Sample Complexity for Topology Estimation in Networks of LTI Systems

Vincent Y. F. Tan and Alan S. Willsky

Abstract— **This paper proposes a consistent and computationally efficient FFT-based algorithm for inferring the network topology where each node in the network is associated to a wide-sense stationary, ergodic, Gaussian process. Each edge of the tree network is characterized by a linear, time-invariant dynamical system and additive white Gaussian noise. The proposed algorithm uses Bartlett's procedure to produce periodogram estimates of cross power spectral densities between processes. Under appropriate assumptions, we prove that the number of vector-valued samples from a single sample path required for consistent estimation is polylogarithmic in the number of nodes in the network. Thus, the sample complexity is low. Our proof uses properties of spectral estimates and analysis for learning tree-structured graphical models.**

I. Introduction

Imagine that there is a large linear electrical circuit consisting of impedances (such as resistors, inductors and capacitors) but its network topology is unknown. We are, however, given noisy measurements of voltages at the nodes in the circuit. Each of these voltages is modelled as a wide-sense stationary (WSS), ergodic, discrete-time stochastic process. Given a finite number of time samples of each node voltage realization, we would like to reconstruct the network topology consistently. But how many samples are required to obtain a "reliable" estimate of the network topology?

The identification of large-scale graphs or networks of systems is an important task in many realms of science and engineering, including control engineering. In the literature, this problem has been studied extensively by the graphical model learning community in which each node is associated to a random variable and the (vector-valued) observations are independent and identically distributed. See [1] for an overview. This differs from the circuit network topology inference problem mentioned above since each node corresponds to a *stationary stochastic process* and we only observe a finite number of samples of *one sample path*.

In this paper, we propose an algorithm to estimate such tree-structured networks, where each edge is characterized by an LTI system plus additive white Gaussian noise. There are two main contributions: Firstly, we show that the proposed algorithm is computationally efficient; it is (up to log factors) quadratic in the number of nodes and linear in the number of observations. Secondly, by using classical results from spectral estimation [2], [3], we prove that under appropriate assumptions, the proposed algorithm has very favorable sample complexity; if the number of samples N exceeds $O(\log^{1+\varepsilon} p)$, where p is the number of nodes in the network, then the undirected network topology can be estimated with high probability as N and p tend to infinity.

Previously, Bach and Jordan [4] used the Bayesian information criterion (BIC) and spectral methods to estimate sparse graphs of stationary time series. In a collection of related works [5]–[7], authors also considered learning models whose topologies are dynamic, i.e., they change over time. Siracusa and Fisher [8] also proposed Bayesian techniques to obtain posterior uncertainties of the underlying topology. Most recently, Materassi and Innocenti [9] proposed a provably consistent algorithm for estimating such tree-structured networks. None of the above works include any sample complexity guarantees. In this work, we combine the results from learning graphical models [10], [11] and spectral estimation [2] to obtain sample complexities for networks of dynamical systems.

II. Preliminaries and System Model

Let $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ be an undirected tree (a connected, acyclic graph) where $\mathcal{V} = \{1, \ldots, p\}$ is the set of nodes and $\mathcal{E} \subset \binom{\mathcal{V}}{2}$ is the set of undirected edges. Let node 1 be labeled as the *root* of the tree. For edge $(i, j) \in \mathcal{E}$, if i is closer to 1 than j, we say that node i is the *parent* of node j and node j is a *child* of node i. Note that by the tree assumption, each node (except the root) has exactly one parent. The root has no parents. Define $\mathfrak{T}$ to be the set of trees with p nodes. We associate to each node $i \in \mathcal{V}$ a WSS, ergodic, discrete-time stochastic process $X_i = \{X_i[n]\}_{n=0}^{\infty}$. Each $X_i[n]$ is a real-valued random variable.

Let $X = (X_1, \ldots, X_p)$ be the vector of stochastic processes. The edge set $\mathcal{E}$ encodes the set of conditional independence relations among the p processes [12]. More precisely, X is *Markov on* $\mathcal{T}$ if for all $i \in \mathcal{V}$, X_i is conditionally independent of all other processes given its parent and its children.

Let the root process X_1 be a Gaussian process. For example, each $X_1[n]$ can be an independent zero-mean, unit-variance Gaussian random variable. For an edge $(i, j) \in \mathcal{E}$, with j being a child of i, we assume that

$$X_j[n] = (h_{j,i} * X_i)[n] + W_j[n], \qquad n = 0, 1, 2, \ldots \quad (1)$$

where $h_{j,i}[n]$ is a (non-zero) causal, stable LTI filter. In addition, for each node j, the process $W_j[n]$ is assumed to be additive white Gaussian noise with power σ_W^2. See Fig. 1. Eqn. (1) says that the child process X_j is a noisy, filtered version of the parent process X_i. Because all filters

V. Tan is with the Dept. of Electrical and Computer Engineering, University of Wisconsin-Madison (vtan@wisc.edu)

A. Willsky is with the Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (willsky@mit.edu)

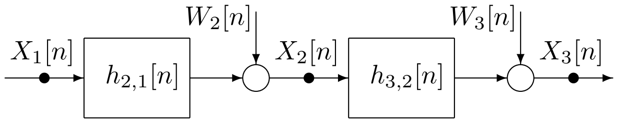

Fig. 1. The nodes are denoted by dark circles. The node set is $\mathcal{V} = \{1,2,3\}$ and the edge set is $\mathcal{E} = \{(1,2),(2,3)\}$. Note the conditional independence relations among the stochastic processes: $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$.

are LTI and the noise is Gaussian, all processes X_j are jointly Gaussian.

In this paper, we are provided with the first N samples of a sample path of the vector-valued stochastic process $\{X[n]\}_{n=0}^{N-1}$. Given this data, we would like to devise an efficient algorithm to obtain a consistent estimate of the tree $\mathcal{T}$. We denote the estimate given N samples $\{X[n]\}_{n=0}^{N-1}$ as $\hat{\mathcal{T}}_N$. In addition, for $\delta > 0$, we consider the following quantity:

$$N(p,\delta) := \inf\{N \in \mathbb{N} : \mathbb{P}(\hat{\mathcal{T}}_N = \mathcal{T}) \geq 1-\delta\}. \quad (2)$$

The quantity $N(p,\delta)$ is the *sample complexity* for estimating the topology of the network. It denotes the number of samples needed to obtain a topology that is the same as the original one with high probability. In general, $N(p,\delta)$ will also be a function of the unknown filters $\{h_{j,i}\}_{(i,j)\in\mathcal{E}}$ and the noise power σ_W^2 but we suppress this dependence. We would like to ensure that the algorithm proposed has low sample complexity, which means that $N(p,\delta)$ increases slowly with $p = |\mathcal{V}|$ as both quantities scale (i.e., tend to infinity).

III. An Algorithm to Estimate the Network Topology Via Spectrum Estimation

In this section, we present a consistent and efficient algorithm to estimate the tree topology $\mathcal{T}$ given the data $\{X[n]\}_{n=0}^{N-1}$. This algorithm is motivated by the Chow-Liu algorithm [13] to approximate arbitrary multivariate distributions with trees. Let $\mathcal{M}(\mathfrak{T})$ be the family of probability measures associated to multivariate processes that are Markov on a tree with p nodes. Consider the following optimization problem:

$$\inf_{\nu\in\mathcal{M}(\mathfrak{T})} D(\mu \,||\, \nu) \quad (3)$$

where μ is an estimate of the probability measure of the underlying multivariate process. In (3), $D(\mu \,||\, \nu)$ is the *relative entropy rate* [14] between the two probability measures μ and ν. Using the same reasoning as in [13], it is straightforward to show the following:

Lemma 3.1 (Chow-Liu for WSS Stochastic Processes): The measure that achieves the minimum in (3) corresponds to a vector-valued process that is Markov on the tree $\mathcal{T}^*$ given by the maximum-weight spanning tree (MWST) problem:

$$\mathcal{T}^* = \operatorname*{argmax}_{\mathcal{T}\in\mathfrak{T}} \sum_{(i,j)\in\mathcal{T}} I_\mu(X_i; X_j), \quad (4)$$

where $I_\mu(X_i;X_j)$ is the *mutual information (MI) rate* of the processes X_i and X_j under μ.

See Appendix for the proof. Lemma 3.1 is similar to the main result in [9] but the derivation using (3) as an information projection is more intuitive. In addition, we obtain the mutual information rate $I_\mu(X_i;X_j)$ as edge weights for the MWST in (4) whereas [9] derived a closely-related quantity using Wiener filtering. Given Lemma 3.1, it remains to estimate the MI rates consistently from the data $\{X[n]\}_{n=0}^{N-1}$. To this end, we first recall from [15] that for two WSS Gaussian processes X_i and X_j,

$$I(X_i;X_j) := -\frac{1}{4\pi}\int_0^{2\pi} \log\left(1-|\gamma_{i,j}(\omega)|^2\right) d\omega \quad (5)$$

where the *magnitude-squared coherence* of the processes X_i and X_j is defined as

$$|\gamma_{i,j}(\omega)|^2 := \frac{|\Phi_{X_i,X_j}(\omega)|^2}{\Phi_{X_i}(\omega)\Phi_{X_j}(\omega)}. \quad (6)$$

In Eqn. (6), $\Phi_{X_i}(\omega)$ and $\Phi_{X_i,X_j}(\omega)$ are the power spectral density (PSD) of X_i and cross PSD of X_i, X_j respectively.[1] By the Cauchy-Schwarz inequality, $0 \leq |\gamma_{i,j}(\omega)|^2 \leq 1$ so the MI rate in (5) is non-negative. We assume that the MI rate for all edges is uniformly bounded away from zero as $p \to \infty$.

At a high level, the algorithm proceeds as follows: Since it is intractable to compute the MI rate directly, we use Bartlett's averaging method [2] to estimate $\Phi_{X_i}(\omega)$ and $\Phi_{X_i,X_j}(\omega)$ from the data before computing the magnitude-squared coherences $|\gamma_{i,j}(\omega)|^2$. We then use a discretized version of (5) to obtain an estimate of the MI rate. Finally we obtain $\hat{\mathcal{T}}_N$ by solving the MWST problem in (4). The details of the algorithm are provided below:

1) Divide each length-N realization $\{X_j[n]\}_{n=0}^{N-1}$ into L non-overlapping segments of length M such that $LM \leq N$, i.e., we form the signal segments:

$$X_j^{(l)}[n] := X_j[lM+n], \quad (7)$$

where $0 \leq n \leq M-1, 0 \leq l \leq L-1$ and $j \in \mathcal{V}$. The choice of L and M is discussed in Section IV-B.

2) Compute the length-M DFT (discrete Fourier transform) for each signal segment:

$$\widetilde{X}_j^{(l)}[k] := \frac{1}{M}\sum_{n=0}^{M-1} X_j^{(l)}[n] e^{-\sqrt{-1}\,2\pi(k+1/2)n/M}. \quad (8)$$

Note that we deliberately sample the DTFT at frequencies $2\pi(k+1/2)/M$ for $k = 0,\ldots,M-1$. The reason for this will become apparent in Lemma 4.3.

3) Estimate the time-averaged periodograms for the PSD and cross PSD using Bartlett's averaging procedure on the L signal segments, i.e.,

$$\hat{\Phi}_{X_i}[k] := \frac{1}{L}\sum_{l=0}^{L-1}\left|\widetilde{X}_i^{(l)}[k]\right|^2, \quad (9a)$$

$$\hat{\Phi}_{X_i,X_j}[k] := \frac{1}{L}\sum_{l=0}^{L-1}\left(\widetilde{X}_i^{(l)}[k]\right)^* \widetilde{X}_j^{(l)}[k]. \quad (9b)$$

[1] For simplicity, we denote the DTFT as $Y(\omega)$ instead of $Y(e^{j\omega})$.

4) Estimate the magnitude-squared coherences:

$$|\hat{\gamma}_{i,j}[k]|^2 := \frac{|\hat{\Phi}_{X_i,X_j}[k]|^2}{\hat{\Phi}_{X_i}[k]\hat{\Phi}_{X_j}[k]}. \qquad (10)$$

5) Estimate the MI rates by using the Riemann sum:

$$\hat{I}(X_i;X_j) := -\frac{1}{2M}\sum_{k=0}^{M-1}\log\left(1-|\hat{\gamma}_{i,j}[k]|^2\right). \qquad (11)$$

6) Solve the MWST problem in (4) with $\{\hat{I}(X_i;X_j)\}_{i,j\in\mathcal{V}}$ as the edge weights to obtain $\hat{\mathcal{T}}_N$.

It is known that unless we average over signal segments as in (9), the periodogram will *not* be a consistent estimate of the PSD [2, Theorem 5.2.4]. However, because we assumed that each process is ergodic and we average across different signal segments, the estimates of the PSD and cross PSD are consistent. In addition, since it is not possible to compute the integral in (5) exactly, we approximate it using a Riemann sum in terms of the DFT values of $\hat{\gamma}_{i,j}[k]$ as in (11). We note that the algorithm presented can be generalized by considering overlapping segments and multiplying a length-M window $v[n]$ to each signal segment $X_j^{(l)}[n]$. This is the well-known Welch's method [2]. We do not consider these generalizations here since the analysis of the sample complexity is much more involved.

Because the DFT operations in Step 2 can be implemented efficiently using the fast Fourier transform (FFT), the proposed algorithm is also computationally efficient.

Proposition 3.2 (Computational Complexity): The computational complexity of the algorithm proposed to estimate the tree network topology $\mathcal{T}$ is bounded above by

$$O(p^2(N\log M+\log p)). \qquad (12)$$

Proof Step 2 requires $O(pLM\log M)$ operations using FFTs. It is easy to see that Steps 1 and 3 - 5 require at most $O(p^2N)$ operations. Finally, the MWST can be implemented using Kruskal's algorithm [16] in $O(p^2\log p)$ operations.

Despite the appealing computational complexity, it is not clear how to choose L and M optimally given each length-N signal $\{X_j[n]\}_{n=0}^{N-1}$. This is an important consideration since there is a fundamental tradeoff between *spectral resolution* (which improves by increasing DFT length M) and reduction of the *variance* of the spectrum estimate (which improves by increasing the number of segments L). In Section IV, we provide intuition on how to choose L and M such that the sample complexity $N(p,\delta)$ is low. In the following, it will be useful to regard M, N and p as functions of L.

IV. Sample Complexity and Proof Outline

In this section, we state our sample complexity result and provide an outline of its proof. Before doing so, we state an overriding assumption: We assume that the arctanh of the estimate of the magnitude-squared coherence[2] satisfies the *normality assumption*, i.e.,

$$\operatorname{arctanh}(|\hat{\gamma}(\omega)|)\sim\mathcal{N}(m_L,\lambda_L). \qquad (13)$$

See [3] and [17] for the details of this normalizing *Fisher z-transform.* The mean in (13) is given as

$$m_L := \operatorname{arctanh}\left(\sqrt{|\gamma(\omega)|^2+\frac{1-|\gamma(\omega)|^2}{2(L-1)}}\right), \qquad (14)$$

and the variance $\lambda_L = 1/(2(L-1))$. Note that this means (by the continuity of arctanh) that the estimate at every ω is asymptotically unbiased and consistent since $m_L\to\operatorname{arctanh}|\gamma(\omega)|$ and $\lambda_L\to 0$.

A. Polylogarithmic Sample Complexity

Theorem 4.1 (Sample Complexity): Assume (13) holds. For appropriately chosen L and M, the sample complexity of the algorithm proposed to estimate the tree network topology $\mathcal{T}$ is

$$N(p,\delta)=O\left(\log^{1+\varepsilon}\left(\frac{p^3}{\delta}\right)\right) \qquad (15)$$

for any $\varepsilon>0$.

The interpretation of this asymptotic result if N and p obey the prescribed scaling law, then the probability of successful estimation $\mathbb{P}(\hat{\mathcal{T}}_N=\mathcal{T})$ can be made arbitrarily close to 1. Note that p is allowed to grow much faster than N, which means that even in the sample-limited regime, we are guaranteed to successfully recover the unknown tree topology $\mathcal{T}$ with relatively few samples.

B. Proof Outline

The proof of Theorem 4.1 is a consequence of four lemmata whose proofs are in the appendices. It is worth noting that standard concentration results from large-deviations theory such as Sanov's Theorem or the Gärtner-Ellis Theorem [18] do not readily apply for deriving concentration results for this problem. Instead, we use the normality assumption [3], [17] to obtain concentration results.

Lemma 4.2 (Concentration of Coherence): Fix $\eta>0$. Define the function $g:[0,1)\to\mathbb{R}^+$ as

$$g(\gamma):=-\frac{1}{4\pi}\log(1-|\gamma|^2). \qquad (16)$$

Then we have the following upper bound:

$$\mathbb{P}\left(|g(\hat{\gamma}(\omega))-g(\gamma(\omega))|>\eta\right)\le e^{-(L-1)\varphi(\gamma(\omega);\eta)}, \qquad (17)$$

where the exponent above $\varphi:[0,1]\times(0,\infty)\to\mathbb{R}^+$ is a continuous function. In addition, $\varphi(\gamma;\eta)$ is monotonically decreasing in γ and monotonically increasing in η.

Lemma 4.2 quantifies the deviation of $|\hat{\gamma}(\omega)|$ from $|\gamma(\omega)|$. Observe that if the tolerance η is large, the exponent φ is also large. Besides, larger coherences values are "harder" to estimate. Recall from Step 2 that we only evaluate the magnitude-squared coherence estimate at M uniformly spaced frequencies $2\pi(k+1/2)/M$ for $k=0,\ldots,M-1$. We now quantify the deviation of estimated MI rate obtained using a Riemann sum from the true MI rate. This is the most technically challenging step in the proof.

[2] We suppress the dependence of γ and $\hat{\gamma}$ on edge (i,j) for brevity.

Lemma 4.3 (Concentration of MI Rate): If the number of DFT points M_L satisfies

$$\lim_{L\to\infty} M_L = \infty, \qquad \lim_{L\to\infty} L^{-1}\log M_L = 0, \tag{18}$$

then for any $\eta > 0$, we have

$$\limsup_{L\to\infty} \frac{1}{L}\log \mathbb{P}\left(|\hat{I} - I| > \eta\right) \leq -\min_{\omega\in[0,2\pi)} \varphi(\gamma(\omega);\eta). \tag{19}$$

Now we use the proof technique in [10], [11] to quantify the error probability of estimating an incorrect topology. For a non-edge (k,l), let $\mathrm{Path}(k,l)$ be the set of edges along the unique path joining nodes k and l. By the data-processing lemma [14] and the fact that the LTI filters are non-zero,

$$\xi := \min_{(k,l)\notin\mathcal{E}} \min_{(i,j)\in\mathrm{Path}(k,l)} I(X_k;X_l) - I(X_i;X_j), \tag{20}$$

is uniformly bounded away from zero for all tree-structured networks. The quantity ξ in (20) is the minimum difference between the MI rate of an edge and the MI rate of any edge along its path.

Lemma 4.4 (Crossover Probability for MI Rates): Let (k,l) be a non-edge and $(i,j) \in \mathrm{Path}(k,l)$. Then assuming that M_L satisfies (18), we have the large deviations upper bound

$$\limsup_{L\to\infty} \frac{1}{L}\log \mathbb{P}\left(\hat{I}(X_k;X_l) \geq \hat{I}(X_i;X_j)\right) \tag{21}$$
$$\leq -\min_{\omega\in[0,2\pi)} \min\left\{\varphi(\gamma_{i,j}(\omega);\xi/2), \varphi(\gamma_{k,l}(\omega);\xi/2)\right\}.$$

This lemma implies that the error in mistaking a non-edge for a true edge decays exponentially fast in L, since the right hand side of (21) is negative. The next lemma utilizes the fact that the number of potential error events as in (21) is $O(p^3)$.

Lemma 4.5 (Error in Topology Estimation): If M_L satisfies (18), then there exists a constant $K > 0$ such that

$$\limsup_{L\to\infty} \frac{1}{L}\log \mathbb{P}\left(\hat{\mathcal{T}}_N \neq \mathcal{T}\right) \leq -K + \limsup_{L\to\infty} \frac{3\log p}{L}. \tag{22}$$

This result shows that if $p = O(1)$, the last term in (22) is zero and the error in network topology estimation decays exponentially fast in the number of signal segments L. Since $M = \lceil L^{\varepsilon}\rceil$ satisfies (18) (for any $\varepsilon > 0$), N is required to be at least $\lceil L^{1+\varepsilon}\rceil$ and hence the error probability in topology estimation can be upper bounded as

$$\mathbb{P}(\hat{\mathcal{T}}_N \neq \mathcal{T}) = O\big(p^3 \exp(-CN^{1/(1+\varepsilon)})\big) \quad \forall\,\varepsilon > 0. \tag{23}$$

The proof of Theorem 4.1 is completed by inverting the relationship $\mathbb{P}(\hat{\mathcal{T}}_N \neq \mathcal{T}) \leq \delta$.

V. Conclusion

We proposed a consistent and efficient algorithm to estimate networks whose edges are characterized by LTI filters and noise. Our main contribution is the asymptotic sample complexity analysis that shows that for very large networks, a relatively small number of samples is required to estimate the network topology reliably. We intend to extend our results in three main directions: Firstly, in place of Bartlett's non-overlapping method for periodogram estimation, we would like to analyze the more accurate Welch's overlapping method [2]. Secondly, we intend to find sufficient conditions that allow for inferring the *directed* tree. Thirdly, we will perform numerical simulations to verify the sample complexity result.

Appendix

Proof This proof extends Chow and Liu's result [13] to stationary stochastic processes. The fact that $\nu \in \mathcal{M}(\mathfrak{T})$ implies that we have the factorization:

$$\nu = \prod_{i\in\mathcal{V}} \nu_i \prod_{(i,j)\in\mathcal{E}} \frac{\nu_{i,j}}{\nu_i \times \nu_j}, \tag{24}$$

where ν_i and $\nu_{i,j}$ are measures corresponding to the the marginal and pairwise processes respectively. Recall that the relative entropy rate is defined as

$$D(\mu\,||\,\nu) = \begin{cases} \int \log \frac{d\mu}{d\nu}\, d\mu & \mu \ll \nu \\ +\infty & \text{o.w.} \end{cases}$$

where $\frac{d\mu}{d\nu}$ is the Radon-Nikodým derivative of μ wrt ν. By stationarity, the relative entropy in (3) can be written as a limit

$$D(\mu\,||\,\nu) = \lim_{N\to\infty} \frac{1}{N} D(\mu[0:N-1]\,||\,\nu[0:N-1]),$$

where $\mu[0:N-1]$ is the probability measure corresponding to the first N time points of the μ process. Hence,

$$D(\mu||\nu) = \lim_{N\to\infty} \frac{1}{N}\int \log\frac{d\mu[0:N-1]}{d\nu[0:N-1]}\, d\mu[0:N-1] \tag{25}$$

Now, substitute (24) into (25) and note that we are optimizing over the measure ν only. Hence, the relative entropy rate is (up to a constant) the limit of the expression

$$\frac{1}{N}\int \sum_{(i,j)\in\mathcal{E}} \log \frac{d\nu_{i,j}[0:N-1]}{d(\nu_i[0:N-1]\times\nu_j[0:N-1])} d\mu[0:N-1]$$

as $N \to \infty$. Each term is minimized by setting $\nu_i = \mu_i$ and $\nu_{i,j} = \mu_{i,j}$ (by non-negativity of relative entropy). By exchanging the sum and integral in the above, we see that each of the terms is the mutual information $I_\mu(X_i[0:N-1]; X_j[0:N-1]) = D(\mu_{i,j}[0:N-1]\,||\,\mu_i[0:N-1]\times\mu_j[0:N-1])$. Now by stationarity, the following limit exists and equals the mutual information rate:

$$\lim_{N\to\infty} \frac{1}{N} I_\mu(X_i[0:N-1]; X_j[0:N-1]) = I_\mu(X_i;X_j).$$

Combining this with the above sum, we see that the minimization problem is (3) is given by the MWST problem in (4) where the edge weights are the mutual information rates in (5).

Proof For simplicity in notation, we drop the dependence of γ on the frequency ω, i.e., denote $\hat{\gamma} = \hat{\gamma}(\omega)$ and $\gamma = \gamma(\omega)$.

Also, recall the definition of g in (16). Then, for $\eta > 0$, we have that $\mathbb{P}(|g(\hat{\gamma}) - g(\gamma)| \geq \eta)$ can be upper bounded as

$$\begin{aligned} &\mathbb{P}\left(\left|\log \frac{1-|\hat{\gamma}|^2}{1-|\gamma|^2}\right| > 4\pi\eta\right) \\ &\leq \mathbb{P}\left(\log \frac{1-|\hat{\gamma}|^2}{1-|\gamma|^2} > 4\pi\eta\right) + \mathbb{P}\left(\log \frac{1-|\hat{\gamma}|^2}{1-|\gamma|^2} < -4\pi\eta\right) \\ &= \mathbb{P}\left(|\hat{\gamma}| < a(\gamma,\eta)\right) + \mathbb{P}\left(|\hat{\gamma}| > a(\gamma,-\eta)\right), \end{aligned} \tag{26}$$

where we applied the union bound and the function $a(\gamma(\omega),\eta)$ in (26) is defined as

$$a(\gamma,\eta) := \sqrt{|\gamma|^2 - (e^{4\pi\eta}-1)(1-|\gamma|^2)}.$$

Because the arctanh function is continuous, we have

$$\begin{aligned} \mathbb{P}(|g(\hat{\gamma}) - g(\gamma)| \geq \eta) &\leq \mathbb{P}\left(\operatorname{arctanh}(|\hat{\gamma}|) < \operatorname{arctanh}(a(\gamma,\eta))\right) \\ &+ \mathbb{P}\left(\operatorname{arctanh}(|\hat{\gamma}|) > \operatorname{arctanh}(a(\gamma,-\eta))\right) \end{aligned}$$

By using the normality assumption as stated in (13) we have that

$$\begin{aligned} \mathbb{P}(|g(\hat{\gamma}) - g(\gamma)| \geq \eta) &\leq Q(\sqrt{2(L-1)}(-b_1(\gamma,\eta) + O(L^{-1}))) \\ &+ Q(\sqrt{2(L-1)}(b_2(\gamma,\eta) + O(L^{-1}))) \end{aligned} \tag{27}$$

where $Q(z) = \int_z^\infty \mathcal{N}(u;0,1)\,du$ and the functions $b_1(\gamma,\eta)$ and $b_2(\gamma,\eta)$ are defined as

$$\begin{aligned} b_1(\gamma,\eta) &:= \operatorname{arctanh}(a(\gamma,\eta) - \operatorname{arctanh}(|\gamma|) < 0 \\ b_2(\gamma,\eta) &:= \operatorname{arctanh}(a(\gamma,-\eta) + \operatorname{arctanh}(|\gamma|) > 0. \end{aligned}$$

We now use the fact that $Q(z) \leq \frac{1}{2}e^{-z^2/2}$ to conclude that (27) can be upper bounded as

$$\begin{aligned} \mathbb{P}(|g(\hat{\gamma}) - g(\gamma)| \geq \eta) \leq \frac{1}{2}\Big(&\exp\left(-2(L-1)b_1(\gamma,\eta)^2\right) \\ &+ \exp\left(-2(L-1)b_2(\gamma,\eta)^2\right)\Big). \end{aligned}$$

The proof is completed by the identification

$$\varphi(\gamma(\omega);\eta) := 2\min_{i=1,2}\left\{b_i(\gamma(\omega),\eta)^2\right\}.$$

This function is positive because b_1^2 and b_2^2 are positive. It can also continuous because $b_i(\gamma(\omega),\eta), i = 1,2$ are continuous. The monotonicity properties also forward straightforwardly from the definitions of a and b_i.

Proof Recall from Riemann integration theory that for a smooth function ϕ, the error in approximating a Riemann integral using its (middle) Riemann sum approximation as in (11) is upper bounded as

$$\left|\int_0^{2\pi} \phi(\tau)\,d\tau - \frac{1}{M}\sum_{k=0}^{M-1} \phi\left(\frac{2\pi(k+1/2)}{M}\right)\right| \leq \frac{B}{M_L^2} \tag{28}$$

where $B := (2\pi)^3 \max_{[0,2\pi)} |\phi''(\tau)|/24$. By using the definition of g in Lemma 4.3, we have

$$\mathbb{P}\left(|\hat{I} - I| > \eta\right) = \mathbb{P}\left(\left|\int_0^{2\pi} g(\hat{\gamma}(\omega)) - g(\gamma(\omega))\,d\omega\right| > \eta\right),$$

where the function g was defined in Lemma 4.2. Next, by the elementary inequality from integration theory $|\int \psi(\tau)\,d\tau| \leq \int |\psi(\tau)|\,d\tau$, we have

$$\mathbb{P}\left(|\hat{I} - I| > \eta\right) \leq \mathbb{P}\left(\int_0^{2\pi} |g(\hat{\gamma}(\omega)) - g(\gamma(\omega))|\,d\omega > \eta\right),$$

Because $\gamma, g \in \mathcal{C}^2$, so is the function $\phi := g\circ\hat{\gamma} - g\circ\gamma \in \mathcal{C}^2$. Hence the constant B in (28) is finite. By using (28), we have the inclusion of the events

$$\begin{aligned} &\left\{\int_0^{2\pi} |g(\hat{\gamma}(\omega)) - g(\gamma(\omega))|\,d\omega > \eta\right\} \subset \\ &\left\{\frac{1}{M_L}\sum_{k=0}^{M_L-1} |g(\hat{\gamma}(\omega_k)) - g(\gamma(\omega_k))| > \eta - \frac{B}{M_L^2}\right\} \end{aligned}$$

Now assuming that $M_L > \sqrt{B/\eta}$, and using monotonicity of measure, we have the upper bound,

$$\begin{aligned} &\mathbb{P}\left(|\hat{I} - I| > \eta\right) \leq \\ &\mathbb{P}\left(\frac{1}{M_L}\sum_{k=0}^{M_L-1} |g(\hat{\gamma}(\omega_k)) - g(\gamma(\omega_k))| > \eta - \frac{B}{M_L^2}\right). \end{aligned} \tag{29}$$

Note that $\{\omega_k = 2\pi(k+1/2)/M : k = 0,1,\ldots,M-1\}$ is the set of frequencies that we sample the DTFT in Step 2. Note that we have replaced the integral over the set $[0,2\pi)$ with a sum over a discrete set of frequencies. The deviation of the sum is much easier to bound. Now, using the fact that the mean of finitely many positive numbers is no greater than the maximum, we can upper bound (29) and hence $\mathbb{P}(|\hat{I} - I| > \eta)$ as follows:

$$\begin{aligned} &\mathbb{P}\left(\max_{0\leq k\leq M_L-1} |g(\hat{\gamma}(\omega_k)) - g(\gamma(\omega_k))| > \eta - \frac{B}{M_L^2}\right) \\ &= \mathbb{P}\left(\bigcup_{k=0}^{M_L-1}\left\{|g(\hat{\gamma}(\omega_k)) - g(\gamma(\omega_k))| > \eta - \frac{B}{M_L^2}\right\}\right) \\ &\leq \sum_{k=0}^{M_L-1} \mathbb{P}\left(|g(\hat{\gamma}(\omega_k)) - g(\gamma(\omega_k))| > \eta - \frac{B}{M_L^2}\right) \qquad (30) \\ &\leq M_L \max_{0\leq k\leq M_L-1} \mathbb{P}\left(|g(\hat{\gamma}(\omega_k)) - g(\gamma(\omega_k))| > \eta - \frac{B}{M_L^2}\right) \\ &\leq M_L \max_{0\leq k\leq M_L-1} \exp\left(-(L-1)\varphi\left(\gamma(\omega_k);\eta - \frac{B}{M_L^2}\right)\right) \qquad (31) \\ &= M_L \exp\left(-(L-1)\min_{0\leq k\leq M_L-1} \varphi\left(\gamma(\omega_k);\eta - \frac{B}{M_L^2}\right)\right) \\ &\leq M_L \exp\left(-(L-1)\min_{\omega\in[0,2\pi]} \varphi\left(\gamma(\omega);\eta - \frac{B}{M_L^2}\right)\right), \qquad (32) \end{aligned}$$

where (30) is from the union bound and (31) follows from Lemma 4.2. Taking the normalized logarithm of (32), we have

$$\begin{aligned} &\frac{1}{L}\log\mathbb{P}\left(|\hat{I} - I| > \eta\right) \leq \\ &\frac{1}{L}\log M_L - \frac{L-1}{L}\min_{\omega\in[0,2\pi]} \varphi\left(\gamma(\omega);\eta - \frac{B}{M_L^2}\right). \end{aligned} \tag{33}$$

Now, by the continuity of φ in (γ, η) and the fact that $[0, 2\pi] \subset \mathbb{R}$ is compact,[3] the assignment $\eta \mapsto \min_{\omega \in [0,2\pi]} \varphi(\gamma(\omega); \eta)$ is continuous. Hence,

$$\lim_{L\to\infty} \min_{\omega\in[0,2\pi]} \varphi\left(\gamma(\omega); \eta - \frac{B}{M_L^2}\right) = \min_{\omega\in[0,2\pi]} \varphi\left(\gamma(\omega); \eta\right),$$

because $M_L \to \infty$. As a result, taking the upper limit of (33) and using the fact that $L^{-1} \log M_L \to 0$ yields the statement of the lemma.

Proof Noting the definition of ξ in (20), we have

$$I(X_k; X_l) \leq I(X_i; X_j) - \xi. \tag{34}$$

for every edge $(i, j) \in \text{Path}(k, l)$ and every non-edge $(k, l) \notin \mathcal{E}$. Eqn. (34) is also due to the data-processing lemma [14, Ch. 1]. Define the event

$$\mathcal{A}_{i,j}(\eta) := \left\{ |\hat{I}(X_i; X_j) - I(X_i; X_j)| > \eta \right\}.$$

Clearly, if the event $\{\hat{I}(X_k; X_l) \geq \hat{I}(X_i; X_j)\}$ occurs, then either $\mathcal{A}_{i,j}(\xi/2)$ or $\mathcal{A}_{k,l}(\xi/2)$ occurs, i.e.,

$$\{\hat{I}(X_k; X_l) \geq \hat{I}(X_i; X_j)\} \subset \mathcal{A}_{i,j}(\xi/2) \bigcup \mathcal{A}_{k,l}(\xi/2). \tag{35}$$

This is because of (34). Following this, we have the upper bound

$$\begin{aligned}\mathbb{P}\left(\hat{I}(X_k; X_l) \geq \hat{I}(X_i; X_j)\right) &\leq \mathbb{P}\left(\mathcal{A}_{i,j}(\xi/2) \bigcup \mathcal{A}_{k,l}(\xi/2)\right) \\ &\leq \mathbb{P}\left(\mathcal{A}_{i,j}(\xi/2)\right) + \mathbb{P}\left(\mathcal{A}_{k,l}(\xi/2)\right), \end{aligned} \tag{36}$$

where (36) is because of the inclusion in (35) and monotonicity of probability measure. This completes the proof since $\mathcal{A}_{i,j}(\eta)$ is precisely the error event in Lemma 4.3.

Proof We use the proof technique in [10], [11]. In particular,

$$\left\{\hat{\mathcal{T}}_N \neq \mathcal{T}\right\} = \bigcup_{(k,l)\notin\mathcal{E}} \bigcup_{(i,j)\in\text{Path}(k,l)} \left\{\hat{I}(X_k; X_l) \geq \hat{I}(X_i; X_j)\right\}.$$

That is, the error event in network topology estimation is equal to the existence of a non-edge (k, l) and an edge (i, j) such that the event $\{\hat{I}(X_k; X_l) \geq \hat{I}(X_i; X_j)\}$ occurs. Using a union bound again, we have that the probability of $\{\hat{\mathcal{T}}_N \neq \mathcal{T}\}$ is upper bounded as

$$\begin{aligned}&\sum_{(k,l)\notin\mathcal{E}} \sum_{(i,j)\in\text{Path}(k,l)} \mathbb{P}\left(\hat{I}(X_k; X_l) \geq \hat{I}(X_i; X_j)\right) \\ &\leq p^3 \max_{(k,l)\notin\mathcal{E}} \max_{(i,j)\in\text{Path}(k,l)} \mathbb{P}\left(\hat{I}(X_k; X_l) \geq \hat{I}(X_i; X_j)\right). \end{aligned} \tag{37}$$

Define the positive number

$$J_{i,j,k,l} := \min_{\omega\in[0,2\pi)} \min\left\{\varphi(\gamma_{i,j}(\omega); \xi/2), \varphi(\gamma_{k,l}(\omega); \xi/2)\right\}.$$

Now take the normalized logarithm and the upper limit in (37) to get

$$\begin{aligned}&\limsup_{L\to\infty} \frac{1}{L} \log \mathbb{P}\left(\hat{\mathcal{T}}_N \neq \mathcal{T}\right) \leq \\ &- \min_{(k,l)\notin\mathcal{E}} \min_{(i,j)\in\text{Path}(k,l)} J_{i,j,k,l} + \limsup_{L\to\infty} \frac{3\log p}{L}.\end{aligned}$$

This yields (22), where the positive constant K is defined as $K := \min_{(k,l)\notin\mathcal{E}} \min_{(i,j)\in\text{Path}(k,l)} J_{i,j,k,l}$. This completes the proof.

Proof Fix $\varepsilon > 0$. Choose $M_L = \lceil L^\varepsilon \rceil$ satisfying (18). Then $N = \lceil L^{1+\varepsilon} \rceil$. We drop the $\lceil\cdot\rceil$ notation from now on for simplicity. Hence $L = N^\alpha$ for $\alpha = \frac{1}{1+\varepsilon} \in (0, 1)$. From (22), we have

$$\limsup_{N\to\infty} \frac{1}{N^\alpha} \log \mathbb{P}\left(\hat{\mathcal{T}}_N \neq \mathcal{T}\right) \leq -K + \limsup_{N\to\infty} \frac{3\log p}{N^\alpha}.$$

which means that for any $C < K$ and N sufficiently large, we have

$$\mathbb{P}\left(\hat{\mathcal{T}}_N \neq \mathcal{T}\right) \leq p^3 \exp(-CN^\alpha).$$

Hence, for the error probability to be less than $\delta > 0$, it suffices to have the number of sample satisfy:

$$N^\alpha = N^{\frac{1}{1+\varepsilon}} \geq O\left(\log \frac{p^3}{\delta}\right).$$

References

[1] D. Heckerman, "A Tutorial on Learning with Bayesian Networks," Tech. Rep. MSR-TR-95-06, Microsoft Research, Mar 1995.

[2] D. R. Brillinger, *Time Series: Data Analysis and Theory*, SIAM, 2001.

[3] A. H. Nuthall and G. C. Carter, "An approximation to the cumulative distribution function of the magnitude-squared coherence estimate," *IEEE Trans. on Acous., Speech and Sig. Proc.*, vol. ASSP-29, no. 4, Aug 1981.

[4] F. Bach and M. I. Jordan, "Learning graphical models for stationary time series," *IEEE. Trans. on Sig. Proc.*, vol. 52, no. 8, Aug 2004.

[5] J. A. Bilmes, "Dynamic Bayesian multinets," in *Proceedings of UAI*, 2000.

[6] S. Kirshner, P. Smyth, and A. W. Robertson, "Conditional Chow-Liu tree structures for modeling discrete-valued time series," in *Proceedings of UAI*, 2004.

[7] X. Xuan and K. Murphy, "Modeling changing dependency structure in multivariate time series," in *Proceedings of ICML*, 2007.

[8] M. R. Siracusa and J. W. Fisher, "Tractable Bayesian inference of time-series dependence structure," in *Proceedings of AISTATS*, 2009.

[9] D. Materassi and G. Innocenti, "Topological identification in networks of dynamical systems," *IEEE Trans. on Auto. Control*, vol. 55, no. 8, Aug 2010.

[10] V. Y. F. Tan, A. Anandkumar, L. Tong, and A. S. Willsky, "A Large-Deviation Analysis for the Maximum Likelihood Learning of Markov Tree Structures," *IEEE Trans. on Inf. Th.*, Mar 2011.

[11] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning Gaussian tree models: Analysis of error exponents and extremal structures," *IEEE Trans. on Signal Processing*, vol. 58, no. 5, May 2010.

[12] S. Lauritzen, *Graphical Models*, Oxford University Press, USA, 1996.

[13] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees.," *IEEE Trans. on Infomation Theory*, vol. 14, no. 3, May 1968.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2nd edition, 2006.

[15] M. S. Pinsker, "Information and information stability of random variables and processes," *Izv. Akad. Nauk*, 1960.

[16] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, McGraw-Hill, 2nd edition, 2003.

[17] R. Bortel and P. Sovka, "Approximation of statistical distribution of magnitude squared coherence estimated with segment overlapping," *Signal Processing*, vol. 87, no. 5, May 2007.

[18] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer, 2nd edition, 1998.

[3] Since $\gamma(0) = \gamma(2\pi)$ (by periodicity of the DTFT), we can either minimize over $[0, 2\pi)$ or $[0, 2\pi]$ in (32).