On the Bias of Directed Information Estimators

Gabriel Schamberg, Student Member, IEEE, Todd P. Coleman, Senior Member, IEEE

Abstract—When estimating the directed information between two jointly stationary Markov processes, it is typically assumed that the recipient of the directed information is itself Markov of the same order as the joint process. While this assumption is often made explicit in the presentation of such estimators, a characterization of when we can expect the assumption to hold is lacking. Using the concept of d-separation from Bayesian networks, we present sufficient conditions for which this assumption holds. We further show that the set of parameters for which the condition is not also necessary has Lebesgue measure zero. Given the strictness of these conditions, we introduce a notion of partial directed information, which can be used to bound the bias of directed information estimates when the directed information recipient is not itself Markov. Lastly we estimate this bound on simulations in a variety of settings to assess the extent to which the bias should be cause for concern.

Index Terms—Directed Information, Estimation, Bias Quantification, Markov

I. INTRODUCTION

The directed information (DI) is a popular measure of asymmetric relationships between two stochastic processes. Since its origination in 1973 [1] and its reemergence in 1990 [2], the DI has been increasingly pervasive throughout science and engineering disciplines. When using the DI to study the inter-process relationships exhibited by real data, i.e. when the true underlying joint statistics are unknown, it is necessary to utilize DI estimation techniques. DI estimators have been studied extensively in the literature using a variety of approaches, including sequential estimation via context tree weighting (CTW) [3], maximum likelihood estimation of generalized linear models for DI between point processes [4], k-NN estimation [5], and plug-in estimation [6]. With the exception of [6], when estimating the DI from \( Y \) to \( X \), all of these estimators work under the assumptions that (i) \( X \) and \( Y \) are jointly stationary ergodic Markov processes and (ii) \( X \) is itself a jointly stationary ergodic Markov process of the same order. For the plug-in estimator studied in [6], it is noted that when assumption (ii) does not hold, the quantity being estimated is in fact not the DI, but rather an upper bound for the DI. Despite the common adoption of assumptions (i) and (ii), the conditions under which they hold and the implications when they do not are not well studied. Our present work seeks to fill this gap in order to ensure that the estimation of DI across scientific disciplines can be conducted in a manner such that the results are reliable.

Relevant discussions regarding the issues surrounding assumption (ii) have been held in the literature on Granger causality (GC) [7]. GC can be viewed as a special case of DI where the processes in question obey a vector autoregressive (VAR) model with Gaussian noise. It is noted in the GC literature that subsets of finite-order VAR processes are in general infinite order autoregressive processes [8]. Thus, estimating a “restricted” model (i.e. one where the candidate influencer is hidden) from data requires estimating a truncated model and induces a bias-variance tradeoff. For the linear Gaussian case, this issue can be avoided by computing the restricted model directly from the full model using the Yule-Walker equations [9]. Unfortunately, there is no clear extension of this approach for arbitrary Markov processes, and other techniques are required.

We employ a Bayesian network perspective to identify when the independence statements required by DI estimators hold. In particular, by representing a collection of interacting processes as a Bayesian network, we can use the d-separation criterion to identify conditional independencies in relevant subsets of the network.

Our contributions are summarized as follows:
- For networks of interacting processes, we provide sufficient conditions for which the conditional independencies needed to obtain unbiased estimates of the directed information hold.
- We show that these conditions are also necessary with the exception of a set of parameters with Lebesgue measure zero.
- We present a bound for the estimation bias that can be estimated reliably under mild conditions.
- To understand the magnitude of the biases in question, we compute the proposed bound for simulated processes in a variety of problem settings.

II. PRELIMINARIES

A. Notation

We will be considering collections of jointly stationary discrete processes \( X, Y, \) and \( Z \), where, at any time \( i \), \( X_i \in \mathcal{X}, Y_i \in \mathcal{Y}, \) and \( Z_i \in \mathcal{Z} \). Without loss of generality, \( Z \) may represent a collection of processes \((Z^{(1)}, \ldots, Z^{(m)}) \in \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_m \). Collections of samples are indicated with superscripts as \( X^{i+k} \triangleq \{X_i, \ldots, X_{i+k}\} \) and \( X^n \triangleq X^n_i \). In general, capital letters will represent random entities and lower case letters will represent their realizations. When a process is Markov of order \( d \) we will refer to it as \( d \)-Markov, unless \( d = 1 \), in which case we will simply refer to it as Markov. We will use \( p \) to represent probability distributions, with the specific distribution being made clear from context.

B. Directed Information

Consider a collection of processes \((X, Y, Z)\). Define the causally conditional DI from \( Y \) to \( X \) given \( Z \) as:

\[
I(Y^n \rightarrow X^n \mid Z^n) = \sum_{i=1}^n I(X_i; Y^i \mid X^{i-1}, Z^i)
\]
and the associated causally conditional DI rate (when it exists) as:

\[
I(Y \rightarrow X \mid Z) = \lim_{n \rightarrow \infty} \frac{1}{n} I(Y^n \rightarrow X^n \mid Z^n).
\]

(3)

If we assume that (i) \((X,Y,Z)\) are jointly \(d\)-Markov, i.e. that \(p(X_i \mid X^{i-1}, Y^i, Z^i) = p(X_i \mid X^{i-1}_{-d}, Y^i_{-d}, Z^i_{-d})\), the second entropy term in (2) can be simplified to \(H(V)\), where \(V\) is the set of nodes \(\{V_1, \ldots, V_m\}\). If the further assumption is made that (ii) \(X_i \perp (X^{i-1}_{-d}, Y^{i-1}_{-d}) \mid (X^{i-1}_{-d}, Z^i_{-d})\) (henceforth “\(X\) is conditionally \(d\)-Markov given \(Z\)”), then the first entropy term can be simplified as \(H(X_i \mid X^{i-1}_{-d}, Z^i_{-d})\). Once this assumption is made, it is clear that the DI can be estimated from data by splitting a stream \((X^n, Y^n, Z^n)\) into a collection of samples \(\{(X^{i-1}_{-d}, Y^{i-1}_{-d}, Z^i_{-d})\}_{i=1}^n\) and estimating the appropriate distributions using a variety of methods [3]–[6].

The goal of this work is to understand when we can expect both of these assumptions to hold and what the consequences are of assuming they both hold when in fact only the first holds. It should be noted that while we consider only a network and the causally conditional DI as above, all of the results demonstrated in the following sections still apply when \(Z = \emptyset\) and the standard DI for a pair of processes is used.

C. Bayesian Networks

To understand the conditions under which the desired independence relationships hold, we can leverage tools from Bayesian networks, which can be used to represent conditional independencies in collections of random variables using a directed acyclic graph (DAG) \(G = (V,E)\), where \(V = \{V_1, \ldots, V_m\}\) is a set of random variables (equivalently nodes or vertices) and \(E \subset V \times V\) is a set of directed edges that do not contain any cycles [10]. The parent set of a node \(V_i\) in a DAG is defined as the set of nodes with arrows going into \(V_i\), \(\mathcal{P}_i \triangleq \{V_j : (V_j \rightarrow V_i) \in E\}\). The defining characteristic of a Bayesian network representation of a joint distribution over the nodes \(V \sim p\) is the ability to factorize the distribution as:

\[
p(V) = \prod_{i=1}^m p(V_i \mid \mathcal{P}_i).
\]

(4)

If this factorization holds for a given \(p\) and \(G\), we say \(G\) is a Bayesian network for \(p\). A key concept when working with Bayesian networks is the \(d\)-separation criterion, which is used to identify subsets of nodes whose conditional independence is implied by the graphical structure. In particular, when given three disjoint subsets of nodes \(A,B,C \subset V\) in a graph \(G\), a straightforward algorithm (shown in Algorithm 1) can be used to determine if \(C\) \(d\)-separates \(A\) and \(B\). When \(C\) \(d\)-separates \(A\) and \(B\), then for any joint distribution \(p(V)\) such that \(G\) is a Bayesian network for \(p\), \(A\) and \(B\) will be conditionally independent given \(C\). While the converse is not true in general (i.e. independence does not imply \(d\)-separation), it has been shown that for specific classes of Bayesian networks, the set of parameters for which the converse does not hold has Lebesgue measure zero [10], [11]. When a graph \(G\) and joint distribution \(p\) are such that \(d\)-separation holds if and only if conditional independence holds for all subsets of nodes, then the distribution \(p\) is called “faithful” to \(G\) [10].

Algorithm 1 d-Separation [12]

**Input**: DAG \(G = (V,E)\) and disjoint sets \(A,B,C \subset V\)

1. Create a subgraph containing only nodes in \(A, B, \text{ or } C\) or with a directed path to \(A, B, \text{ or } C\)
2. Connect with an undirected edge any two variables that share a common child
3. For each \(c \in C\), remove \(c\) and any edge connected to \(c\)
4. Make every edge an undirected edge
5. Conclude that \(A\) and \(B\) are \(d\)-separated by \(C\) if and only if there is no path connecting \(A\) and \(B\)

III. CHARACTERIZATION OF PROCESSES WITH CONDITIONAL MARKOVITY

A. Network Representation of Markov Processes

A Bayesian network is a very natural representation for collections of Markov processes. In particular, using the chain rule to factorize the joint distribution over \(n\) time steps of the processes \((X,Y,Z)\) yields:

\[
p(X^n, Y^n, Z^n) = \prod_{i=1}^n p(X_i, Y_i, Z_i \mid X^{i-1}_{-d}, Y^{i-1}_{-d}, Z^{i-1}_{-d}).
\]

(5)

We next make the additional assumption (A1) that \(X_i, Y_i, \text{ and } Z_i\) are pairwise conditionally independent given the past \(\{X^{i-1}_{-d}, Y^{i-1}_{-d}, Z^{i-1}_{-d}\}\). This assumption facilitates construction of a Bayesian network, as we can rely on the arrow of time to determine the direction of arrows in the network. In the absence of (A1), we cannot construct a unique Bayesian network representation of Markov processes without making alternative assumptions (the details of which will be discussed in future work). This is similar reasoning to that of [13], where (A1) is used for establishing the equivalence between DI graphs and minimal generative model graphs. Under (A1), we can further simplify (5) as:

\[
p(X^n, Y^n, Z^n) = \prod_{i=1}^n \prod_{S \in \{X_i, Y_i, Z_i\}} p(S \mid X^{i-1}_{-d}, Y^{i-1}_{-d}, Z^{i-1}_{-d}).
\]

(6)

Comparing (4) and (6), it is clear that we can represent a collection of processes as a Bayesian network by letting each node be a single time point of a process (i.e. \(X_i, Y_i, \text{ or } Z_i\)) with parents \(\mathcal{P}_X, \mathcal{P}_Y, \mathcal{P}_Z, \subseteq \{X^{i-1}_{-d}, Y^{i-1}_{-d}, Z^{i-1}_{-d}\}\). Given that there may be multiple valid Bayesian networks for a particular distribution, we note that \(X_i, Y_i, \text{ and } Z_i\) may not be conditionally dependent on the entire set \(\{X^{i-1}_{-d}, Y^{i-1}_{-d}, Z^{i-1}_{-d}\}\). Thus, when constructing a Bayesian network for \((X,Y,Z)\) we include an edge \(S_{i-k} \rightarrow S_i'\) for \(S, S' \in \{X,Y,Z\}\) and \(k = 1, \ldots, d\) only if:

\[
I(S_{i-k} ; S_i' \mid \{X^{i-1}_{-d}, Y^{i-1}_{-d}, Z^{i-1}_{-d}\} \setminus S_{i-k}) > 0.
\]

(7)
B. Necessary and Sufficient Conditions for d-Separation

Using the Bayesian network construction given by (7), we can leverage the d-separation criterion to gain a better understanding of the types of conditions which give rise to the conditional independence relationships needed for DI estimation. To start, we identify necessary and sufficient conditions for which \(X_i\) will be d-separated from \((X^{i-l-1}, Z^{i-l-1})\) by \((X^{i-1}_1, Z^{i-1}_1)\):

**Theorem 1.** Let \((X, Y, Z)\) be a collection of jointly stationary d-Markov processes satisfying (A1). If \(I(Y^n \to X^n \mid Z^n) = 0\) then \(X\) is conditionally d-Markov given \(Z\). If \(I(Y^n \to X^n \mid Z^n) > 0\), \(X\) is conditionally Markov given \(Z\) of order 2d or less if:

\[
I(Y_j; Y_k \mid X^i, Z^i) = 0 \quad \forall j \leq k \leq i
\]

If \(I(Y^n \to X^n \mid Z^n) > 0\) but (8) is not satisfied, there will not exist any positive integer \(l\) such that \((X^{i-1}_1, Z^{i-1}_1)\) d-separates \(X_i\) from \((X^{i-l-1}, Z^{i-l-1})\) in the Bayesian network generated according to (7).

**Proof.** The first statement of the theorem follows trivially from the removal of \(Y_i\) from \(p(X_i \mid X^{i-1}_1, Y^{i-1}_1, Z^{i-1}_1)\). Moving on, we will show that if \(I(Y_j; Y_k \mid X^i, Z^i) = 0\) for all \(j < k \leq i\), \(X\) is conditionally Markov of order at most 2d given \(Z\). Note that:

\[
p(X_i \mid X^{i-1}_1, Z^{i-1}_1) = \sum_{y^{i-1}_1, Z^{i-1}_1} p(X_i \mid X^{i-1}_1, y^{i-1}_1, Z^{i-1}_1) \prod_{j=1}^{i-1} p(y_j \mid X^{i-1}_1, Z^{i-1}_1)
\]

(9)

\[
= \sum_{y^{i-1}_1} p(X_i \mid X^{i-1}_1, y^{i-1}_1, Z^{i-1}_1) \prod_{j=1}^{i-1} p(y_j \mid X^{i-1}_1, Z^{i-1}_1)
\]

(10)

\[
= \sum_{y^{i-1}_1} p(X_i \mid X^{i-1}_1, y^{i-1}_1) \prod_{j=1}^{i-1} p(y_j \mid X^{i-1}_1)
\]

(11)

where (9) follows from the chain rule and the conditional independence of \(y^{i-1}_1\) given \((X^{i-1}_1, Z^{i-1}_1)\), (10) follows from the joint Markovity of \(X\) and \(Y\) and the conditional independence of \(y^{i-1}_1\), and (11) follows from the conditional independence of the past and the future given the present for Markov processes.

Next we will show that if there is some \(j < k \leq i\) such that \(I(Y_j; Y_k \mid X^i, Z^i) > 0\), there is no positive integer \(l\) such that \((X^{i-1}_1, Z^{i-1}_1)\) d-separates \((X^{i-l-1}_1, Z^{i-l-1}_1)\) from \(X_i\). To do this, we first note that \((X^i, Z^i)\) does not d-separate \(Y_j\) and \(Y_k\), because if it did, they would be conditionally independent. As such, when performing the d-separation algorithm given by Algorithm[1] \(Y_j\) and \(Y_k\) will be connected by an undirected edge after completing step 4. Furthermore, if we let \(\tau_l = k - j\), then by the joint stationarity of \((X, Y, Z)\), every \(Y_i\) will be connected to \(Y_{i-\tau_l}\) at the end of step 4. Furthermore, we know that \(I(Y^n \to X^n \mid Z^n) > 0\) implies that for some \(q \leq m\), there is a directed edge from \(Y_q\) to \(X_m\). Letting \(\tau_2 = m - q\), we know from the joint stationarity of \((X, Y, Z)\) that for every \(X_i\), there is an incoming directed edge from \(Y_{i-\tau_2}\). As such, at the end of step 4, every \(X_i\) will be part of an undirected path connecting \(Y_{i-\tau_2}, Y_{i-\tau_2-\tau_1}, Y_{i-\tau_2-2\tau_1}, \ldots\). Thus, for any \(l \geq 1\) this path can be followed \(r\) steps such that \(r\tau_l > d\). Then we know that \(Y_{i-\tau_2-\tau_1}\) is connected via an undirected edge to \(Y_{i-\tau_2-\tau_1-\tau_2} = X_{i-\tau_l}\). Recalling that in step 3 of the d-separation algorithm, \((X^{i-1}_1, Z^{i-1}_1)\) have been removed from the graph, we note that since \(i - r\tau_1 < i - l\), \(X_{i-\tau_l}\) is in the graph. Thus, there is an undirected path connecting \(X_{i-\tau_l} \in (X^{i-l-1}, Z^{i-l-1})\) and \(X_i\) for any \(l\). □

Theorem[1] uses d-separation to provide us with a characterization of networks of processes that are guaranteed to have the conditional independence relations required by DI estimators. With regard to the processes for which we cannot demonstrate d-separation (i.e. those not satisfying (8)), the only distributions that will have the desired conditional independence relations are those that are *unfaithful* to their graphs. While there is ample discussion in the literature noting that these distributions are typically not seen in practice (see [10] and citations therein), a formal characterization within the present context is desired.

C. Completeness of d-Separation

For a DAG \(G = (V, E)\), define \(\Gamma_G \subset \mathbb{R}^M\) to represent the set of all discrete distributions \(p(V)\) such that the \(G\) is a Bayesian network for \(p\). Further define \(\Omega_G \subset \Gamma_G\) to be the subset of those distributions that are unfaithful to \(G\). Then, it was shown in [11] the \(\Omega_G\) has Lebesgue measure zero with respect to \(\mathbb{R}^M\), where \(M\) is the number of parameters needed to specify the joint distribution \(p\). Unfortunately, this result cannot be directly applied to our problem. To see why, let \(\Theta_G \subset \mathbb{R}^N\) represent the set of parameters defining discrete jointly stationary d-Markov processes satisfying (A1) for which \(G\) gives the Bayesian network constructed according [7]. Defining \(\theta_{X,Y,Z} = p(S_i \mid X^{i-1}_1, Y^{i-1}_1, Z^{i-1}_1)\) for \(S \in \{X, Y, Z\}\) and \(\theta = (\theta_{X,Y,Z}) : a \in X \cup Y \cup Z, b \in X^d \times Y^d \times Z^d\), we can see that there are \(N \approx (|X| + |Y| + |Z| - 3)|X|^d|Y|^d|Z|^d\) many parameters uniquely defining such a process. Next define \(\Theta_G' \subset \Theta_G\) to be the subset of parameters such that the induced distribution \(p\) is unfaithful to \(G\). Then, it is clear that, due to the stationarity constraint, \(N \ll M\), and the Lebesgue measure of \(\Theta_G'\) with respect to \(\mathbb{R}^M\) does not tell us what the Lebesgue measure of \(\Theta_G\) is with respect to \(\mathbb{R}^N\). Returning to the question at hand, we seek to know when we can expect \(X\) to be conditionally d-Markov given \(Z\) despite the conditional independence not being implied by d-separation. Using a similar technique to [11], the following theorem states that, when \(d = 1\), the set of such parameters has Lebesgue measure zero:
Theorem 2. The set of parameters defining a collection $(X, Y, Z)$ of jointly stationary irreducible aperiodic Markov processes such that there exists a positive integer $l$ where $X$ is conditionally $l$-Markov given $Z$ but $(X^{i-1}, Z^{i-1})$ does not $d$-separate $X_i$ from $(X^{i-1}, Z^{i-1})$ in the Bayesian network constructed by [7] has Lebesgue measure zero with respect to $\mathbb{R}^N$.

Proof. We will show that the statement holds for a fixed $l$, noting that a countably infinitely union of measure zero sets has measure zero. First note that, if $X$ is conditionally $l$-Markov given $Z$, then for any $x_{i-1}^{l-1}, x_{i-1}^{l-1} \in X$ and $z_{i-1}^{l-1}, z_{i-1}^{l-1} \in Z$ the following equality must hold:

$$p(x_i | x_{i-1}^{l-1}, z_{i-1}^{l-1}) = \frac{p(x_i | x_{i-1}^{l-1}, z_{i-1}^{l-1})}{p(x_{i-1}^{l-1}, z_{i-1}^{l-1})} \quad (12)$$

where we define $x_{i-1}^{l-1} = \{x_{i-1}^{l-1}, x_{i-1}^{l-1}\}$ and $z_{i-1}^{l-1} = \{z_{i-1}^{l-1}, z_{i-1}^{l-1}\}$. We will demonstrate that the equation given by (12) amounts to solving a polynomial function of the parameters $\theta$. It is shown in [14] that the set of solutions to a non-trivial polynomial (i.e. one that is not solved by all of $\mathbb{R}^N$) will have Lebesgue measure zero with respect to $\mathbb{R}^N$. Focusing on the left side of (12), we see that:

$$p(x_i | x_{i-1}^{l-1}, x_{i-1}^{l-1}) = \sum_{y_{i-1}^{l-1}} \theta_{x,y,z} p(y_{i-1}^{l-1} | x_{i-1}^{l-1}, z_{i-1}^{l-1})$$

$$= \sum_{y_{i-1}^{l-1}} \theta_{x,y,z} \frac{p(x_{i-1}^{l-1}, y_{i-1}^{l-1}, z_{i-1}^{l-1})}{p(x_{i-1}^{l-1}, z_{i-1}^{l-1})} \prod_{j=1}^l \theta_{x,y,z}$$

$$= \frac{\sum_{y_{i-1}^{l-1}} \theta_{x,y,z} \pi(x_{i-1}^{l-1}, y_{i-1}^{l-1}, z_{i-1}^{l-1}) \prod_{j=1}^l \theta_{x,y,z}}{\sum_{y_{i-1}^{l-1}} \pi(x_{i-1}^{l-1}, y_{i-1}^{l-1}, z_{i-1}^{l-1}) \prod_{j=1}^l \theta_{x,y,z}} \quad (13)$$

where $\pi : |X| \times |Y| \times |Z| \rightarrow [0, 1]$ is the invariant distribution and $\theta_{x,y,z} = \theta_{x,y,z} \theta_{y,z}$. Next, define a matrix $A \in \mathbb{R}^{|X| |Y| |Z|}$ containing the transition probabilities, i.e. $A_{j,k} = \theta_{x,y,z}$ where $R$ is some enumeration over the $|X| |Y| |Z|$. Possible values taken by $(X, Y, Z)$. Then we can represent $\pi$ in vector form $\pi \in [0, 1]^{|X| |Y| |Z|}$ as the unique solution to $\pi = \pi A$. Let $\bar{\pi}$ be an arbitrary vector satisfying $(A^T - I)\bar{\pi} = 0$, and note that for any $\bar{\pi}$ there is a constant $C$ such that $C\bar{\pi} = \pi$. Such a vector $\bar{\pi}$ can be found by performing Gauss-Jordan elimination on $(A^T - I)$, and as a result, each element $\bar{\pi}$ can be written as fractions of polynomial functions of $\theta$. Replacing $\pi$ with $C\bar{\pi}$ in its functional form $\pi : |X| \times |Y| \times |Z| \rightarrow \mathbb{R}$ in (13) we see that $C$ cancels in the numerator and denominator and thus each side of (12) can be written entirely as fractions of polynomial functions of $\theta$. Next, repeat the process on the right hand side of (12) with $x_{i-1}^{l-1} \rightarrow y_{i-1}^{l-1}$ and $x_{i-1}^{l-1} \rightarrow z_{i-1}^{l-1}$. Then, for any term that appears as a fraction, we can multiply both sides of (12) by the denominator and repeat until (12) is a polynomial function of $\theta$. Finally, we note that the polynomial given by (12) is trivial only if every process is a solution. Though omitted here for brevity, one can easily show that the polynomial is non-trivial by constructing a counterexample.

IV. QUANTIFYING ESTIMATION BIAS

We have shown that DI estimator are reliant upon a condition that is unlikely to be satisfied. Thus, we now define two augmented notions of DI that do not require X to be conditionally Markov in order to be accurately estimated.

Definition 1. The $k^{th}$-order causally conditioned truncated directed information (TDI) from Y to X given Z is defined as:

$$I^{(k)}_T(Y^n \rightarrow X^n | Z^n) \triangleq \sum_{i=1}^n I(X_i; Y_{i-k}^{i-1}, Z_{i-k}^{i-1})$$

(14)

The TDI in its unconditional form is discussed in [6] in the context of plug-in estimators of DI. Should both Markovicity and conditional Markovicity hold for a collection of processes, then the TDI and the DI are equivalent. However, having shown that conditional Markovicity is unlikely to hold, we here name the TDI to emphasize that it is a fundamentally different measure from the traditional DI.

Definition 2. The $k^{th}$-order causally conditioned partial directed information (PDI) from Y to X given Z is defined as:

$$I^{(k)}_P(Y^n \rightarrow X^n | Z^n) \triangleq \sum_{i=1}^n I(X_i; Y_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1})$$

(15)

The PDI can be thought of as measuring the unique influence of the $k$ most recent samples of Y on X. It is important to note that, under the assumption that $(X, Y, Z)$ are jointly d-Markov, we have that:

$$I(X_i; Y_{i-k}^{i-1} | X_{i-k}^{i-1}, Y_{i-k}^{i-1}) = H(X_i | X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1}) - H(X_i | X_{i-k}^{i-1}, Z_{i-k}^{i-1})$$

Thus, it is clear that estimators of DI can be extended to estimate the PDI without the additional requirement of conditional Markovicity, though the details of these estimators are postponed for future work. Defining the TDI and PDI rates $I^{(k)}_T$ and $I^{(k)}_P$ to be the normalized limits analogous with the DI rate given by (3), we are able to bound the DI rate from above and below as follows:

Theorem 3. Let $(X, Y, Z)$ be jointly stationary d-Markov. For $k_1 \geq 1$ and $k_2 \geq d$, the causally conditional TDI and PDI rates bound the DI rate as:

$$I^{(k)}_T(Y \rightarrow X || Z) \leq I(Y \rightarrow X || Z) \leq I^{(k_2)}_T(Y \rightarrow X || Z)$$

(16)

with both bounds becoming equalities as $k_1, k_2 \rightarrow \infty$.

Proof. Note that for any $k_1 \geq 1$ and $k_2 \geq d$:

$$H(X_i | X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1})$$

$$H(X_i | X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1})$$

$$H(X_i | X_{i-k}^{i-1}, Z_{i-k}^{i-1})$$

$$H(X_i | X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1})$$

$$H(X_i | X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1})$$

$H(X_i | X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1})$
\[
\begin{align*}
\leq H(X_i \mid X_{i-k}^{i-1}, Z_{i-k}^{i-1}) - H(X_i \mid X_{i-k}^{i-1}, Y_{i-k}^{i-1}, Z_{i-k}^{i-1})
\end{align*}
\]

where (18), (19), and (21) follow from conditioning reduces entropy and (20) follows from joint d-Markovity of \((X, Y, Z)\). Taking the sum over \(i = 1, \ldots, n\) and the normalized limit as \(n \to \infty\) gives the desired result, noting that (17), (18), and (21) become the PDI, DI, and TDI rates, respectively.

V. SIMULATIONS

In the above sections we have demonstrated that while one cannot reasonably expect to data to satisfy the necessary assumptions for obtaining unbiased estimates of DI, the TDI and PDI can be used to provide upper and lower bounds for the true DI. A natural next question is, how significant is the difference between PDI and TDI? To address this question, we simulate a pair of jointly stationary Markov discrete processes in four settings, each characterized by a particular simplification of the generative distribution \(p(X_i, Y_i \mid X_{i-1}^{i-k}, Y_{i-1}^{i-k})\):

\[
\begin{align*}
p(X_i \mid Y_{i-1}) & p(Y_i \mid Y_{i-1}) \quad (S1) \\
p(X_i \mid X_{i-1}, Y_{i-1}) & p(Y_i \mid Y_{i-1}) \quad (S2) \\
p(X_i \mid X_{i-1}, Y_{i-1}) & p(Y_i \mid X_{i-1}, Y_{i-1}) \quad (S3) \\
p(X_i \mid X_{i-2}^{i-1}, Y_{i-2}^{i-1}) & p(Y_i \mid X_{i-2}^{i-1}, Y_{i-2}^{i-1}) \quad (S4)
\end{align*}
\]

For each of these graphical structures, we conducted 100 experiments with \(|X| = |Y| = 4\) for (S1)-(S3) and \(|X| = |Y| = 3\) for (S4). In each experiment, the parameters were sample as independent exponential random variables and then appropriately normalized, yielding parameters drawn uniformly from the probability simplex \([15]\). Using the sampled parameters, sequences \((x^n, y^n)\) were generated with \(n = 300000\) large enough to ensure that accurate estimates of the TDI and PDI could be obtained. \(\hat{I}^{(k)}_T(Y \to X)\) and \(\hat{I}^{(k)}_P(Y \to X)\) were estimated using CTW estimators in the style of \(\hat{I}_3\) in \([6]\) for \(k = d, d+1, \ldots, d+10\). Figure 1 shows boxplots representing \(\hat{I}^{(k)}_T(Y \to X) - \hat{I}^{(k)}_P(Y \to X)\) for varying values of \(k\) along with the mean (across trials) DI rate, which was determined by the value converged upon by the TDI and PDI. We can see that while the bound on the bias quickly converges to zero as \(k\) increases, there are many examples in every setting when \(k = d\) for which the bound on the bias is rather large relative to the mean DI rate. Furthermore, we can see that as the structures get more complex, the bound on the bias tends to be larger. This suggests that while (S4) is not covered by Theorem 2, alternative proof techniques may exist for demonstrating that the results hold for \(d > 1\). Thus, when working with real data, it may be prudent to use the TDI and PDI to upper and lower bound the DI rate rather than simply relying on the TDI as a proxy for DI.

REFERENCES

[1] H. Marko, “The bidirectional communication theory—a generalization of information theory,” IEEE Trans. on Comm., 1973.
[2] J. Massey, “Causality, feedback and directed information,” in Proc. Int. Symp. Inf. Theory Applic., 1990.
[3] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, “Universal estimation of directed information,” IEEE Trans. on Inf. Theory, 2013.
[4] 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,” J. of Comp. Neuroscience, 2011.
[5] Y. Murin, “K-nn estimation of directed information,” ArXiv preprint arXiv:1711.08516, 2017.

![Figure 1](https://example.com/figure1.png)

Fig. 1. Difference between truncated and partial directed information (y-axis) for different values of \(k\) (x-axis) under different process structures (panels).

[1] Code can be found in the following repository: [https://github.com/gabeschamberg/directed_info_bias](https://github.com/gabeschamberg/directed_info_bias)

[12] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer, “Independence properties of directed Markov fields,” Networks, 1990.
[13] C. J. Quinn, N. Kiyavash, and T. P. Coleman, “Directed information graphs,” IEEE Trans. on Inf. Theory, 2015.
[14] M. Okamoto, “Distinctness of the eigenvalues of a quadratic form in a multivariate sample,” The Annals of Statistics, 1973.
[15] L. Devroye, Non-uniform random variate generation. Springer-Verlag, 1986.