed information can be estimated based on prior assumptions on the processes' structure, such as Markov properties, and empirically calculating probabilities [1], [3]. On the other hand, Jiao *et al.* [5] propose a universal estimator of directed information which is not restricted to any Markov assumption. Nonetheless, in the core of their technique, they consider a context tree weighting algorithm with different depths, which intuitively resembles a learning algorithm for estimating the order of a Markov structure. Other assumptions used in the study of DIGs, that constrain the structure of the underlying graph, are tree structures [9] or a limit on the nodes' degree [3].

The estimation performance on the detection of edges on a DIG is crucial since it allows to characterize, for instance, the optimum test for detection, or the minimum number of samples needed to reliably obtain the underlying model, i.e., the sample complexity. In [3], the authors derive a bound on the sample complexity using total variation when the directed information between two nodes is empirically estimated. Following that work, Kontoyiannis *et al.* [10] investigate the performance of a test for causality between two nodes, and they show that the convergence rate of the empirical directed information can be improved if it is calculated conditioned on the true relationship between the nodes. In other words, the underlying structure of the true model has an effect on the detection performance of the whole graph. Motivated by this result, in this paper, we study a hypothesis test over a complete graph (not just a link between two nodes) when the directed information is empirically estimated, and we provide interesting insights into the problem. Moreover, we show that for every existing edge in the true graph, the estimation converges with $\mathcal{O}(1/\sqrt{n})$, while if there is no edge in the true model, convergence is of the order of $\mathcal{O}(1/n)$.

The rest of the paper is organized as follows. In Section II, notations and definitions are introduced. In particular, the directed information is reviewed and the definition of an edge in a DIG is presented. The main results of our work are then shown in Section III. First, the performance of a hypothesis test for a single edge is studied, where we

This work was supported in part by the Knut and Alice Wallenberg Foundation.

The authors are with the ACCESS Linnaeus Centre, School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden (e-mails: sinmo@kth.se, germanb@kth.se, skoglund@kth.se).

analyze the asymptotic behavior of estimators based on the knowledge about the true edges. Then, we demonstrate how the detection of the whole graph relies on the test for each edge. Finally, in the last section, the paper is concluded.

II. Preliminaries

Assume a network with m nodes representing processes $\{\mathbf{X}_1, \dots, \mathbf{X}_m\}$. The observation of the l-th process in the discrete time interval t_1 to t_2 is described by the random variable $X_{l,t_1}^{t_2}$, which at each time takes values on the discrete alphabet $\mathcal{X}$. With a little abuse of notation Y_i and Y_1^n represent the observations of the process $\mathbf{Y}$ at instance i and in the interval 1 to n, respectively.

The metric used to describe the causality relationship of these processes is the directed information, as suggested previously, since it can describe more general structures without further assumptions (such as linearity). Directed information is mainly used in information theory to characterize channels with feedback and it is defined based on *causally* conditioned probabilities.

Definition 1. *The probability distribution of* Y_1^n causally *conditioned on* X_1^n *is defined as*

$$P_{Y_1^n \,||\, X_1^n} = \prod_{i=1}^{n} P_{Y_i|X_1^i, Y_1^{i-1}}.$$

The entropy rate of the process $\mathbf{Y}$ causally conditioned on $\mathbf{X}$ is then defined as:

$$\begin{aligned} H(\mathbf{Y} \,||\, \mathbf{X}) &\triangleq \lim_{n\to\infty} \frac{1}{n} H(Y_1^n \,||\, X_1^n) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y_1^{i-1}, X_1^i). \end{aligned}$$

Consequently, the directed information rate of $\mathbf{X}$ to $\mathbf{Y}$ causally conditioned on $\mathbf{Z}$ is expressed as below:

$$\begin{aligned} I(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z}) &\triangleq H(\mathbf{Y} \,||\, \mathbf{Z}) - H(\mathbf{Y} \,||\, \mathbf{X}, \mathbf{Z}) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(Y_i\,;\, X_1^i | Y_1^{i-1}, Z_1^i). \end{aligned} \quad (1)$$

Pairwise directed information does not unequivocally determine the one-step causal influence among nodes in a network. Instead, the history of the other remaining nodes should also be considered. Similarly as introduced in [1], [8], an edge from $\mathbf{X}_i$ to $\mathbf{X}_j$ exists in a directed information graph iff

$$I(\mathbf{X}_i \to \mathbf{X}_j \,||\, \mathbf{X}_{[m]\setminus\{i,j\}}) > 0, \quad (2)$$

where $[m] \triangleq \{1, 2, \dots, m\}$. Having observed the output of every process, the edges of the graph can be estimated which results in a weighted directed graph. However, when only the existence of a directed edge is investigated the performance of the detection can be improved. This is presented in Section III.

There exist several methods to estimate information theoretic values which most of them intrinsically deal with counting possible events to estimate distributions. One such method is the empirical distribution, which we define as follows.

Definition 2. *Let* $x_{[m]}^n = (x_{1,1}^n, x_{2,1}^n, \dots, x_{m,1}^n)$ *be a realization of the random variables* $X_{[m]}^n = (X_{1,1}^n, X_{2,1}^n, \dots, X_{m,1}^n)$. *The joint empirical distribution of* $k' \triangleq k+1$ *consecutive time instances of all nodes is then defined as:*

$$\hat{P}^{(n)}_{X_{[m]}^{k'}}(\mathbb{o}_{[m]}^{k'}) = \frac{1}{n-k} \sum_{t=1}^{n-k} \prod_{i=1}^{m} \mathbb{1}[x_{i,t}^{t+k} = \mathbb{o}_{i,1}^{k'}], \quad \forall \mathbb{o}_{i,1}^{k'} \in \mathcal{X}^{k'}. \quad (3)$$

The joint distribution of any subset of nodes is then a marginal distribution of (3).

By plugging in the empirical distribution we can derive estimators for information theoretic quantities such as the entropy H, where we use the notation $\hat{H}$ to distinguish the empirical estimator, i.e.,

$$\hat{H}(X^{k'}) = -\sum_{\mathbb{o}^{k'} \in \mathcal{X}^{k'}} \hat{P}^{(n)}_{X^{k'}}(\mathbb{o}^{k'}) \log\left(\hat{P}^{(n)}_{X^{k'}}(\mathbb{o}^{k'})\right). \quad (4)$$

A causal influence in the network implies that the past of a group of nodes affects the future of some other group or themselves. This motivates us to focus on a network of joint Markov processes in this paper, since it characterizes a state dependent operation for nodes, although we may put further assumptions to make calculations more tractable. For simplicity, we assume a three-node network (i.e., $m = 3$) and the processes to be $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ in the rest of the work, since the extension of the results for $m > 3$ is straightforward.

Assumption 1. $\{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$ *is a jointly stationary Markov process of order* k.

Let us define the $|\mathcal{X}|^{3k} \times |\mathcal{X}|^3$ transition probability matrix Q with elements

$$Q(X_{k+1}, Y_{k+1}, Z_{k+1} | X_1^k, Y_1^k, Z_1^k).$$

Then, the next assumption prevents further complexities in the steps of the proof of our main result.

Assumption 2. *All transition probabilities are positive, i.e.,* $Q > \mathbf{0}$.

This condition provides ergodicity for the joint Markov process and results in the joint empirical distribution asymptotically converging to the stationary distribution[1], i.e.,

$$\hat{P}^{(n)}_{X_1^{k+1}, Y_1^{k+1}, Z_1^{k+1}} \to P_{\bar{X}_1^{k+1}, \bar{Y}_1^{k+1}, \bar{Z}_1^{k+1}}$$

In general, the directed information rate $I(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z})$ cannot be expressed with the stationary random variables

[1]The stationary distribution is denoted either as $P_{\bar{X}_1^{k+1}, \bar{Y}_1^{k+1}, \bar{Z}_1^{k+1}}$ or as $\bar{P}_{X_1^{k+1}, Y_1^{k+1}, Z_1^{k+1}}$ in the sequel.

$\bar{X}_1^{k+1}, \bar{Y}_1^{k+1}$, and $\bar{Z}_1^{k+1}$, since a good estimator requires unlimited samples for perfect estimation. To see this,

$$\begin{aligned} I(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z}) &= I(\bar{Y}_{k+1}; \bar{X}_1^{k+1} | \bar{Y}_1^k, \bar{Z}_1^{k+1}) \\ &\quad - I(\bar{Y}_{k+1}; \bar{Y}_{-\infty}^0, \bar{Z}_{-\infty}^0 | \bar{Y}_1^k, \bar{Z}_1^{k+1}) \\ &\leq I(\bar{Y}_{k+1}; \bar{X}_1^{k+1} | \bar{Y}_1^k, \bar{Z}_1^{k+1}), \end{aligned} \tag{5}$$

where we use the Markov property in the first equation, and the inequality holds since the mutual information is non-negative. Thus, with a limited sampling interval, an upper bound would be derived. The next assumption makes (5) hold with equality.

Assumption 3. *For processes **Y** and **Z**, the Markov chain*

$$\bar{Y}_{k+1} - (\bar{Y}_1^k \bar{Z}_1^{k+1}) - (\bar{Y}_{-\infty}^0 \bar{Z}_{-\infty}^0)$$

holds.

Note that the above assumption should hold for every other two pairs of processes if we are interested in studying the whole graph and not only a single edge.

III. Hypothesis test for Directed Information Graphs

Consider a graph $\mathcal{G}$, where the edge from node i to node j is denoted by v_{ij}; we say that $v_{ij} = 1$ if the node i causally influences the node j, otherwise, $v_{ij} = 0$. A hypothesis test to identify the graph is performed on the adjacency matrix V, whose elements are the v_{ij}s, and the performance of the test is studied through its false alarm and detection probabilities

$$P_F = P(\hat{V} = V^* | V \neq V^*), \tag{6}$$

$$P_D = P(\hat{V} = V^* | V = V^*), \tag{7}$$

where $\hat{V}$ is the estimation of V (properly defined later), and V^* is the hypothesis model to test. In Theorem 1 below, both an upper bound on P_F and a lower bound on P_D are derived.

Theorem 1. *For a directed information graph with adjacency matrix V of size $m \times m$, if Assumptions* 1–3 *hold, the performance of the test for the hypothesis V^* is bounded as:*

$$P_F \leq 1 - P_G\left(\frac{r}{2}, I_{th}\right), \tag{8}$$

$$P_D \geq \max\left\{1 - N_0\left[1 - P_G\left(\frac{r}{2}, I_{th}\right)\right], 0\right\}, \tag{9}$$

using the plug-in estimation of n samples with $n \to \infty$. The function P_G is the regularized gamma function*, and $N_0 = m(m-1) - N_1$ with N_1 denoting the number of directed edges in the hypothesis graph, and $r = |\mathcal{X}|^{mk}(|\mathcal{X}|^m - 1)$. Finally, I_{th} is the threshold value used to decide the existence of an edge, and its order is $\mathcal{O}(1)$.*

The proof of Theorem 1 consists of two steps. First, the asymptotic behavior of the test for a single edge is derived in Section III-A. Afterwards, the hypothesis test over the whole graph is studied based on the tests for each single edge. It can be seen later on that, by Remark 1 and Corollary 1, the performance of testing the graph remains as good as a test for causality of a single edge.

TABLE I
Dimensions of Index Sets for $m = 3$.

Set	Dimension
Θ	$r = \lvert\mathcal{X}\rvert^{3k}(\lvert\mathcal{X}\rvert^3 - 1)$
Γ	$d = \lvert\mathcal{X}\rvert^{3k}(\lvert\mathcal{X}\rvert^2 - 1)$
Γ'	$d' = \lvert\mathcal{X}\rvert^{2k+1}(\lvert\mathcal{X}\rvert - 1)$

Remark 1. *Note that by increasing I_{th} while remaining of order $\mathcal{O}(1)$, $P_G\left(\frac{r}{2}, I_{th}\right)$ gets arbitrarily close to one, which results in the probability of detection converging to one as the probability of false alarm tends to zero.*

A. Asymptotic Behavior of a Single Edge Estimation

In general, every possible probability transition matrix Q can be parametrized with $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^r$ (see Table I). The vector θ is formed by concatenating the elements of Q in a row-wise manner excluding the last (linearly dependent) column. However, if the transition probability could be factorized due to a Markov property among its variables, the matrix might thus be addressed with a lower dimension parameter.

To see this, let us concentrate in our 3-node network $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$. If $v_{xy} = 0$, or equivalently $I(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z}) = 0$, then by Assumption 3, the transition probability can be factorized as follows,

$$\begin{aligned} &Q_{\phi_{xy}}(X_{k+1}, Y_{k+1}, Z_{k+1} | X_1^k, Y_1^k, Z_1^k) = \\ &Q_{\gamma_{xy}}(X_{k+1}, Z_{k+1} | X_1^k, Y_1^k, Z_1^k) Q_{\gamma'_{xy}}(Y_{k+1} | Y_1^k, Z_1^{k+1}). \end{aligned} \tag{10}$$

Here the transition matrix is parametrized by $\phi_{xy} \in \Phi_{xy}$ where ϕ_{xy} has two components: $\gamma_{xy} \in \Gamma$ and $\gamma'_{xy} \in \Gamma'$, and $\Phi_{xy} = \Gamma \times \Gamma'$. The dimensions of the sets are shown in Table I; note that $r > d + d'$. The vectors γ_{xy} and γ'_{xy} are also formed by concatenating the elements of their respective matrices as in the case of θ. More details are found in the proof of Theorem 2 in Appendix A.

Now consider the Neyman-Pearson criteria to test the hypothesis Φ_{xy} within Θ.

Definition 3. *The log-likelihood is defined as*

$$\begin{aligned} &L_n^\theta(X_1^n, Y_1^n, Z_1^n) \\ &\quad = \log\left(Q_\theta(X_{k+1}^n, Y_{k+1}^n, Z_{k+1}^n | X_1^k, Y_1^k, Z_1^k)\right) \\ &\quad = \log\left(\prod_{i=k+1}^{n} Q_\theta(X_i, Y_i, Z_i | X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1})\right). \end{aligned}$$

Let $\theta^\star$ and $\phi_{xy}^\star$ be the most likely choice of transition matrix with general and null hypothesis $v_{xy} = 0$, respectively,

i.e.,

$$\theta^\star = \arg\max_{\Theta} L_n^\theta(X_1^n, Y_1^n, Z_1^n),$$
$$\phi_{xy}^\star = \arg\max_{\Phi_{xy}} L_n^{\phi_{xy}}(X_1^n, Y_1^n, Z_1^n). \quad (11)$$

As a result, the test for causality boils down to check the difference between likelihoods, i.e., the log-likelihood ratio:

$$\Lambda_{xy,n} = L_n^{\theta^\star}(X_1^n, Y_1^n, Z_1^n) - L_n^{\phi_{xy}^\star}(X_1^n, Y_1^n, Z_1^n), \quad (12)$$

which is the Neyman-Pearson criteria for testing Φ_{xy} within Θ. Then, in the following theorem, $\Lambda_{xy,n}$ is shown to converge to a χ^2 distribution of finite degree. The proof follows from standard results in [11, Th. 6.1].

Theorem 2. *Consider a network with three nodes $\{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$ and arbitrarily choose two nodes $\boldsymbol{X}$ and $\boldsymbol{Y}$. Suppose Assumptions 1–3 hold, then*

$$2\Lambda_{xy,n} \xrightarrow{\mathcal{L}} \chi^2_{r-d-d'},$$

if $v_{xy} = 0$ as $n \to \infty$.

Proof: The conditions of the theorem imply that the true underlying structure for the transition matrix is from Φ_{xy} which is required as in [11, Th. 6.1]. The rest of the proof follows similar steps as in [10]. See Appendix A for further details. ∎

Remark 2. *Note that the asymptotic result from Theorem 2 depends only on the dimensions of the sets and not in the particular pair of nodes involved. Furthermore, the result also holds for a network with more than three nodes by properly defining the dimensions of the sets.*

Remark 3. *Knowledge about the absence of edges other than v_{xy} in the network results in $\Lambda_{xy,n}$ converging to a χ^2 distribution of higher degree since* (10) *could be further factorized. To see this, assume $v_{xy} = 0$ and consider that a knowledge S about the links (for example, the whole adjacency matrix V) was given. Then, let the transition probability be factorized as much as possible, so it can be parametrized by Φ'_{xy} which has lower or equal dimension than Φ_{xy}. Take*

$$\Lambda'_{xy,n} = L_n^{\theta^\star}(X_1^n, Y_1^n, Z_1^n) - L_n^{\phi'^\star_{xy}}(X_1^n, Y_1^n, Z_1^n),$$

where

$$\phi'^\star_{xy} = \arg\max_{\Phi'_{xy}} L_n^{\phi'_{xy}}(X_1^n, Y_1^n, Z_1^n).$$

Intuitively, by following similar steps as in the proof of Theorem 2, we obtain that $\Lambda'_{xy,n}$ behaves as a χ^2_q random variable, where $r > q > r - d - d'$. Since the cumulative distribution function of the χ^2_q is a decreasing function with respect to the degree q then,

$$\begin{aligned} P_G\left(\frac{r}{2}, a\right) &\le P_G\left(\frac{q}{2}, a\right) \\ &= P(\Lambda'_{xy,n} < a | S, v_{xy} = 0) \\ &\le P(\Lambda_{xy,n} < a | v_{xy} = 0) \\ &= P_G\left(\frac{r-d-d'}{2}, a\right), \end{aligned} \quad (13)$$

for sufficiently large n and any $a > 0$. The lower bound in (13) *allows us to ignore the knowledge about other nodes.*

Consider now the estimation of the directed information defined as plugging in the empirical distribution (instead of the true distribution) into $I(Y_{k+1}; X_1^{k+1} | Y_1^k, Z_1^{k+1})$, i.e.,

$$\hat{I}_n^{(k)}(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z}) \triangleq I(\hat{Y}_{k+1}; \hat{X}_1^{k+1} | \hat{Y}_1^k, \hat{Z}_1^{k+1}).$$

Then, the following lemma states that $\hat{I}_n^{(k)}(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z})$, is proportional to $\Lambda_{xy,n}$ with an $\mathcal{O}(n)$ factor.

Lemma 1. *$\Lambda_{xy,n} = (n-k)\hat{I}_n^{(k)}(\boldsymbol{X} \to \boldsymbol{Y} \,||\, \boldsymbol{Z})$, which is the plug-in estimator of the directed information.*

Proof: The proof follows from standard definitions and noting that the KL-divergence is positive and minimized by zero. See Appendix B for the complete proof. ∎

Now, let us define the decision rule for checking the existence of an edge in the graph as follows:

$$\hat{v}_{i,j} \triangleq \begin{cases} 1 & \text{if } (n-k)\hat{I}_n^{(k)}(\mathbf{X}_i \to \mathbf{X}_j \,||\, \mathbf{X}_{[m]\setminus\{i,j\}}) \ge I_{th} \\ 0 & \text{o.w.}, \end{cases}$$

where I_{th} is of order $\mathcal{O}(1)$. Then for any knowledge S about states of edges in the true graph, as long as $v_{xy} = 0$ we have:

$$\begin{aligned} &P(\hat{v}_{xy} = 1 | S, v_{xy} = 0) \\ &\quad = P((n-k)\hat{I}_n^{(k)}(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z}) > I_{th} | S, v_{xy} = 0) \\ &\quad \le 1 - P_G\left(\frac{r}{2}, I_{th}\right), \end{aligned} \quad (14)$$

where the inequality is due to Remark 3.

From Theorem 2 and Lemma 1, it is inferred that when in the true adjacency matrix $v_{xy} = 0$, then the empirical estimation of the directed information converges to zero with a χ^2 distribution at a rate $\mathcal{O}(1/n)$. The asymptotic behavior of $\hat{I}_n^{(k)}(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z})$ is different if the edge is present, i.e., $v_{xy} = 1$, which is addressed in the following theorem.

Theorem 3. *Consider a network with three nodes $\{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$ and arbitrarily choose two nodes $\boldsymbol{X}$ and $\boldsymbol{Y}$. Suppose Assumptions 1–3 hold and let $\bar{I}_{xy} \triangleq I(\bar{Y}_{k+1}; \bar{X}_1^{k+1} | \bar{Y}_1^k, \bar{Z}_1^{k+1})$, then,*

$$\sqrt{n-k}\left[\hat{I}_n^{(k)}(\boldsymbol{X} \to \boldsymbol{Y} \,||\, \boldsymbol{Z}) - \bar{I}_{xy}\right] \to \mathcal{N}(0, \sigma^2), \quad (15)$$

with a finite σ^2 as $n \to \infty$, if $v_{xy} = 1$.

Proof: The empirical distribution can be decomposed in two parts, where the first one vanishes at a rate faster than $\mathcal{O}(1/\sqrt{n})$ and the second part converges at a rate $\mathcal{O}(1/\sqrt{n})$.

The condition $v_{xy} = 1$ keeps the second part positive so it determines the asymptotic convergence of $\hat{I}_n^{(k)}(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z})$. Refer to Appendix C for further details. ∎

Remark 4. *Knowledge about the state of other edges in the true graph model does not affect the asymptotic behavior presented in Theorem 3, given that the condition $v_{xy} = 1$ makes the convergence of the estimator independent of all other nodes. This can be seen by following the steps of the proof, where we only use the fact that if the true edge exists then $\bar{I}_{xy} > 0$ and* (10) *does not hold.*

We can use Remark 4 to conclude that:

$$\begin{aligned} &P(\hat{v}_{xy} = 0 | S, v_{xy} = 1) \\ &\quad = P((n-k)\hat{I}_n^{(k)}(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z}) < I_{th} | S, v_{xy} = 1) \\ &\quad = P((n-k)\hat{I}_n^{(k)}(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z}) < I_{th} | v_{xy} = 1) \\ &\quad = 1 - Q\left(\frac{I_{th} - (n-k)\bar{I}_{xy}}{\sqrt{n-k}\,\sigma} \right), \end{aligned} \tag{16}$$

for sufficiently large n, where $Q(\cdot)$ is the Q-function, and where the last equality is due to Theorem 3. Note that if $v_{xy} = 1$ then $\bar{I}_{xy} > 0$.

B. Hypothesis Test over an Entire Graph

The performance of testing a hypothesis V^* for a graph is studied by means of the false alarm and detection probabilities defined in (6) and (7), respectively. The results may be considered as an extension of the hypothesis test over a single edge in the graph.

First, let the false alarm probability be upper-bounded as

$$P_F = P(\hat{V} = V^* | V \neq V^*) \leq \min_{i,j} P(\hat{v}_{ij} = v_{ij}^* | V \neq V^*).$$

If $V \neq V^*$, there exist nodes τ and ρ such that $v_{\tau\rho} \neq v_{\tau\rho}^*$. Hence,

$$\begin{aligned} P_F &\leq \min_{i,j} P(\hat{v}_{ij} = v_{ij}^* | V \neq V^*) && (17) \\ &\leq P(\hat{v}_{\tau\rho} = v_{\tau\rho}^* | V \neq V^*) && (18) \\ &= P(\hat{v}_{\tau\rho} = v_{\tau\rho}^* | V \neq V^*, v_{\tau\rho} \neq v_{\tau\rho}^*) \\ &= \begin{cases} P(\hat{v}_{\tau\rho} = 0 | V \neq V^*, v_{\tau\rho} = 1) \\ P(\hat{v}_{\tau\rho} = 1 | V \neq V^*, v_{\tau\rho} = 0) \end{cases} \\ &\leq \begin{cases} 1 - Q\left(\frac{I_{th} - (n-k)\bar{I}_{\tau\rho}}{\sqrt{n-k}\,\sigma} \right) & \text{if } v_{\tau\rho} = 1 \\ 1 - P_G\left(\frac{r}{2}, I_{th}\right) & \text{if } v_{\tau\rho} = 0 \end{cases} && (19) \end{aligned}$$

where the last inequality is due to (14) and (16).

On the other hand, the complement of the detection probability may be upper-bounded using the union bound:

$$\begin{aligned} 1 - P_D &= P(\hat{V} \neq V^* | V = V^*) \leq \sum_{i,j} P(\hat{v}_{ij} \neq v_{ij}^* | V = V^*) \\ &= \sum_{\substack{i,j \\ v_{ij}=1}} P(\hat{v}_{ij} \neq v_{ij}^* | V = V^*) + \sum_{\substack{i,j \\ v_{ij}=0}} P(\hat{v}_{ij} \neq v_{ij}^* | V = V^*) \\ &= \sum_{\substack{i,j \\ v_{ij}=1}} P(\hat{v}_{ij} = 0 | V = V^*, v_{ij} = 1) \\ &\quad + \sum_{\substack{i,j \\ v_{ij}=0}} P(\hat{v}_{ij} = 1 | V = V^*, v_{ij} = 0) \\ &\leq N_1 \left(1 - Q\left(\frac{I_{th} - (n-k)\bar{I}}{\sqrt{n-k}\,\sigma} \right) \right) + N_0 \left(1 - P_G\left(\frac{r}{2}, I_{th} \right) \right), \end{aligned} \tag{20}$$

where N_0 and N_1 are the number of off-diagonal 0s and 1s in the true matrix V, i.e., $N_0 + N_1 = m(m-1)$, and $\bar{I} \triangleq \min_{\substack{i,j \\ \text{s.t. } v_{ij}=1}} \bar{I}_{ij}$. The last inequality holds due to (14) and (16).

Since $\bar{I}_{ij} = I(\bar{X}_{j,k+1}; \bar{X}_{i,1}^{k+1} | \bar{X}_{j,1}^{k}, \bar{X}_{[m]\setminus\{i,j\},1}^{k+1}) > 0$ and it is of order $\mathcal{O}(1)$, then as $n \to \infty$ and noting that

$$\lim_{a \to \infty} 1 - Q(-a) = 0,$$

we have that,

$$P_F \leq 1 - P_G\left(\frac{r}{2}, I_{th} \right) \tag{21}$$

$$1 - P_D \leq N_0 \left[1 - P_G\left(\frac{r}{2}, I_{th} \right) \right]. \tag{22}$$

This concludes the proof of Theorem 1.

Corollary 1. *In the special case the hypothesis test is performed on a single edge, for the false alarm probability,* (17) *and* (18) *become equal and we have*

$$P_F' \triangleq P(\hat{v}_{xy} = 1 | v_{xy} = 0) = 1 - P_G\left(\frac{r - d - d'}{2}, I_{th} \right),$$

and for the detection probability,

$$P_D' \triangleq P(\hat{v}_{xy} = 1 | v_{xy} = 1) = 1,$$

as $n \to \infty$, which is in the same lines as the argument in [10, Sec. III-C] for $m = 2$.

C. Numerical Results

The bounds derived in Theorem 1 state that the detection probability can be desirably close to one while the false alarm probability can remain near zero with a proper threshold test. In Fig. 1, these bounds are depicted with respect to different values of I_{th} for a network with $m = 5$ nodes. The joint process is assumed to be a Markov process of order $k = 2$, and the random variables take values on a binary alphabet ($|\mathcal{X}| = 2$).

It can be observed in the figure that, for fixed I_{th}, P_D improves as N_0 decreases, i.e., when the graph becomes sparser. Furthermore, by a proper choice of I_{th}, we can reach to optimal performance of the hypothesis test, i.e., $P_D = 1$ and $P_F = 0$.

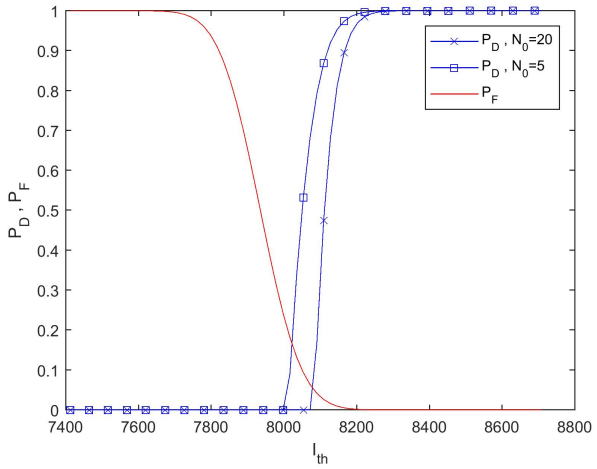

Fig. 1. Lower bound for detection probability P_D and upper bound for P_F, derived by varying the threshold of the test I_{th}, and with $k = 2$, $m = 5$ and binary alphabet. Since $P_F \geq 0$ and $P_D \leq 1$ by increasing I_{th}, $P_F \to 0$ and $P_D \to 1$.

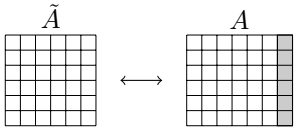

Fig. 2. The matrix $\tilde{A}$ is formed by removing the last column of A.

IV. Summary and Remarks

In this paper, we investigated the performance of a hypothesis test for detecting the underlying directed graph of a network of stochastic processes, which represents the causal relationship among nodes, by empirically calculating the directed information as the measure. We showed that the convergence rate of the directed information estimator relies on the existence or not of the link in the real structure. We further showed that with a proper adjustment of the threshold test for single edges, the overall hypothesis test is asymptotically optimal.

This work may be expanded by considering a detailed analysis on the sample complexity of the hypothesis test. Moreover, we assumed in this work that the estimator has access to samples from the whole network while in practice this might not be the case (see e.g. [12]).

Appendix A
Proof of Theorem 2

For any right stochastic matrix A of dimensions $n_a \times m_a$, let the matrix $\tilde{A}$ denote the first $m_a - 1$ linearly independent columns of A, as depicted in Fig. 2.

Without loss of generality, consider $\mathcal{X}$ to be the set of integers $\{1, 2, \ldots, |\mathcal{X}|\}$ which simplifies the indexing of elements in the alphabet. Let $u_{x,1}^k$ denote $(u_{x,1}, u_{x,2}, \ldots, u_{x,k}) \in \mathcal{X}^k$, and $u'_x, u'_y, u'_z \in \mathcal{X}$ excluding $(u'_x, u'_y, u'_z) = (|\mathcal{X}|, |\mathcal{X}|, |\mathcal{X}|)$. Next define the $3k + 3$ vector

$$\vec{u} \triangleq (u_{x,1}^k, u_{y,1}^k, u_{z,1}^k, u'_x, u'_y, u'_z) \tag{23}$$

which is associated with an element of $\tilde{Q}$ (the sub-matrix of the transition probability matrix Q).

Every $\vec{u}$ can be addressed via the pair $(l_{\vec{u}}, g_{\vec{u}})$ where $l_{\vec{u}} \in [1 : |\mathcal{X}|^{3k}]$ and $g_{\vec{u}} \in [1 : |\mathcal{X}|^3 - 1]$ indicate the row and column of that element, respectively. Also, let $f_{\vec{u}} \triangleq (l_{\vec{u}} - 1)(|\mathcal{X}|^3 - 1) + g_{\vec{u}}$, which denotes the index of that element when vectorizing $\tilde{Q}$. Any possible transition matrix can then be indexed with a vector

$$\theta = (\theta_1, \theta_2, \ldots, \theta_r) = (\theta_{f_{\vec{u}}}) \in \Theta$$

as Q_θ, where Θ has dimension r (see Table I) and θ is constructed by concatenation of rows in $\tilde{Q}_\theta$.

Suppose now that $v_{xy} = 0$ or equivalently, by definition (2), $I(\mathbf{X} \to \mathbf{Y} \,||\, \mathbf{Z}) = 0$. Then

$$\begin{aligned} &Q(X_{k+1}, Y_{k+1}, Z_{k+1} | X_1^k, Y_1^k, Z_1^k) \\ &= P(X_{k+1}, Z_{k+1} | X_1^k, Y_1^k, Z_1^k) P(Y_{k+1} | Y_1^k, Z_1^{k+1}). \end{aligned} \tag{24}$$

Thus, the transition matrix Q is determined by the elements of two other matrices T_1 and T_2 given by (24). Define the vectors

$$\begin{aligned} \vec{w} &\triangleq (i_{x,1}^k, i_{y,1}^k, i_{z,1}^k, i'_x, i'_z), \\ \vec{w}' &\triangleq (i_{y,1}^k, (i_{z,1}^k, i'_z), i'_y), \end{aligned}$$

which are associated with an element in $\tilde{T}_1$ and $\tilde{T}_2$, such that $(i'_x, i'_z) \neq (|\mathcal{X}|, |\mathcal{X}|)$ in $\vec{w}$ and $i'_y \neq |\mathcal{X}|$ in $\vec{w}'$. Then

$$\begin{aligned} f_{\vec{w}} &\triangleq (l_{\vec{w}} - 1)(|\mathcal{X}|^2 - 1) + g_{\vec{w}}, \\ f_{\vec{w}'} &\triangleq (l_{\vec{w}'} - 1)(|\mathcal{X}| - 1) + g_{\vec{w}'}, \end{aligned}$$

where the pairs of row and column indices for each element in $\tilde{T}_1$ and $\tilde{T}_2$ are then $(l_{\vec{w}}, g_{\vec{w}})$ and $(l_{\vec{w}'}, g_{\vec{w}'})$, respectively.

A matrix Q such as the one in (24) can be parametrized by a vector $\phi_{xy} \in \Phi_{xy}$, where $\Phi_{xy} = \Gamma \times \Gamma'$ has dimension $d \cdot d'$ (see Table I). Then

$$\begin{aligned} &Q_{\phi_{xy}} = \\ &\quad Q_{\gamma_{xy}}(X_{k+1}, Z_{k+1} | X_1^k, Y_1^k, Z_1^k) Q_{\gamma'_{xy}}(Y_{k+1} | Y_1^k, Z_1^{k+1}), \end{aligned}$$

where

$$\gamma_{xy} = (\gamma_{f_{\vec{w}}}) \in \Gamma \quad \text{and} \quad \gamma'_{xy} = (\gamma'_{f_{\vec{w}'}}) \in \Gamma'$$

determine ϕ_{xy}, are vectors of length d and d', and are constructed by concatenating the rows of $\tilde{Q}_{\gamma_{xy}}$ and $\tilde{Q}_{\gamma'_{xy}}$, respectively. There exists then the mapping $h : \Phi_{xy} \to \Theta$ such that component-wise:

$$(h(\phi_{xy}))_{f_{\vec{u}}} = \gamma_{f_{\vec{w}}} \cdot \gamma'_{f_{\vec{w}'}}. \tag{25}$$

Consider the matrix $K(\phi_{xy})$ of size $(r+1) \times (d+d')$ such that for every element:

$$(K(\phi_{xy}))_{f_{\vec{u}}, f} = \begin{cases} \frac{\partial Q_{h(\phi_{xy})}(u'_x, u'_y, u'_z | u_{x,1}^k, u_{y,1}^k, u_{z,1}^k)}{\partial \gamma_f} & f \leq d \\ \frac{\partial Q_{h(\phi_{xy})}(u'_x, u'_y, u'_z | u_{x,1}^k, u_{y,1}^k, u_{z,1}^k)}{\partial \gamma'_{f-d}} & f > d. \end{cases} \tag{26}$$

In other words, every row of the matrix $K(\phi_{xy})$ is a derivative of an element of $Q_{h(\phi_{xy})}$ with respect to all elements of γ_{xy} followed by the derivatives with respect to γ'_{xy}.

According to [11, Th. 6.1], and by the definition of $\theta^\star$ and $\phi^\star_{xy}$ in (11),

$$2\left\{L_n^{\theta^\star}(X_1^n,Y_1^n,Z_1^n)-L_n^{\phi^\star_{xy}}(X_1^n,Y_1^n,Z_1^n)\right\}\xrightarrow{\mathcal{L}}\chi^2_{r-d-d'},$$

if $Q_{h(\phi_{xy})}$ has continuous third order partial derivatives and $K(\phi_{xy})$ is of rank $d+d'$. The first condition holds according to the definition of h in (25). To verify the second condition we can observe that there exist four types of rows in $K(\phi_{xy})$:

- Type 1: Take the rows $\vec{u}_1=(i_1^k,j_1^k,l_1^k,i',j',l')$ in (26) such that $(i',l')\neq(|\mathcal{X}|,|\mathcal{X}|)$ jointly and $j'\neq|\mathcal{X}|$. This means that in the $(f_{\vec{u}_1})$-th row of K, the derivatives are taken from

$$Q_{h(\phi_{xy})}(i',j',l'|i_1^k,j_1^k,l_1^k)=\gamma_{f_{\vec{w}_1}}\cdot\gamma'_{f_{\vec{w}'_1}},$$

  where $\vec{w}_1=(i_1^k,j_1^k,l_1^k,i',l')$ and $\vec{w}'_1=(j_1^k,(l_1^k,l'),j')$. So all elements in such rows are zero except in the columns $f_{\vec{w}_1}$ and $(d+f_{\vec{w}'_1})$, which take the values $\gamma'_{f_{\vec{w}'_1}}$ and $\gamma_{f_{\vec{w}_1}}$, respectively.
- Type 2: Now consider the rows $\vec{u}_2=(i_1^k,j_1^k,l_1^k,i',j',l')$ such that $(i',l')=(|\mathcal{X}|,|\mathcal{X}|)$ and $j'\neq|\mathcal{X}|$. This means that in the $(f_{\vec{u}_2})$-th row of K, the derivatives are taken from

$$Q_{h(\phi_{xy})}(i',j',l'|i_1^k,j_1^k,l_1^k)=\left(1-\sum_{(a,b)\neq(|\mathcal{X}|,|\mathcal{X}|)}\gamma_{f_{\vec{w}_2(a,b)}}\right)\gamma'_{f_{\vec{w}'_2}},$$

  where we define $\vec{w}_2(a,b)=(i_1^k,j_1^k,l_1^k,a,b)$ and $\vec{w}'_2=(j_1^k,(l_1^k,l'),j')$. So all elements in such rows are zero except in the $(|\mathcal{X}|^2-1)$ columns (among the first d columns) from $f_{\vec{w}_2(1,1)}$ to $f_{\vec{w}_2(|\mathcal{X}|,|\mathcal{X}|-1)}$ which are equal to $-\gamma'_{f_{\vec{w}'_2}}$, and the column $(d+f_{\vec{w}'_2})$ which is equal to

$$1-\sum_{(a,b)\neq(|\mathcal{X}|,|\mathcal{X}|)}\gamma_{f_{\vec{w}_2(a,b)}}.$$

- Type 3: Consider the rows $\vec{u}_3=(i_1^k,j_1^k,l_1^k,i',j',l')$ such that $(i',l')\neq(|\mathcal{X}|,|\mathcal{X}|)$ and $j'=|\mathcal{X}|$. Also let $\vec{w}_3=(i_1^k,j_1^k,l_1^k,i',l')$ and $\vec{w}'_3(a)=(j_1^k,(l_1^k,l'),a)$. Then, all elements of such rows are zero except in the column $f_{\vec{w}_3}$ which takes the value

$$1-\sum_{a\neq|\mathcal{X}|}\gamma'_{f_{\vec{w}'_3(a)}},$$

  and the $(|\mathcal{X}|-1)$ columns (among the last d' columns) from $d+f_{\vec{w}'_3(1)}$ to $d+f_{\vec{w}'_3(|\mathcal{X}|-1)}$ that are equal $-\gamma_{f_{\vec{w}_3}}$.
- Type 4: Lastly, consider rows $\vec{u}_4=(i_1^k,j_1^k,l_1^k,i',j',l')$ such that $(i',l')=(|\mathcal{X}|,|\mathcal{X}|)$ and $j'=|\mathcal{X}|$. Assume vectors $\vec{w}_4(a,b)=(i_1^k,j_1^k,l_1^k,a,b)$ and $\vec{w}'_4(a)=(j_1^k,(l_1^k,l'),a)$. Then , the only non-zero elements belong to the $(|\mathcal{X}|^2-1)$ columns from $f_{\vec{w}_4(1,1)}$ to $f_{\vec{w}_4(|\mathcal{X}|,|\mathcal{X}|)}$ (among the first d columns) which are equal to

$$-\left(1-\sum_{a\neq|\mathcal{X}|}\gamma'_{f_{\vec{w}'_4(a)}}\right),$$

  and the $(|\mathcal{X}|-1)$ columns from $d+f_{\vec{w}'_4(1)}$ to $d+f_{\vec{w}'_4(|\mathcal{X}|-1)}$ (among the last d' columns) which are equal to

$$-\left(1-\sum_{(a,b)\neq(|\mathcal{X}|,|\mathcal{X}|)}\gamma_{f_{\vec{w}_4(a,b)}}\right).$$

We show now that if a linear combination of all columns equals the vector zero, then all coefficients should be zero as well. Let c_f be the f-th column of $K(\phi_{xy})$ then if

$$\sum_{f=1}^{d+d'}\alpha_f c_f=\vec{0},\tag{27}$$

then, $\alpha_f=0,\forall f$. To see this, consider the Type 1 row with $i_1^k=j_1^k=l_1^k=1^k$ and $i'=l'=1$. Since it only has two non-zero elements, we have that

$$\forall j'\in[1:|\mathcal{X}|-1]:\quad \alpha_1\gamma'_{j'}+\alpha_{j'}\gamma_1=0.\tag{28}$$

Then, take the Type 3 row with $i_1^k=j_1^k=l_1^k=1^k$ and $i'=l'=j'=1$, where we have that

$$\alpha_1\left(\sum_{a\neq|\mathcal{X}|}1-\gamma'_a\right)-\sum_{b\neq|\mathcal{X}|}\alpha_b\gamma_1=0.\tag{29}$$

From (28) and noting that we have assumed $Q>\mathbf{0}$, if $\alpha_1>0$ then $\alpha_{j'}<0$ for all $j'\in[1:|\mathcal{X}|-1]$. Hence, the left-hand side of (29) is strictly positive and not zero. An analogous result is found assuming $\alpha_1<0$. By contradiction, we conclude that $\alpha_1=0$, and from (28),

$$\forall j'\in[1:|\mathcal{X}|-1]:\quad \alpha_{j'}=0.$$

By varying (i',l') and for all combinations of (i_1^k,j_1^k,l_1^k) we derive that all α_fs are zero, and as a result, $K(\phi_{xy})$ has $d+d'$ linearly independent columns which meets the second condition. The proof of Theorem 2 is thus complete. ■

Appendix B
Proof of Lemma 1

The proof follows similar steps as the one in [10, Prop. 9]. Using the definition of log-likelihood,

$$\begin{aligned}
&L_n^{\theta^\star}(X_1^n,Y_1^n,Z_1^n)\\
&=\max_{\theta\in\Theta}\sum_{i=k+1}^{n}\log(Q_\theta(X_i,Y_i,Z_i|X_{i-k}^{i-1},Y_{i-k}^{i-1},Z_{i-k}^{i-1}))\\
&=\max_{\theta\in\Theta}\sum_{x_1^{k+1}y_1^{k+1}z_1^{k+1}}(n-k)\hat{P}_{X_1^{k+1}Y_1^{k+1}Z_1^{k+1}}(x_1^{k+1}y_1^{k+1}z_1^{k+1})\\
&\qquad\times\log(Q_\theta(x_{k+1}y_{k+1}z_{k+1}|x_1^ky_1^kz_1^k))\\
&=-(n-k)\Bigg[\min_{\theta\in\Theta}\Big\{D\big(\hat{P}_{X_1^{k+1}Y_1^{k+1}Z_1^{k+1}}\;||\;Q_\theta\otimes\hat{P}_{X_1^kY_1^kZ_1^k}\big)\Big\}\\
&\quad+\sum_{x_1^{k+1}y_1^{k+1}z_1^{k+1}}\hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})\log\left(\frac{\hat{P}(x_1^ky_1^kz_1^k)}{\hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})}\right)\Bigg],
\end{aligned}\tag{30}$$

where

$$\begin{aligned}
&(Q_\theta\otimes\hat{P}_{X_1^kY_1^kZ_1^k})(x_1^{k+1}y_1^{k+1}z_1^{k+1})\triangleq\\
&\qquad\hat{P}_{X_1^kY_1^kZ_1^k}(x_1^ky_1^kz_1^k)Q_\theta(x_{k+1}y_{k+1}z_{k+1}|x_1^ky_1^kz_1^k).
\end{aligned}$$

Since the KL-divergence is minimized by zero, then

$$L_n^{\theta^\star}(X_1^n, Y_1^n, Z_1^n) = (n-k)\cdot \\ [\hat{H}(X_1^k, Y_1^k, Z_1^k) - \hat{H}(X_1^{k+1}, Y_1^{k+1}, Z_1^{k+1})]. \quad (31)$$

On the other hand, for the second log-likelihood, we have:

$$\begin{aligned} &L_n^{\phi^\star}(X_1^n, Y_1^n, Z_1^n) \\ &= \max_{\phi\in\Phi} \sum_{i=k+1}^{n} \log(Q_\phi(X_i, Y_i, Z_i | X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1})) \\ &= \underbrace{\max_{\phi^{xz}} \sum_{i=k+1}^{n} \log(Q_{\phi^{xz}}(X_i, Z_i | X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1}))}_{A_1} \\ &\quad + \underbrace{\max_{\phi^{y}} \sum_{i=k+1}^{n} \log(Q_{\phi^{y}}(Y_i | Y_{i-k}^{i-1}, Z_{i-k}^{i}))}_{A_2}. \end{aligned}$$

With a similar approach as in (30), we can expand A_1 and A_2 as it is shown in (32) at the bottom of the page. As a result,

$$\begin{aligned} &L_n^{\phi^\star}(X_1^n, Y_1^n, Z_1^n) \\ &= (n-k)\Big[\hat{H}(X_1^k, Y_1^k, Z_1^k) - \hat{H}(X_1^{k+1}, Y_1^k, Z_1^{k+1}) \\ &\quad + \hat{H}(Y_1^k, Z_1^{k+1}) - \hat{H}(Y_1^{k+1}, Z_1^{k+1})\Big]. \end{aligned} \quad (33)$$

Finally, combining (31) and (33), we obtain

$$\begin{aligned} \Lambda_{xy,n} &= L_n^{\theta^\star}(X_1^n, Y_1^n, Z_1^n) - L_n^{\phi_{xy}^\star}(X_1^n, Y_1^n, Z_1^n) \\ &= (n-k)[\hat{H}(Y_{k+1}|Y_1^k, Z_1^{k+1}) - \hat{H}(Y_{k+1}|X_1^{k+1}, Y_1^k, Z_1^{k+1})] \\ &= (n-k)\hat{I}_n^{(k)}(\mathbf{X}\to\mathbf{Y}\,||\,\mathbf{Z}), \end{aligned} \quad (34)$$

which concludes the proof of Lemma 1. ∎

Appendix C
Proof of Theorem 3

We begin by expanding the expression $\hat{I}_n^{(k)}(\mathbf{X}\to\mathbf{Y}\,||\,\mathbf{Z})$ using the definition of the empirical distribution in (3) and we obtain (35), found at the bottom of the next page. We then proceed to analyze the asymptotic behavior of the estimator.

The first four terms in (35), i.e., the KL-divergence terms, decay faster than $\mathcal{O}(1/\sqrt{n})$. This is shown later in the proof. On the other hand, since $v_{xy} = 1$, $I(\bar{Y}_{k+1}; \bar{X}_1^{k+1}|\bar{Y}_1^k, \bar{Z}_1^{k+1}) > 0$ due to (5) and Assumption 3, and thus, the last term in (35) is non-zero and dominates the convergence of the estimator, as we see next. Here, one observes that conditioning on $v_{xy} = 1$ is sufficient to analyze the convergence of $\hat{I}_n^{(k)}(\mathbf{X}\to\mathbf{Y}\,||\,\mathbf{Z})$ and further knowledge about other edges is irrelevant (see Remark 4). We then conclude that,

$$\lim_{n\to\infty} \sqrt{n-k}\,\hat{I}_n^{(k)}(\mathbf{X}\to\mathbf{Y}\,||\,\mathbf{Z}) = \lim_{n\to\infty} \frac{1}{\sqrt{n-k}} \sum_{i=k+1}^{n} S_i,$$

where

$$\begin{aligned} S_i \triangleq\ &\log P_{\bar{Y}_{k+1}\bar{X}_1^{k+1}|\bar{Y}_1^k\bar{Z}_1^{k+1}}(y_i x_{i-k}^i | y_{i-k}^{i-1} z_{i-k}^i) \\ &- \log P_{\bar{Y}_{k+1}|\bar{Y}_1^k\bar{Z}_1^{k+1}}(y_i | y_{i-k}^{i-1} z_{i-k}^i) \\ &- \log P_{\bar{X}_1^{k+1}|\bar{Y}_1^k\bar{Z}_1^{k+1}}(x_{i-k}^i | y_{i-k}^{i-1} z_{i-k}^i). \end{aligned}$$

We note that S_i is a functional of the chain $\{(X_{i-k}^i, Y_{i-k}^i, Z_{i-k}^i)\}$ and its mean is $\mathbb{E}[S] = I(\bar{Y}_{k+1}; \bar{X}_1^{k+1}|\bar{Y}_1^k, \bar{Z}_1^{k+1})$. The chain is ergodic and we can thus apply the central limit theorem [13, Sec. I.16] to the partial sums to obtain

$$\sqrt{n-k}\left(\frac{1}{n-k}\sum_{i=k+1}^{n-1} S_i - \mathbb{E}[S]\right) \to \mathcal{N}(0,\sigma^2), \quad (36)$$

where σ^2 is bounded.

Now, to complete the proof, it only remains to show that the KL-divergence terms in (35) multiplied by a $\sqrt{n-k}$ factor converge to zero as $n\to\infty$. We present the proof for one term and the others follow a similar approach. We first recall the Taylor expansion with Lagrange remainder form,

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(x^*)(x-a)^2}{2!},$$

for some $x^* \in (a,x)$. Then, let us define $\rho \triangleq \frac{\bar{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})}{\hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})}$, so we can expand the first KL-divergence term as:

$$\begin{aligned} &D(\hat{P}_{X_1^{k+1}Y_1^{k+1}Z_1^{k+1}}\,||\,\bar{P}_{X_1^{k+1}Y_1^{k+1}Z_1^{k+1}}) \\ &= -\sum\nolimits_{x_1^{k+1}y_1^{k+1}z_1^{k+1}} \hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})\log\rho \\ &= -\sum \hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})\Big[(\rho-1) - \frac{(\rho-1)^2}{2!\tau^2}\Big] \\ &= \sum \hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})(\rho-1)^2\frac{1}{2\tau^2} \quad (37) \\ &= \sum \Big(\hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1}) - \bar{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})\Big)^2 C, \quad (38) \end{aligned}$$

for some $\tau\in(1,\rho)$, where

$$C \triangleq \frac{1}{2\hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})\tau^2},$$

$$A_1 = -(n-k)\left[\min_{\phi^{xz}}\left\{D\big(\hat{P}_{X_1^{k+1},Y_1^k,Z_1^{k+1}}\,||\,Q_{\phi^{xz}}\otimes\hat{P}_{X_1^k,Y_1^k,Z_1^k}\big)\right\} + \sum_{x_1^{k+1},y_1^k,z_1^{k+1}} \hat{P}(x_1^{k+1},y_1^k,z_1^{k+1})\log\left(\frac{\hat{P}(x_1^k,y_1^k,z_1^k)}{\hat{P}(x_1^{k+1},y_1^k,z_1^{k+1})}\right)\right]$$

$$A_2 = -(n-k)\left[\min_{\phi^{y}}\left\{D\big(\hat{P}_{Y_1^{k+1},Z_1^{k+1}}\,||\,Q_{\phi^{y}}\otimes\hat{P}_{Y_1^k,Z_1^{k+1}}\big)\right\} + \sum_{y_1^{k+1},z_1^{k+1}} \hat{P}(y_1^{k+1},z_1^{k+1})\log\left(\frac{\hat{P}(y_1^k,z_1^{k+1})}{\hat{P}(y_1^{k+1},z_1^{k+1})}\right)\right]. \quad (32)$$

and (37) follows due to

$$\sum \hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})(\rho-1) = \sum \bar{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1}) - \hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1}) = 0.$$

Since the Markov model is assumed to be ergodic (Assumption 2), $\hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1}) \nrightarrow 0$, and therefore C is bounded. Now $\forall i \in [1:n-k]$ consider the sequence

$$T_i(x_1^{k+1}y_1^{k+1}z_1^{k+1}) \triangleq \mathbb{1}[X_i^{k+i}Y_i^{k+i}Z_i^{k+i} = x_1^{k+1}y_1^{k+1}z_1^{k+1}]$$

with mean $\bar{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})$. According to the law of iterated logarithms,

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n-k}(T_i - \bar{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1}))}{\sqrt{(n-k)\log\log(n-k)}} = \sqrt{2} \quad \text{a.s.}$$

Using the definition of the empirical distribution, this implies

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{(n-k)(\hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1}) - \bar{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1}))}{\sqrt{(n-k)\log\log(n-k)}} \\
&= \limsup_{n\to\infty} \frac{\hat{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1}) - \bar{P}(x_1^{k+1}y_1^{k+1}z_1^{k+1})}{\sqrt{\log\log(n-k)}/\sqrt{n-k}} \\
&= \sqrt{2} \quad \text{a.s.} \qquad (39)
\end{aligned}$$

As a result we can rewrite (38) and conclude that

$$\begin{aligned}
&\limsup_{n\to\infty} \sqrt{n-k}\, D(\hat{P}_{X_1^{k+1}Y_1^{k+1}Z_1^{k+1}} \,||\, \bar{P}_{X_1^{k+1}Y_1^{k+1}Z_1^{k+1}}) \\
&= \limsup_{n\to\infty} \frac{\log\log(n-k)}{\sqrt{n-k}} \sum_{x_1^{k+1}y_1^{k+1}z_1^{k+1}} 2C = 0,
\end{aligned}$$

given that each term in the finite sum is bounded. Therefore, as $n \to \infty$, the four KL-divergence terms in (35) multiplied by a $\sqrt{n-k}$ factor tend to zero and the proof of Theorem 3 is thus complete. ∎

REFERENCES

[1] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the Directed Information to Infer Causal Relationships in Ensemble Neural Spike Train Recordings," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 17–44, Feb. 2011.

[2] K. Friston, L. Harrison, and W. Penny, "Dynamic Causal Modelling," *NeuroImage*, vol. 19, no. 4, pp. 1273–1302, Aug. 2003.

[3] 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.

[4] W. M. Lord, J. Sun, N. T. Ouellette, and E. M. Bollt, "Inference of Causal Information Flow in Collective Animal Behavior," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 2, no. 1, pp. 107–116, Jun. 2016.

[5] J. Jiao, H. H. Permuter, L. Zhao, Y. H. Kim, and T. Weissman, "Universal Estimation of Directed Information," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6220–6242, Oct. 2013.

[6] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, Aug. 1969.

[7] J. Massey, "Causality, Feedback and Directed Information," in *Proc. Int. Symp. Inf. Theory Applic. (ISITA), Honolulu, HI, USA*, Nov. 1990, pp. 303–305.

[8] P.-O. Amblard and O. J. J. Michel, "On Directed Information Theory and Granger Causality Graphs," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 7–16, Feb. 2011.

[9] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Efficient Methods to Compute Optimal Tree Approximations of Directed Information Graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3173–3182, Jun. 2013.

[10] I. Kontoyiannis and M. Skoularidou, "Estimating the Directed Information and Testing for Causality," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6053–6067, Nov. 2016.

[11] P. Billingsley, *Statistical Inference for Markov Processes*. Chicago, IL, USA: Univ. of Chicago Press, 1961.

[12] J. Scarlett and V. Cevher, "Lower Bounds on Active Learning for Graphical Model Selection," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, USA, 20–22 Apr. 2017, pp. 55–64.

[13] K. L. Chung, *Markov Chains: With Stationary Transition Probabilities*, 2nd ed., ser. A Series of Comprehensive Studies in Mathematics. New York, NY, USA: Springer-Verlag, 1967, vol. 104.

$$\begin{aligned}
\hat{I}_n^{(k)}(\mathbf{X}\to\mathbf{Y}\,||\,\mathbf{Z}) &= \sum_{\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}} \hat{P}(\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}) \log \frac{\hat{P}(\mathbb{y}_{k+1}\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})}{\hat{P}(\mathbb{y}_{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})\hat{P}(\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})} \\
&= \sum_{\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}} \hat{P}(\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}) \log \left[\frac{\hat{P}(\mathbb{y}_{k+1}\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})\bar{P}(\mathbb{y}_{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})\bar{P}(\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})}{\hat{P}(\mathbb{y}_{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})\hat{P}(\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})\bar{P}(\mathbb{y}_{k+1}\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})}\right] \\
&\quad + \sum_{\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}} \hat{P}(\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}) \log \frac{\bar{P}(\mathbb{y}_{k+1}\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})}{\bar{P}(\mathbb{y}_{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})\bar{P}(\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})} \\
&= \sum_{\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}} \hat{P}(\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}) \log \left[\frac{\hat{P}(\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1})\hat{P}(\mathbb{y}_1^k\mathbb{z}_1^{k+1})}{\hat{P}(\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1})\hat{P}(\mathbb{x}_1^{k+1}\mathbb{y}_1^k\mathbb{z}_1^{k+1})} \cdot \frac{\bar{P}(\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1})\bar{P}(\mathbb{x}_1^{k+1}\mathbb{y}_1^k\mathbb{z}_1^{k+1})}{\bar{P}(\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1})\bar{P}(\mathbb{y}_1^k\mathbb{z}_1^{k+1})}\right] \\
&\quad + \sum_{\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}} \hat{P}(\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}) \log \frac{\bar{P}(\mathbb{y}_{k+1}\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})}{\bar{P}(\mathbb{y}_{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})\bar{P}(\mathbb{x}_1^{k+1}|\mathbb{y}_1^k z_1^{k+1})} \\
&= D(\hat{P}_{X_1^{k+1}Y_1^{k+1}Z_1^{k+1}} \,||\, \bar{P}_{X_1^{k+1}Y_1^{k+1}Z_1^{k+1}}) + D(\hat{P}_{Y_1^kZ_1^{k+1}} \,||\, \bar{P}_{Y_1^kZ_1^{k+1}}) \\
&\quad - D(\hat{P}_{Y_1^{k+1}Z_1^{k+1}} \,||\, \bar{P}_{Y_1^{k+1}Z_1^{k+1}}) - D(\hat{P}_{X_1^{k+1}Y_1^kZ_1^{k+1}} \,||\, \bar{P}_{X_1^{k+1}Y_1^kZ_1^{k+1}}) \\
&\quad + \sum_{\mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}} \left(\frac{1}{n-k}\sum_{i=k+1}^{n} \mathbb{1}[x_{i-k}^i y_{i-k}^i z_{i-k}^i = \mathbb{x}_1^{k+1}\mathbb{y}_1^{k+1}\mathbb{z}_1^{k+1}]\right) \log \frac{\bar{P}(\mathbb{y}_{k+1}\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})}{\bar{P}(\mathbb{y}_{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})\bar{P}(\mathbb{x}_1^{k+1}|\mathbb{y}_1^k\mathbb{z}_1^{k+1})}. \qquad (35)
\end{aligned}$$